ISP Architecture – Deploy virtualized public BGP routers with CHR for large scale transit peering.

KEVIN MYERS, NETWORK ARCHITECT / MANAGING PARTNER

MTCINE #1409

MIKROTIK CERTIFIED TRAINER
Kevin Myers, Network Architect

Jackson, Mississippi – United States

18 + years in IT, Network Architecture and Engineering

Areas of Design Focus:
- MikroTik integration with large multi-vendor networks
- Design/Implement/Operate BGP/MPLS/OSPF Wireline and WISP service provider networks
- Design/Implement/Operate Data Center (Enterprise and Cloud) networks

Certifications
- MTCINE #1409 & MikroTik Certified Trainer
- MikroTik – MTCWE, MTCUME, MTCRE, MTCTCE, MTCNA
- Cisco/Microsoft – CCNP, CCNA, MCP
www.iparchitechs.com

Global Leaders in MikroTik Design and Engineering

#1 ranked MikroTik consulting firm in North America

The most successful MikroTik global integrator – we bill thousands of hours in MikroTik engineering across 6 continents.

The first consulting firm to offer 24/7 MikroTik technical assistance with enterprise level SLAs

Operate at large scale supporting networks with tens of thousands or routers, switches, firewalls, etc
• www.iparchitechs.com

• Our Services

• **Global Professional Services** – Consulting for Design, Engineering, Integration and Operations

• **Fully Managed Network Services** - per rack unit support for full network management and monitoring

• **24/7 support contracts per device** – support all MikroTik devices with 24/7 TAC support and 4 hour SLAs.

• **MultiLingual Support in:** English, Français, Polski, Español
**Objectives**

- Identify the use case for virtualizing public BGP routers and providing full table peerings and transit.

- Discuss a practical design with a small number of upstream BGP providers.

- Discuss larger scale applications with many upstream BGP providers.

- Overview of using the CHR in VMWARE ESXi with 10 Gbps or more of traffic.
• **Definitions for Virtualization**

• **Hypervisor** – A hypervisor or virtual machine monitor (VMM) is a piece of computer software, firmware or hardware that creates and runs **virtual machines**.

• **Paravirtualized NIC** – Paravirtual drivers are ones where the virtualization platform does not have to emulate another device, such as an Intel E1000 NIC. These paravirtual drivers cut the extra overhead out by ditching the emulation layer, which usually results in significant performance increases.

• **vSwitch** – virtual software switch in the hypervisor that handles VLAN tagging and VM to VM communication
• **What problem are we trying to solve?**

• **Resource Utilization** – Currently, RouterOS only utilizes one core for BGP which can become a bottleneck when there are many peerings and routes. CHR can run on a more powerful CPU and make better use of one core.

• **Scaling Transit** – Using virtual routers to provide transit and peering allows an ISP to install hardware much less often to serve new customers. Peering CHR routers can be brought up once a current CHR is full.

• **Cost** – By using the same Hypervisor platform, new CHRs can be deployed much cheaper than adding a CCR without a waste of resources for this use case.

• **Redundancy** – Multiple hypervisors allows for a single router instance to become highly available across multiple hardware platforms.
Design Overview – simple topology:

CHR BGP PE – Simple Topology - Overview

- IX-Europe Amsterdam
- IX-NorthAmerica New York

- IPA-MUM-BGP-1
  CCR1036-8G-2S+
- IPA-MUM-BGP-2
  CCR1036-8G-2S+

- IPA-MUM-ESXI-1
  HP ProLiant DL360
- IPA-MUM-AGGREGATION-1
  CCR1009-8G-1S-1S+PC
- IPA-MUM-CORE-1
  CCR1072-1G-8S+
- IPA-MUM-BGP-CE-1
  RB3011UiAS

- 1 Gbps Copper
- 10 Gbps Fiber
• Virtualization – CHR vs x86

• Why use the CHR instead of the traditional x86 VM?

  • **Paravirtualized NIC** – Using the CHR allows us to use the a paravirtualized NIC such as VMWARE’s VMXNET3 which is capable of speeds beyond 10 Gbps. The E1000 NIC used in the x86 VM is only capable of 1 Gbps.

  • **Optimized for Virtualization** –
    • 64 bit support
    • Fastpath support
    • Driver support

  • **Future enhancements** – The CHR will continue to be developed and improved
Virtualization – Deploying CHR in ESXi

Currently, the CHR has to be deployed in another hypervisor and then exported to be used in ESXi.
Virtualization – ESXi -

Use the VMXNET3 paravirtualized NIC for the best performance and 10 Gpbs + performance
Design Overview – logical topology:

![Logical Topology Diagram](diagram.png)
• Using VPLS to deliver a direct L2 handoff for transit

• Why not advertise the full BGP table throughout the network or use MPLS L3 VPN?

• Resource utilization
  • Router memory available affects supported routing table size
  • Performance – convergence will be much slower than dedicated VMs once you add more customers

• Isolation/Security
  • Allows completely segregated public transport without exposing the underlying MPLS core
  • Allows for more granular segregation of customers using VLANs

• EOIP is a viable alternative for non-MPLS networks.
• Upstream Provider #1 – IX Europe Amsterdam

• Using BGP VM for full IPv4 table from www.stubarea51.net

```
[admin@IPA-MUM-BGP-1] > routing bgp peer print status
Flags: X - disabled, E - established
0 E name="IX-Europe-Amsterdam" instance=default remote-address=100.99.98.1 remote-as=65000 tcp-md5-key="" nexthop-choice=default multihop=no route-reflect=no hold-time=30m keepalive-time=3m ttl=default in-filter="" out-filter=BGP-OUT address-families=ip default-originate=never remove-private-as=no as-override=no passive=no use-bfd=no remote-id=100.99.98.1 local-address=100.99.98.2 uptime=2h5m54s prefix-count=293364 updates-sent=0 updates-received=3696861 withdrawn-sent=0 withdrawn-received=0 remote-hold-time=30m used-hold-time=30m used-keepalive-time=3m state=established

1 E name="IPA-MUM-BGP-2" instance=default remote-address=1.1.1.2 remote-as=65530 tcp-md5-key="" nexthop-choice=default multihop=no route-reflect=no hold-time=3m ttl=default in-filter="" out-filter="" address-families=ip update-source=Lol default-originate=never remove-private-as=no as-override=no passive=no use-bfd=no remote-id=100.99.98.6 local-address=1.1.1.1 uptime=22h53m43s prefix-count=175749 updates-sent=3811350 updates-received=2329918 withdrawn-sent=3568244 withdrawn-received=2139684 remote-hold-time=3m used-hold-time=3m used-keepalive-time=1m refresh-capability=yes as4-capability=yes state=established

2 E name="IPA-MUM-BGP-PE-1" instance=default remote-address=1.1.1.5 remote-as=65530 tcp-md5-key="" nexthop-choice=default multihop=no route-reflect=no hold-time=3m ttl=default in-filter="" out-filter="" address-families=ip update-source=Lol default-originate=never remove-private-as=no as-override=no passive=no use-bfd=no remote-id=1.1.1.5 local-address=1.1.1.1 uptime=1dh43m33s prefix-count=2 updates-sent=3936313 updates-received=2 withdrawn-sent=3689815 withdrawn-received=0 remote-hold-time=3m used-hold-time=3m used-keepalive-time=1m refresh-capability=yes as4-capability=yes state=established
```
• Upstream Provider #2 – IX North America NYC

• Using BGP VM for full IPv4 table from www.stubarea51.net
• **BGP PE VM – BGP Routes**

• PE Router takes in a full table from each provider and advertises the best routes to the CE router
• BGP CE Router – BGP Routes

• Full BGP Table across both upstreams is advertised to the transit customer without carrying a full BGP table throughout the network.

```
[admin@IPA-MUM-BGP-CE-1] > routing bgp peer print status
Flags: X - disabled, E - established
0 E name="IPA-MUM-BGP-PE-1" instance=default remote-address=100.101.102.1 remote-as=65530 tcp-md5-key="" nexthop-choice=default multihop=no route-reflect=no hold-time=3m ttl=default in-filter="" out-filter=""
address-families=ip default-originate=never remove-private-as=no as-override=no passive=no use-bfd=no remote-id=1.1.1.5 local-address=100.101.102.2 uptime=23h41m28s prefix-count=400335
updates-sent=1 updates-received=6354378 withdrawn-sent=0 withdrawn-received=4599799 remote-hold-time=3m used-hold-time=3m used-keepalive-time=1m refresh-capability=yes as4-capability=yes
state=established
```
• Core Router – # of Routes

• CE router is receiving 400,000+ routes but the network core has a small routing table which improves convergence speed and performance

• Core has 14 routes to transport 400,000 routes!!
• **BGP PE - Scaling**

• **How to scale using the BGP PE?**

  • Add more peerings to the CHR BGP PE
    • Depending on the hardware used, we can use approximately 5 to 10 full table peerings per CHR BGP PE

  • Add more CHR BGP PE routers
    • Adding BGP PE routers allows for more customer transit peers
    • Each CHR BGP router must peer back to both BGP edge routers
    • Be careful not to add too many full table peerings to the edge routers...this can drastically affect the performance.

• **Route Reflection**

  • Use RRs to feed multiple PEs
Design Overview – Large scale topology:

CHR BGP PE – Large Scale Topology - Overview

- **1 Gbps Copper**
- **10 Gbps Fiber**

**IX-Europe Amsterdam**
- **IPA-MUM-BGP-1**
  - CCR1036-8G-2S+

**IPA-MUM-ESXi-1**
- HP ProLiant DL360

**IPA-MUM-CORE-1**
- CCR1072-1G-8S+

**IPA-MUM-BGP-2**
- CCR1036-8G-2S+

**IX-NorthAmerica New York**
- **IPA-MUM-BGP-CE-1**
  - RB3011UiAS

- **IPA-MUM-BGP-CE-2**
  - RB3011UiAS

- **IPA-MUM-BGP-CE-3**
  - RB3011UiAS

- **IPA-MUM-BGP-CE-4**
  - RB3011UiAS

- **IPA-MUM-BGP-CE-5**
  - RB3011UiAS

- **IPA-MUM-BGP-CE-6**
  - RB3011UiAS

**IPA-MUM-AGGREGATION-1**
- CCR1009-8G-1S-1S+PC

**Powered by MikroTik**
- IPA-MUM-BGP-RR-1
  - CHR 6.34.2
- IPA-MUM-BGP-PE-1
  - CHR 6.34.2
- IPA-MUM-BGP-PE-2
  - CHR 6.34.2
- IPA-MUM-BGP-PE-3
  - CHR 6.34.2

**Virtualized by VMware**
Design Overview – Large scale logical topology:

CHR BGP PE – Logical Topology - Overview

- IX-Europe
  - Amsterdam
  - BGP AS 65000
  - 100.99.98.0/30

- IX-NorthAmerica
  - New York
  - BGP AS 65001
  - 100.99.98.4/30

- IPA-MUM-BGP-1
  - CCR1036-8G-2S+
  - 1.1.1.1

- IPA-MUM-BGP-RR-1
  - CHR 6.34.2

- IPA-MUM-BGP-2
  - CCR1036-8G-2S+
  - 1.1.1.2

- IPA-MUM-BGP-3
  - 1.1.1.3

- IPA-MUM-CORE-1
  - CCR1072-1G-8S+

- IPA-MUM-BGP-PE-1
  - CHR 6.34.2

- IPA-MUM-BGP-PE-2
  - CHR 6.34.2

- IPA-MUM-BGP-PE-3
  - CHR 6.34.2

- VLAN 101 (VPLS)
- VLAN 102 (VPLS)
- VLAN 103 (VPLS)
- VLAN 104 (VPLS)
- VLAN 105 (VPLS)
- VLAN 106 (VPLS)

- BGP AS 65530

- BGP AS Multiple
• Adding BGP Route Reflectors for scalability

Reduce the number of peerings to the BGP border routers and thus resource utilization – in this model, we have a 6 to 1 ratio...can scale even further to 12 to 2, 21 to 3 and beyond!

• Route Reflectors do not change the next hop learned by default so they can be out of path and do not need to carry traffic. Typically deployed with OSPF/BGP and loopback peering

• Multiple RRs can peer into the BGP border routers to distribute resource utilization

• Route Reflection
  • Use RRs to feed multiple PEs – MIkroTik CCR can send 1 million routes to RR clients in under 2 minutes. Virtualized CHR on Intel CPU will be slightly faster. RouterOS v7 will improve even more..
  • Scale new RRs as needed
Design Overview – MikroTik CHR vs Cisco ASR1000V

- Cisco ASR1000V has a very expensive cafeteria licensing model for cloud operators and ISP
- MikroTik CHR has more performance potential for a mere fraction of the cost
- Many other use cases for CHR – Firewall, Core Router, Hosted Router

<table>
<thead>
<tr>
<th>Specifications</th>
<th>MikroTik CHR</th>
<th>Cisco ASR1000V</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Throughput</td>
<td>10 Gbps +</td>
<td>Limited to 10 Gbps</td>
</tr>
<tr>
<td>MPLS Throughput</td>
<td>10 Gbps +</td>
<td>Limited to 5 Gbps</td>
</tr>
<tr>
<td>IPSEC license</td>
<td>Included</td>
<td>Separate license ($6500)</td>
</tr>
<tr>
<td>Firewall license</td>
<td>Included</td>
<td>Separate license ($3000)</td>
</tr>
<tr>
<td>License</td>
<td>1 Gbps, 10 Gbps, Unlimited</td>
<td>1 Gbps, 5 Gbps, 10 Gbps</td>
</tr>
<tr>
<td>Cost</td>
<td>$250 for Unlimited</td>
<td>Upwards of $30,000 for up to 10 Gbps with advanced services</td>
</tr>
</tbody>
</table>
• Come by the IP ArchiTechs booth and get Tik Tacs!

Mikro Tik TAC = MikroTik Technical Assistance Center
Questions?

The content of this presentation will be available at iparchitechs.com

Please come see us at the IP ArchiTechs booth in the Exhibitor Hall

Email: kevin.myers@iparchitechs.com
Office: (303) 590-9943
Web: www.iparchitechs.com

Thank you for your time and enjoy the MUM!!