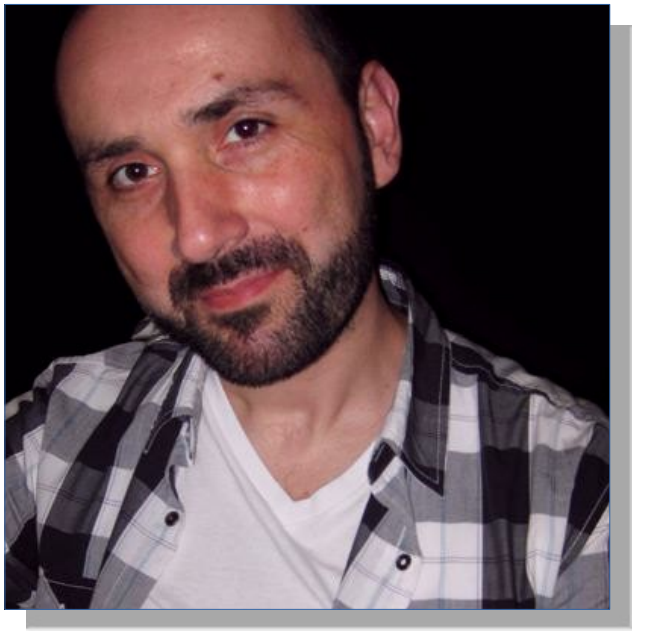
A może VXLAN?

30.09.2023

Jacek Rokicki

- sys/net/dev ops,
- w IT na poważniej od 1998,
- entuzjasta systemów operacyjnych z rodziny *nix,
- odkryłem MikroTika w 2011,
- architekt wysoko dostępnych rozwiązań z wykorzystaniem Libre/Open Source,
- na co dzień wspieram rozwój kilku popularnych platform OTT.



GitHub yacq/MBUM

jacek-rokicki

MikroTik
MTCNA
MTCRE
MTCINE
MTCSWE
MTCSE
MTCTCE
MTCIPv6E
MTCUME
MTCWE
MTCEWE

Agenda

- Kilka słów o historii rozwiązania
- Charakterystyka
- Porównanie z innymi rozwiązaniami L2
- Zalety i wady
- Jak to właściwie działa
- Jakie systemy/urządzenia wspierają
- Implementacja w ROS
- Live demo
- Zakończenie

Jak powstawał VXLAN

- Xsigo Systems w 2011 wprowadza własnościowe rozwiązanie podobne do VXLAN
- VMware w sierpniu 2011 na konferencji VMworld prezentuje VXLAN – owoc współpracy z Cisco i Aristą. Tworzy się draft standardu.
- Po 3 latach w sierpniu 2014 zostaje zatwierdzony standard opisany w RFC7348

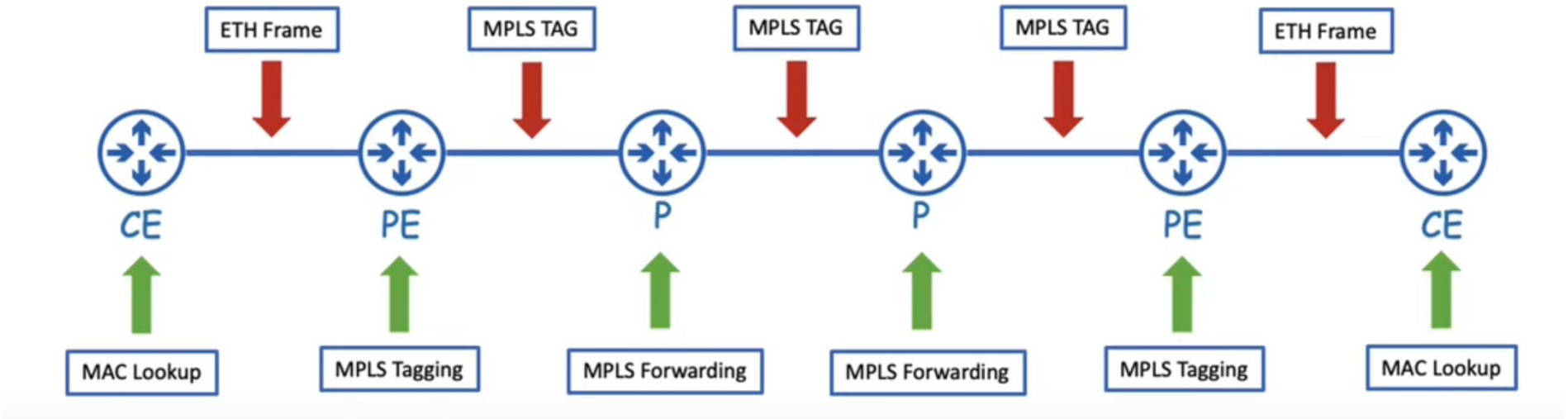
Charakterystyka

- Używa warstwy 3 jako sieci transportowej
- Wykorzystuje UDP i standardowo port 4789 (w ROS 8472)
- Identyfikator sieci VXLAN ID (VNI) zajmuje 24bity - >16mln możliwości
- Dodaje 50 bajtów do pakietu IPv4 oraz 70 dla IPv6
- Pozwala tworzyć połączenia jeden do wielu
- Zakończenia tuneli VTEP (VXLAN tunnel endpoint) odpowiadają za encapsulacje/decapsulacje i wysłanie pakietu w odpowiednie miejsce

Porównanie z VLAN

- Rozmiar identyfikatora: 24 vs 12bit
- Działa w L3 vs L2
- Redundancja poprzez rozwiązania L3 (BGP, ECMP) vs STP oraz 802.3ad
- Wysoka elastyczność, VTEP-y są przenaszalne w ramach sieci
- Wyższy koszt sprzętu niż w przypadku VLAN
- Wymaga większego MTU, lekko spada wydajność przez większy narzut w nagłówkach

Porównanie z MPLS



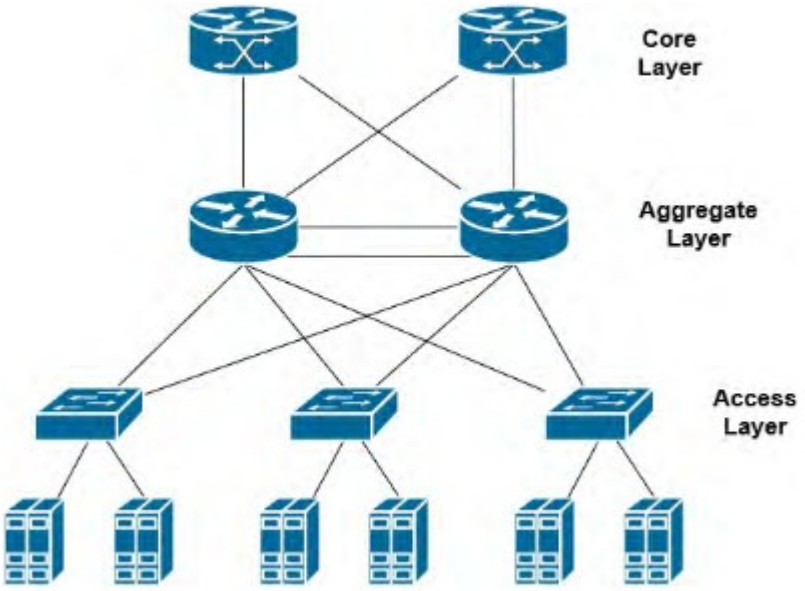
- Każdy router musi obsługiwać znacznik, nie można szybko wdrożyć na istniejącej sieci IP

Zalety

- Rozwiązuje problem dużej ilości ruchu w szkielecie sieci
- Eliminuję prowadzenie dużej ilości połączeń L2 w szkielecie sieci – brak STP, brak niewykorzystanych łączy, brak propagacji awarii
- Rozwiązuje problem pojemności tablic mac address przełączników
- Separuje architekturę sieci koniecznych do pracy wirtualek od architektury sieci data center – zespoły administratorów mogą pracować oddzielnie, zmniejszone ryzyko pomyłek
- Umożliwia tworzenie odseparowanych sieci L2 dla dużej ilości klientów

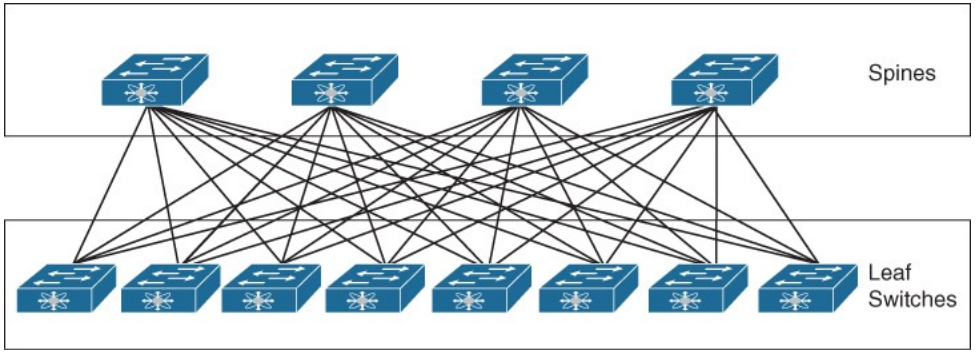
Wady

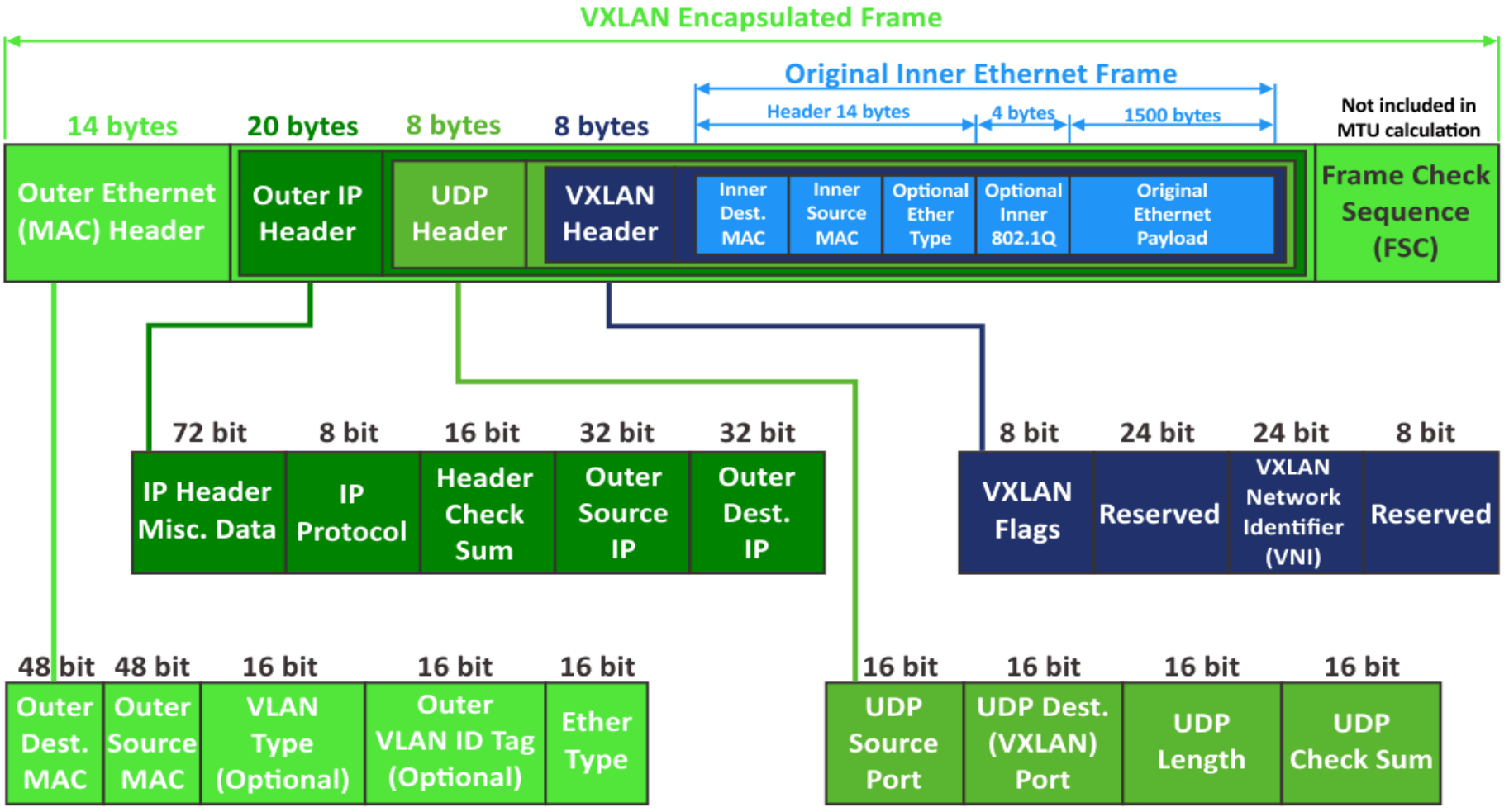
- Wdrożenie i wstępna konfiguracja może być skomplikowana
- Potrzebuję wsparcia na przełącznikach w warstwie dostępowej lub/i przełączniku wirtualnym
- Sieć szkieletowa powinna obsługiwać ramki jumbo aby uniknąć degradującej wydajność fragmentacji (nagłówek VXLAN to co najmniej dodatkowe 50bajtów)
- Sieć transportowa powinna obsługiwać muticasty



- Duża ilość ruchu N-S
- L2 w warstwie dystrybucyjnej
- STP i niewykorzystane łącza
- Propagacja awarii

- Tylko L3 w szkielecie sieci
- Do komunikacji w L2 potrzeba sieci overlay (VXLAN)




VXLAN Encapsulated Frame
Original Inner Ethernet Frame
Header 14 bytes
4 bytes
1500 bytes
Not included in MTU calculation
14 bytes
20 bytes
8 bytes
8 bytes
Outer Ethernet (MAC) Header
Outer IP Header
UDP Header
VXLAN Header
Inner Dest. MAC
Inner Source MAC
Optional Ether Type
Optional Inner 802.1Q
Original Ethernet Payload
Frame Check Sequence (FSC)
72 bit
8 bit
16 bit
32 bit
32 bit
IP Header Misc. Data
IP Protocol
Header Check Sum
Outer Source IP
Outer Dest. IP
8 bit
24 bit
24 bit
8 bit
VXLAN Flags
Reserved
VXLAN Network Identifier (VNI)
Reserved
48 bit
48 bit
16 bit
16 bit
16 bit
Outer Dest. MAC
Outer Source MAC
VLAN Type (Optional)
Outer VLAN ID Tag (Optional)
Ether Type
16 bit
16 bit
16 bit
16 bit
UDP Source Port
UDP Dest. (VXLAN) Port
UDP Length
UDP Check Sum

Nagłówek VXLAN składa się z czterech pól:

- Pole 1: zarezerwowane na potrzeby przyszłych zastosowań, 8 bitów, wszystkie wyzerowane
- Pole 2: VNI, 24 bity
- Pole 3: zarezerwowane na potrzeby przyszłych zastosowań, 24 bity
- Pole 4: flagi, 8 bitów ustawionych na 0 z wyjątkiem bitu trzeciego, który jest ustawiony na binarne 1 i oznacza poprawny nagłówek VXLAN

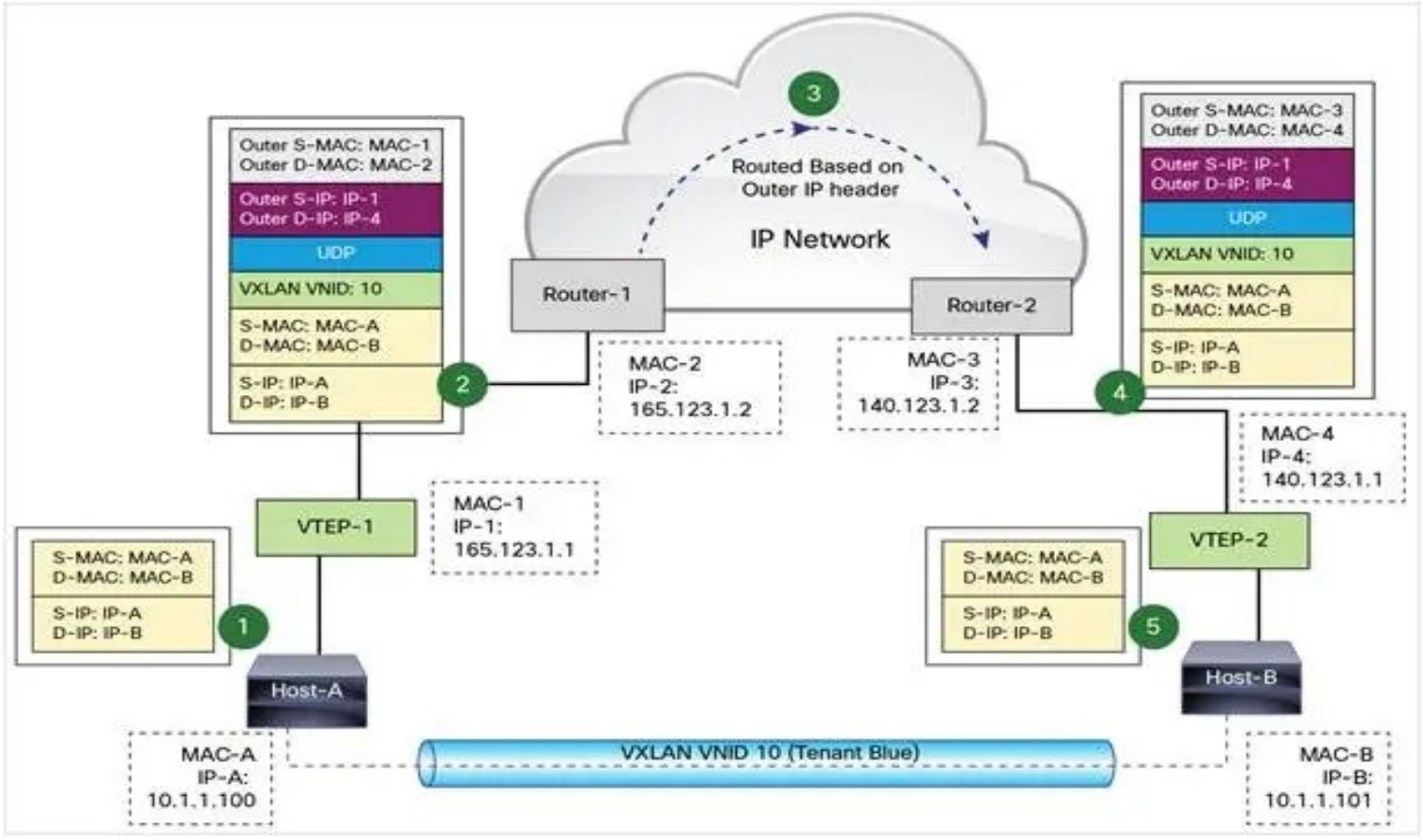
Narzut podczas korzystania z VXLAN:

- IPv4

  1514 (wewnętrzna ramka) + 4 (wewnętrzny znacznik VLAN) + 50 (VXLAN) + 4 (VXLAN Transport VLAN Tag) = 1572 bajty

- IPv6

  1514 (ramka wewnętrzna) + 4 (wewnętrzny znacznik VLAN) + 70 (IPv6 VXLAN) + 4 (znacznik VXLAN Transport VLAN) = 1592 bajty


3
Routed Based on
Outer IP header
IP Network
Router-1
Router-2
Outer S-MAC: MAC-1
Outer D-MAC: MAC-2
Outer S-IP: IP-1
Outer D-IP: IP-4
UDP
VXLAN VNID: 10
S-MAC: MAC-A
D-MAC: MAC-B
S-IP: IP-A
D-IP: IP-B
2
MAC-2
IP-2:
165.123.1.2
MAC-3
IP-3:
140.123.1.2
Outer S-MAC: MAC-3
Outer D-MAC: MAC-4
Outer S-IP: IP-1
Outer D-IP: IP-4
UDP
VXLAN VNID: 10
S-MAC: MAC-A
D-MAC: MAC-B
S-IP: IP-A
D-IP: IP-B
4
MAC-4
IP-4:
140.123.1.1
VTEP-1
MAC-1
IP-1:
165.123.1.1
VTEP-2
S-MAC: MAC-A
D-MAC: MAC-B
S-IP: IP-A
D-IP: IP-B
1
S-MAC: MAC-A
D-MAC: MAC-B
S-IP: IP-A
D-IP: IP-B
5
Host-A
Host-B
MAC-A
IP-A:
10.1.1.100
VXLAN VNID 10 (Tenant Blue)
MAC-B
IP-B:
10.1.1.101

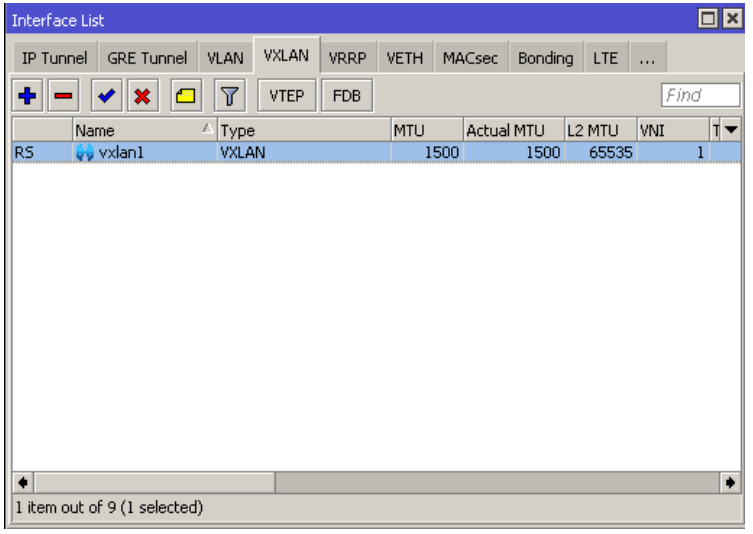
- Zakładamy, że obie strony znają już mapowania między MAC a VTEP
- Host A tworzy ramkę z MAC Host B i przesyła do VTEP-1
- VTEP-1 posiada mapowanie MAC Host B do VTEP-2 i przeprowadza proces encapsulacji. W zew. nagłówku IP źródłem jest VTEP-1 a celem VTEP-2
- Pakiet jest przesyłany przez sieć transportową do VTEP-2
- VTEP-2 dekapsuluje pakiet zdejmując kolejne warstwy i przekazuję ramkę do Hosta B bazując na docelowym adresie MAC.

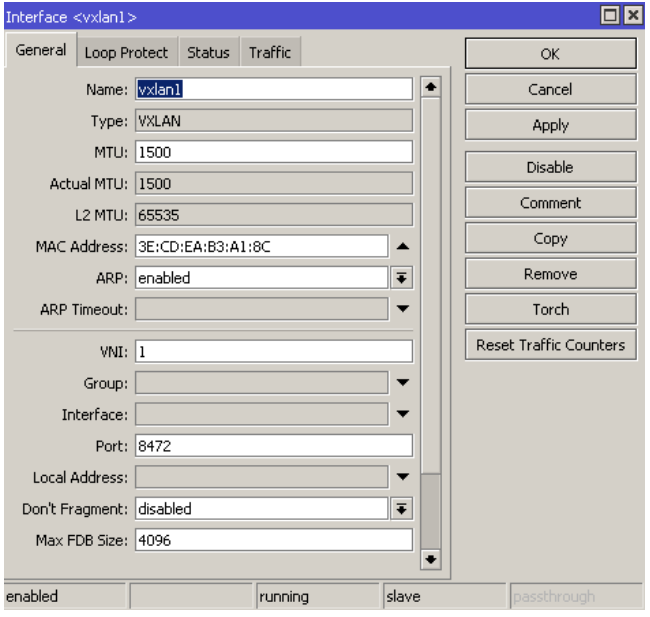
Wsparcie w sprzęcie i oprogramowaniu

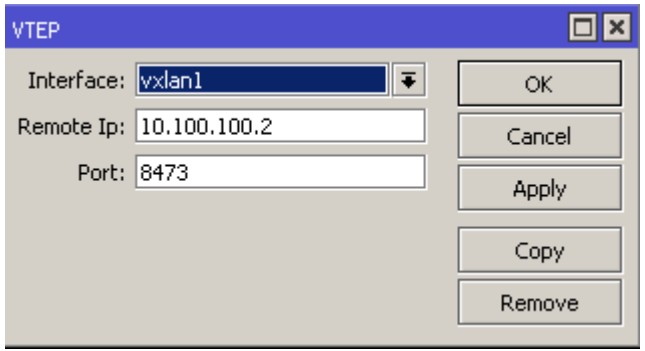
- MikroTik! :)
- Przełączniki vendorów standaryzujących rozwiązanie (Cisco, Arista) oraz większość urządzeń klasy datacenter
- VMware ESX
- MS Hyper-V
- OpenStack (Neutron)
- Linux w wersji >3.7
- Windows za pomocą opensource-owego UBridge-a

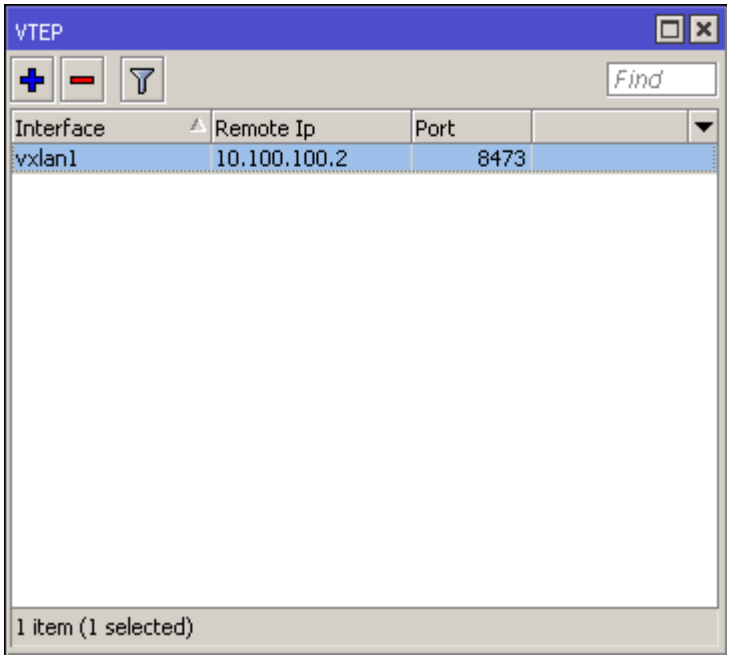
- /interface vxlan

  add name=vxlan1 port=8472 vni=1





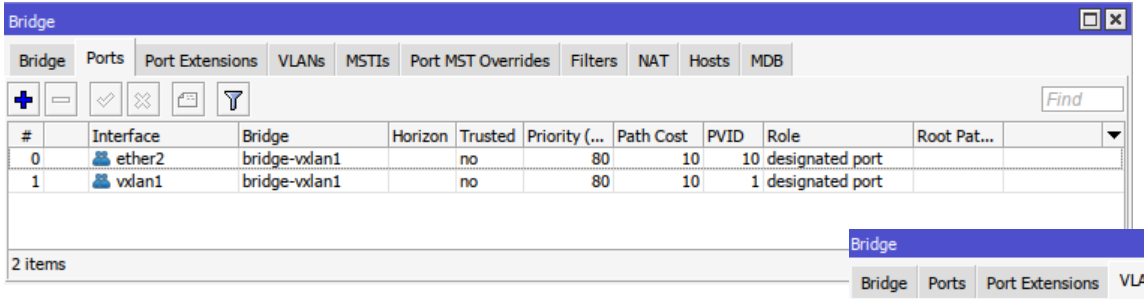


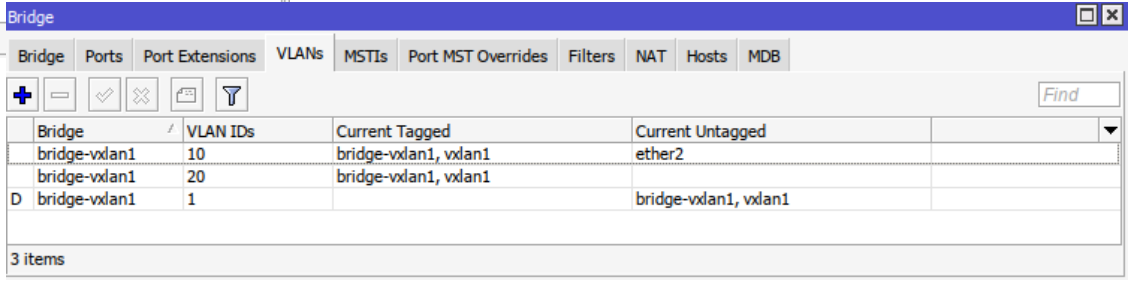


- /interface vxlan vteps

  add interface=vxlan1 remote-ip=10.100.100.2

- Powstał wirtualny interfejs, który możemy potraktować w zależności od potrzeb:
  - zmostkować z innymi (np. ether2, wlan1 itd.)
  - wykorzystać powstały interfejs jako nośnik dla vlan-ów. Wystarczy włączyć obsługę w mostku oraz odpowiednio skonfigurować dodane do niego interfejsy w zakładce VLAN.





Live demo
/
pytania

Dziękuję za uwagę