NFVRG
Internet-Draft
Intended status: Informational
Expires: September 22, 2016

R. Szabo, Ed.
Z. Qiang
Ericsson
M. Kind
Deutsche Telekom AG

March 21, 2016

Recursive virtualization and programming for network and cloud resources

draft-irtf-nfvrg-unify-recursive-programming-00

Abstract

The introduction of Network Function Virtualization (NFV) in carrier-grade networks promises improved operations in terms of flexibility, efficiency, and manageability. NFV is an approach to combine network and compute virtualizations together. However, network and compute resource domains expose different virtualizations and programmable interfaces. In [I-D.unify-nfvrg-challenges] we argued for a joint compute and network virtualization by looking into different compute abstractions.

In this document we analyze different approaches to orchestrate a service graph with transparent network functions relying on a public telecommunication network and ending in a commodity data center. We show that a recursive compute and network joint virtualization and programming has clear advantages compared to other approaches with separated control between compute and network resources. In addition, the joint virtualization will have cost and performance advantages by removing additional virtualization overhead. The discussion of the problems and the proposed solution is generic for any data center use case; however, we use NFV as an example.

Status of This Memo

This Internet-Draft will expire on September 22, 2016.

Table of Contents

1. Introduction

To a large degree there is agreement in the research community that rigid network control limits the flexibility of service creation. In [I-D.unify-nfvrg-challenges]

- we analyzed different compute domain abstractions to argue that joint compute and network virtualization and programming is needed for efficient combination of these resource domains;

- we described challenges associated with the combined handling of compute and network resources for a unified production environment.

Our goal here is to analyze different approaches to instantiate a service graph with transparent network functions into a commodity Data Center (DC). More specifically, we analyze

- two black box DC set-ups, where the intra-DC network control is limited to some generic compute only control programming interface;

- a white box DC set-up, where the intra-DC network control is exposed directly to for a DC external control to coordinate forwarding configurations;

- a recursive approach, which illustrates potential benefits of a joint compute and network virtualization and control.

The discussion of the problems and the proposed solution is generic for any data center use case; however, we use NFV as an example.

2. Terms and Definitions

We use the terms compute and "compute and storage" interchangeably throughout the document. Moreover, we use the following definitions, as established in [ETSI-NFV-Arch]:

NFV: Network Function Virtualization - The principle of separating network functions from the hardware they run on by using virtual hardware abstraction.

NFVI: NFV Infrastructure - Any combination of virtualized compute, storage and network resources.

VNF: Virtualized Network Function - a software-based network function.

MANO: Management and Orchestration - In the ETSI NFV framework [ETSI-NFV-MANO], this is the global entity responsible for management and orchestration of NFV lifecycle.

Further, we make use of the following terms:

NF: a network function, either software-based (VNF) or appliance-based.

SW: a (routing/switching) network element with a programmable control plane interface.

DC: a data center is an interconnection of Compute Nodes (see below) with a data center controller, which offers programmatic resource control interface to its clients.

CN: a server, which is controlled by a DC control plane and provides execution environment for virtual machine (VM) images such as VNFs.

3. Use Cases

Service Function Chaining (SFC) looks into the problem how to deliver end-to-end services through the chain of network functions (NFs). Many of such NFs are envisioned to be transparent to the client, i.e., they intercept the client connection for adding value to the services without the knowledge of the client. However, deploying network function chains in DCs with Virtualized Network Functions (VNFs) are far from trivial [I-D.ietf-sfc-dc-use-cases]. For example, different exposures of the internals of the DC will imply different dynamisms in operations, different orchestration complexities and may yield for different business cases with regards to infrastructure sharing.

We investigate different scenarios with a simple NF forwarding graph of three VNFs (o->VNF1->VNF2->VNF3->o), where all VNFs are deployed within the same DC. We assume that the DC is a multi-tier leaf and spine (CLOS) and that all VNFs of the forwarding graph are bump-in-the-wire NFs, i.e., the client cannot explicitly access them.

3.1. Black Box DC

In Black Bock DC set-ups, we assume that the compute domain is an autonomous domain with legacy (e.g., OpenStack) orchestration APIs. Due to the lack of direct forwarding control within the DC, no native L2 forwarding can be used to insert VNFs running in the DC into the forwarding graph. Instead, explicit tunnels (e.g., VxLAN) must be used, which need termination support within the deployed VNFs. Therefore, VNFs must be aware of the previous and the next hops of the forwarding graph to receive and forward packets accordingly.

3.1.1. Black Box DC with L3 tunnels

Figure 1 illustrates a set-up where an external VxLAN termination point in the SDN domain is used to forward packets to the first NF (VNF1) of the chain within the DC. VNF1, in turn, is configured to forward packets to the next SF (VNF2) in the chain and so forth with VNF2 and VNF3.

In this set-up VNFs must be capable of handling L3 tunnels (e.g., VxLAN) and must act as forwarders themselves. Additionally, an operational L3 underlay must be present so that VNFs can address each other.

Furthermore, VNFs holding chain forwarding information could be untrusted user plane functions from 3rd party developers. Enforcement of proper forwarding is problematic.

Additionally, compute only orchestration might result in sub-optimal allocation of the VNFs with regards to the forwarding overlay, for example, see back-forth use of a core switch in Figure 1.

In [I-D.unify-nfvrg-challenges] we also pointed out that within a single Compute Node (CN) similar VNF placement and overlay optimization problem may reappear in the context of network interface cards and CPU cores.

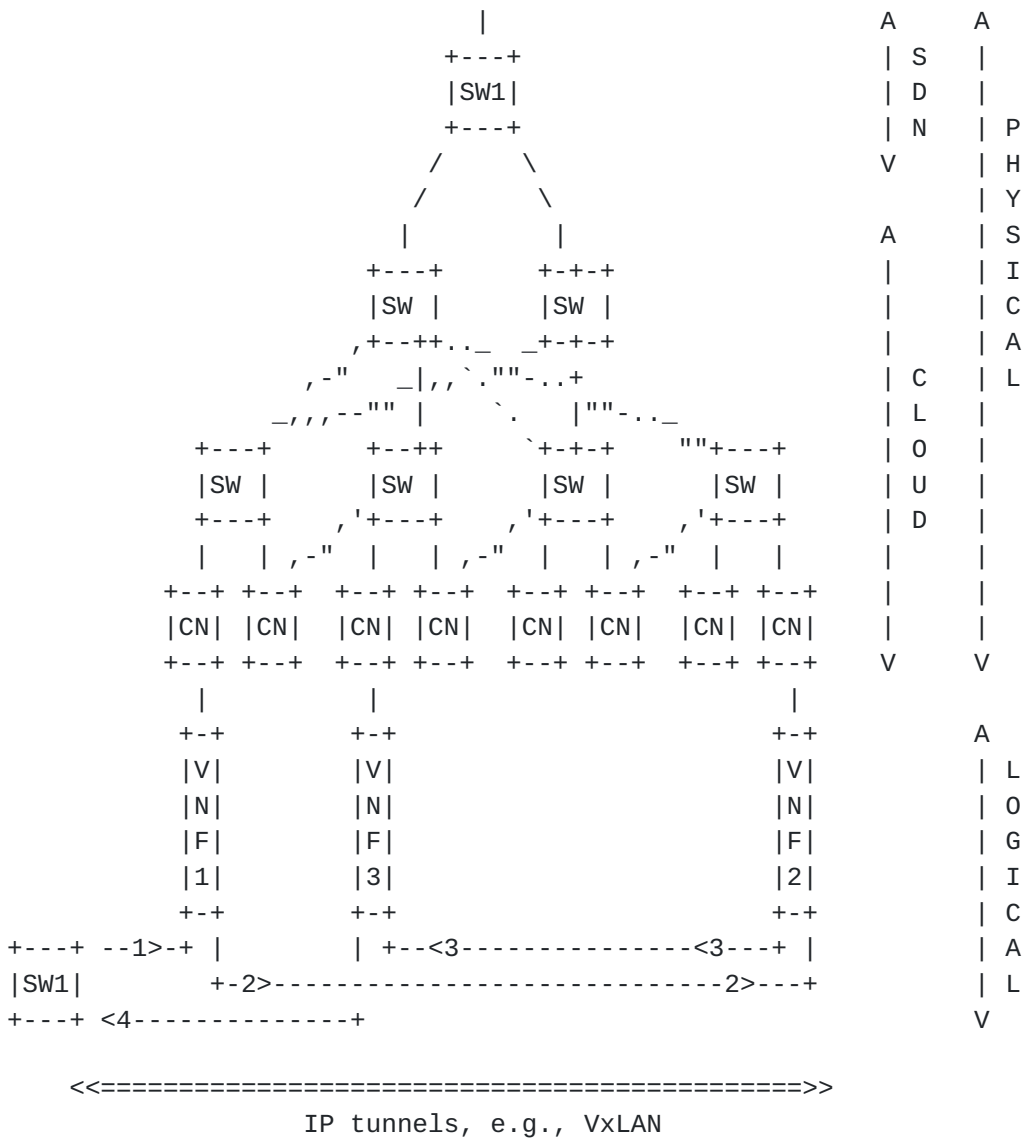

Figure 1: Black Box Data Center with VNF Overlay

3.1.2. Black Box DC with external steering

Figure 2 illustrates a set-up where an external VxLAN termination point in the SDN domain is used to forward packets among all the SFs (VNF1-VNF3) of the chain within the DC. VNFs in the DC need to be configured to receive and send packets between only the SDN endpoint, hence are not aware of the next hop VNF address. Shall any VNFs need to be relocated, e.g., due to scale in/out as described in [I-D.zu-nfvrg-elasticity-vnf], the forwarding overlay can be transparently re-configured at the SDN domain.

Note however, that traffic between the DC internal SFs (VNF1, VNF2, VNF3) need to exit and re-enter the DC through the external SDN switch. This, certainly, is sub-optimal an results in ping-pong traffic similar to the local and remote DC case discussed in [I-D.zu-nfvrg-elasticity-vnf].

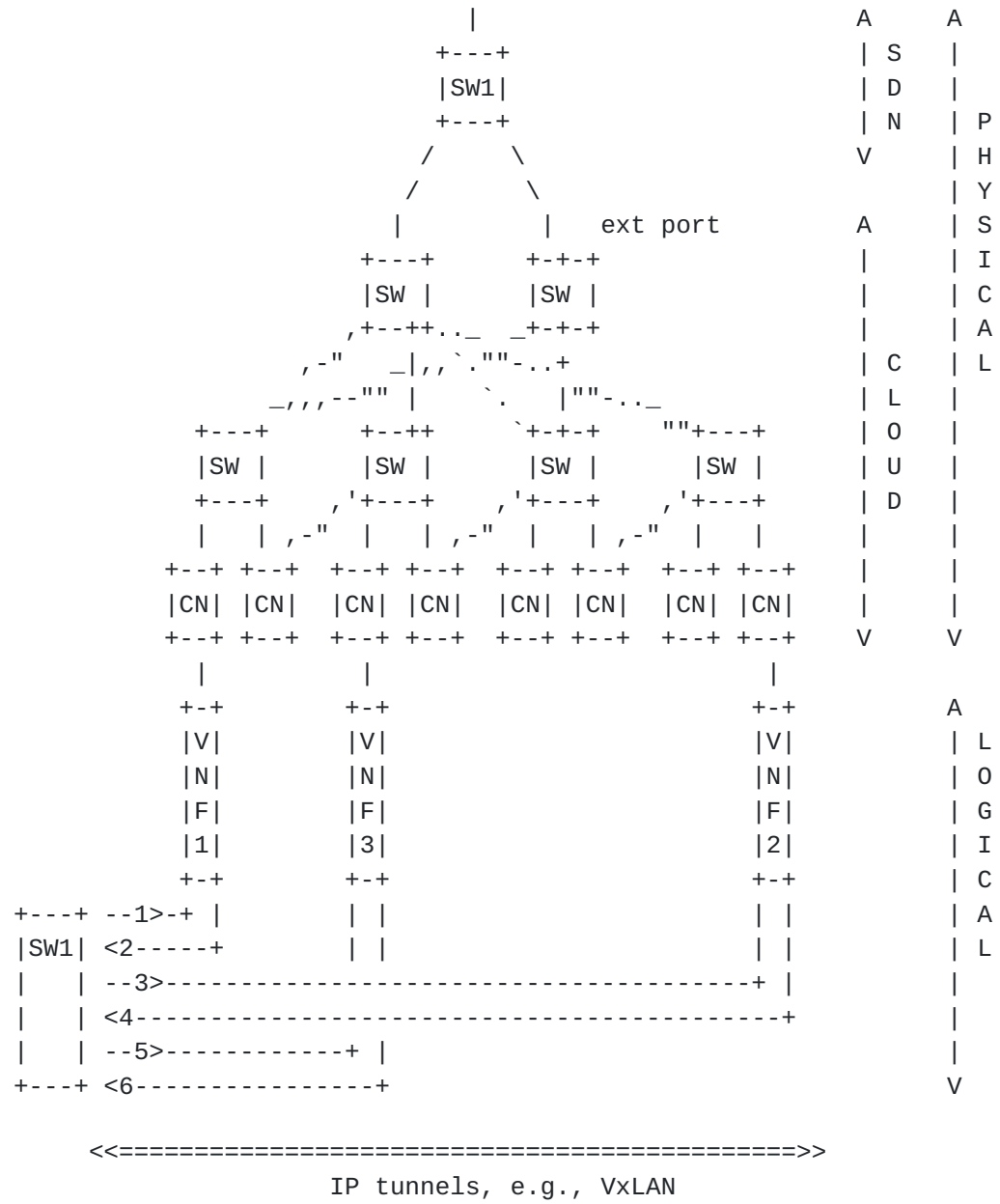

Figure 2: Black Box Data Center with ext Overlay

3.2. White Box DC

Figure 3 illustrates a set-up where the internal network of the DC is exposed in full details through an SDN Controller for steering control. We assume that native L2 forwarding can be applied all through the DC until the VNFs' port, hence IP tunneling and tunnel termination at the VNFs are not needed. Therefore, VNFs need not be forwarding graph aware but transparently receive and forward packets. However, the implications are that the network control of the DC must be handed over to an external forwarding controller (see that the SDN domain and the DC domain overlaps in Figure 3). This most probably prohibits clear operational separation or separate ownerships of the two domains.

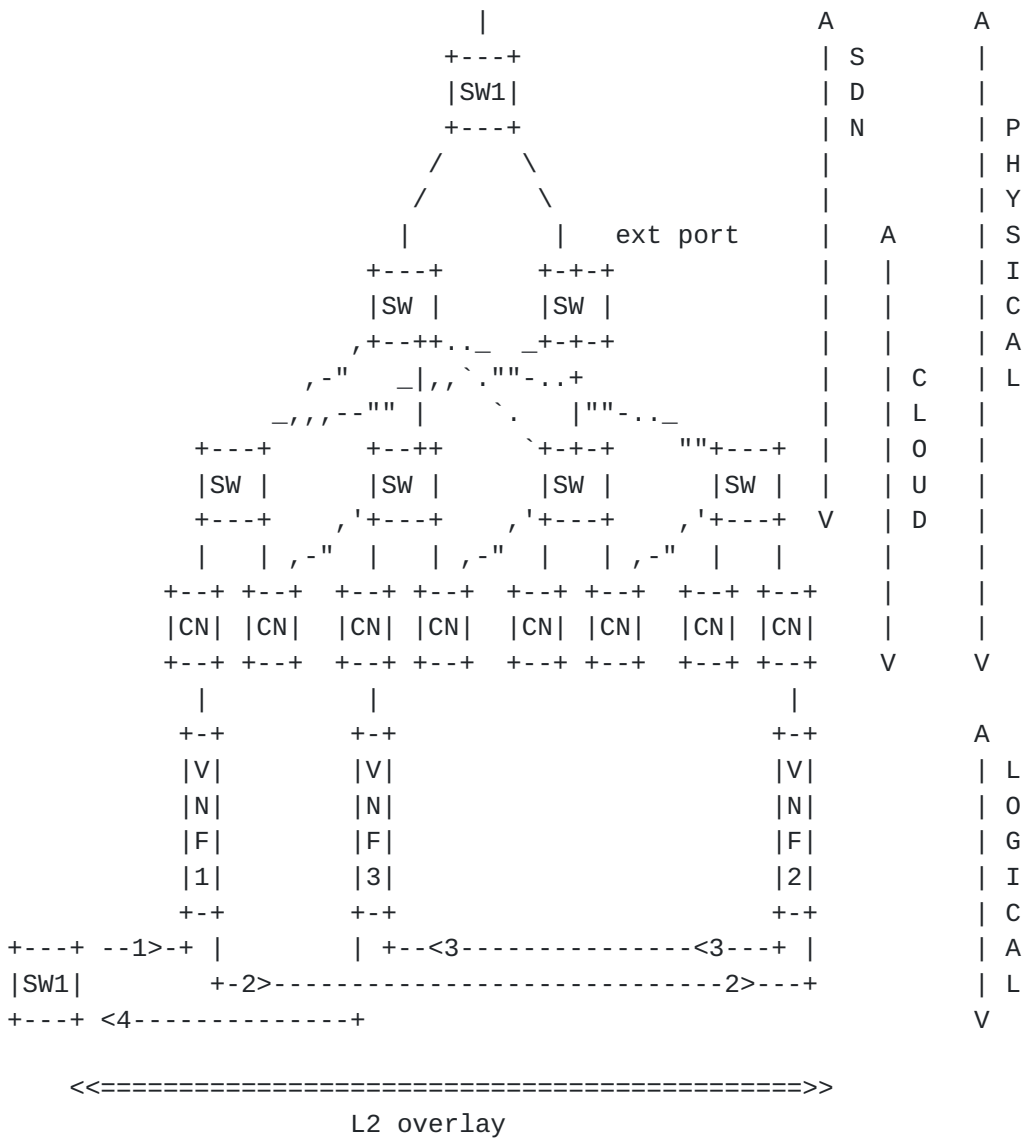

Figure 3: White Box Data Center with L2 Overlay

3.3. Conclusions

We have shown that the different solutions imply different operation and management actions. From network operations point of view, it is not desirable to run and manage similar functions several times (L3 blackbox DC case) - especially if the networking overlay can be easily managed upfront by using a programmatic interface, like with the external steering in black and whitebox DC scenarios.

4. Recursive approach

We argued in [I-D.unify-nfvrg-challenges] and [I-D.caszpe-nfvrg-orchestration-challenges] for a joint software and network programming interface. Consider that such joint software and network abstraction (virtualization) exists around the DC with a corresponding resource programmatic interface. A software and network programming interface could include VNF requests and the definition of the corresponding network overlay. However, such programming interface is similar to the top level services definition, for example, by the means of a VNF Forwarding Graph.

Figure 4 illustrates a joint domain virtualization and programming setup. In Figure 4 "[x]" denotes ports of the virtualized data plane while "x" denotes port created dynamically as part of the VNF deployment request. Over the joint software and network virtualization VNF placement and the corresponding traffic steering could be defined in an atomic, which is orchestrated, split and handled to the next levels (see Figure 5) in the hierarchy for further orchestration. Such setup allows clear operational separation, arbitrary domain virtualization (e.g., topology details could be omitted) and constraint based optimization of domain wide resources.

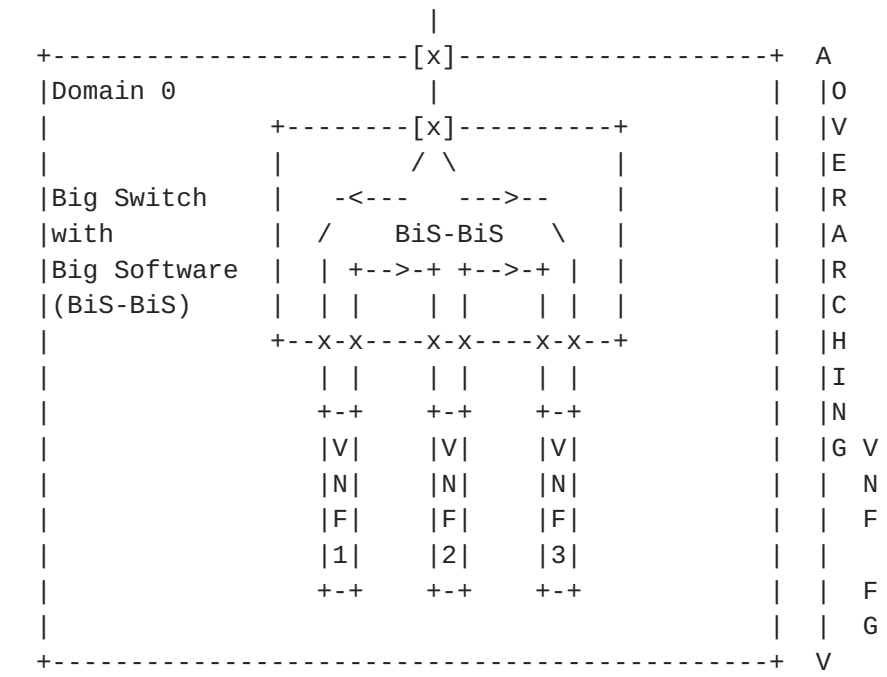

Figure 4: Recursive Domain Virtualization and Joint VNF FG programming: Overarching View

```
       +-------------------------|-----------------------------+ A
       | +----------------------[x]---------------------+  AV  | |
       | | Domain 1             /  \                    |  |N  | |
       | |                     |    A                   |  |F  | |
       | | Big Switch (BS)     |    |                   |  |   | |O
       | |                     V    |                   |  |F  | |V
       | |                    /      \                  |  |G  | |E
       | +-----------------[x]--------[x]---------------+  V1  | |R
       |                    |          |                       | |A
       | +------------------|----------|----------------+  A   | |R
       | |Domain 2          |          A                |  |   | |C
       | |                  V          |                |  |   | |H
       | |             +---[x]--------[x]----+          |  |V  | |I
       | |Big Switch   |   /   BiS-BiS  \    |          |  |N  | |N
       | |with         |  /              \   |          |  |F  | |G
       | |Big Software |  | +-->-+ +-->-+ |  |          |  |   | |
       | |(BiS-BiS)    |  | |    | |    | |  |          |  |F  | |V
       | |             +--x-x----x-x----x-x--+          |  |G  | |N
       | |                | |    | |    | |             |  |2  | |F
       | |                +-+    +-+    +-+             |  |   | |
       | |                |V|    |V|    |V|             |  |   | |F
       | |                |N|    |N|    |N|             |  |   | |G
       | |                |F|    |F|    |F|             |  |   | |
       | |                |1|    |2|    |3|             |  |   | |
       | |                +-+    +-+    +-+             |  |   | |
       | +----------------------------------------------+  V   | |
       +-------------------------------------------------------+ V
```

Figure 5: Recursive Domain Virtualization and Joint VNF FG programming: Domain Views

4.1. Virtualization

Let us first define the joint software and network abstraction (virtualization) as a Big Switch with Big Software (BiS-BiS). A BiS-BiS is a node abstraction, which incorporates both software and networking resources with an associated joint software and network control API (see Figure 6).

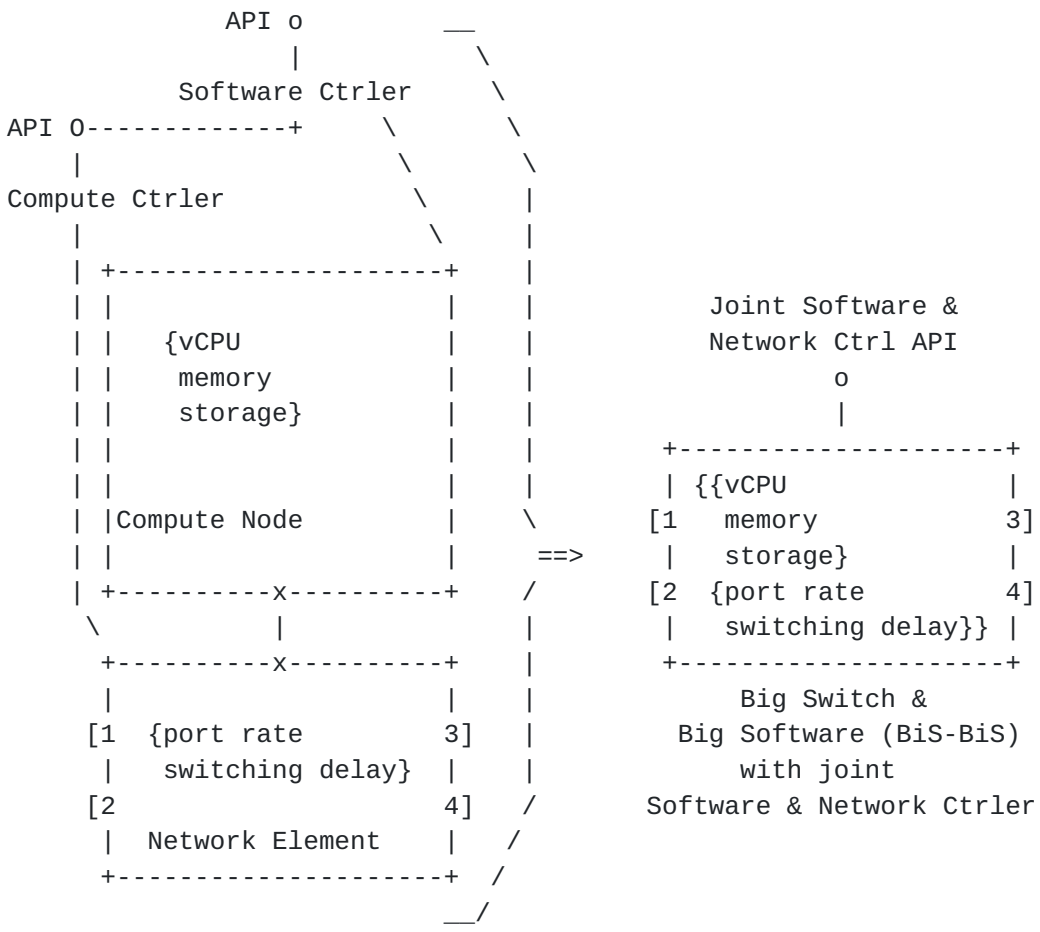

Figure 6: Big Switch with Big Software definition

The configuration over a BiS-BiS allows the atomic definition of NF placements and the corresponding forwarding overlay as a Network Function - Forwarding Graph (NF-FG). The embedment of NFs into a BiS-BiS allows the inclusion of NF ports into the forwarding overlay definition (see ports a, b, ...,f in Figure 7). Ports 1,2, ..., 4 are seen as infrastructure ports while NF ports are created and destroyed with NF placements.

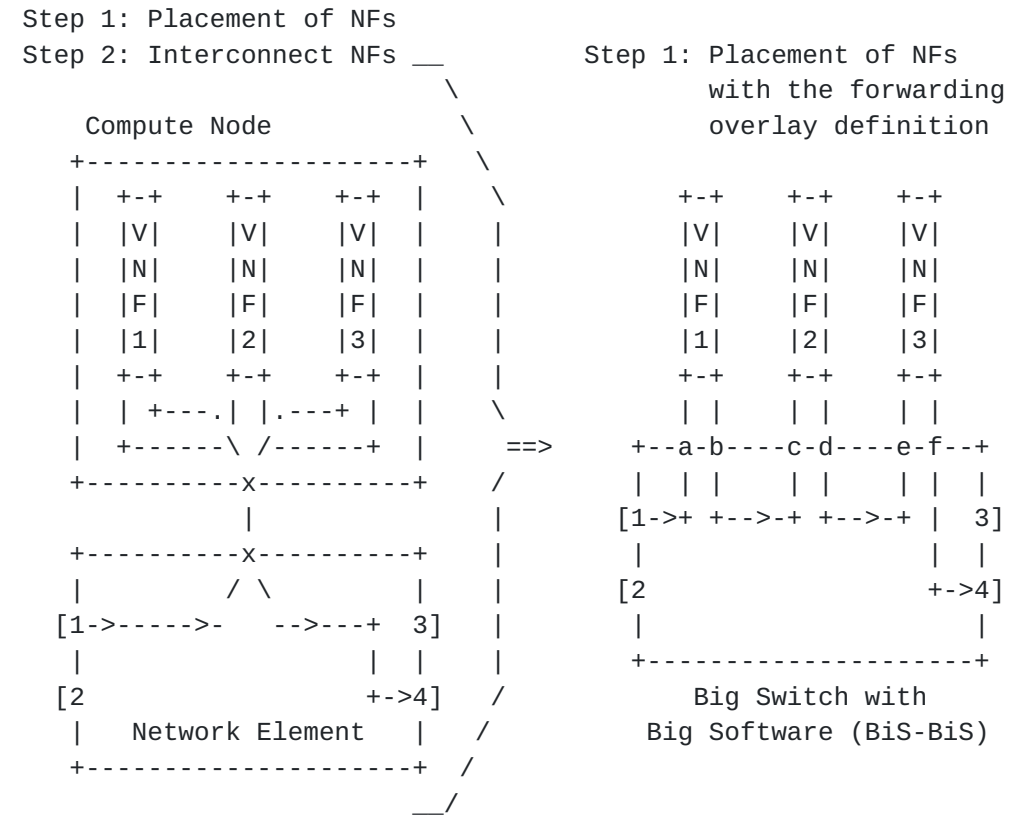

Figure 7: Big Switch with Big Software definition with a Network Function - Forwarding Graph (NF-FG)

4.1.1. The virtualizer's data model

4.1.1.1. Tree view

```
module: virtualizer
   +--rw virtualizer
      +--rw id          string
      +--rw name?       string
      +--rw nodes
      |  +--rw node* [id]
      |     +--rw id              string
      |     +--rw name?           string
      |     +--rw type            string
      |     +--rw ports
      |     |  +--rw port* [id]
      |     |     +--rw id            string
      |     |     +--rw name?         string
      |     |     +--rw port_type?    string
      |     |     +--rw capability?   string
      |     |     +--rw sap?          string
      |     |     +--rw sap_data
```

```
|     |     |  +--rw technology?   string
|     |     |  +--rw resources
|     |     |     +--rw delay?       string
|     |     |     +--rw bandwidth?   string
|     |     |     +--rw cost?        string
|     |     +--rw control
|     |     |  +--rw controller?     string
|     |     |  +--rw orchestrator?   string
|     |     +--rw addresses
|     |     |  +--rw l2?   string
|     |     |  +--rw l3* [id]
|     |     |  |  +--rw id           string
|     |     |  |  +--rw name?        string
|     |     |  |  +--rw configure?   string
|     |     |  |  +--rw client?      string
|     |     |  |  +--rw requested?   string
|     |     |  |  +--rw provided?    string
|     |     |  +--rw l4?   string
|     |     +--rw metadata* [key]
|     |        +--rw key      string
|     |        +--rw value?   string
|     +--rw links
|     |  +--rw link* [id]
|     |     +--rw id           string
|     |     +--rw name?        string
|     |     +--rw src?         ->
|     |     +--rw dst?         ->
|     |     +--rw resources
|     |        +--rw delay?       string
|     |        +--rw bandwidth?   string
|     |        +--rw cost?        string
|     +--rw resources
|     |  +--rw cpu        string
|     |  +--rw mem        string
|     |  +--rw storage    string
|     |  +--rw cost?      string
|     +--rw metadata* [key]
|     |  +--rw key      string
|     |  +--rw value?   string
|     +--rw NF_instances
|     |  +--rw node* [id]
|     |     +--rw id           string
|     |     +--rw name?        string
|     |     +--rw type?        string
|     |     +--rw ports
|     |     |  +--rw port* [id]
|     |     |     +--rw id            string
|     |     |     +--rw name?         string
```

```
|     |     |     +--rw port_type?    string
|     |     |     +--rw capability?   string
|     |     |     +--rw sap?          string
|     |     |     +--rw sap_data
|     |     |     |  +--rw technology?   string
|     |     |     |  +--rw resources
|     |     |     |     +--rw delay?       string
|     |     |     |     +--rw bandwidth?   string
|     |     |     |     +--rw cost?        string
|     |     |     +--rw control
|     |     |     |  +--rw controller?     string
|     |     |     |  +--rw orchestrator?   string
|     |     |     +--rw addresses
|     |     |     |  +--rw l2?   string
|     |     |     |  +--rw l3* [id]
|     |     |     |  |  +--rw id           string
|     |     |     |  |  +--rw name?        string
|     |     |     |  |  +--rw configure?   string
|     |     |     |  |  +--rw client?      string
|     |     |     |  |  +--rw requested?   string
|     |     |     |  |  +--rw provided?    string
|     |     |     |  +--rw l4?   string
|     |     |     +--rw metadata* [key]
|     |     |        +--rw key      string
|     |     |        +--rw value?   string
|     |     +--rw links
|     |     |  +--rw link* [id]
|     |     |     +--rw id           string
|     |     |     +--rw name?        string
|     |     |     +--rw src?         ->
|     |     |     +--rw dst?         ->
|     |     |     +--rw resources
|     |     |        +--rw delay?       string
|     |     |        +--rw bandwidth?   string
|     |     |        +--rw cost?        string
|     |     +--rw resources
|     |     |  +--rw cpu        string
|     |     |  +--rw mem        string
|     |     |  +--rw storage    string
|     |     |  +--rw cost?      string
|     |     +--rw metadata* [key]
|     |        +--rw key      string
|     |        +--rw value?   string
|     +--rw capabilities
|     |  +--rw supported_NFs
|     |     +--rw node* [id]
|     |        +--rw id           string
|     |        +--rw name?        string
```

```
|     |        +--rw type?        string
|     |        +--rw ports
|     |        |  +--rw port* [id]
|     |        |     +--rw id            string
|     |        |     +--rw name?         string
|     |        |     +--rw port_type?    string
|     |        |     +--rw capability?   string
|     |        |     +--rw sap?          string
|     |        |     +--rw sap_data
|     |        |     |  +--rw technology?   string
|     |        |     |  +--rw resources
|     |        |     |     +--rw delay?       string
|     |        |     |     +--rw bandwidth?   string
|     |        |     |     +--rw cost?        string
|     |        |     +--rw control
|     |        |     |  +--rw controller?     string
|     |        |     |  +--rw orchestrator?   string
|     |        |     +--rw addresses
|     |        |     |  +--rw l2?   string
|     |        |     |  +--rw l3* [id]
|     |        |     |  |  +--rw id           string
|     |        |     |  |  +--rw name?        string
|     |        |     |  |  +--rw configure?   string
|     |        |     |  |  +--rw client?      string
|     |        |     |  |  +--rw requested?   string
|     |        |     |  |  +--rw provided?    string
|     |        |     |  +--rw l4?   string
|     |        |     +--rw metadata* [key]
|     |        |        +--rw key      string
|     |        |        +--rw value?   string
|     |        +--rw links
|     |        |  +--rw link* [id]
|     |        |     +--rw id           string
|     |        |     +--rw name?        string
|     |        |     +--rw src?         ->
|     |        |     +--rw dst?         ->
|     |        |     +--rw resources
|     |        |        +--rw delay?       string
|     |        |        +--rw bandwidth?   string
|     |        |        +--rw cost?        string
|     |        +--rw resources
|     |        |  +--rw cpu        string
|     |        |  +--rw mem        string
|     |        |  +--rw storage    string
|     |        |  +--rw cost?      string
|     |        +--rw metadata* [key]
|     |           +--rw key      string
|     |           +--rw value?   string
```

```
|     +--rw flowtable
|        +--rw flowentry* [id]
|           +--rw id           string
|           +--rw name?        string
|           +--rw priority?    string
|           +--rw port         ->
|           +--rw match        string
|           +--rw action       string
|           +--rw out?         ->
|           +--rw resources
|              +--rw delay?       string
|              +--rw bandwidth?   string
|              +--rw cost?        string
+--rw links
|  +--rw link* [id]
|     +--rw id           string
|     +--rw name?        string
|     +--rw src?         ->
|     +--rw dst?         ->
|     +--rw resources
|        +--rw delay?       string
|        +--rw bandwidth?   string
|        +--rw cost?        string
+--rw metadata* [key]
|  +--rw key      string
|  +--rw value?   string
+--rw version?    string
```

Figure 8: Virtualizer's YANG data model: tree view

4.1.1.2. YANG Module

```
<CODE BEGINS> file "virtualizer.yang"
module virtualizer {
  namespace "urn:unify:virtualizer";
  prefix "virtualizer";
  organization "ETH";
  contact "Robert Szabo <robert.szabo@ericsson.com>";

  revision "2016-02-24" {
    description "V5.0: Common port configuration were added to the yang model
from the metadata fields";
  }

  revision "2016-02-19" {
    description "Added port/control (for Cf-Or interface); port/resources;
link-resources/cost and sofware-resource/cost for administrative metric;
clarifications for port/capability";
```

```
  }

  revision "2016-01-28" {
```

```
    description "Metadata added to infra_node and virtualizer level;
Virtualizer's revised data model based on virtualizer3; changes: link key is
set to id";
  }

  //======================== REUSABLE GROUPS ================================

  grouping id-name {
    leaf id { type string; }
    leaf name { type string;}
  }

  grouping id-name-type {
    uses id-name;
    leaf type {
      type string;
      // for infrastructue view: mandatory true; --> refined in infrastrucutre
view
      mandatory false;
    }
  }

  grouping metadata {
    list metadata {
      min-elements 0;
      key key;
      leaf key{
        type string;
        mandatory true;
      }
      leaf value{
        type string;
        mandatory false;
      }
    }
  }

  grouping link-resource {
    leaf delay {
      type string;
      mandatory false;
    }
    leaf bandwidth {
      type string;
      mandatory false;
    }
    leaf cost {
      description "Administrative metric.";
```

```
      type string;
      mandatory false;
    }
```

```
  }

  grouping l3-address {
    uses id-name;
    leaf configure {
      description "True: this is a configuration request; False: this is fyi";
      type string;
    }
    leaf client {
      description "Configuration service support at the client: {'dhcp-client',
'pre-configured'}; if not present it is left to the infrastructure to deal with
it.";
      type string;
    }
    leaf requested {
      description "To request port configuration, options: {'public', 'ip/
mask'}, where public means the request of public IP address and private ip/mask
a given address/mask configuration";
      type string;
    }
    leaf provided {
      description "The provided L3 configuration in response to the requested
field.";
      type string;
    }
  }
  // ------------ PORTS -------
  grouping port {
    uses id-name;
    leaf port_type {
      description "{port-abstract, port-sap} port-sap is to represent UNIFY
domain boundary; port-abstract is to represent UNIFY native port. Technology
specific attributes of a SAP is in the metadata.";
      type string;
    }
    leaf capability {
      description "To describe match and action capabilities associated with
the port, e.g., match=port,tag,ip,tcp,udp,mpls,of1.0, where port: based
forwarding; tag: unify abstract tagging; ip: ip address matching etc.";
      type string;
    }
    leaf sap {
      type string;
    }
    container sap_data {
      leaf technology {
        description "e.g., ('IEEE802.1q': '0x00c', 'MPLS': 70, 'IEEE802.1q')";
        type string;
```

```
      }
      container resources{
        description "Only used for domain boundary ports (port-sap type), where
this is used to derive interconnection link characteristics.";
        uses link-resource;
      }
    }
    container control {
      description "Used to connect this port to a UNIFY orchestrator's Cf-Or
reference point. Support controller - orchestrator or orchestrator - controller
connection establishment.";
      leaf controller{
```

```
        description "URI of the local controller service at this NF, e.g.,
http://*:8080/cf-or/";
        type string;
      }
      leaf orchestrator{
        description "URI of the scoped orchestration service offered to this NF
specifically, e.g., http://192.168.1.100:8080/cf-or/";
        type string;
      }
    }
    container addresses {
      leaf l2 {
        description "Requested or provided";
        type string;
      }
      list l3 {
        key "id";
        uses l3-address;
      }
      leaf l4 {
        description "e.g., request: {tcp/22, tcp/8080}; response {tcp/22:
(192.168.1.100, 1001)";
        type string;
      }
    }
    uses metadata;
  }

  // ------------ FLOW CONTROLS -------

  grouping flowentry {
    description "The flowentry syntax will follow ovs-ofctrl string format. The
UNIFY general tagging mechanism will be use like 'mpls'-> 'tag', i.e.,
push_tag:tag; pop_tag:tag...";
    uses id-name;
    leaf priority {
      type string;
    }
    leaf port {
      type leafref {
        path "";
      }
      mandatory true;
    }
    leaf match {
      description "The match syntax will follow ovs-ofctrl string format with
'mpls'->'tag', e.g.,: in_port=port, dl_tag=A, where port is the leafref above";
      type string;
```

```
      mandatory true;
    }
    leaf action {
      description "The action syntax will follow ovs-ofctrl string format with
'mpls'->'tag', e.g.,: push_tag:A, set_tag_label:A, output:out, where out is the
leafref below";
      type string;
      mandatory true;
```

```
    }
    leaf out {
      type leafref {
        path "";
      }
    }
    container resources{
      uses link-resource;
    }

  }

  grouping flowtable {
    container flowtable {
      list flowentry {
        key "id";
        uses flowentry;
      }
    }
  }

  // ------------ LINKS  -------


  grouping link {
    uses id-name;
    leaf src {
      type leafref {
        path "";
      }
    }
    leaf dst {
      type leafref {
        path "";
      }
    }
    container resources{
      uses link-resource;
    }
  }

  grouping links {
    container links {
      list link {
        key "id";
        uses link;
      }
    }
```

```
  }

  // ---------- NODE -------------------

  grouping software-resource {
    leaf cpu {
      type string;
      mandatory true;
    }
    leaf mem {
      type string;
      mandatory true;
    }
    leaf storage {
      type string;
      mandatory true;
    }
    leaf cost {
      description "Administrative metric.";
      type string;
      mandatory false;
    }
  }

  grouping node {
    description "Any node: infrastructure or NFs";
    uses id-name-type;
    container ports {
      list port{
        key "id";
        uses port;
      }
    }
    uses links;
    container resources{
      uses software-resource;
    }
    uses metadata;
  }

  grouping nodes {
    list node{
      key "id";
      uses node;
    }
  }

  grouping infra-node { // they can contain other nodes (as NFs)
```

```
    uses node {
      refine type {
        mandatory true;
      }
    }
    container NF_instances {
      uses nodes;
    }
    container capabilities {
      container supported_NFs { // if supported NFs are enumerated
        uses nodes;
      }
    }
    uses flowtable;
  }



  //======================== NF-FG: Virtualizer and the Mapped request
================================

  container virtualizer {
    description "Container for a single virtualizer";
    uses id-name {
      refine id {
        mandatory true;
      }
    }
    container nodes{
      list node{ // infra nodes
        key "id";
        uses infra-node;
      }
    }
    uses links; // infra links
    uses metadata;
    leaf version {
      description "yang and virtualizer library version";
      type string;
    }
  }
}
<CODE ENDS>
```

Figure 9: Virtualizer's YANG data model

5. Relation to ETSI NFV

According to the ETSI MANO framework [ETSI-NFV-MANO], an NFVO is split into two functions:

- The orchestration of NFVI resources across multiple VIMs, fulfilling the Resource Orchestration functions. The NFVO uses the Resource Orchestration functionality to provide services that support accessing NFVI resources in an abstracted manner independently of any VIMs, as well as governance of VNF instances sharing resources of the NFVI infrastructure

- The lifecycle management of Network Services, fulfilling the network Service Orchestration functions.

Similarly, a VIM is split into two functions:

- Orchestrating the allocation/upgrade/release/reclamation of NFVI resources (including the optimization of such resources usage), and

- managing the association of the virtualised resources to the physical compute, storage, networking resources.

The functional split is shown in Figure 14.

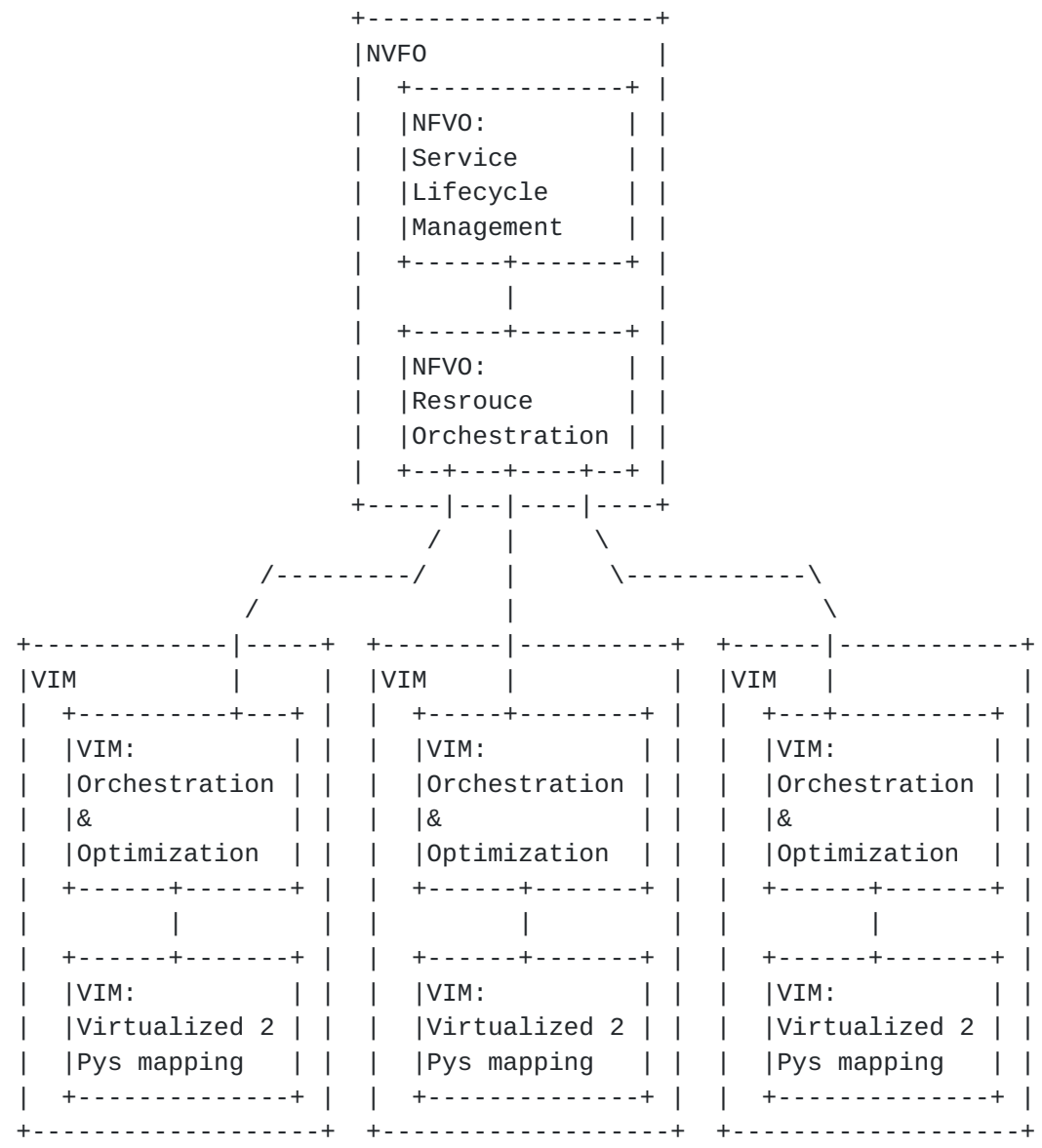

Figure 10: Functional decomposition of the NFVO and the VIM according to the ETSI MANO

If the Joint Software and Network Control API (Joint API) could be used between all the functional components working on the same abstraction, i.e., from the north of the VIM Virtualized to physical mapping component to the south of the NFVO: Service Lifecycle Management as shown in Figure 11, then a more flexible virtualization programming architecture could be created as shown in Figure 12.

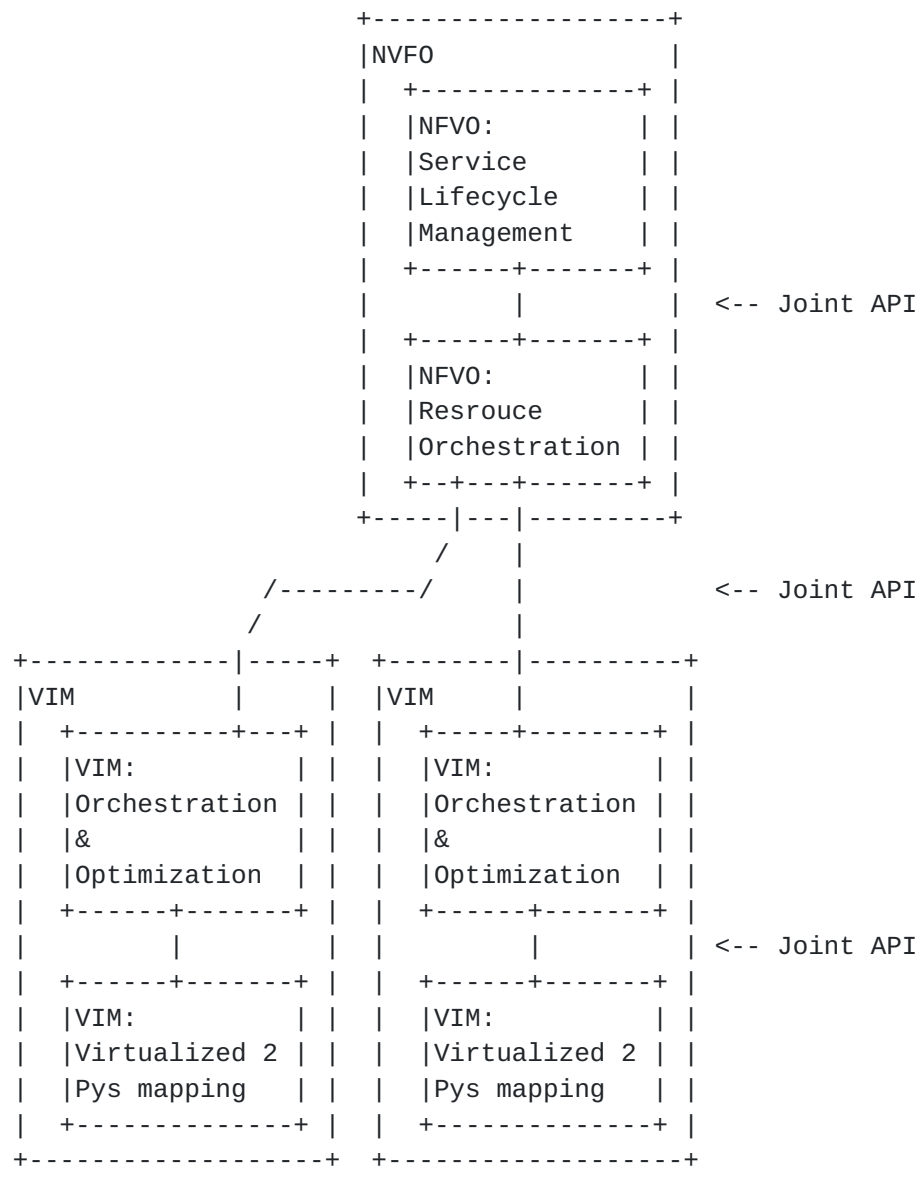

Figure 11: Functional decomposition of the NFVO and the VIM with the Joint Software and Network control API

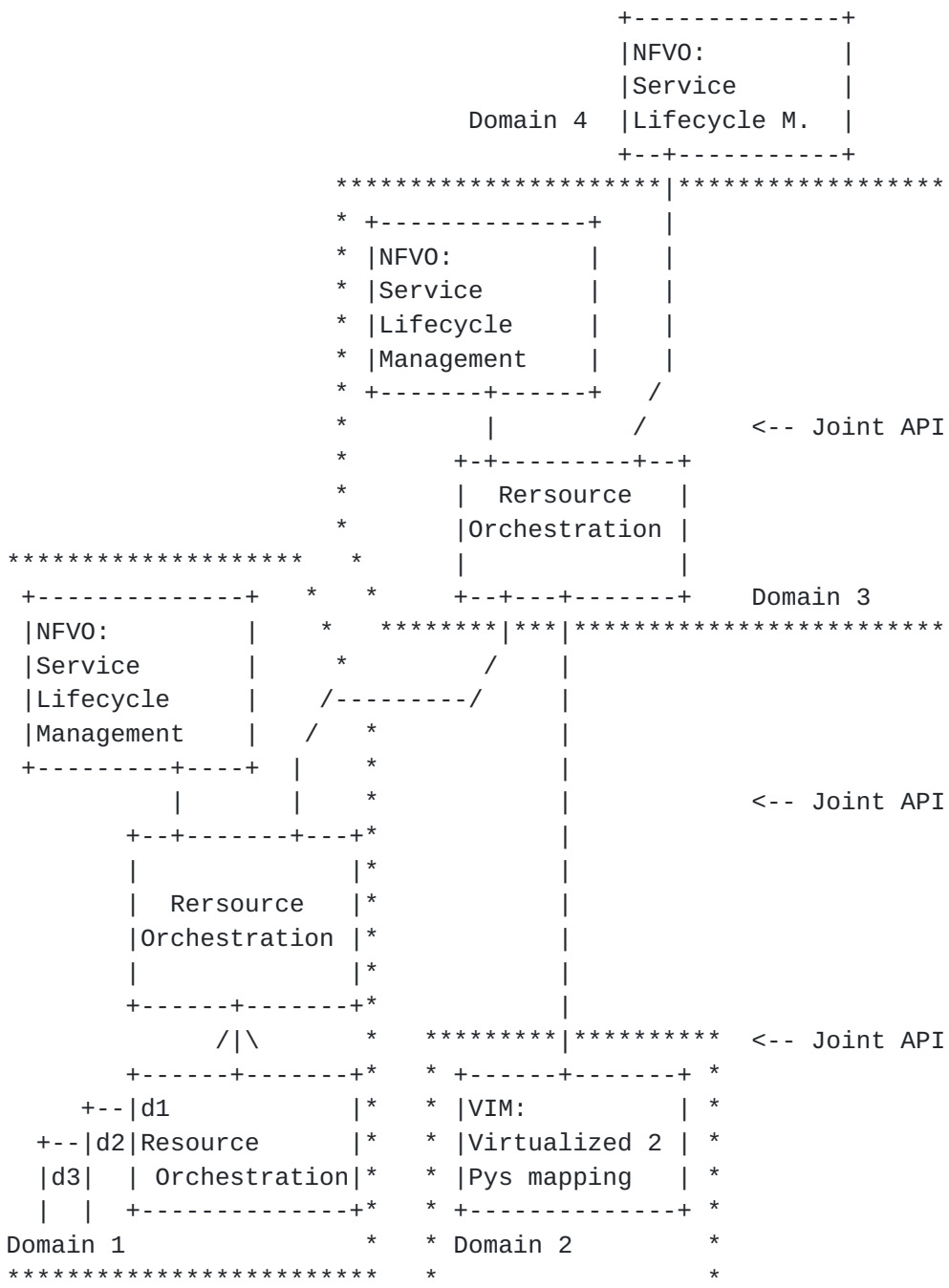

Figure 12: Joint Software and Network Control API: Recurring Flexible Architecture

5.1. Policy based resource management

In Figure 13 we show various policies mapped to the MANO architecture:

- Tenant Policies: Tenant policies exist whenever a domain offers a virtualization service to more than one consumer. User tenants may exists at the northbound of the NFVO. Additionally, if a VIM expose resource services to more than one NFVO, then each NFVO may appear as a tenant (virtualization consumer) at the northbound of the VIM.

- Wherever virtualization services are produced or consumed corresponding export and import policies may exist. Export policies govern the details of resources, capabilities, costs, etc. exposed to consumers. In turn, consumers (tenants) apply import policies to filter, tweak, annotate resources and services received from their southbound domains. An entity may at the same time consume and produce virtualization services hence apply both import and export policies.

- Operational policies support the business logic realized by the domain's ownership. They are often associated with Operations or Business Support Systems (OSS or BSS) and frequently determine operational objectives like energy optimization, utilization targets, offered services, charing models, etc. Operational policies may be split according to different control plane layers, for example, i) lifecycle and ii) resource management layers within the NFVO.

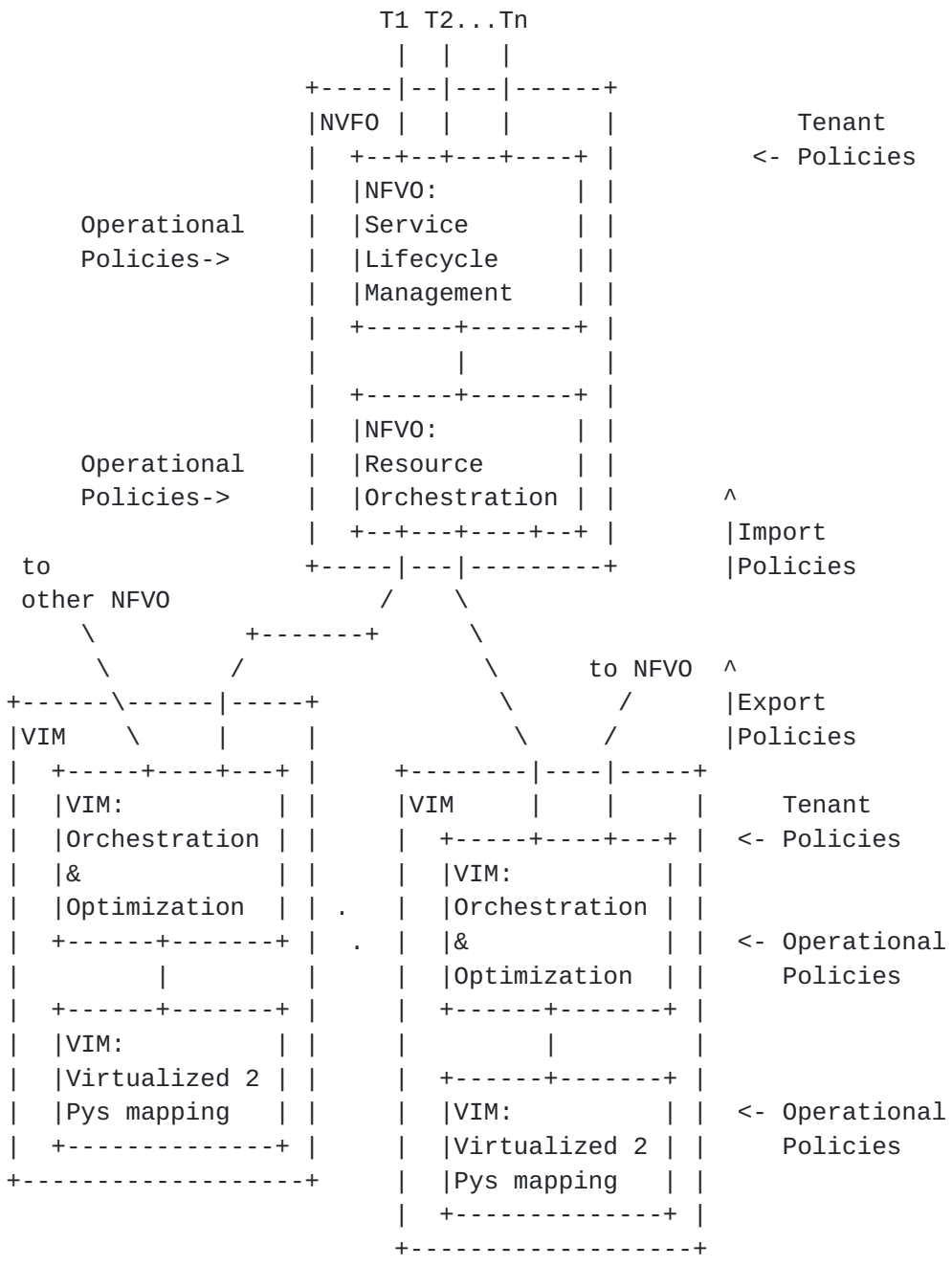

Figure 13: Policies within the MANO framework

6. Examples

6.1. Infrastructure reports

Figure 14 and Figure 15 show a single node infrastructure report. The example shows a BiS-BiS with two ports, out of which Port 0 is also a Service Access Point 0 (SAP0).

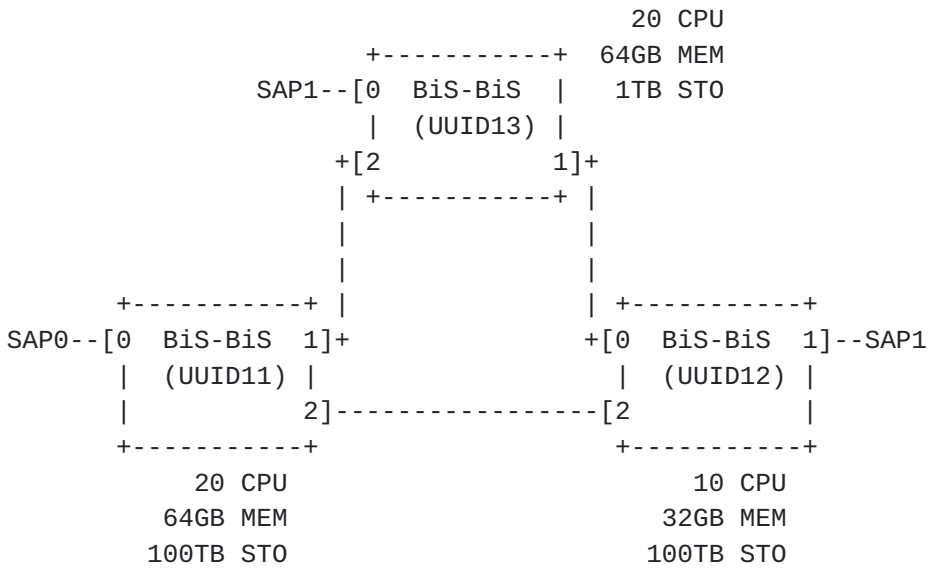

Figure 14: Single node infrastructure report example: Virtualization view

```
<virtualizer xmlns="http://fp7-unify.eu/framework/virtualizer">
    <id>UUID001</id>
    <name>Single node simple infrastructure report</name>
    <nodes>
        <node>
            <id>UUID11</id>
            <name>single Bis-Bis node</name>
            <type>BisBis</type>
            <ports>
                <port>
                    <id>0</id>
                    <name>SAP0 port</name>
                    <port_type>port-sap</port_type>
                    <vxlan>...</vxlan>
                </port>
                <port>
                    <id>1</id>
                    <name>North port</name>
                    <port_type>port-abstract</port_type>
                    <capability>...</capability>
                </port>
                <port>
                    <id>2</id>
                    <name>East port</name>
                    <port_type>port-abstract</port_type>
                    <capability>...</capability>
                </port>
            </ports>
            <resources>
                <cpu>20</cpu>
                <mem>64 GB</mem>
                <storage>100 TB</storage>
            </resources>
        </node>
    </nodes>
</virtualizer>
```

Figure 15: Single node infrastructure report example: xml view

Figure 16 and Figure 17 show a 3-node infrastructure report with 3 BiS-BiS nodes. Infrastructure links are inserted into the virtualization view between the ports of the BiS-BiS nodes.

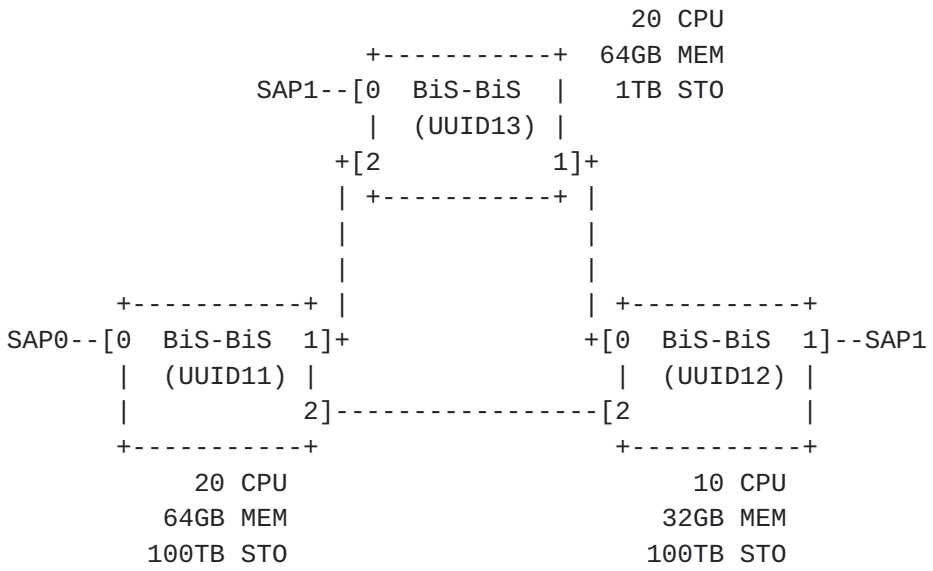

Figure 16: 3-node infrastructure report example: Virtualization view

```
 <virtualizer xmlns="http://fp7-unify.eu/framework/virtualizer">
     <id>UUID002</id>
     <name>3-node simple infrastructure report</name>
     <nodes>
         <node>
             <id>UUID11</id>
             <name>West Bis-Bis node</name>
             <type>BisBis</type>
             <ports>
                 <port>
                     <id>0</id>
                     <name>SAP0 port</name>
                     <port_type>port-sap</port_type>
                     <vxlan>...</vxlan>
                 </port>
                 <port>
                     <id>1</id>
                     <name>North port</name>
                     <port_type>port-abstract</port_type>
                     <capability>...</capability>
                 </port>
                 <port>
                     <id>2</id>
                     <name>East port</name>
                     <port_type>port-abstract</port_type>
                     <capability>...</capability>
                 </port>
             </ports>
             <resources>
```

```
                    <cpu>20</cpu>
                    <mem>64 GB</mem>
                    <storage>100 TB</storage>
                </resources>
            </node>
            <node>
                <id>UUID12</id>
                <name>East Bis-Bis node</name>
                <type>BisBis</type>
                <ports>
                    <port>
                        <id>1</id>
                        <name>SAP1 port</name>
                        <port_type>port-sap</port_type>
                        <vxlan>...</vxlan>
                    </port>
                    <port>
                        <id>0</id>
                        <name>North port</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                    <port>
                        <id>2</id>
                        <name>West port</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                </ports>
                <resources>
                    <cpu>10</cpu>
                    <mem>32 GB</mem>
                    <storage>100 TB</storage>
                </resources>
            </node>
            <node>
                <id>UUID13</id>
                <name>North Bis-Bis node</name>
                <type>BisBis</type>
                <ports>
                    <port>
                        <id>0</id>
                        <name>SAP2 port</name>
                        <port_type>port-sap</port_type>
                        <vxlan>...</vxlan>
                    </port>
                    <port>
                        <id>1</id>
```

```
                    <name>East port</name>
                    <port_type>port-abstract</port_type>
                    <capability>...</capability>
                </port>
                <port>
                    <id>2</id>
                    <name>West port</name>
                    <port_type>port-abstract</port_type>
                    <capability>...</capability>
                </port>
            </ports>
            <resources>
                <cpu>20</cpu>
                <mem>64 GB</mem>
                <storage>1 TB</storage>
            </resources>
        </node>
    </nodes>
    <links>
        <link>
            <id>0</id>
            <name>Horizontal link</name>
            <src>../../nodes/node[id=UUID11]/ports/port[id=2]</src>
            <dst>../../nodes/node[id=UUID12]/ports/port[id=2]</dst>
            <resources>
                <delay>2 ms</delay>
                <bandwidth>10 Gb</bandwidth>
            </resources>
        </link>
        <link>
            <id>1</id>
            <name>West link</name>
            <src>../../nodes/node[id=UUID11]/ports/port[id=1]</src>
            <dst>../../nodes/node[id=UUID13]/ports/port[id=2]</dst>
            <resources>
                <delay>5 ms</delay>
                <bandwidth>10 Gb</bandwidth>
            </resources>
        </link>
        <link>
            <id>2</id>
            <name>East link</name>
            <src>../../nodes/node[id=UUID12]/ports/port[id=0]</src>
            <dst>../../nodes/node[id=UUID13]/ports/port[id=1]</dst>
            <resources>
                <delay>2 ms</delay>
                <bandwidth>5 Gb</bandwidth>
            </resources>
```

```
          </link>
        </links>
    </virtualizer>
```

Figure 17: 3-node infrastructure report example: xml view

6.2. Simple requests

Figure 18 and Figure 19 show the allocation request for 3 NFs (NF1: Parental control B.4, NF2: Http Cache 1.2 and NF3: Stateful firewall C) as instrumented over a BiS-BiS node. It can be seen that the configuration request contains both the NF placement and the forwarding overlay definition as a joint request.

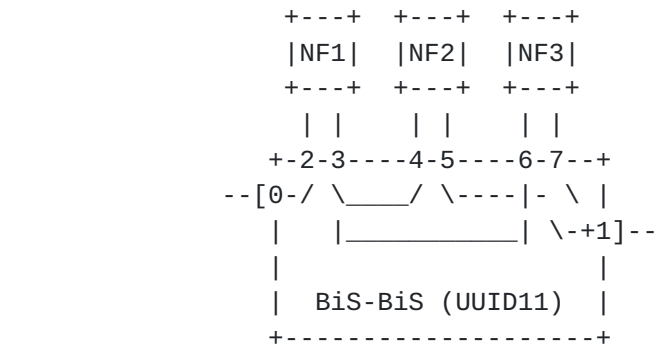

Figure 18: Simple request of 3 NFs on a single BiS-BiS: Virtualization view

```
<virtualizer xmlns="http://fp7-unify.eu/framework/virtualizer">
    <id>UUID001</id>
    <name>Single node simple request</name>
    <nodes>
        <node>
            <id>UUID11</id>
            <NF_instances>
                <node>
                    <id>NF1</id>
                    <name>first NF</name>
                    <type>Parental control B.4</type>
                    <ports>
                        <port>
                            <id>2</id>
                            <name>in</name>
                            <port_type>port-abstract</port_type>
                            <capability>...</capability>
                        </port>
                        <port>
                            <id>3</id>
```

```
                        <name>out</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                </ports>
            </node>
            <node>
                <id>NF2</id>
                <name>cache</name>
                <type>Http Cache 1.2</type>
                <ports>
                    <port>
                        <id>4</id>
                        <name>in</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                    <port>
                        <id>5</id>
                        <name>out</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                </ports>
            </node>
            <node>
                <id>NF3</id>
                <name>firewall</name>
                <type>Stateful firewall C</type>
                <ports>
                    <port>
                        <id>6</id>
                        <name>in</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                    <port>
                        <id>7</id>
                        <name>out</name>
                        <port_type>port-abstract</port_type>
                        <capability>...</capability>
                    </port>
                </ports>
            </node>
        </NF_instances>
        <flowtable>
            <flowentry>
                <port>../../ports/port[id=0]</port>
```

```
                        <match>*</match>
                        <action>output:../../NF_instances/node[id=NF1]
                          /ports/port[id=2]</action>
                    </flowentry>
                    <flowentry>
                        <port>../../NF_instances/node[id=NF1]
                          /ports/port[id=3]</port>
                        <match>fr-a</match>
                        <action>output:../../NF_instances/node[id=NF2]
                          /ports/port[id=4]</action>
    rpcre                </flowentry>
                    <flowentry>
                        <port>../../NF_instances/node[id=NF1]
                          /ports/port[id=3]</port>
                        <match>fr-b</match>
                        <action>output:../../NF_instances/node[id=NF3]
                          /ports/port[id=6]</action>
                    </flowentry>
                    <flowentry>
                        <port>../../NF_instances/node[id=NF2]
                          /ports/port[id=5]</port>
                        <match>*</match>
                        <action>output:../../ports/port[id=1]</action>
                    </flowentry>
                    <flowentry>
                        <port>../../NF_instances/node[id=NF3]
                          /ports/port[id=7]</port>
                        <match>*</match>
                        <action>output:../../ports/port[id=1]</action>
                    </flowentry>
                </flowtable>
            </node>
        </nodes>
    </virtualizer>
```

Figure 19: Simple request of 3 NFs on a single BiS-BiS: xml view

7. Experimentations

We have implemented the proposed recursive control plane architecture with joint software and network virtualization and control. We used a Python based open source implementation [virtualizer-library] of the virtualizer data structure for the orchestration API. We used the Extensible Service ChAin Prototyping Environment (ESCAPE) [ESCAPE] as the general orchestration platform with various technology specific domain adapters like OpenStack, Docker and Ryu SDN controller. A detailed service function chaining report is available at [I-D.unify-sfc-control-plane-exp].

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

TBD

10. Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 619609 - the UNIFY project. The views expressed here are those of the authors only. The European Commission is not liable for any use that may be made of the information in this document.

We would like to thank in particular David Jocha and Janos Elek from Ericsson for the useful discussions.

11. Informative References

[ESCAPE] BME, "Extensible Service ChAin Prototyping Environment (open source)", Mar. 2016, <http://sb.tmit.bme.hu/mediawiki/index.php/ESCAPE>.

[ETSI-NFV-Arch] ETSI, "Architectural Framework v1.1.1", Oct 2013, <http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.01.01_60/gs_NFV002v010101p.pdf>.

[ETSI-NFV-MANO] ETSI, "Network Function Virtualization (NFV) Management and Orchestration V0.6.1 (draft)", Jul. 2014, <http://docbox.etsi.org/ISG/NFV/Open/Latest_Drafts/NFV-MAN001v061-%20management%20and%20orchestration.pdf>.

[I-D.caszpe-nfvrg-orchestration-challenges] Carrozzo, G., Szabo, R., and K. Pentikousis, "Network Function Virtualization: Resource Orchestration Challenges", draft-caszpe-nfvrg-orchestration-challenges-00 (work in progress), November 2015.

[I-D.ietf-sfc-dc-use-cases] Surendra, S., Tufail, M., Majee, S., Captari, C., and S. Homma, "Service Function Chaining Use Cases In Data Centers", draft-ietf-sfc-dc-use-cases-04 (work in progress), January 2016.

[I-D.unify-nfvrg-challenges]
Szabo, R., Csaszar, A., Pentikousis, K., Kind, M., Daino, D., Qiang, Z., and H. Woesner, "Unifying Carrier and Cloud Networks: Problem Statement and Challenges", draft-unify-nfvrg-challenges-03 (work in progress), January 2016.

[I-D.unify-sfc-control-plane-exp]
Szabo, R. and B. Sonkoly, "SFC Control Plane Experiment: UNIFYed Approach", March 2016, <draft-unify-sfc-control-plane-exp>.

[I-D.zu-nfvrg-elasticity-vnf]
Qiang, Z. and R. Szabo, "Elasticity VNF", draft-zu-nfvrg-elasticity-vnf-01 (work in progress), March 2015.

[virtualizer-library]
Ericsson, "Python based virtualizer library for Netconf protocol (open source)", Mar. 2016, <https://github.com/Ericsson/unify-virtualizer>.

Authors' Addresses

Robert Szabo (editor)
Ericsson Research, Hungary
Irinyi Jozsef u. 4-20
Budapest 1117
Hungary

Email: robert.szabo@ericsson.com
URI: http://www.ericsson.com/

Zu Qiang
Ericsson
8400, boul. Decarie
Ville Mont-Royal, QC 8400
Canada

Email: zu.qiang@ericsson.com
URI: http://www.ericsson.com/

Mario Kind
Deutsche Telekom AG
Winterfeldtstr. 21
10781 Berlin
Germany

Email: mario.kind@telekom.de