Nový networking pro nové IT

Tomáš Kubica, Enterprise Architect + Daniel Prchal
tomas.kubica@hpe.com
Twitter: @tkubica

cloudsvet.cz **netsvet.cz**

Úspěšné budou firmy, které využijí třetí platformu: Social, Mobile, Big Data, Cloud

amazon

NETFLIX

Spotify

airbnb

KICKSTARTER

140M vs. 500M

Lean Enterprise, Inspirace výrobou

Agile
Fluid
DevOps

Waste
Value stream
Unplanned work


Infrastructure as code
PaaS
CI/CD
Kontejner
Dev
Ops
Tato zeď musí pryč !
Inovace (ne šetření) jsou klíčem k úspěchu

Zajistěte vašim aplikacím přenositelnost

DevOps

Jen kontejner nestačí

Dev

Buildpack
A/B testing
Services
Late binding

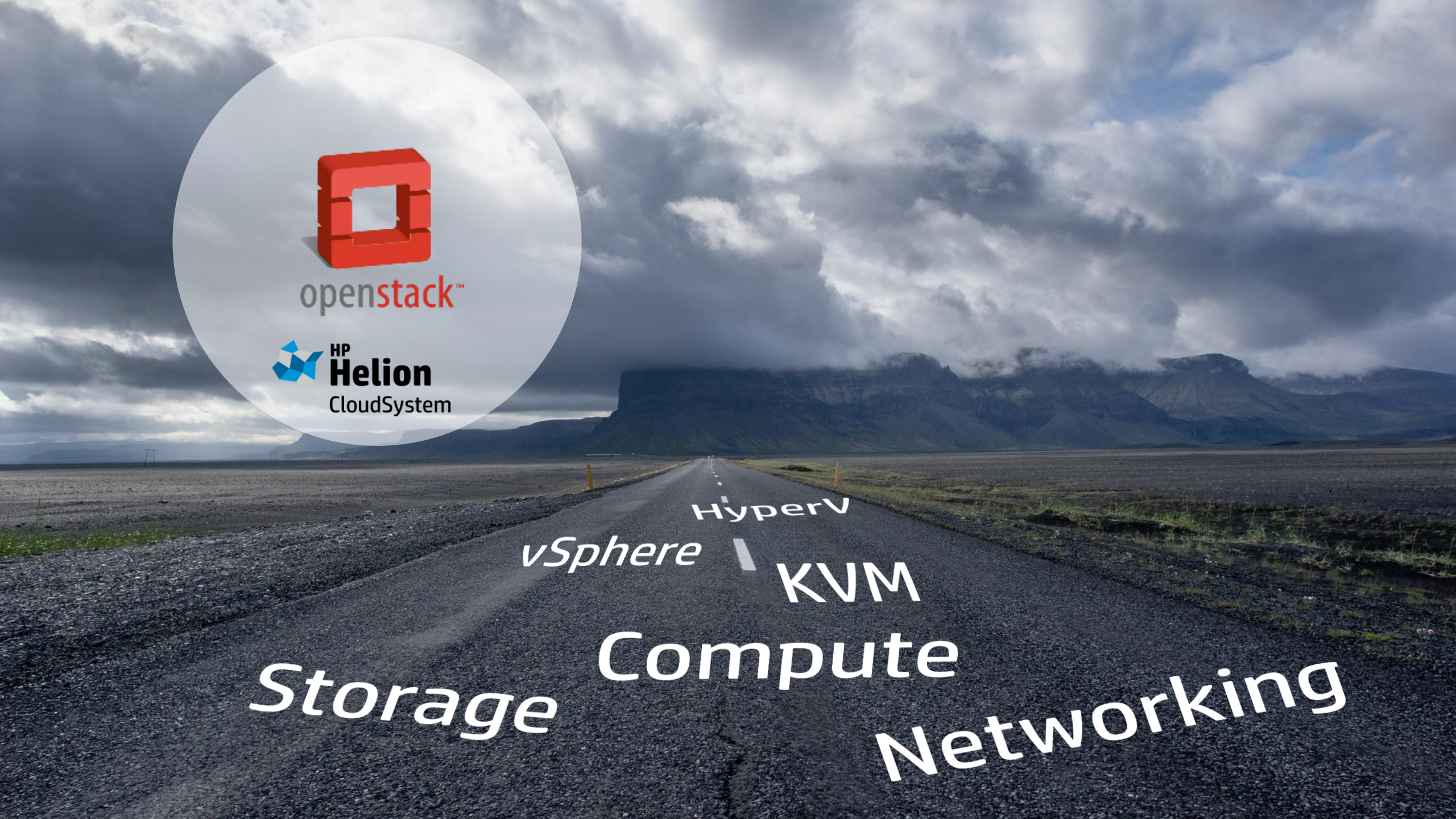
CLOUD FOUNDRY

docker

HP Helion Development Platform

Ops

Balancing
Autoscaling
HA
Rollback
Green/Blue
DBaaS
MSGaaS


openstack™
HP Helion CloudSystem
HyperV
vSphere
KVM
Compute
Storage
Networking

Zvolte vás přístup ke zrychlení IT

Změňte IT

Bimodal IT (dvourychlostní)

Innovation labs


Kam jedou sítě?
Nové sítě pro nové IT

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

Network Function Virtualization

Disagregace

Open source

DevOps

Mluvme možná o něčem nepříjemném

- Zapomeňte na SNMP a CLI
- Nesmíte se přihlásit do prvku přes SSH a udělat změnu
- Víc funkcí a protokolů? Naopak – většinu pryč
- Učte se CI/CD nástroje, zejména Ansible, Git, RESTful API, YAML
- Vaše infrastruktura = definice desired state, ne skript
- Vaše politika = intent-based networking, ne VLAN a ACL
- Vaše dokumentace = „kód“ nebo manifest
- Programovat OpenFlow by měli umět 3 lidi na planetě

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

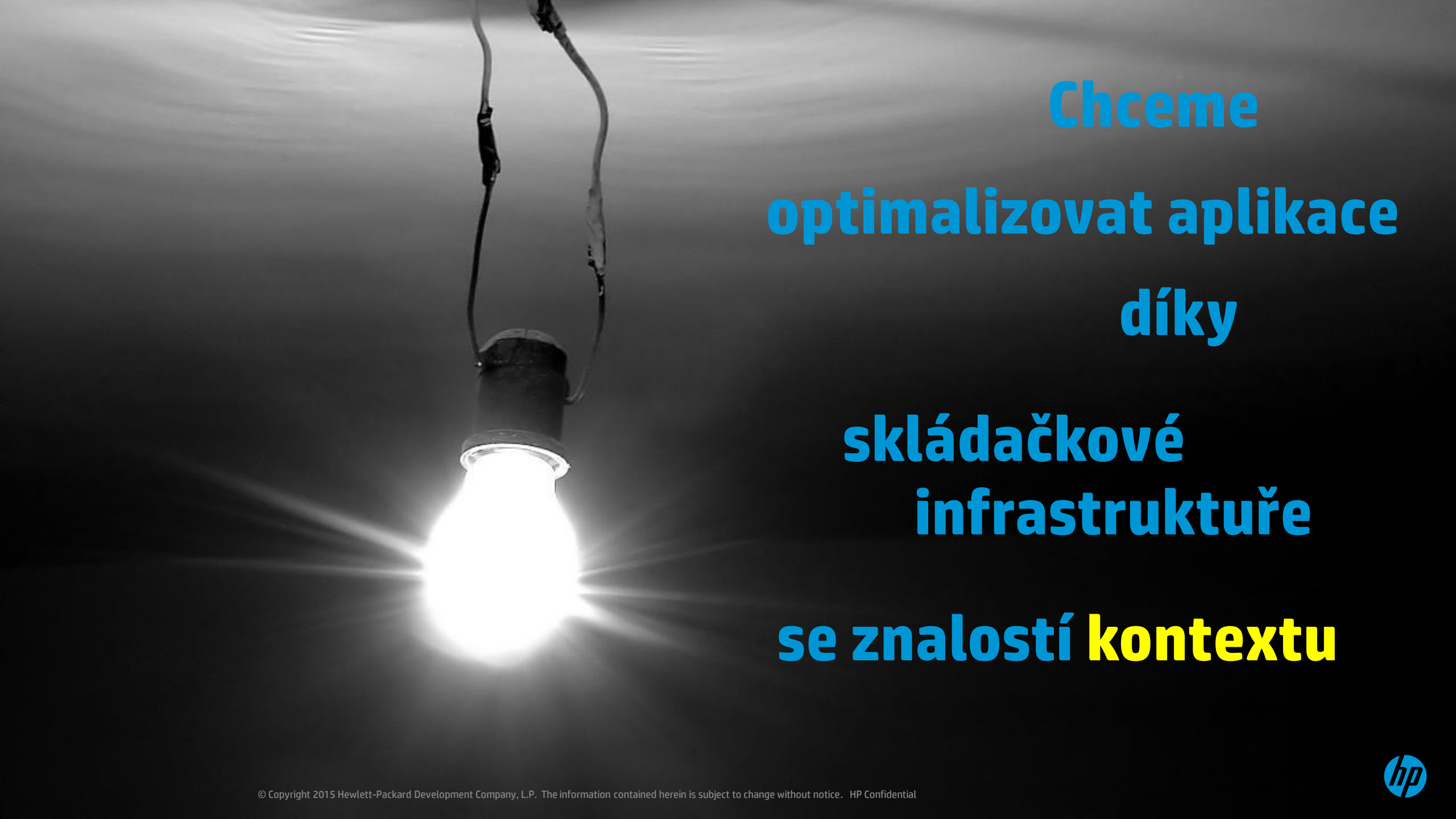
Network Function Virtualization

Disagregace

Open source

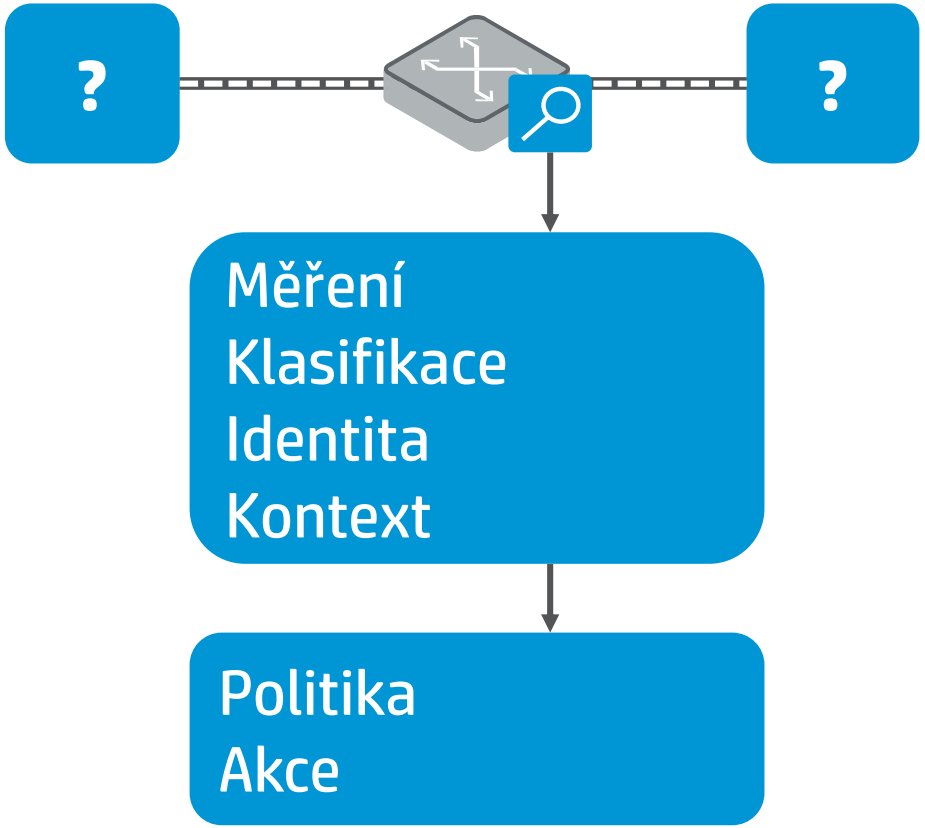
DevOps

Proč SDN?



Tradičně jsme odkázáni jen na to, co vidíme procházet sítí…

Odhadovat – Application ‘Aware’ Network



Drahá infrastruktura s důrazem na hardware

Opožděná integrace nových aplikací

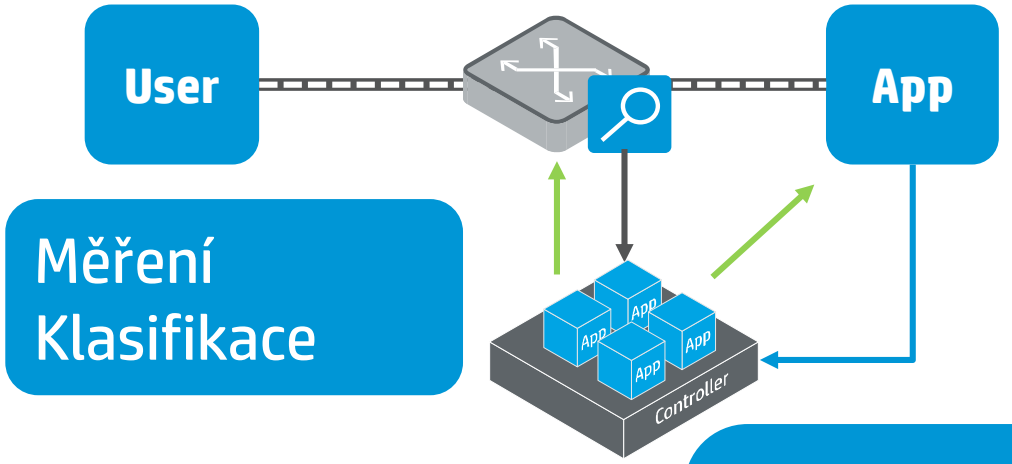
Je obtížné získat data z omezených a šifrovaných zdrojů

Vědět – Context-Driven Network

Otevřená programovatelná architektura

Otevřenost umožňuje samostatnou až živelnou integraci

Vhled do aplikací, událostí, kontextu i telemetrie



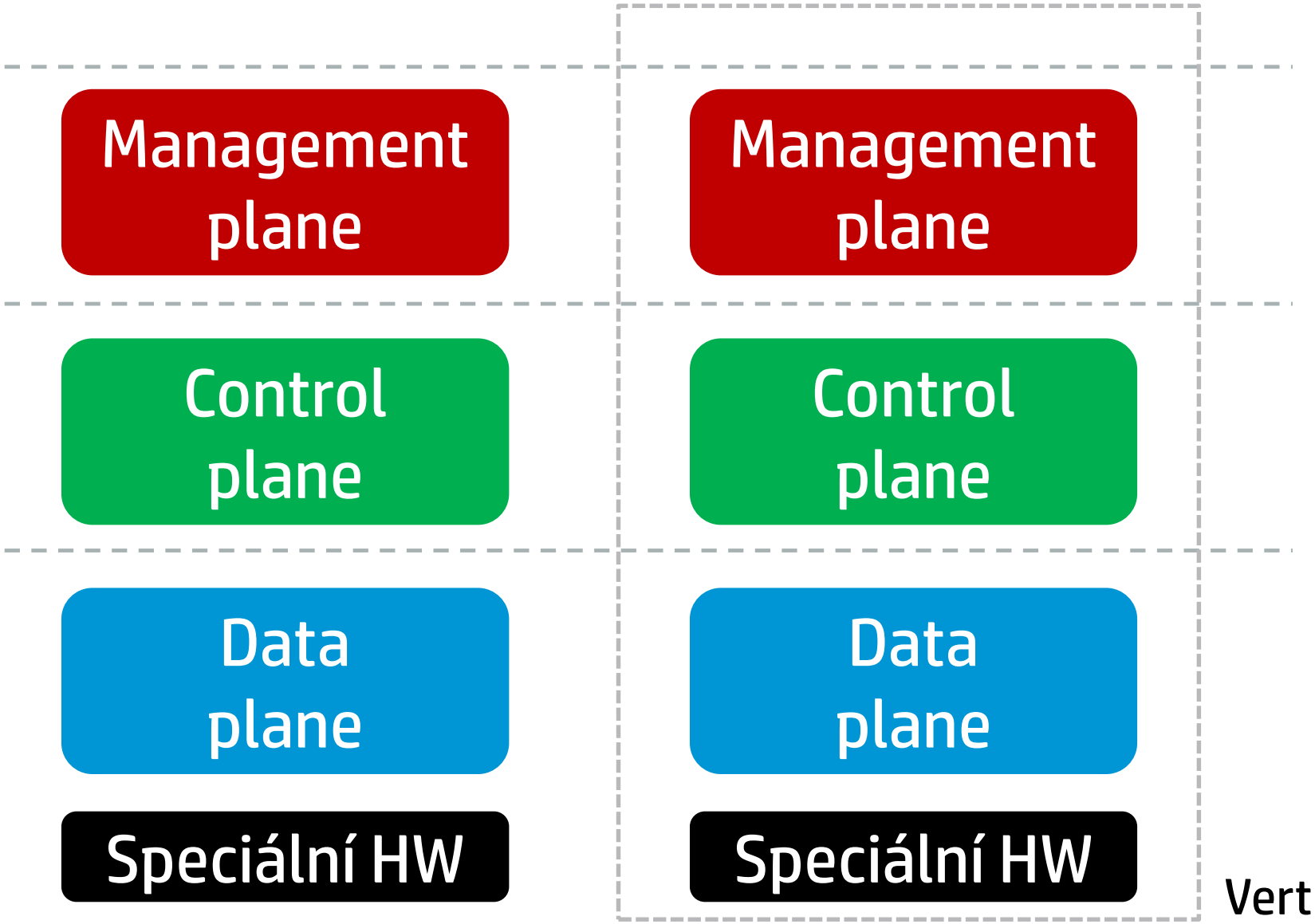
Měření
Klasifikace

Politika
Koordinovaná akce

Identita
Událost
Kontext
Požadavek


SDN
NABÍZÍ ZPŮSOB
JAK ZÍSKAT KONTEXT
A APLIKOVAT POLITIKU

Tradiční síťové prvky



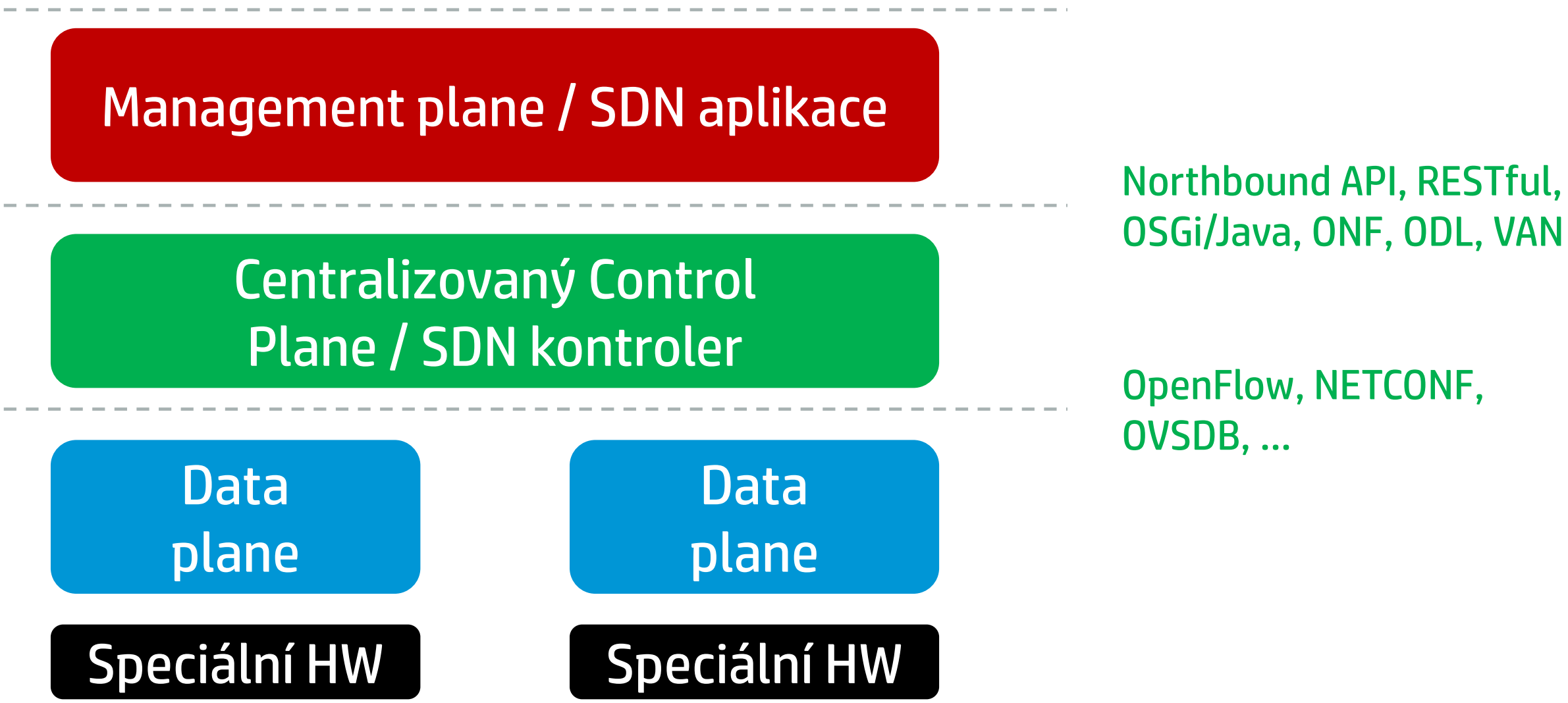
CLI, SNMP
Python, NETCONF, XMPP, Puppet, Chef

Interní, uzavřené

Interní, uzavřené

Vertikálně integrováno

Software-defined Networking



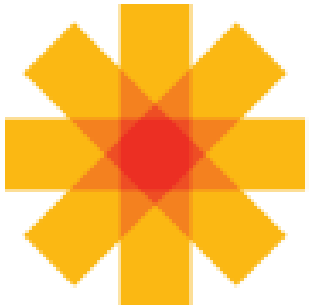
Kontrolery

Jaký si mám vybrat kontroler?

Ryu
Floodlight
POX
NOX
Beacon

Intent-driven SDN

OPEN NETWORKING
FOUNDATION



DAYLIGHT

Software-defined Networking:
Vyjadřujte záměry, dávejte kontext

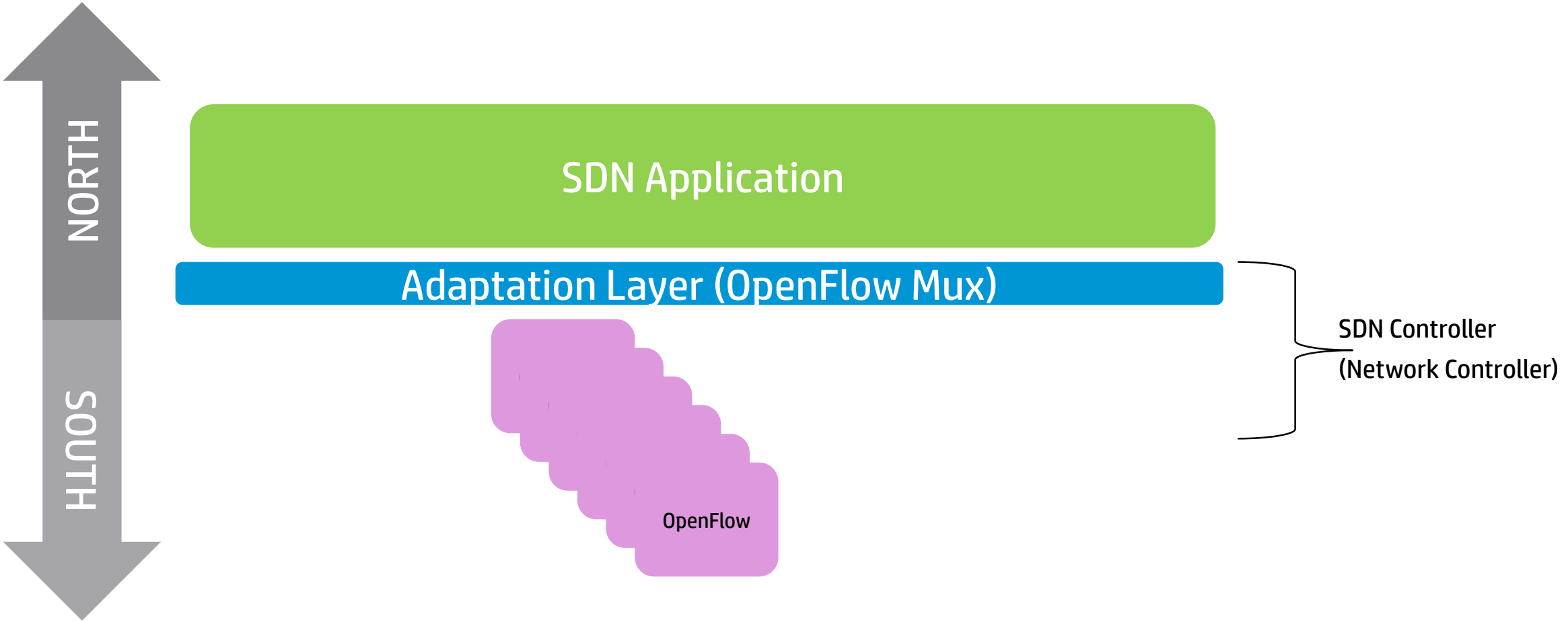
SDN App Store
Powered by HP

OPEN NETWORKING
FOUNDATION

OPEN
DAYLIGHT

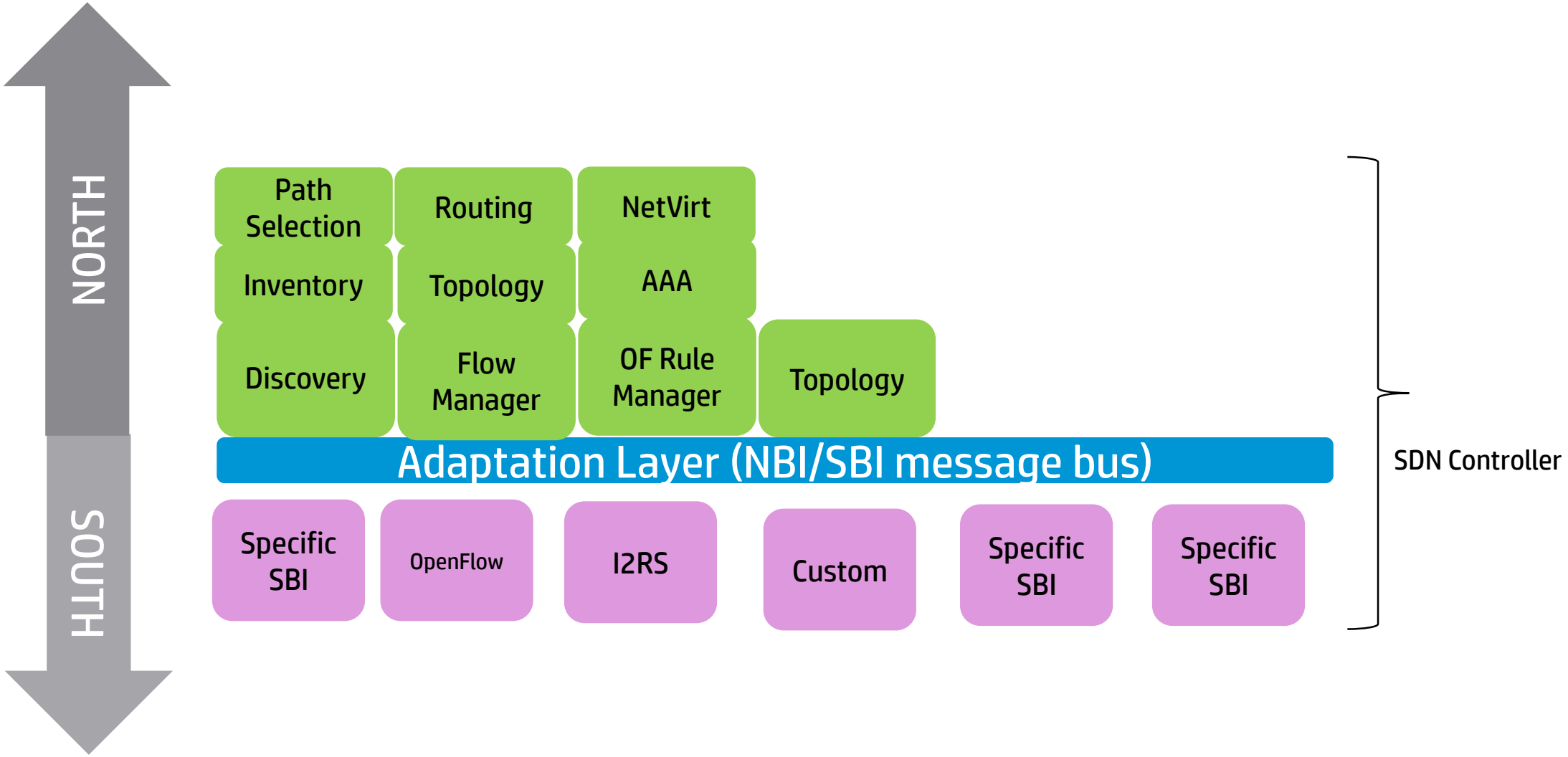
První pokus

OpenFlow == assembler, switch je flow tabulka



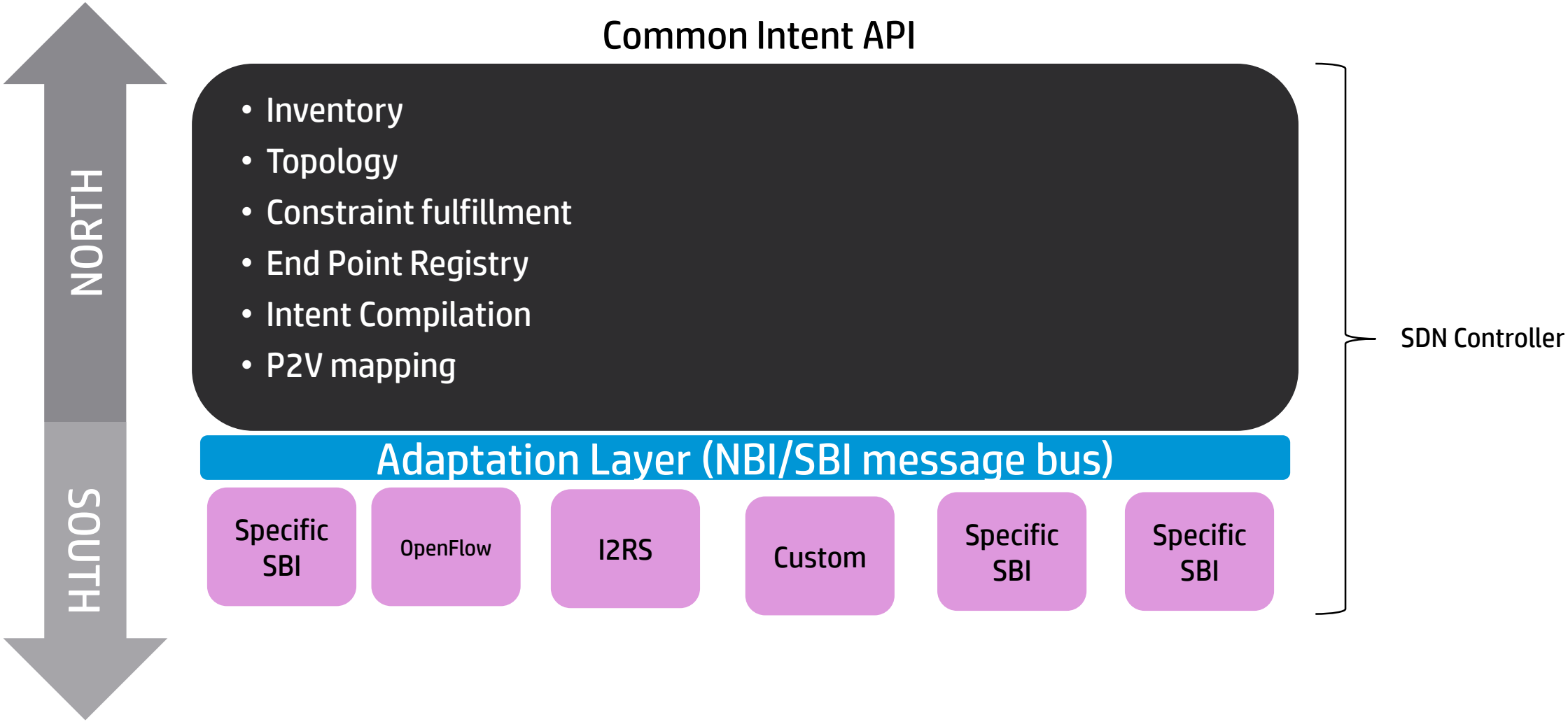
Druhý pokus - SDK

Vývojář může programovat síť, ale příliš pozdě na standardizáci



Doufejme poslední pokus – Intent API driven Black Box

Dovoluje milionům lidí ovládat sítě



Intent je

- Co, ne jak
- Univerzální jazyk
- Přenositelný
- Škálovatelný
- Přináší kontext
- Menší prostor pro útok
- Příklady:
  - Martin má v zasedčce přístup na Internet
  - Jana může na HR systémy v pracovní době
  - Tomáš, pokud komunikuje se R&D aplikací, musí přistupovat šifrovanou cestou

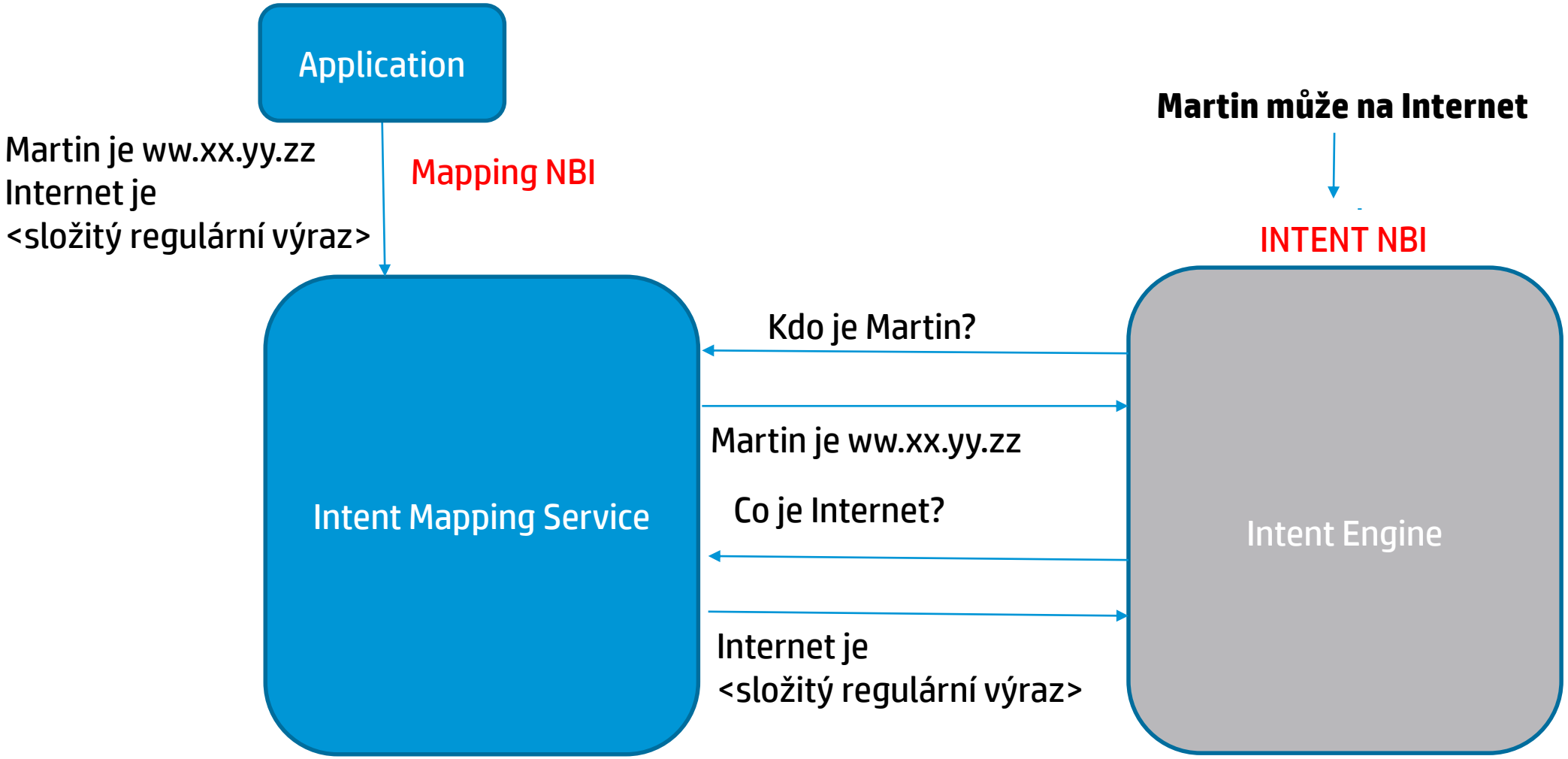
Mapování záměru (nepřenosné)

- Zařízení
- Protokoly
- Výrobci
- Rozhraní
- Adresy
- Lokace
- 5-tuples
- 12-typles

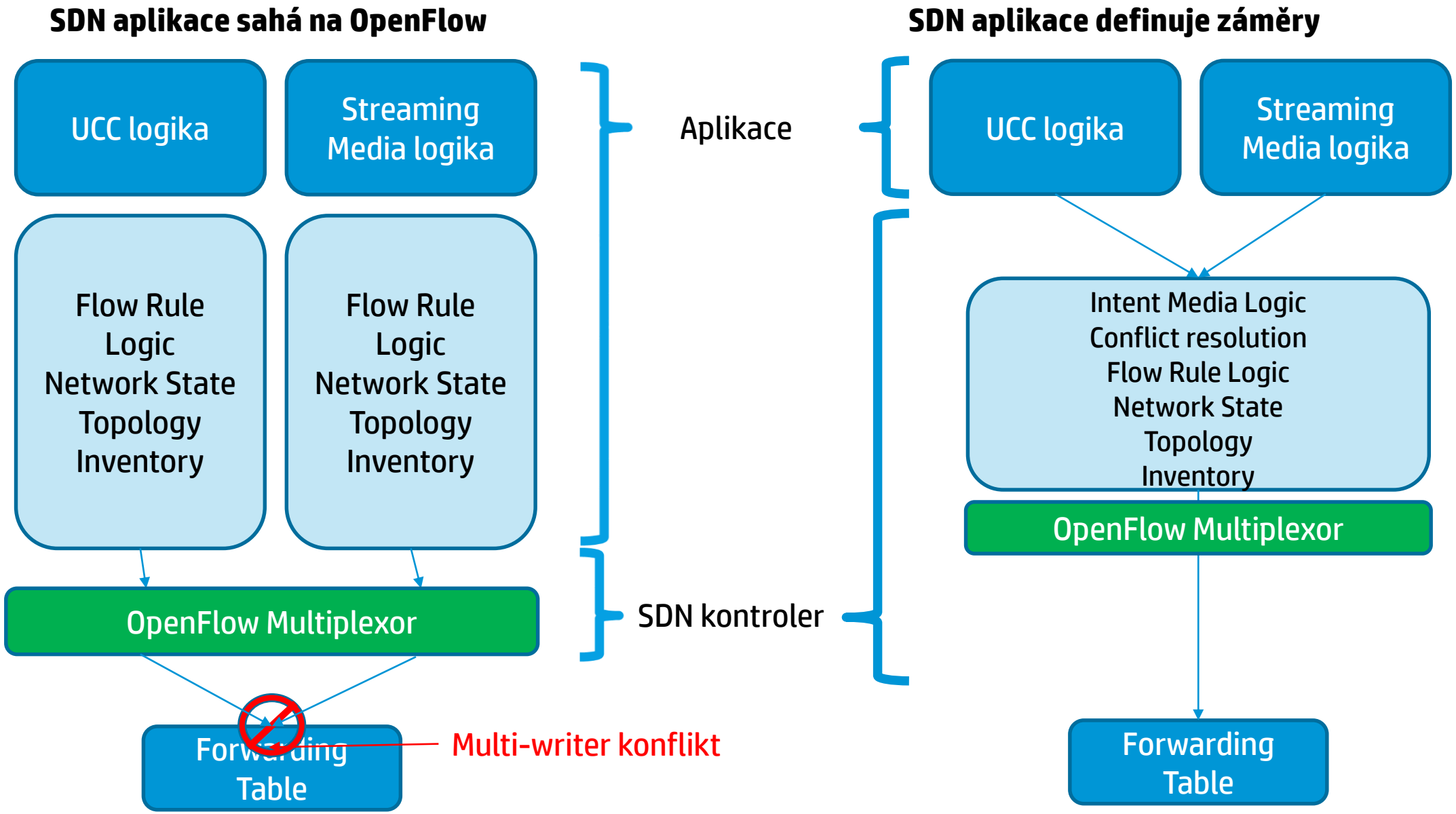
Záměr (přenosný)

- Vztahy mezi skupinami koncových bodů
- Nálepky/objekty (např. Internet, HR, Franta)
- Záměrová slovesa (např. povol, zakaž, přesměruj)
- Omezení a vlastnosti (QoS, izolace,...)

Martin chce internet

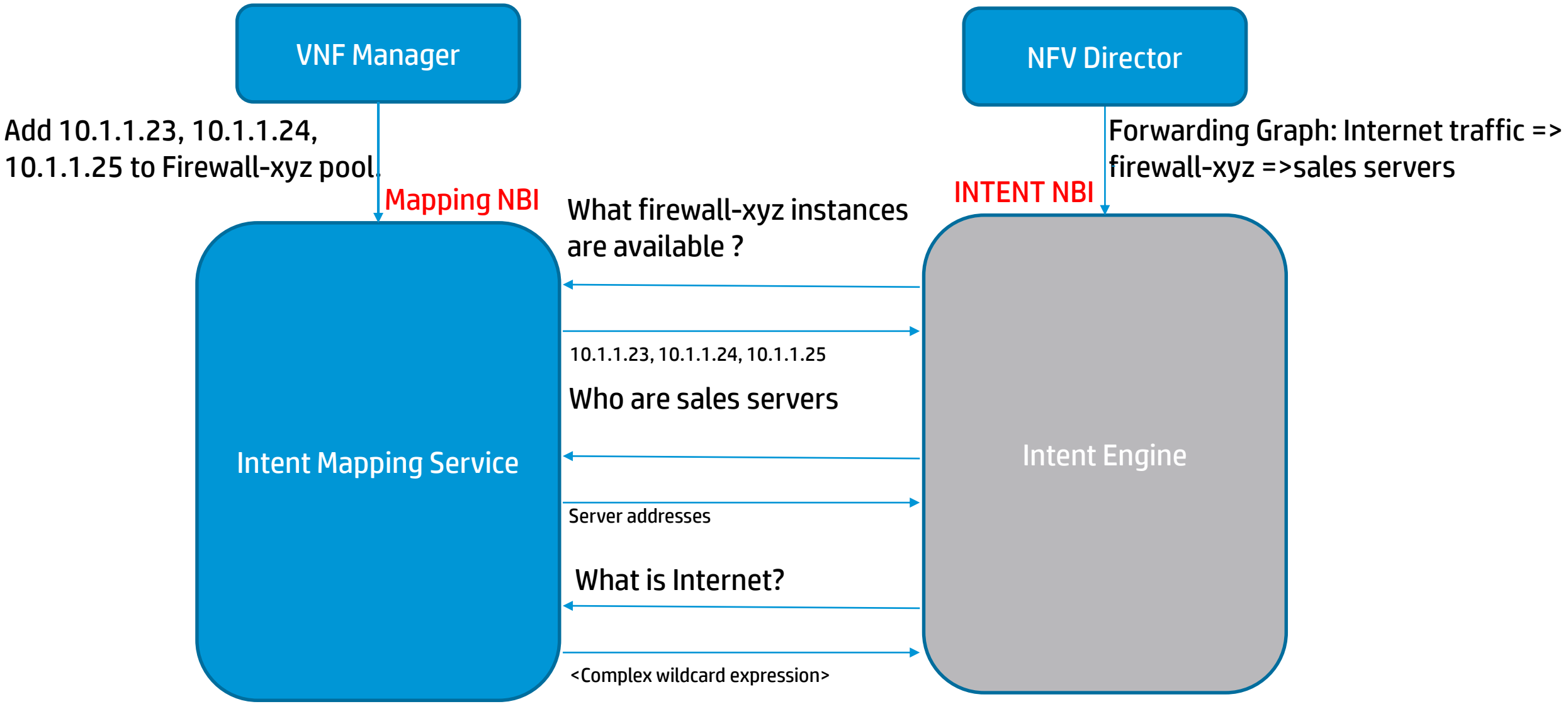


Řešení pro Multi-Writer situaci (jinak prakticky neřešitelné)



Pozn.: Ano, použití více tabulek se zkoušelo, ale nedopadlo lépe…a v hardware není možné.

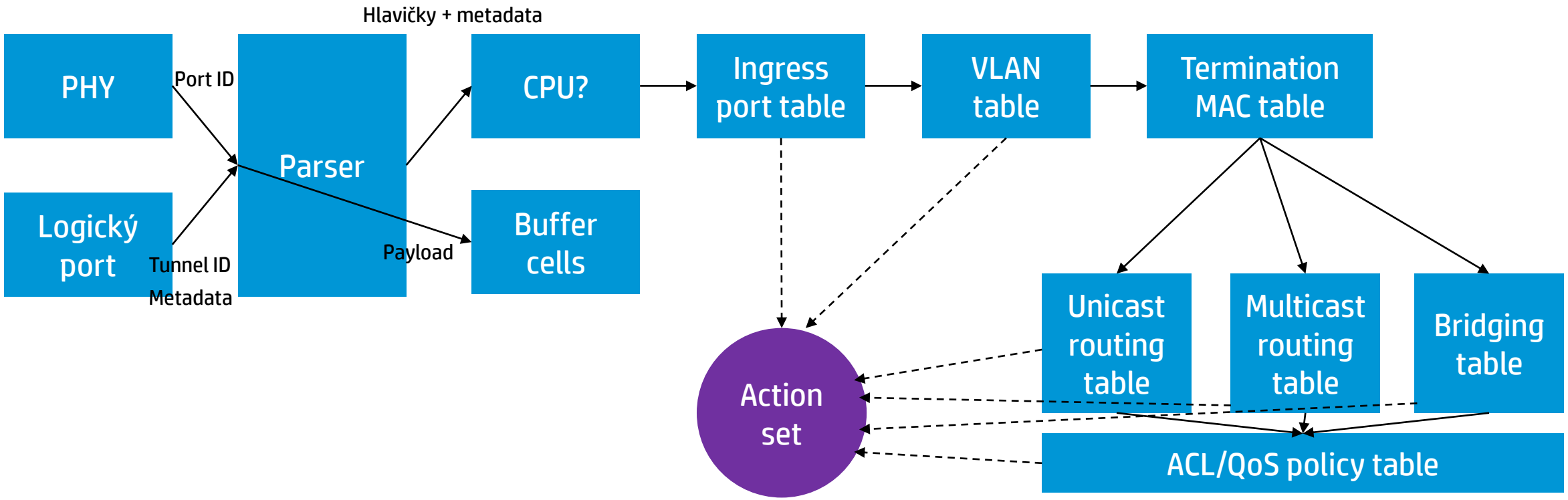
Service chaining (pro NFV) s využitím intent-based SDN



Hardware

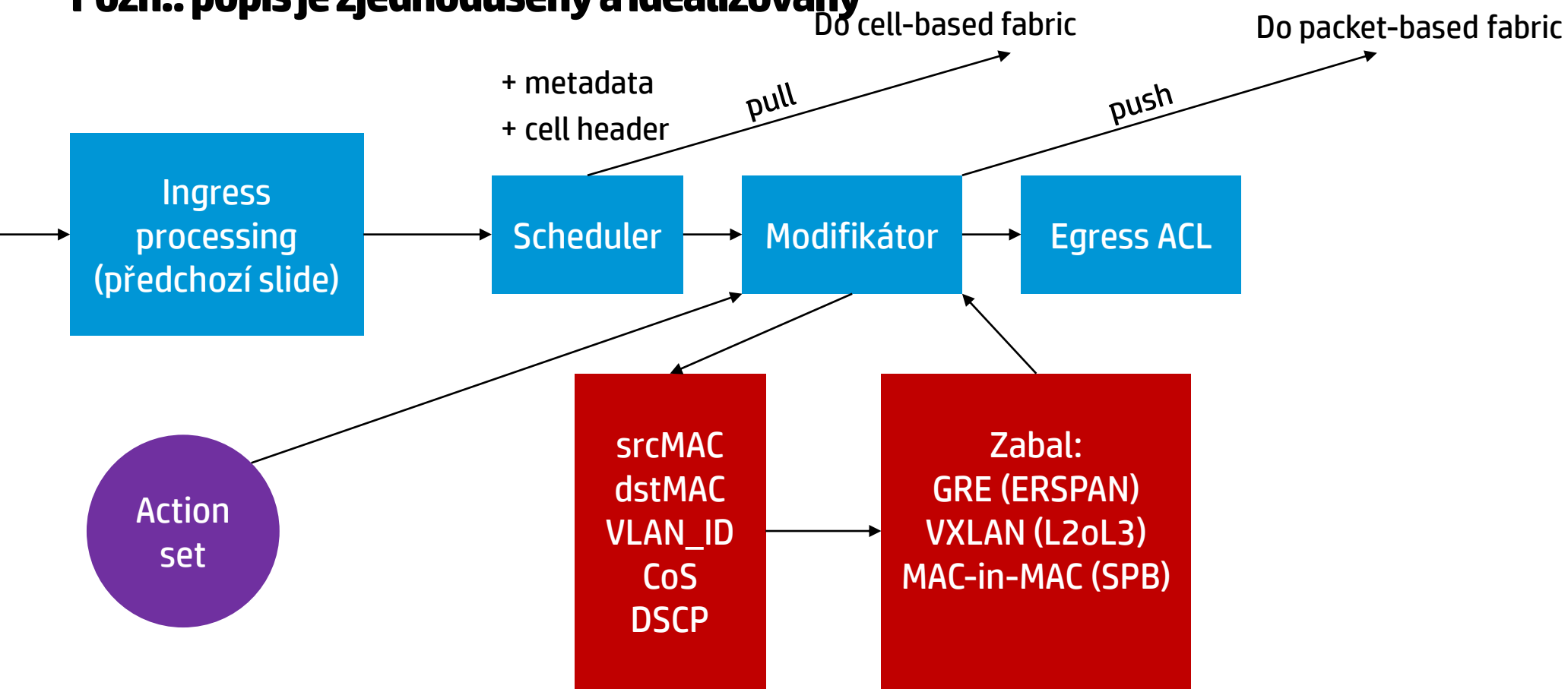
Jak funguje switch ASIC?

Pozn.: popis je zjednodušený a idealizovaný



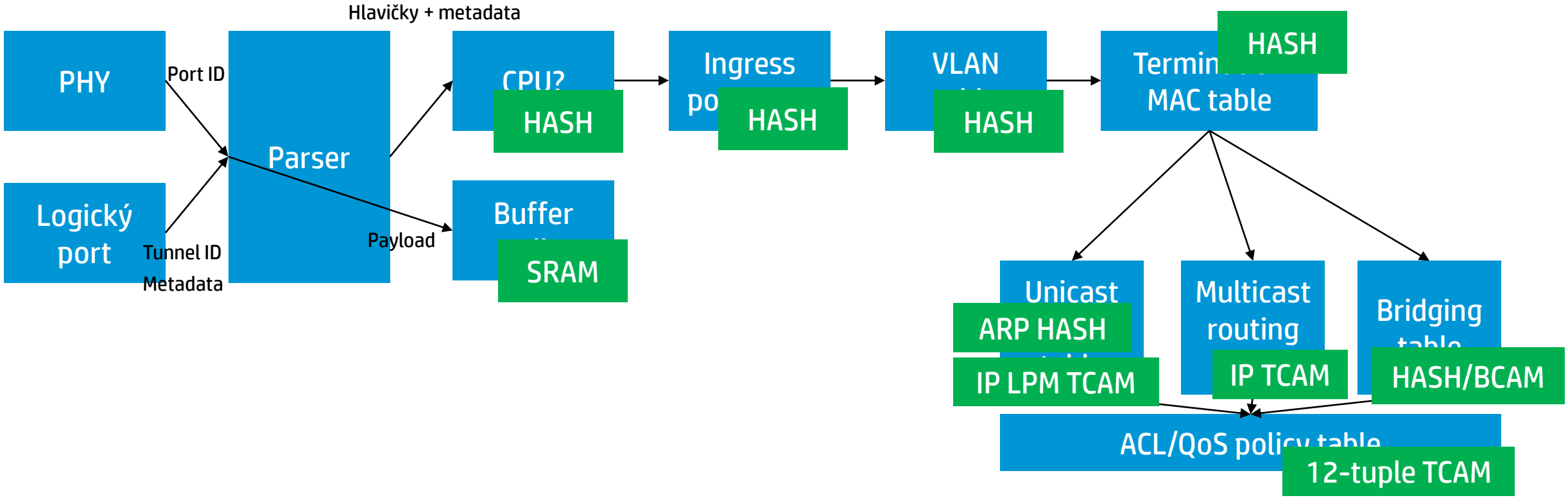
Jak funguje switch ASIC?

Pozn.: popis je zjednodušený a idealizovaný



Jak funguje switch ASIC?

Pozn.: popis je zjednodušený a idealizovaný



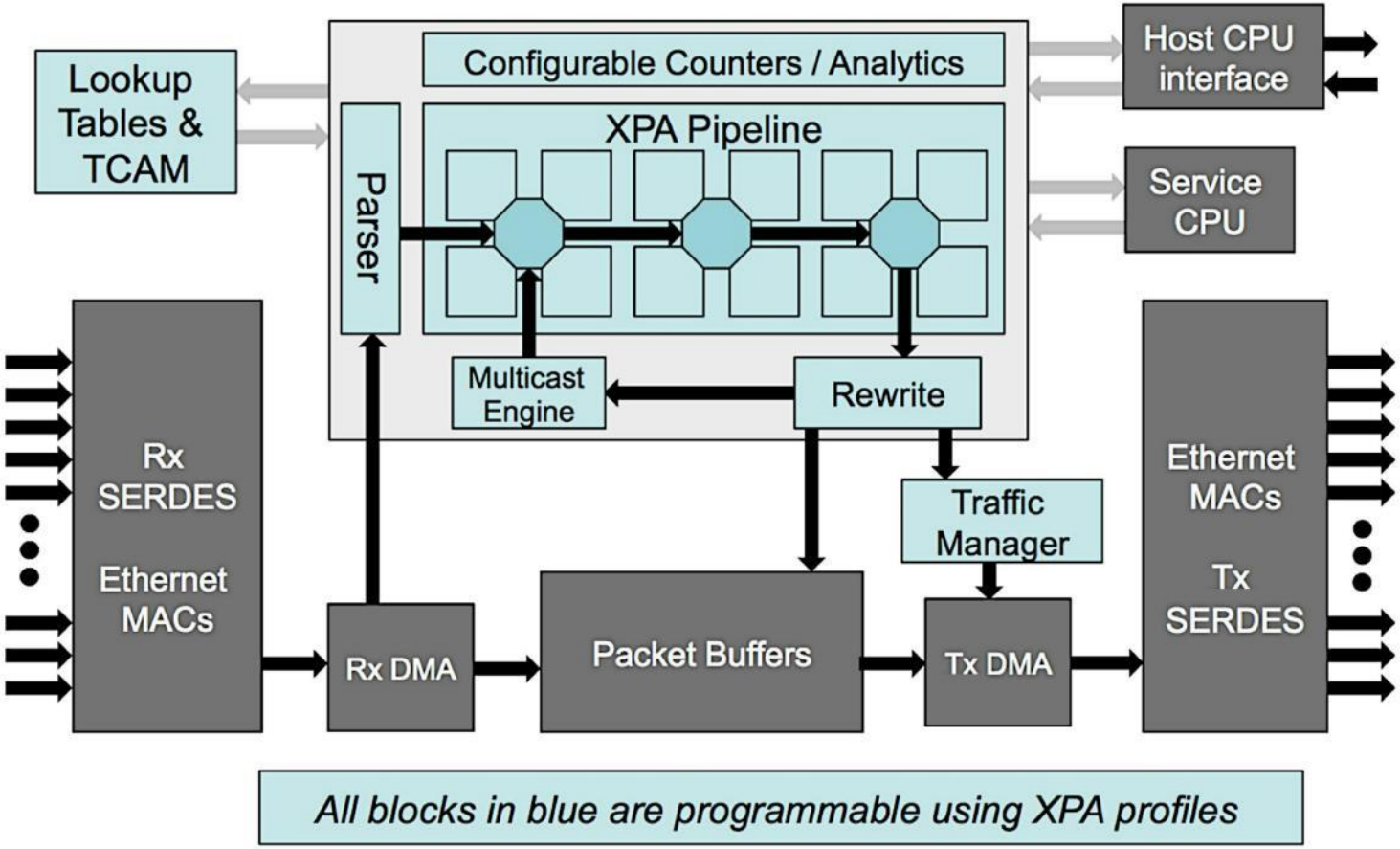
V čem je potíž?

- Průtokovost, nelze se vracet (např. problém routingu VXLAN provozu)
- Všechny datové konstrukty pevně dané (např. nelze přidat data-plane hlavičku bez změny ASIC jako je VXLAN, 802.1BR, FCoE)
- Modifikační funkce pevně dané (měnit můžete jen hlavičky, které se mění při switchingu/routingu)
- Počet a pořadí tabulek je pevně dané
- Malé možnosti realokace kapacit tabulek (ale čím dál tím lepší viz. Broadcom Trident II)

HP ProVision ASIC pro campus

- Definujete si typ a pořadí tabulek (až 32)
- 10x microcode engine pro manipulaci s pakety
- FPGA pro pattern match (jakékoli bity)
- FPGA pro modifikaci odchozího provozu (bity v offsetu určité délky)
- Programovatelnost tabulek přímo přes OpenFlow (můžete např. vytvořit tabulku !)

Cavium XPliant ASIC do DC



XPliant Packet Architecture (XPA) Block Diagram

Co na to Broadcom?

Zatím nemá

- Bit pattern match (tzn. jakýkoli i nový protokol)
- Flexibilní engine (jakékoli operace)
- Modifikace offset/pattern (jakýkoli protokol)

Tamahawk (2015)

- Vysoká hustota
- Flexibilní pořadí operací
- Vysokokapacitní flow countery

Trident II+ (2014)

- VXLAN routing
- Nativní 100G
- Menší spotřeba

Trident II (2013)

- Vysoká hustota
- Realokovatelné kapacity tabulek (např. forwarding vs. routing)
- VXLAN enkapsulace

Trident I (2010)

- Vysoká hustota

Alternativní přístupy

Ani nejmodernější ASIC neumí:

- Match na variable field lenght (aka TLV)
- Zásadní změnu logiky při zachování hardwarového zpracování
- Držení složitějšího state (např. NAT/PAT)
- Vysoko-kapacitní pattern match (TCAM nebo FPGA uvnitř ASIC je velmi malé, drahé a žere)

FPGA

Např. Pre-procesor v routeru (QoS)

- Výkon blížící se ASIC
- Mikrokód určuje logiku
- Programuje se stejně jako design ASIC (nesmírně obtížné)
- Při vyšší kapacitě začně být extrémně drahý

Multi-core CPU

Např. Pobočkový router

- Naprostá flexibilita
- Logiku určuje kód (příjemné API)
- Relativně nižší výkon
- Nutná paralelizace nefunguje pro všechny situace
- Nepředvídatelný výkon/latence

Network Processor

Např. Výkonný router

- Velmi dobrá flexibilita
- Logiku určuje kód (použitelná komplexita)
- Přístup aka GPU (masivní paralelizace) nemusí fungovat pro všechny situace
- Vysoká cena

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

Network Function Virtualization

Disagregace

Open source

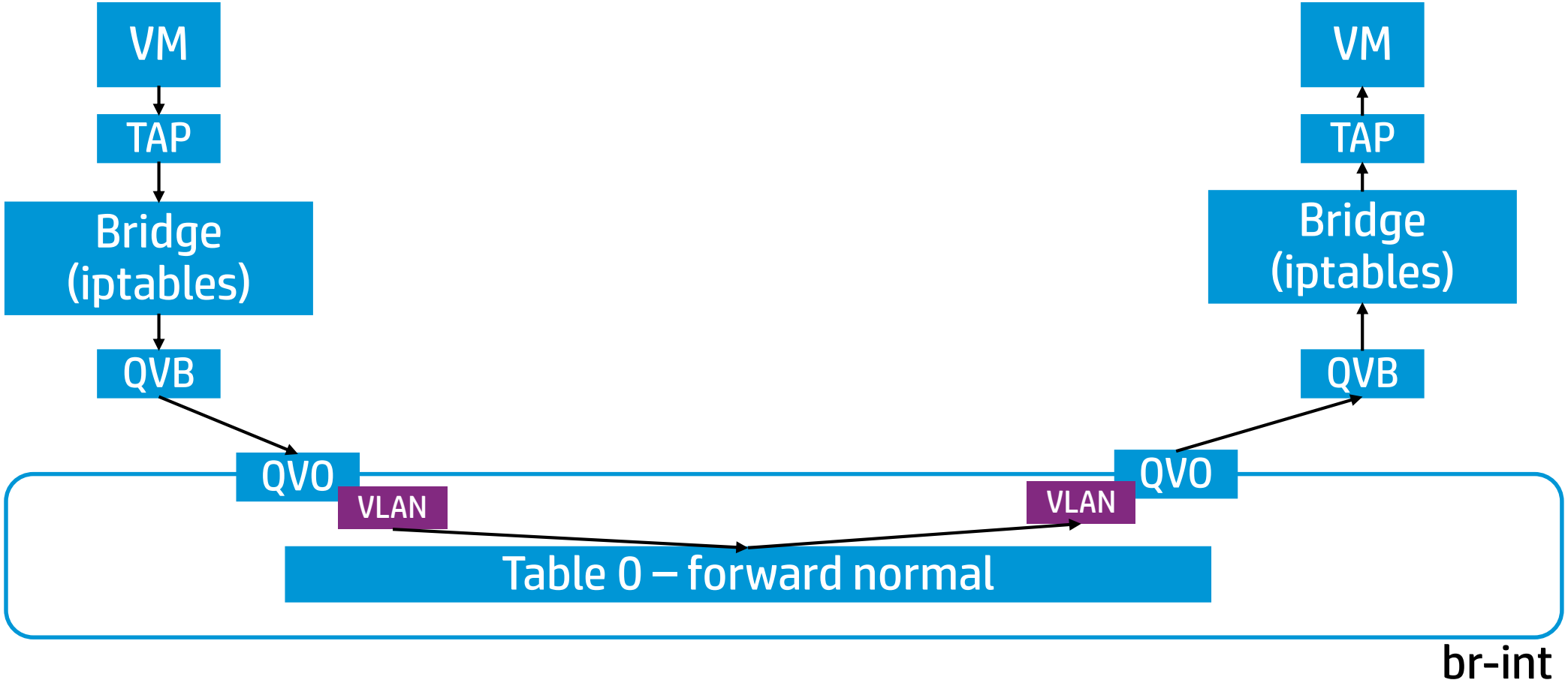
DevOps

Chcete vědět jak funguje Neutron implementace s OVS?

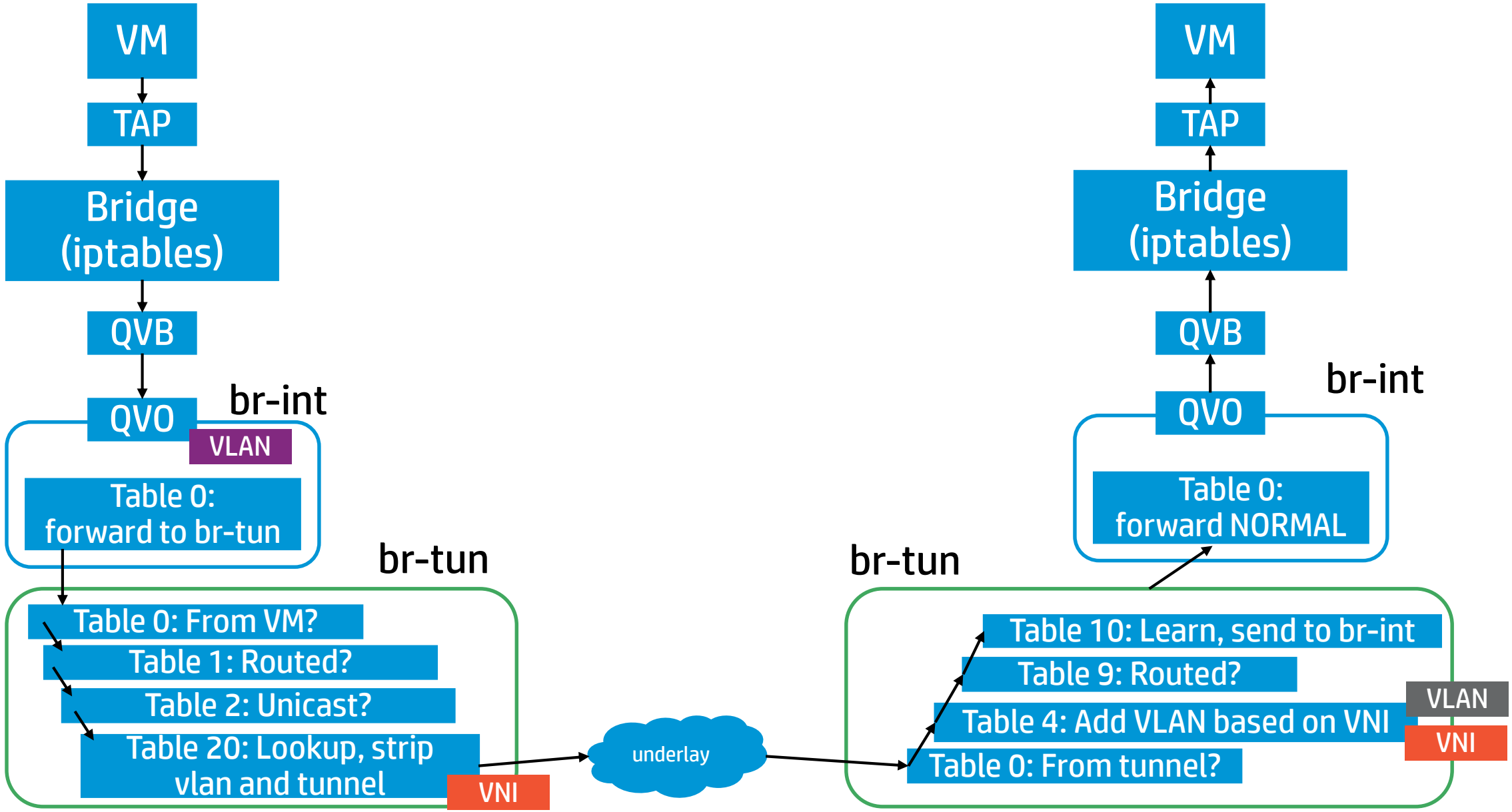
Stáhněte si lab guide 3 na:

http://www.cloudsvet.cz/?page_id=10

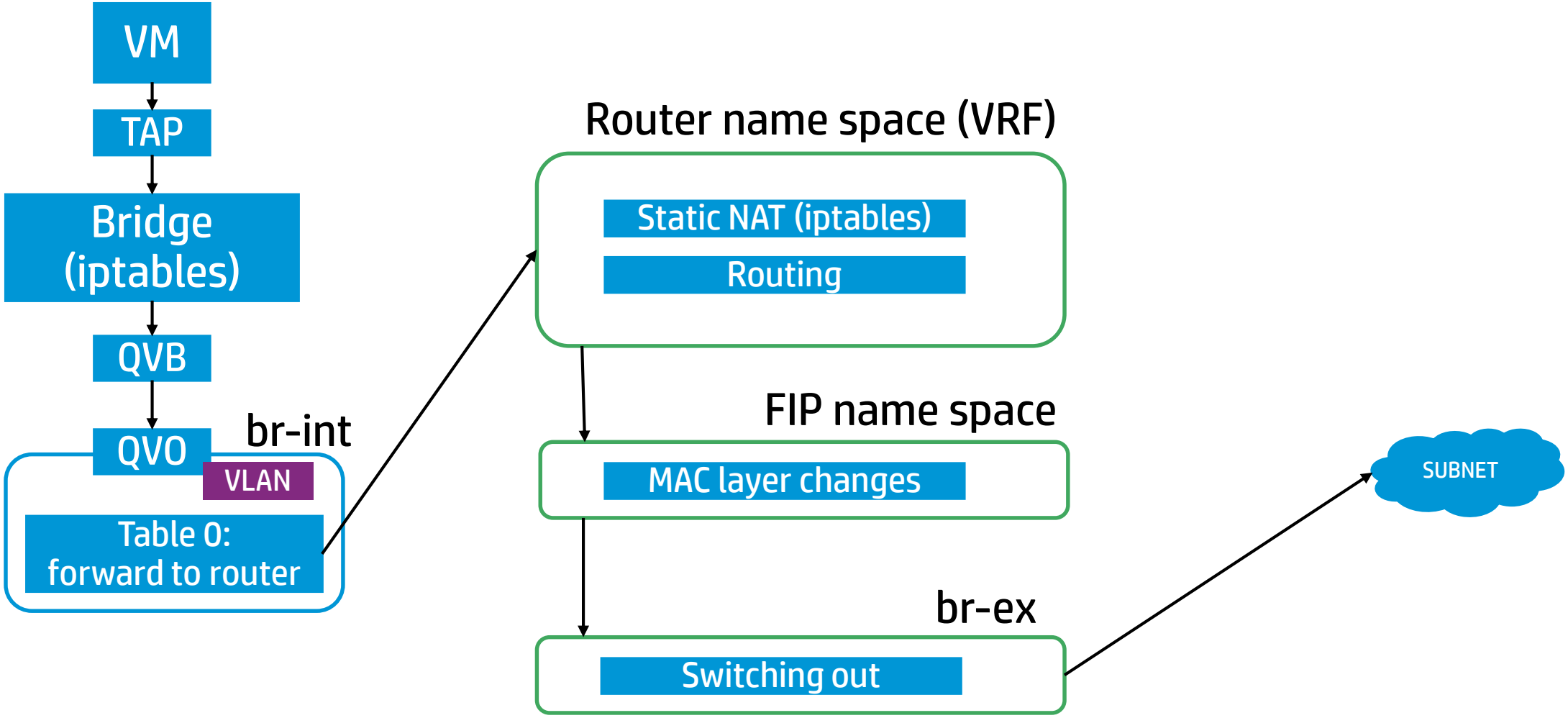
Komunikace mezi VM ve stejné síti na stejném compute node



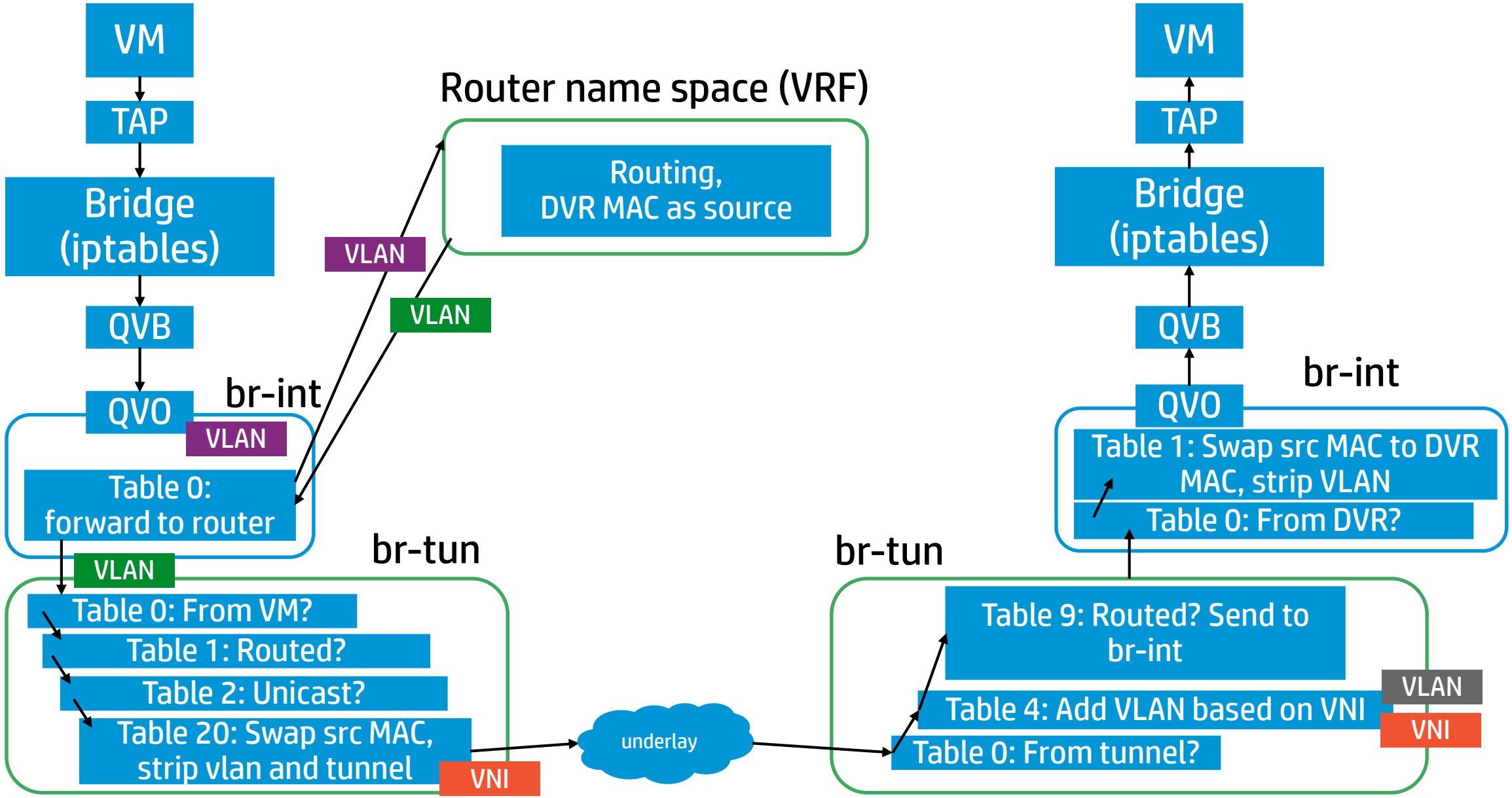
Komunikace mezi VM ve stejné síti na odlišných compute node



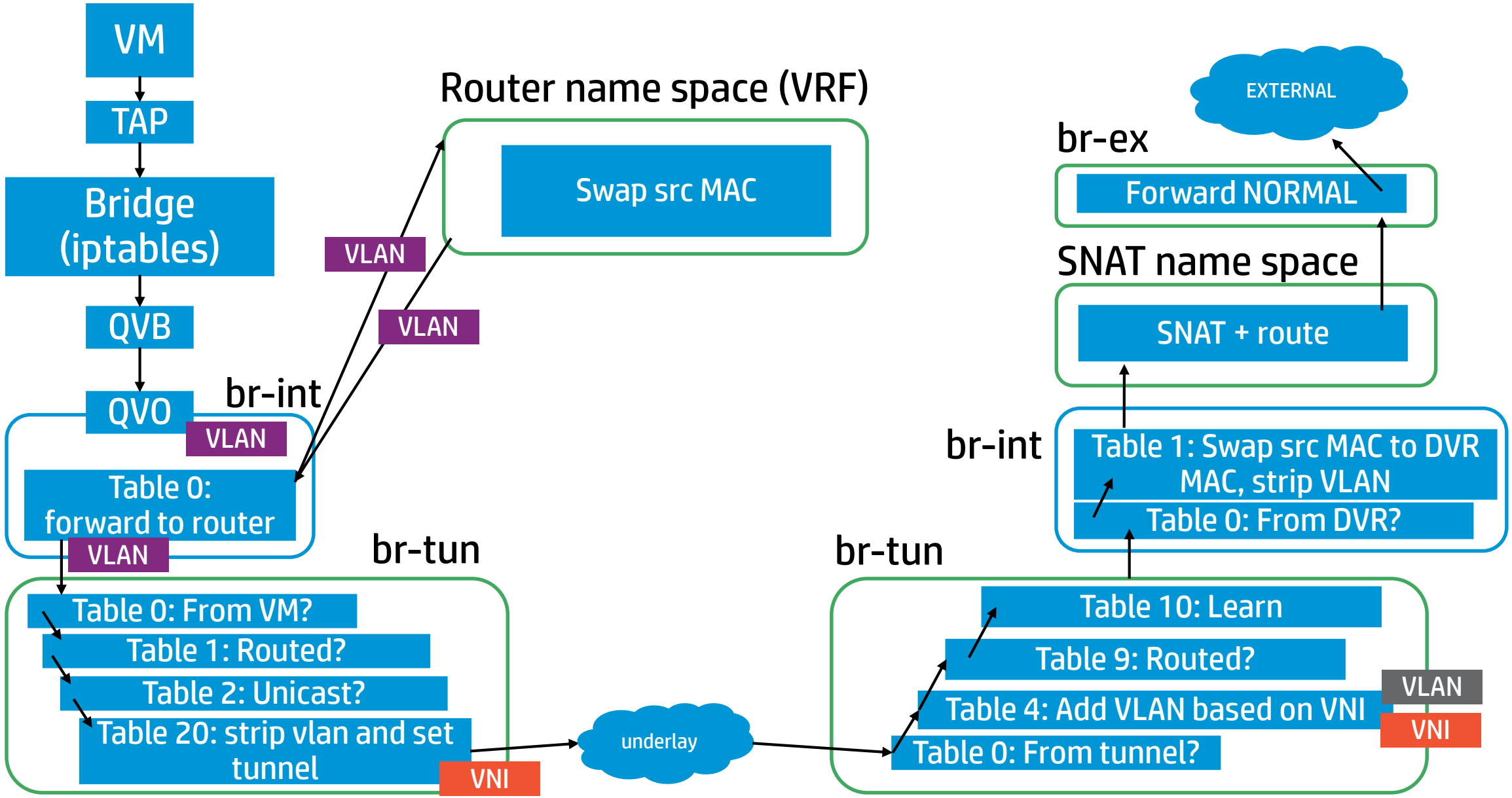
Komunikace z VM do světa s Floating IP



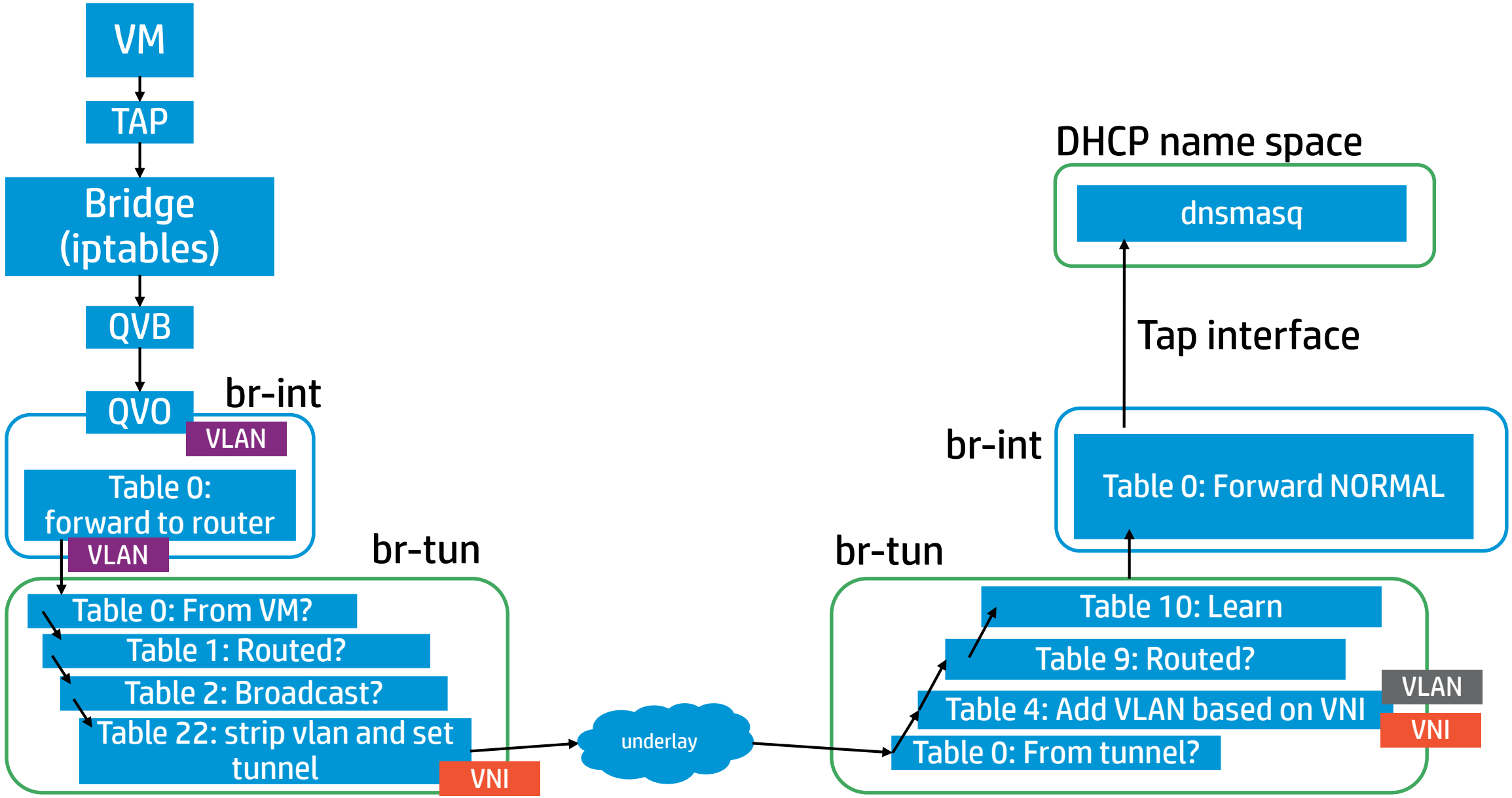
Routing mezi subnety v rámci tenant na různých compute node



Komunikace do externí sítě s použitím source NAT



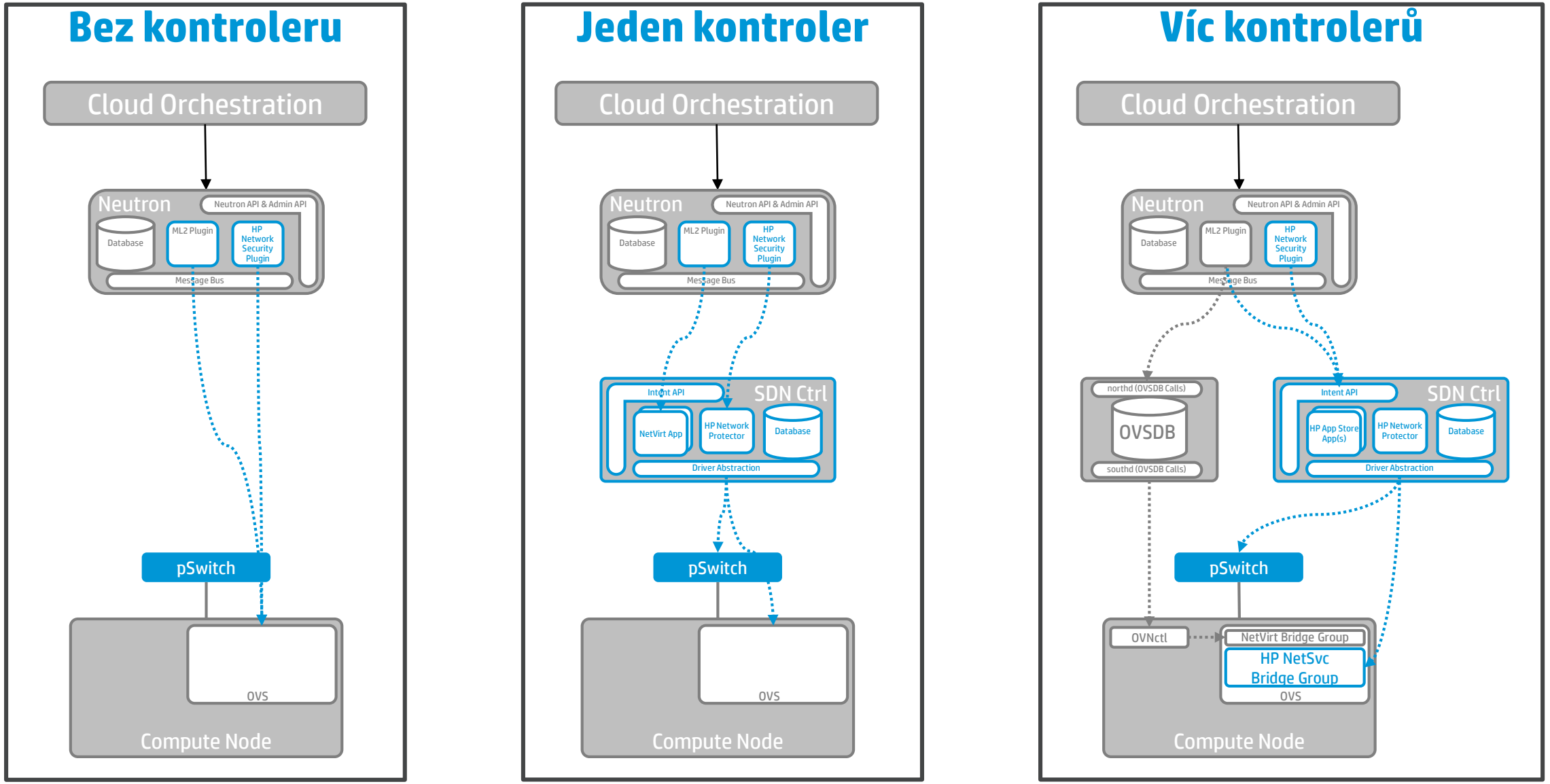
DHCP


VM
TAP
Bridge (iptables)
QVB
QVO
br-int
VLAN
Table 0: forward to router
VLAN
br-tun
Table 0: From VM?
Table 1: Routed?
Table 2: Broadcast?
Table 22: strip vlan and set tunnel
VNI
underlay
DHCP name space
dnsmasq
Tap interface
br-int
Table 0: Forward NORMAL
br-tun
Table 10: Learn
Table 9: Routed?
Table 4: Add VLAN based on VNI
VLAN
VNI
Table 0: From tunnel?

Jaká je potíž s virtualizací sítě?

- Moc zaměření na netvirt, málo na rámec (= nerozšiřitelnost aka NSX nebo OVN)
- Rozmělnění vývoje díky bezkontrolerovým a vícekontrolerovým řešením (viz dále)
- Málo koordinace Neutron a non-Neutron řešení (např. SDN pro overlay + SDN pro underlay)

Soupeřící modely



Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

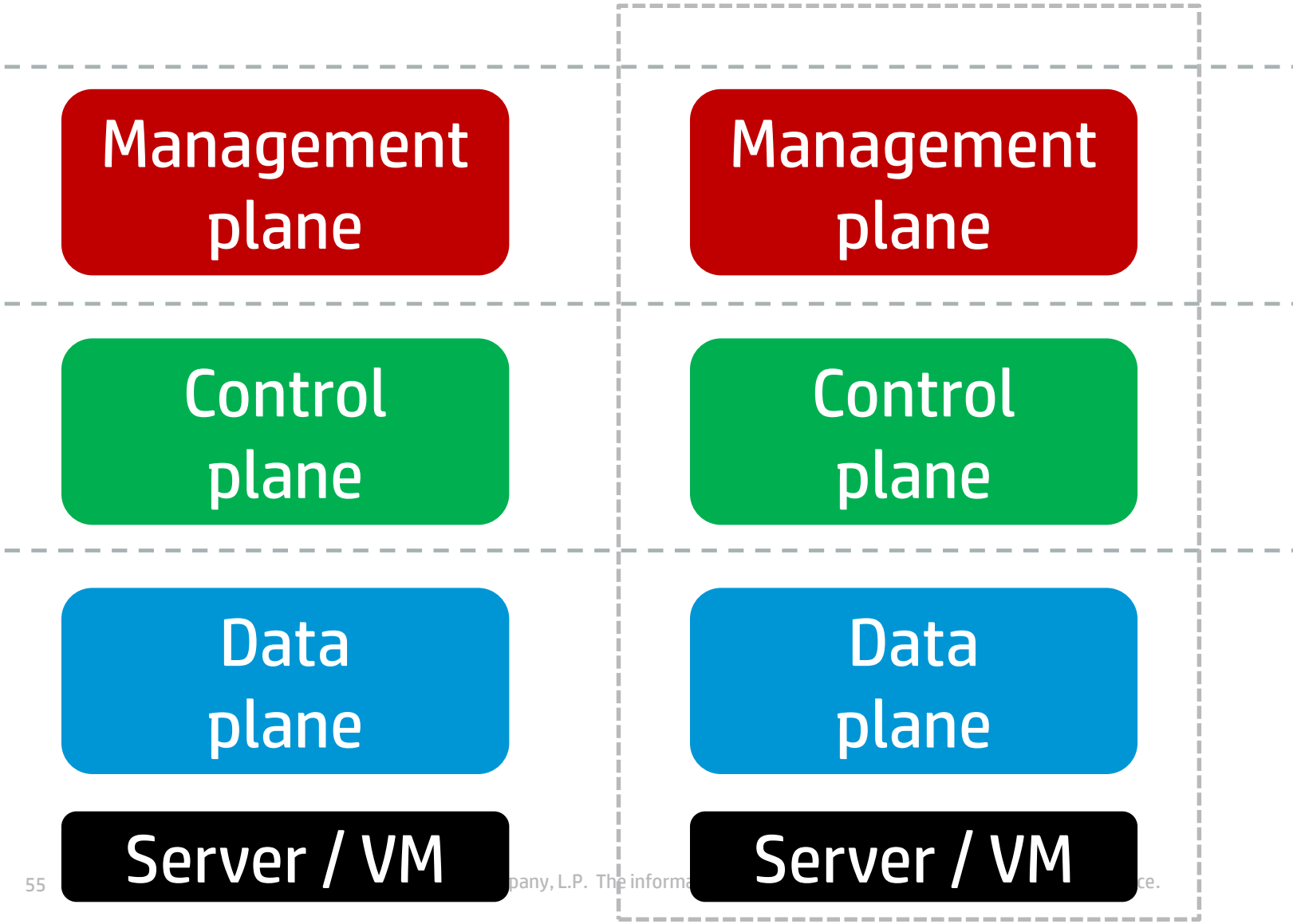
Network Function Virtualization

Disagregace

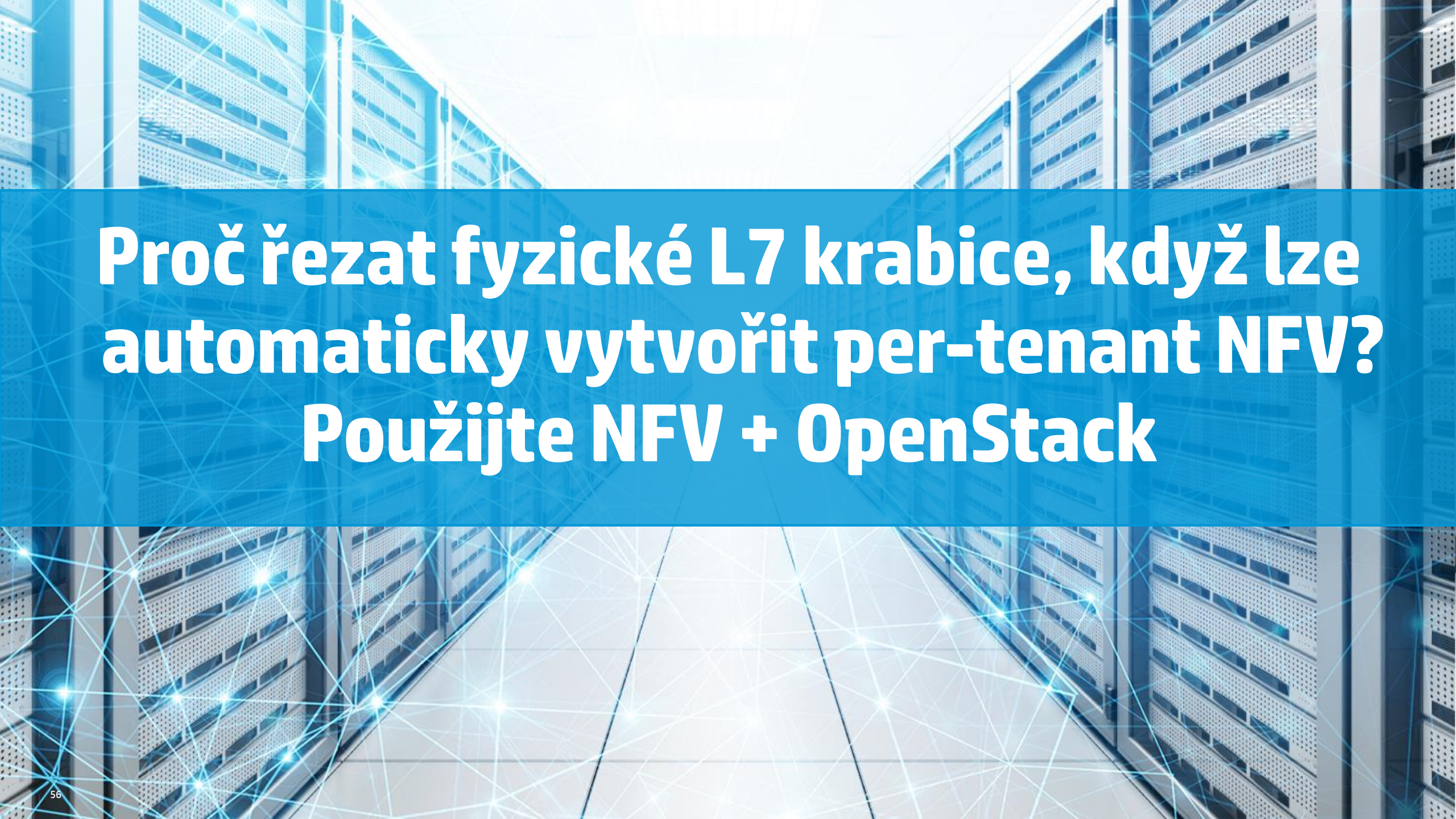
Open source

DevOps

Virtualizace síťových funkcí (NFV)



Přesun síťových funkcí ze specializovaného hardware do obyčejného serveru nebo VM

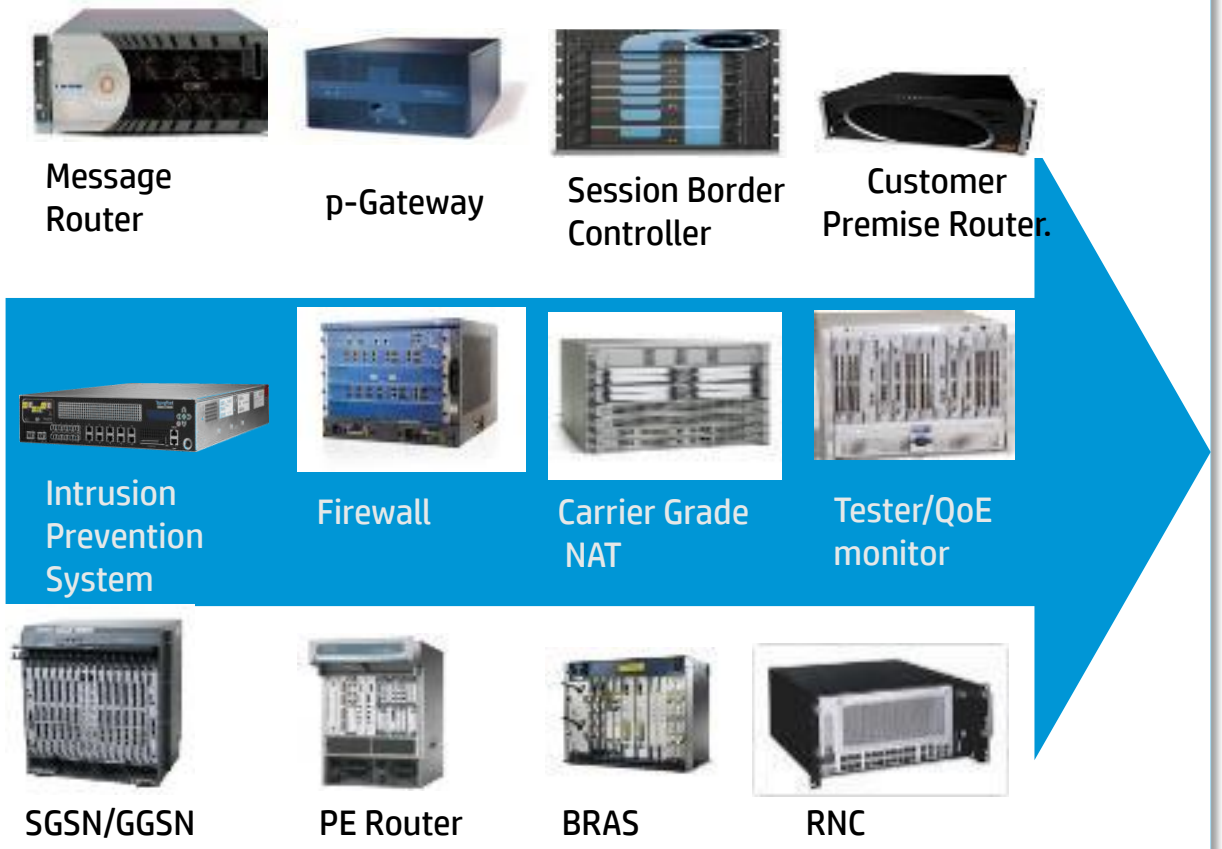


Proč řezat fyzické L7 krabice, když lze automaticky vytvořit per-tenant NFV? Použijte NFV + OpenStack

Jak poslat provoz uživatele do různých L7 služeb ve správném pořadí aniž byste potřebovali vědět, kde služba běží? Použijte NFV + SDN

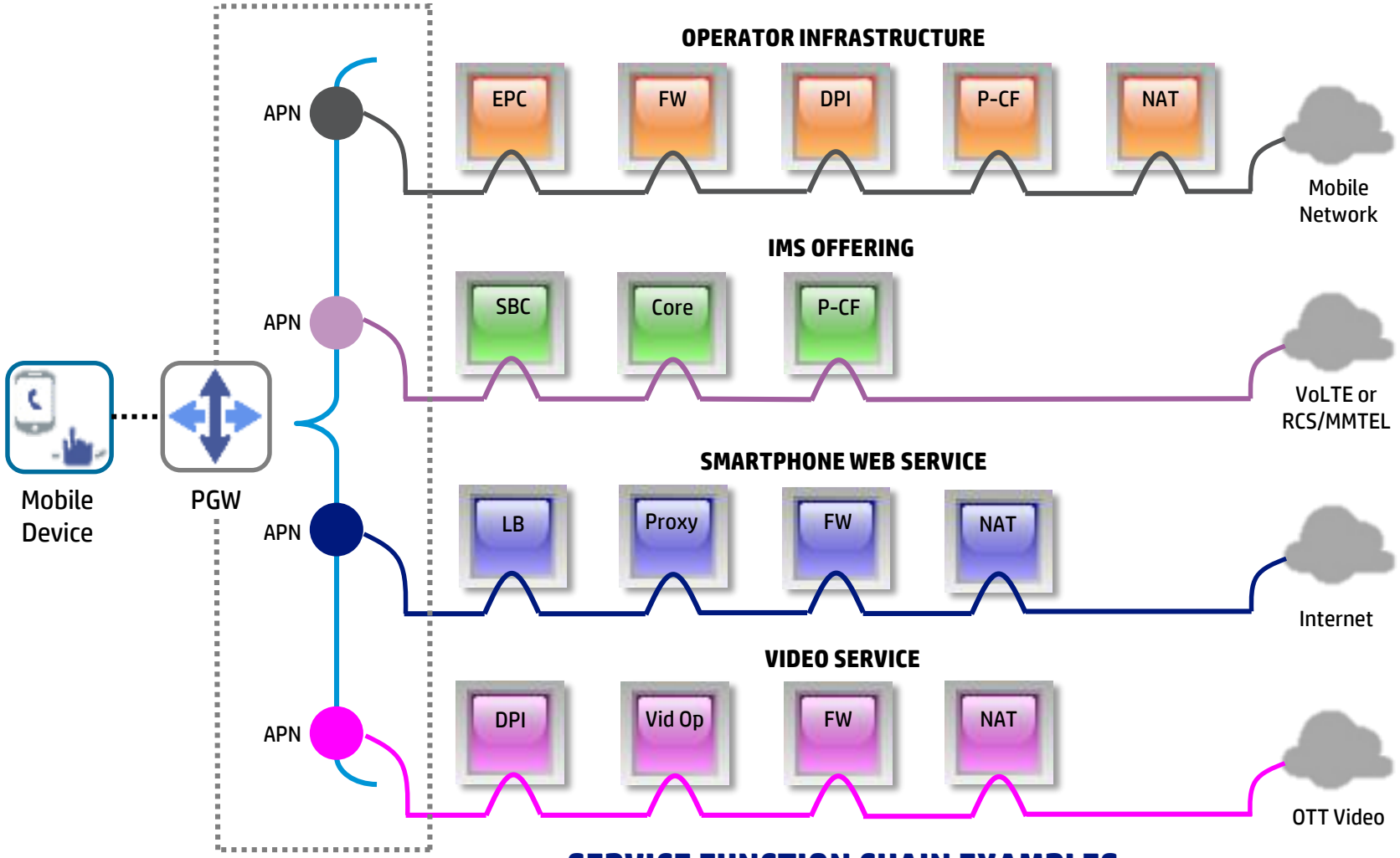
Jak vznikl požadavek na NFV?

Proprietary Network Appliances



Software Appliances

Praktické použití service chaining



SERVICE FUNCTION CHAIN EXAMPLES

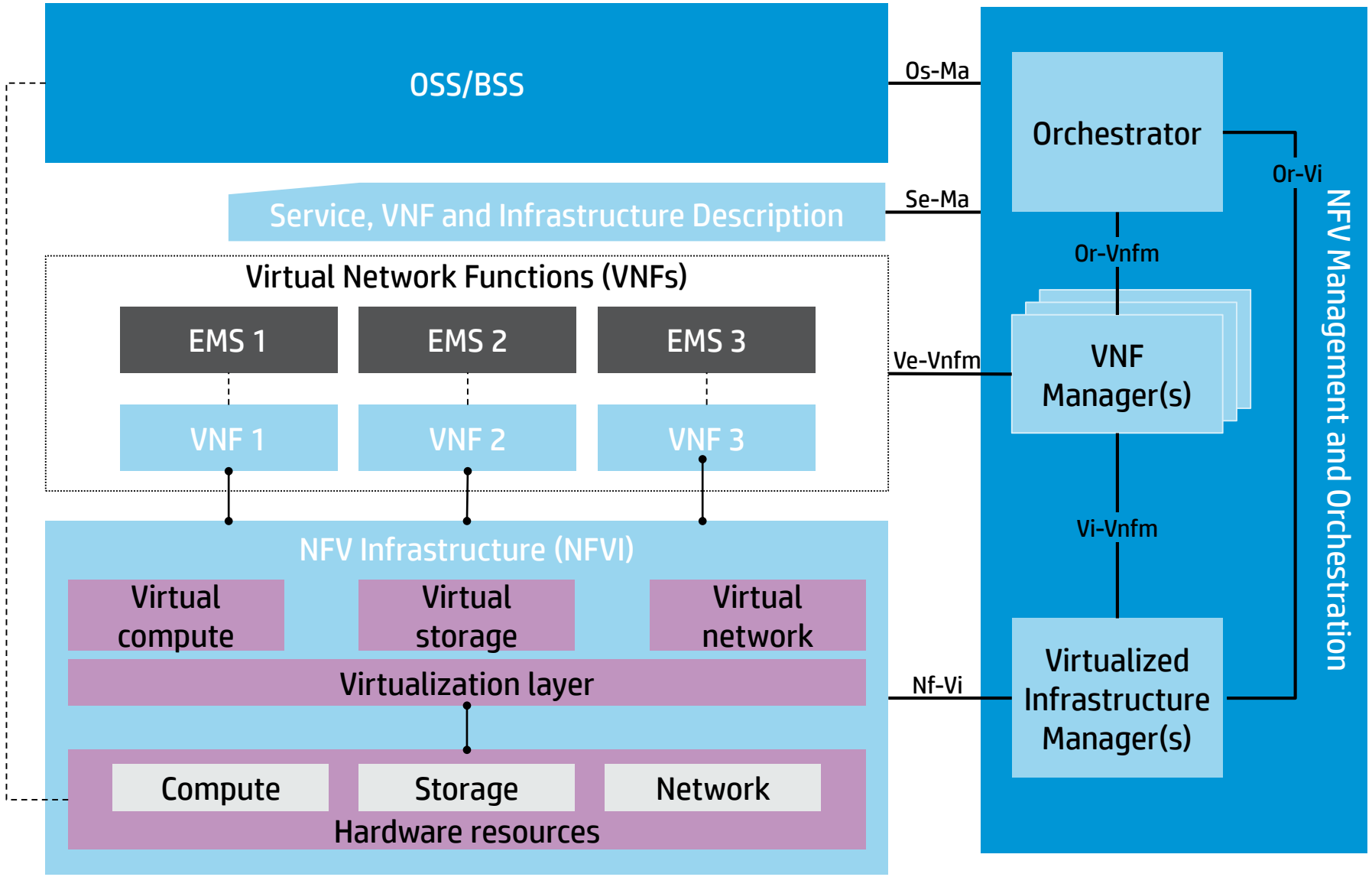
Middleboxes via VNF / VNFC Forwarding Graphs (Package)

APN: Access Point Name
LB: Load Balancer
FW: Firewall
SBC: Session Border Controller

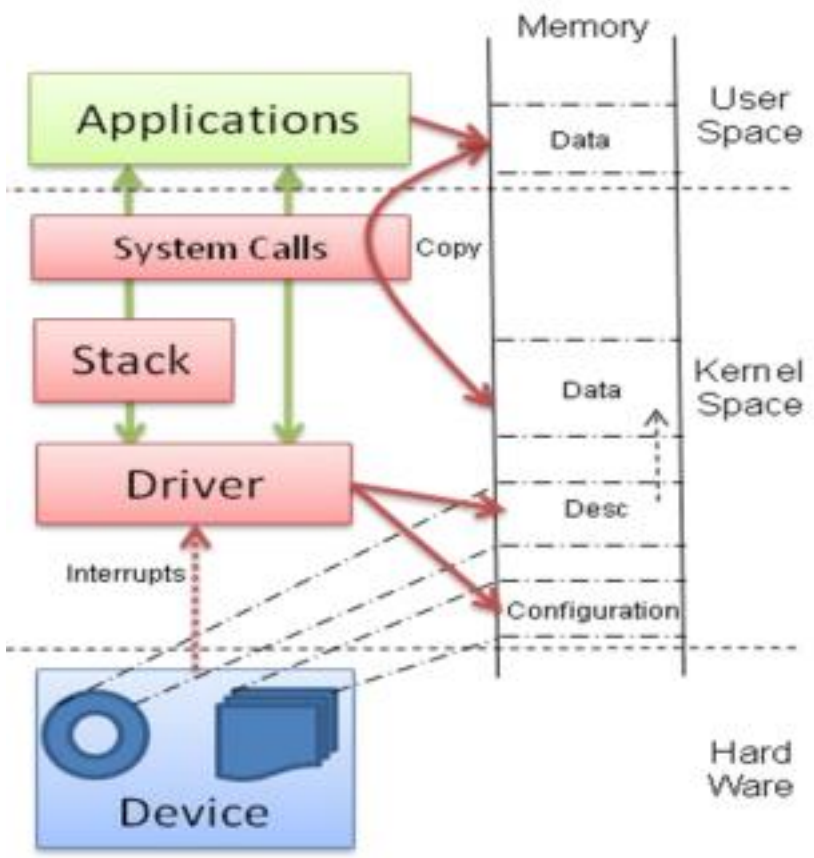
Core: IMS Core Components
P-CF: Policy & Charging Functions
Vid-Op: Video Optimizer
NAT: Network Address Translator

EPC: Evolved Packet Core
Proxy: Web Proxy
DPI: Deep Packet Inspection

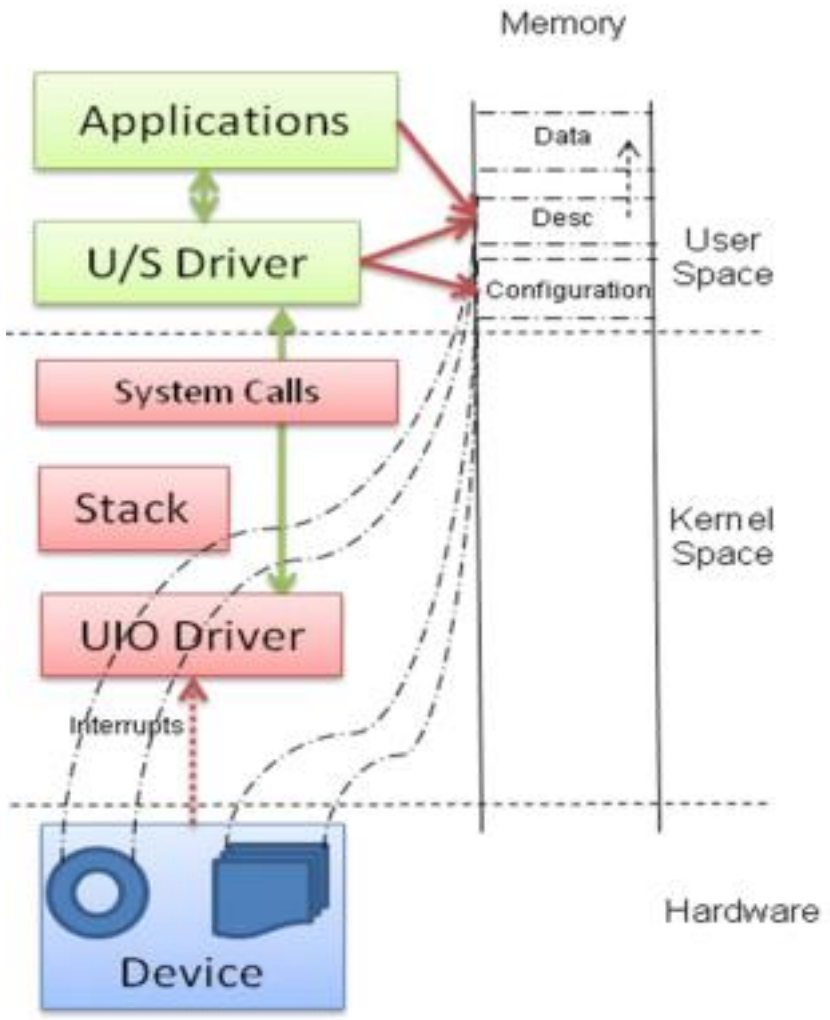
NFV architektura u operátorů



Přechody mezi kernel space a user space bolí

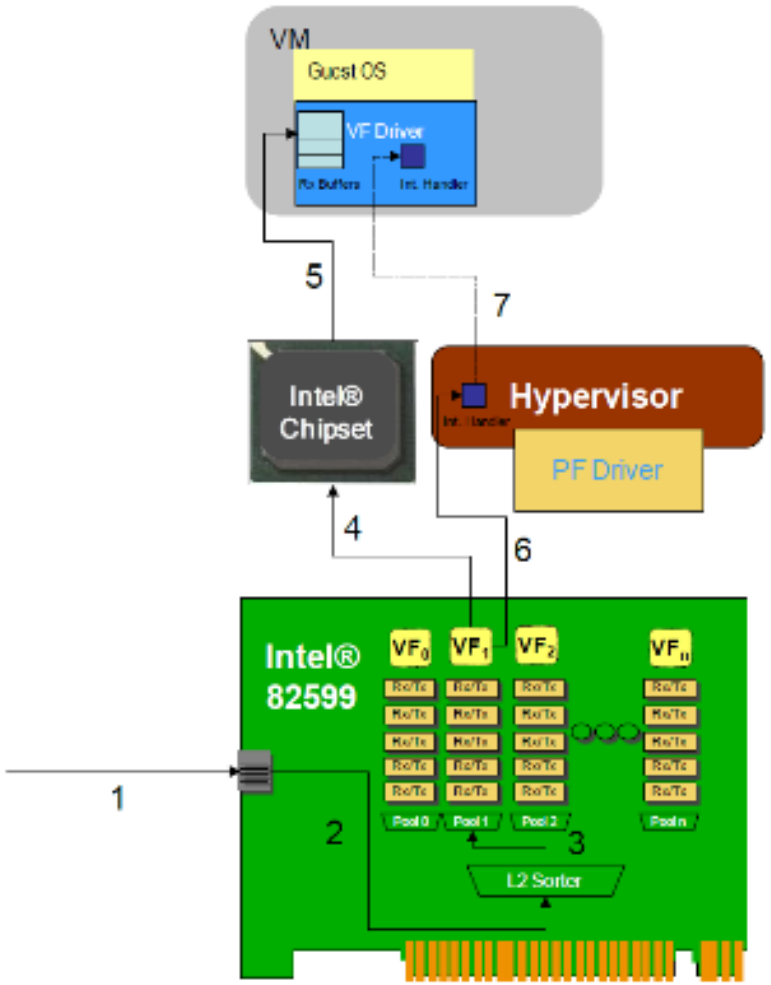


Kernel space network driver



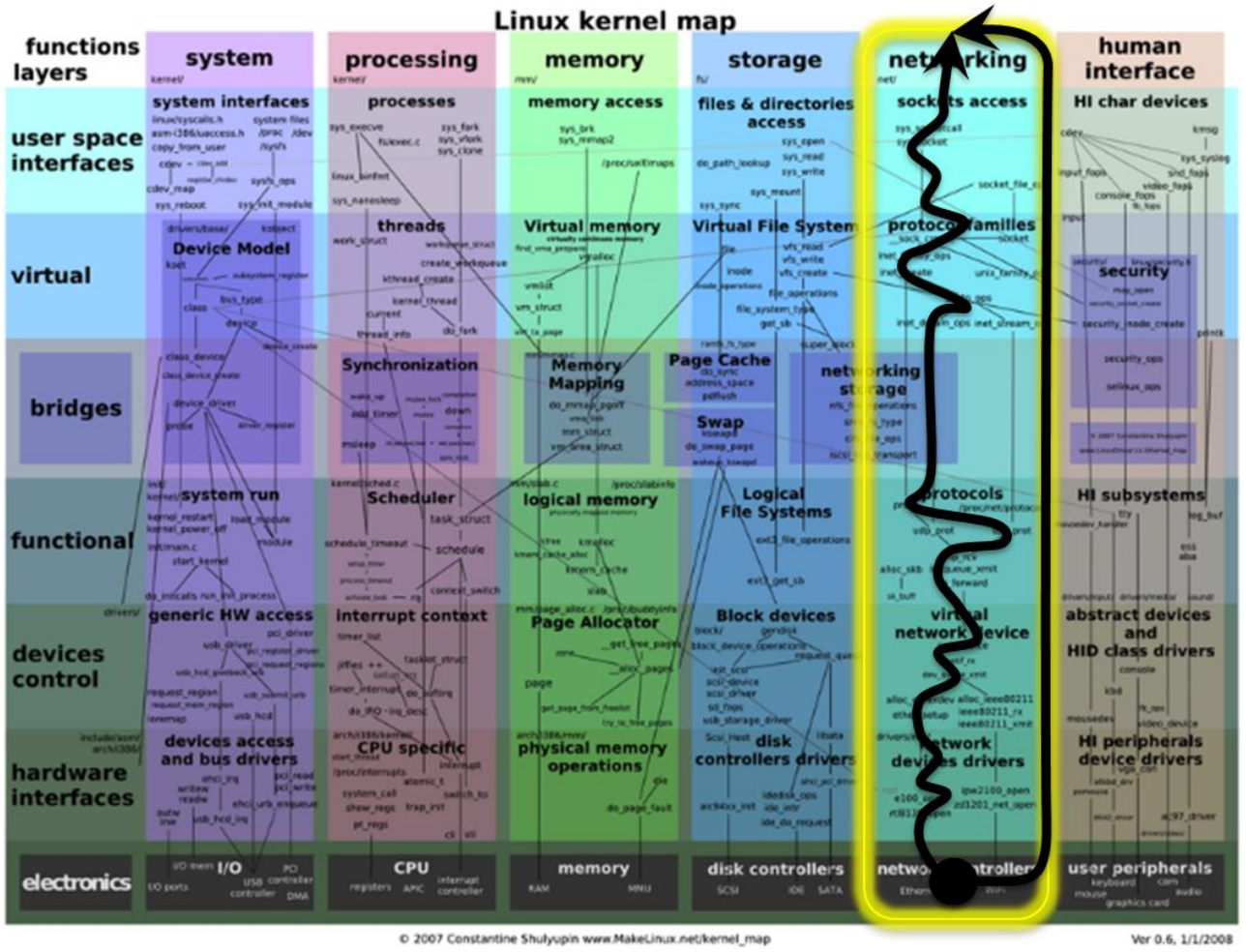
User space network driver

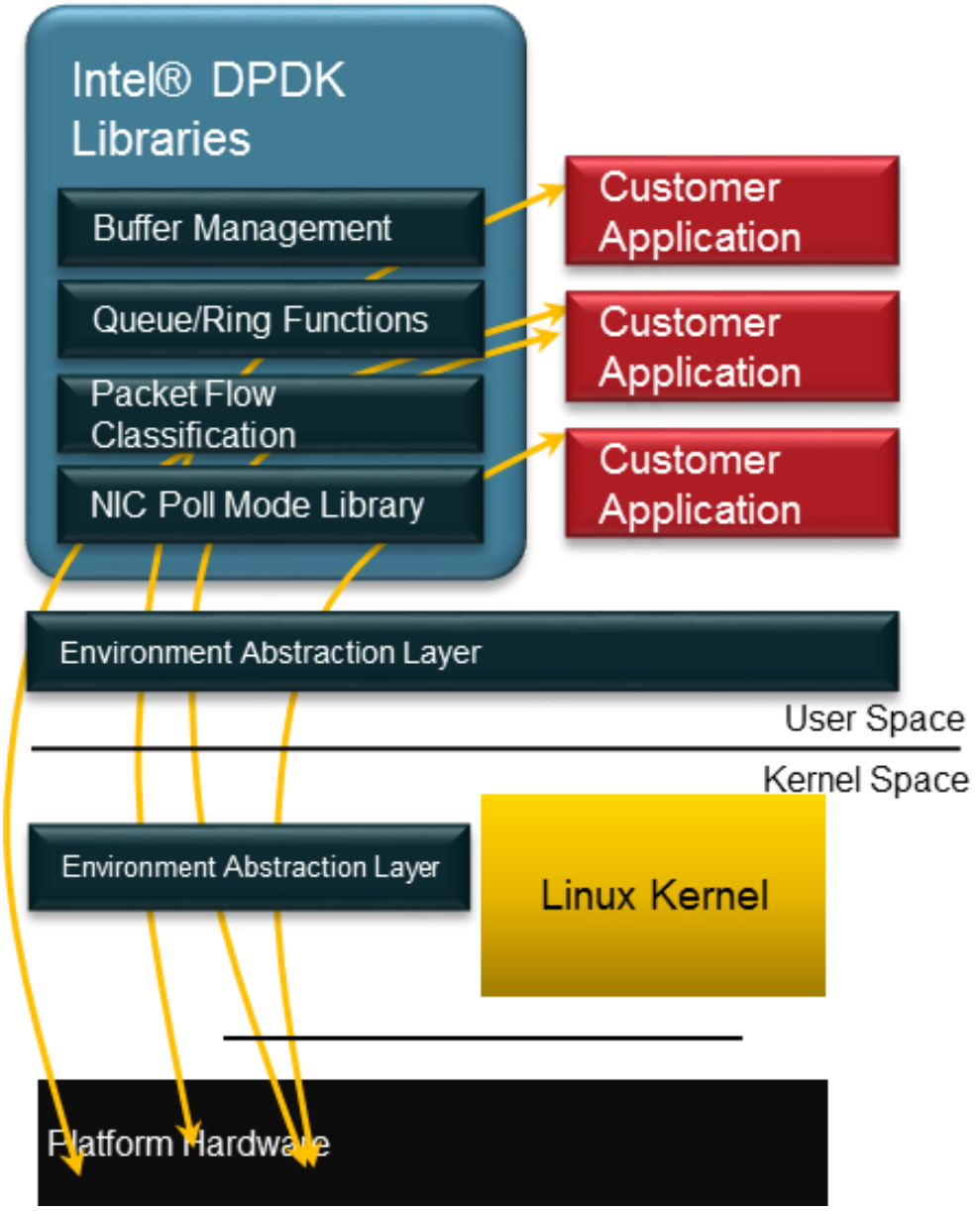
Hardwarová akcelerace – SR-IOV



- **Step 1 & 2:** Packet Arrives, sent to the L2 Sorter/Switch.
- **Step 3:** Packet is sorted based upon destination MAC address; in this case, it matches Pool/VF 1.
- **Step 4:** NIC initiates DMA action to move packet to VM
- **Step 5:** DMA action hits the Intel Chipset, where VT-d (configured by the Hypervisor) performs the required Address Translation, for the DMA operation; resulting in the packet being DMA'd into the VM's VF Driver buffers.
- **Step 6:** NIC posts MSI-X interrupt indicating a Rx transaction has been completed. This interrupt is received by the Hypervisor.
- **Step 7:** The Hypervisor injects a virtual interrupt to the VM indicating a Rx transaction has been completed, the VM's VF Driver then processes the packet.

Intel DPDK





Víte, že se s paketem v Linux jádře dějí docela věci?

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

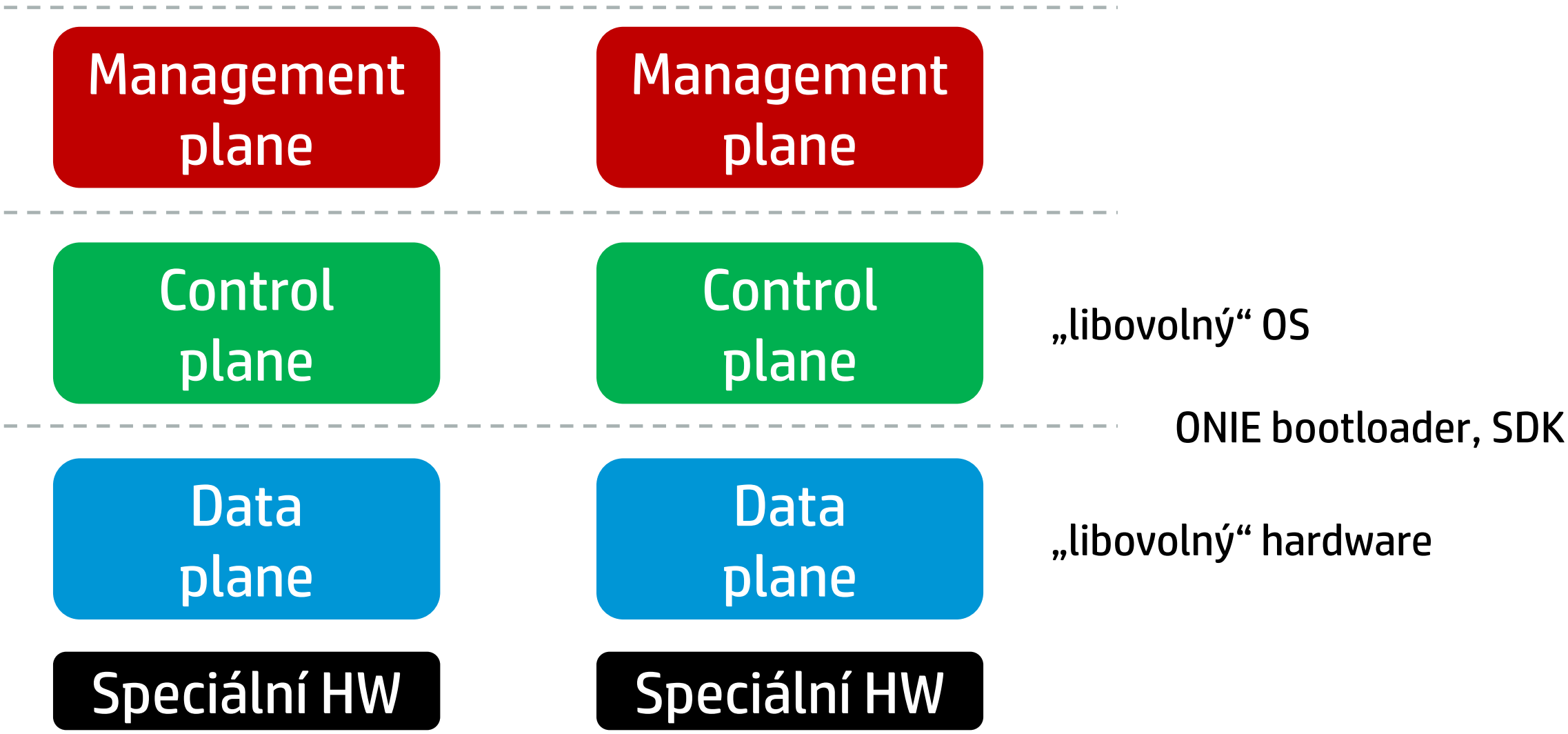
Network Function Virtualization

Disagregace

Open source

DevOps

Disagregované prvky (brite-box / white-box)



Nástroje a správa vendora

L2/L3 síťové protokoly

Operační systém vendora

Značkový integrovaný box

ASIC nebo merchant silicon

Integrované sítě

Disagregované sítě

Co trh nabízí?

Brite-box

DNI

Accton

Quanta

White-box (ODM)

OS

ASIC

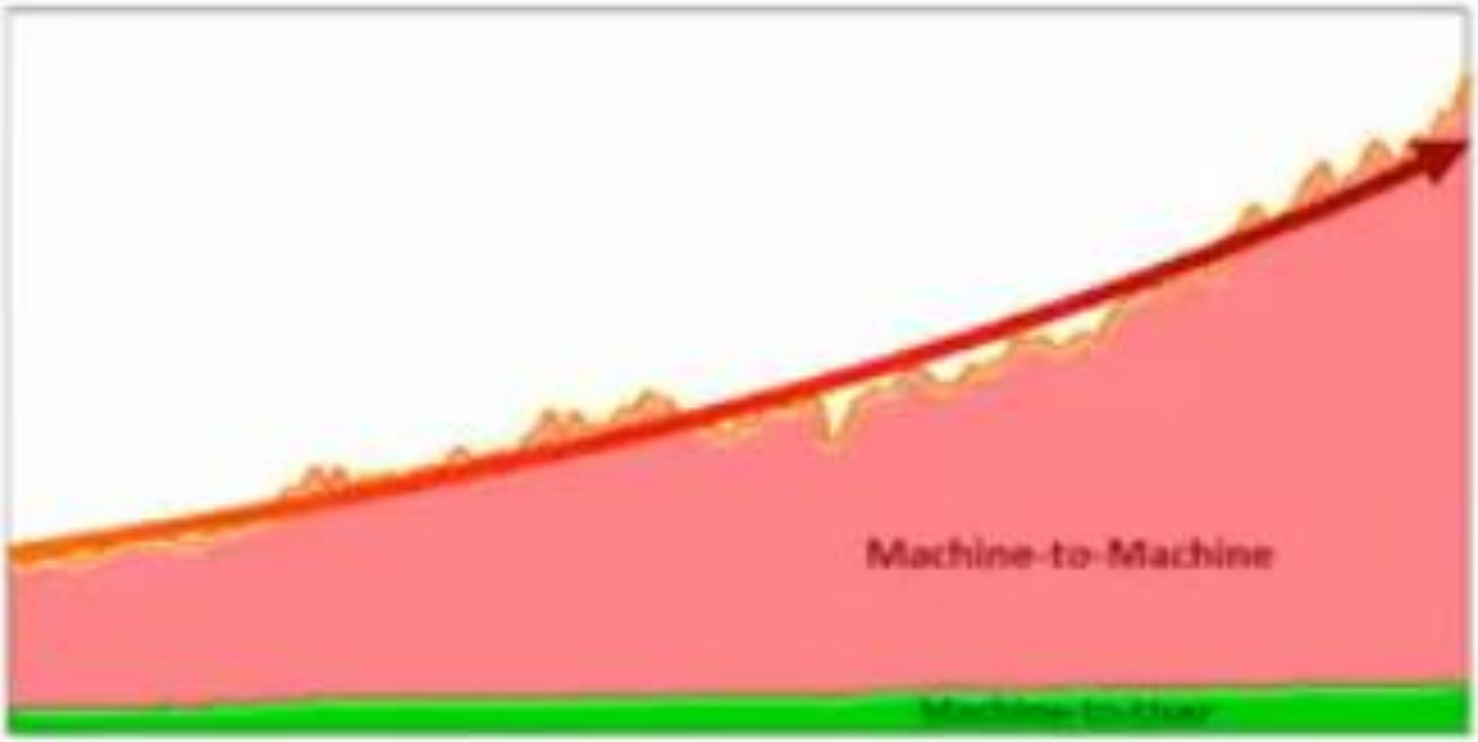
Web-scale / hyper-scale sítě jsou jiné

Source: Google and CBS.

Source: Facebook

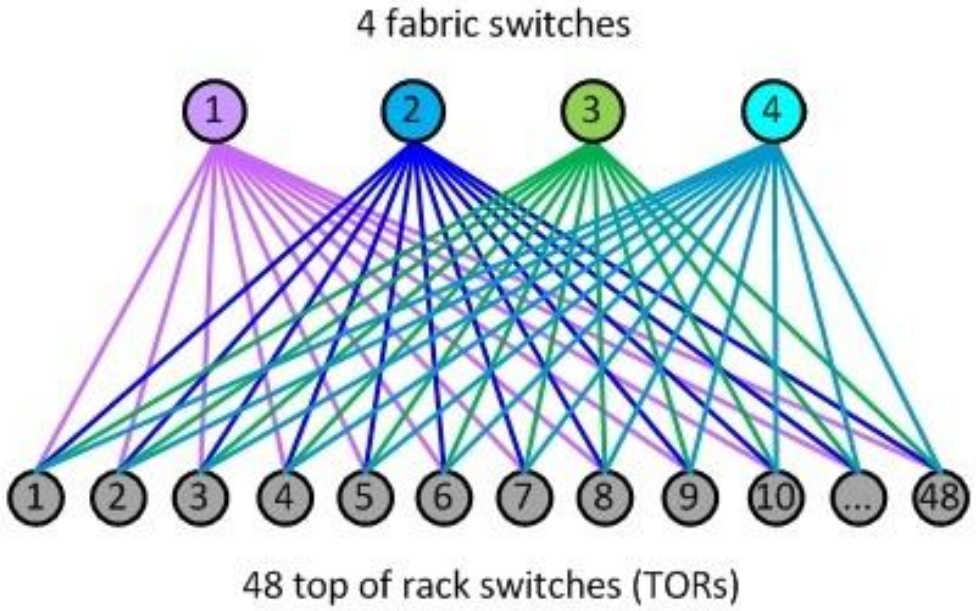
- Mnohem jednodušší sítě (CLOS, L3)
- Jednotný hardware (výnosy z rozsahu)
- Velmi modulární architektura pro rychlý růst
- Omezený set aplikací, vyladění sítě namíru
- Velmi časté změny, DevOps
- Extrémně školení a zdatní lidé

Naprostá většina provozu zůstává uvnitř DC

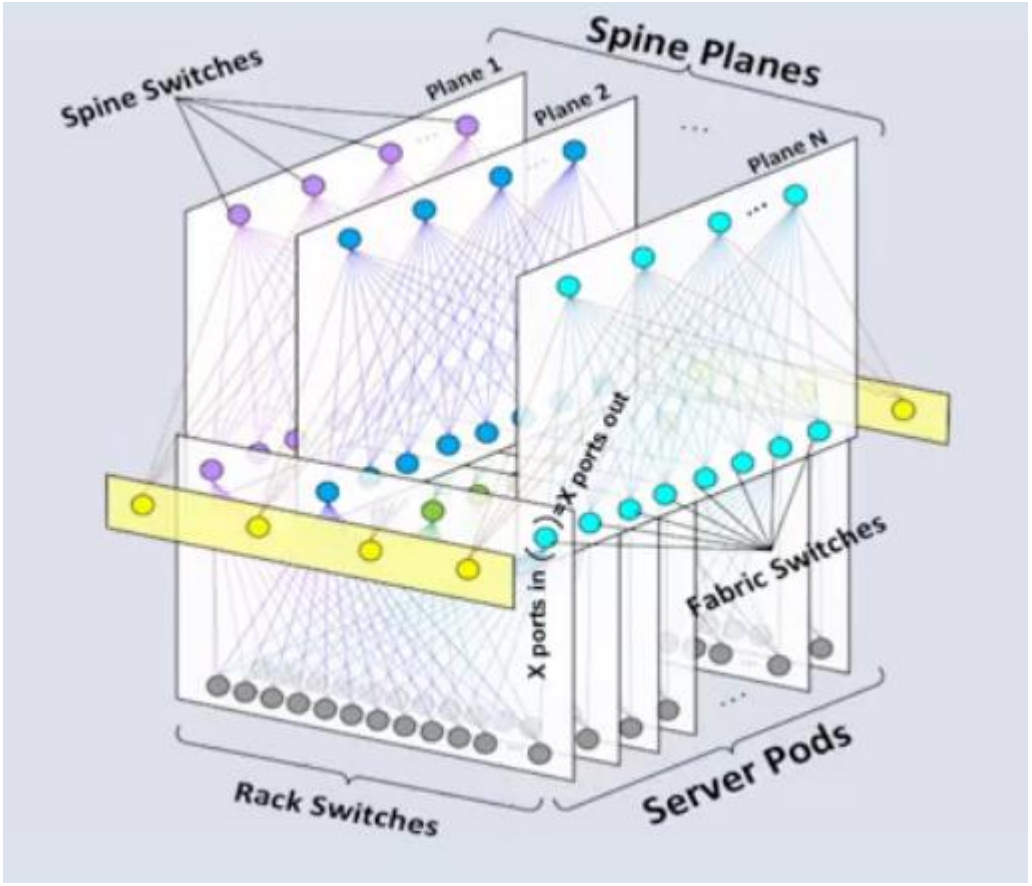


Příklad: Facebook

Facebook

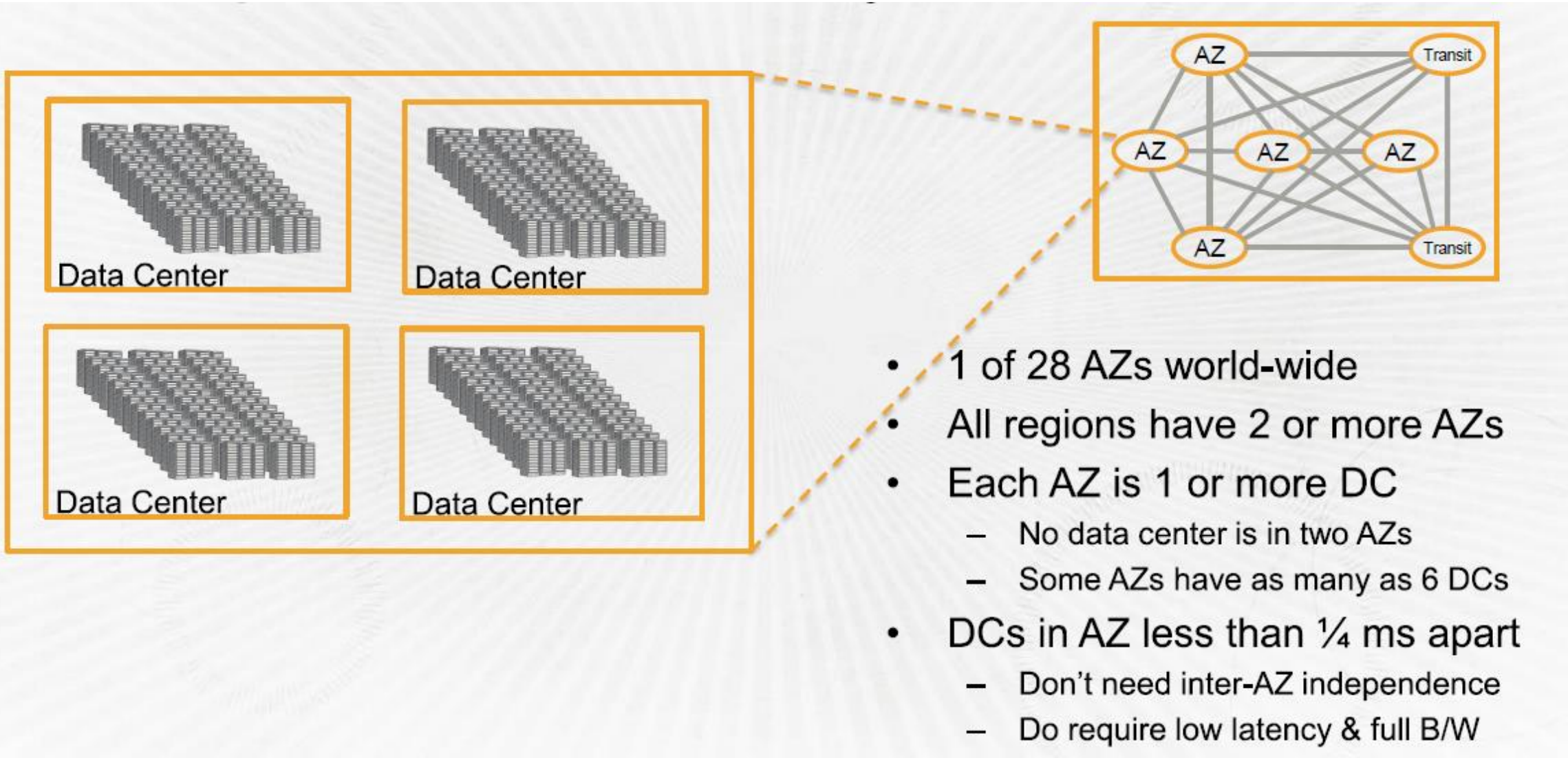


Source: Facebook.



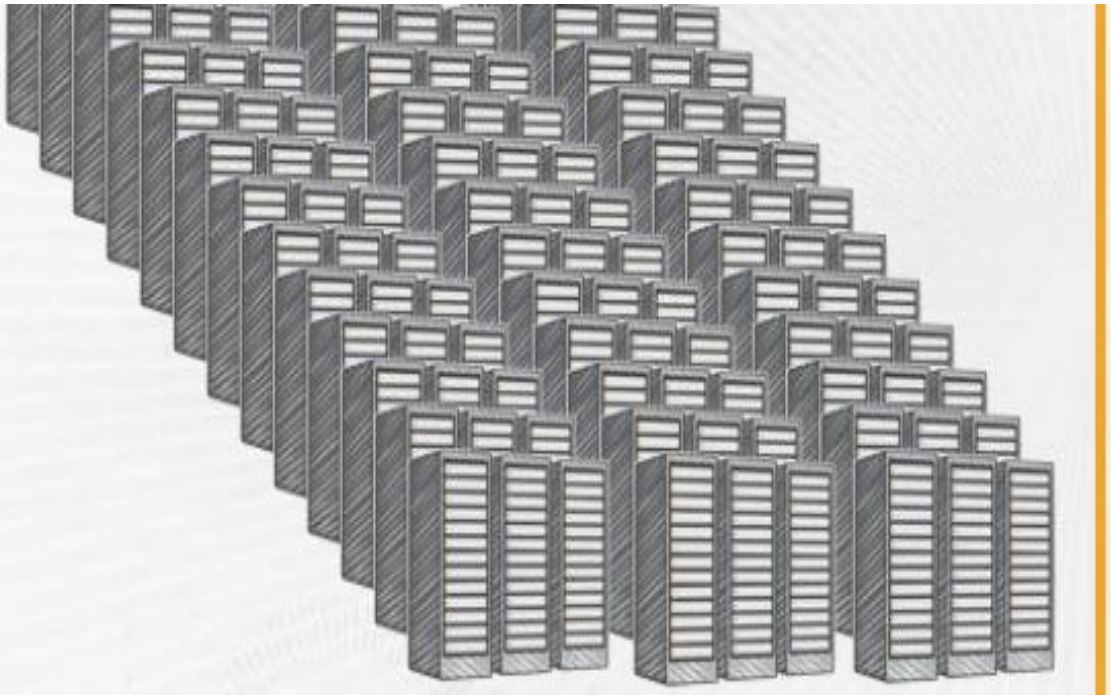
Stovky tisíc 10G portů, non-blocking rack-to-rack

Datové centrum Amazon AWS



Source: AWS.

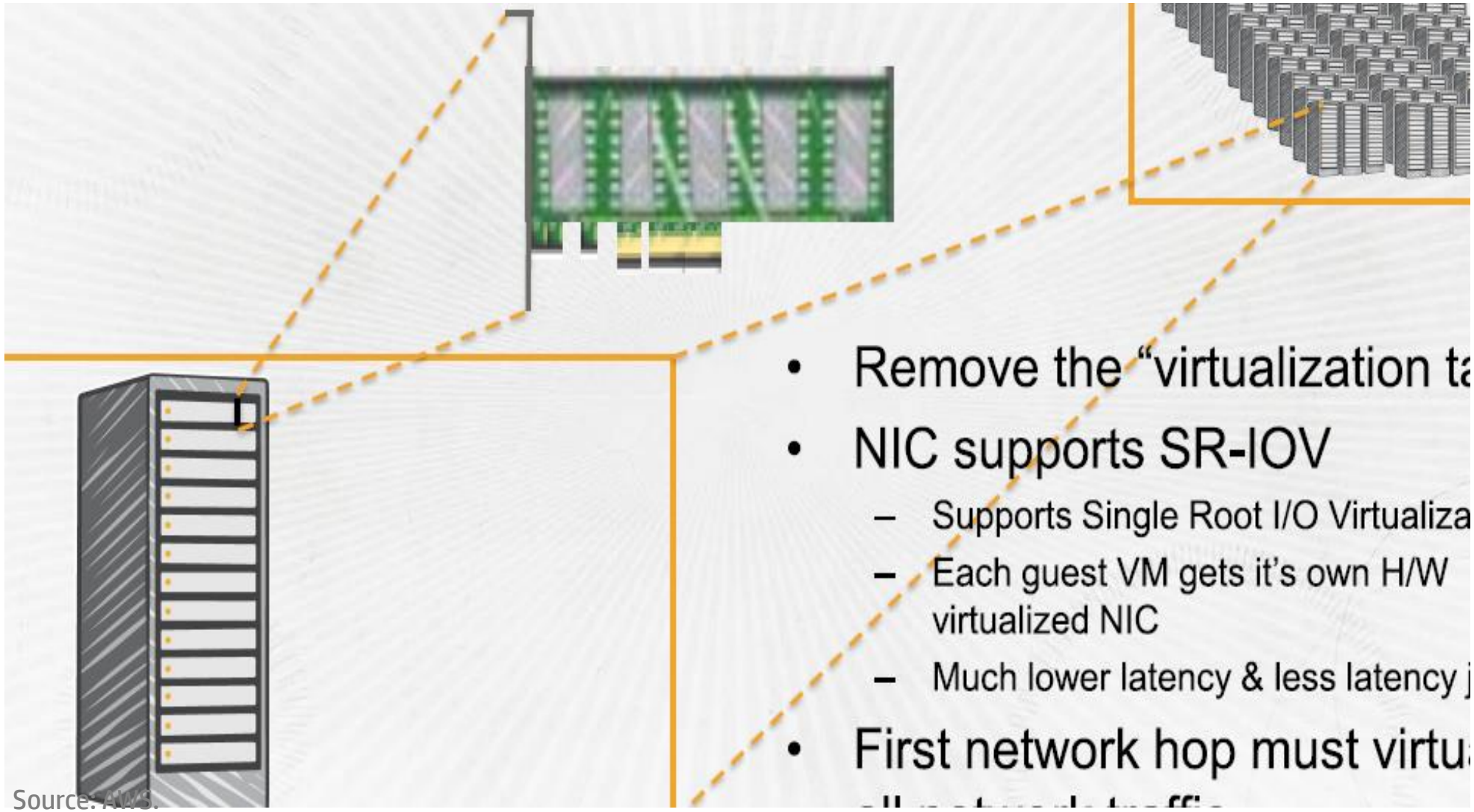
Datové centrum Amazon AWS



- Single DC typically over 50,000 servers & often over 80,000
  - Larger DCs undesirable (blast radius)
- Up to 102Tbps provisioned to a single DC
- AWS custom network equipment:
  - Multi-ODM sourced

Source: AWS.

Datové centrum Amazon AWS



Source: AWS.

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

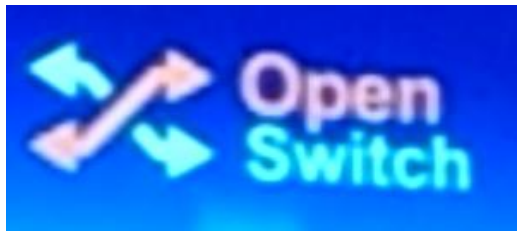
Network Virtualization

Network Function Virtualization

Disagregace

Open source

DevOps



Open Switch

První open source OS

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking

Network Virtualization

Network Function Virtualization

Disagregace

Open source

DevOps

CI/CD a Infrastructure as code
Co to znamená pro síť?

Síť uvádíte do požadovaného stavu jeho popisem
Např. YAML playbook v Ansible

Dokumentací sítě je Infrastructure as code
Místo obrázku manifest/YAML, který lze spustit a stav vynutit

Manifest je ve version control systému
Všechno musí být například v Git nebo SVN

Nejsou přípustné jiné ruční změny konfigurace
Change se implementuje změnou manifestu, ne přímo přes SSH

CI/CD a Infrastructure as code
Proč?

Daleko méně chyb a nekonzistencí
Je dokázáno, že člověk ve stejné roli udělá daleko víc chyb

Skutečný stav vždy odpovídá záměru a dokumentaci
Dokumentace není na čtení, není špatně pochopena nebo neprovedena, je živá

Bezpečák je naprosto nadšený
Všechny změny jsou evidované včetně osob a na jednom místě

Vždy to dopadne stejně (Dobře? Špatně? Opakovatelně.)
Pokud se desired state rozhodí (výměna odumřelého boxu), snadno ho vrátíte zpět bez chyb

CI/CD a Infrastructure as code
Co je co a co je vhodné pro networking?

Configuration Management
Ansible
Salt
Chef
Puppet

Interface prvků
RESTful
NETCONF
OpenFlow
SNMP
CLI

NetVirt
OpenStack Heat
Docker SocketPlane
Docket libnetwork
DCN/Nuage
NSX
Contrail

Version Control
Git
SVN

Review Control
Gerrit

CI/CD automation
Jenkins

Ukažme si Ansible v praxi …
… na „nejmenovaném“ prvku

Pozn.: potřebujete od výrobce moduly, tedy jakési drivery a příkazy

Soubor s proměnnými, abych měl důležité informace na jednom místě

```
---
  vlans:
    - id: 10
      name: Finance
      leftip: 10.10.0.1
      rightip: 10.10.0.2
      mask: 255.255.255.0
      vip: 10.10.0.254

    - id: 20
      name: HR
      leftip: 10.20.0.1
      rightip: 10.20.0.2
      mask: 255.255.255.0
      vip: 10.20.0.254

    - id: 30
      name: Printers
      leftip: 10.30.0.1
      rightip: 10.30.0.2
      mask: 255.255.255.0
      vip: 10.30.0.254

    - id: 40
      name: Guests
      leftip: 10.40.0.1
      rightip: 10.40.0.2
      mask: 255.255.255.0
      vip: 10.40.0.254
```

```
[all:vars]
username=admin
password=admin

[core]
10.0.0.11
10.0.0.12

[leftcore]
10.0.0.11

[rightcore]
10.0.0.12

[leftcore:vars]
vrrp_priority=250
ISL_port=GigabitEthernet2/0

[rightcore:vars]
vrrp_priority=100
ISL_port=GigabitEthernet2/0
```

Hosts file, tedy seznam prvků v síti, jejich role, login

Můj playbook

```
---

  - name: Configure basics on all devices
    hosts: all
    gather_facts: no
    connection: local
    vars_files:
      - vlans.yml

    tasks:

      - xxx_facts: username={{ username }} password={{ password }}
hostname={{ inventory_hostname }}

      - name: Turn on LLDP
        xxx_command:
          command:
          - lldp global enable
          type: config
          username: "{{ username }}"
          password: "{{ password }}"
          hostname: "{{ inventory_hostname }}"

      - name: ensure VLANs exist
        xxx_vlan:
          vlanid: "{{ item.id }}"
          state: present
          name: "{{ item.name }}"
          username: "{{ username }}"
          password: "{{ password }}"
          hostname: "{{ inventory_hostname }}"
        with_items:
          "{{ vlans }}"
```

```
  - name: Configure core devices
    hosts: core
    gather_facts: no
    connection: local
    vars_files:
      - vlans.yml

    tasks:

      - name: ensure L3 VLAN interfaces exist
        xxx_interface:
          state: present
          name: Vlan-interface{{ item.id }}
          username: "{{ username }}"
          password: "{{ password }}"
          hostname: "{{ inventory_hostname }}"
        with_items:
          "{{ vlans }}"

      - name: Configure inter-switch link port type
        xxx_interface:
          name: "{{ ISL_port}}"
          type: bridged
          username: "{{ username }}"
          password: "{{ password }}"
          hostname: "{{ inventory_hostname }}"

      - name: Enable all VLANs on ISL
        xxx_switchport:
          name: "{{ ISL_port}}"
          link_type: trunk
          permitted_vlans: "1-1000"
          username: "{{ username }}"
          password: "{{ password }}"
          hostname: "{{ inventory_hostname }}"
```

Můj playbook

```
- name: Configure LEFT core device
  hosts: leftcore
  gather_facts: no
  connection: local
  vars_files:
    - vlans.yml

  tasks:

    - name: ensure IP information is configured
      xxx_ipinterface:
        state: present
        name: Vlan-interface{{ item.id }}
        addr: "{{ item.leftip }}"
        mask: "{{ item.mask }}"
        username: "{{ username }}"
        password: "{{ password }}"
        hostname: "{{ inventory_hostname }}"
      with_items:
        "{{ vlans }}"

    - name: configure VRRP instances
      xxx_vrrp:
        vrid: "{{ item.id }}"
        vip: "{{ item.vip }}"
        priority: "{{ vrrp_priority }}"
        interface: Vlan-interface{{ item.id }}
        username: "{{ username }}"
        password: "{{ password }}"
        hostname: "{{ inventory_hostname }}"
      with_items:
        "{{ vlans }}"
```

```
- name: Configure RIGHT core device
  hosts: rightcore
  gather_facts: no
  connection: local
  vars_files:
    - vlans.yml

  tasks:

    - name: ensure IP information is configured
      xxx_ipinterface:
        state: present
        name: Vlan-interface{{ item.id }}
        addr: "{{ item.rightip }}"
        mask: "{{ item.mask }}"
        username: "{{ username }}"
        password: "{{ password }}"
        hostname: "{{ inventory_hostname }}"
      with_items:
        "{{ vlans }}"

    - name: configure VRRP instances
      xxx_vrrp:
        vrid: "{{ item.id }}"
        vip: "{{ item.vip }}"
        priority: "{{ vrrp_priority }}"
        interface: Vlan-interface{{ item.id }}
        username: "{{ username }}"
        password: "{{ password }}"
        hostname: "{{ inventory_hostname }}"
      with_items:
        "{{ vlans }}"
```

Můj playbook

```
- name: Save configurate on all devices
  hosts: all
  gather_facts: no
  connection: local
  vars_files:
    - vlans.yml

  tasks:

    - name: save configuration to flash
      xxx_save:
        username: "{{ username }}"
        password: "{{ password }}"
        hostname: "{{ inventory_hostname }}"
```

```
root@ubuntu:~/ansible-demo# ansible-playbook -i hosts tomas.yml
```

Nemluvme jen o SDN....
... Mluvme o nové síťařině

Software-defined Networking
Network Virtualization
Network Function Virtualization
Disagregace
Open source
DevOps

+

Nemluvme jen o SDN....
... Mluvme o tom, jak na to

- Na nic nečekejte, „Uber“ pro váš trh může přijít kdykoli
- Založte innovation labs nebo bimodal IT
- Kontinuita a stávající předpoklady jsou důležité pro provoz, ale svazující pro design a rozvoj váš i vaší sítě
- A především: čtěte cloudsvet.cz ☺

Tomáš Kubica a Daniel Prchal
Hewlett-Packard Enterprise