The Slovak Internet eXchange

TRILL Deployment in SIX

Marian Ďurkovič

www.six.sk

Basic Facts

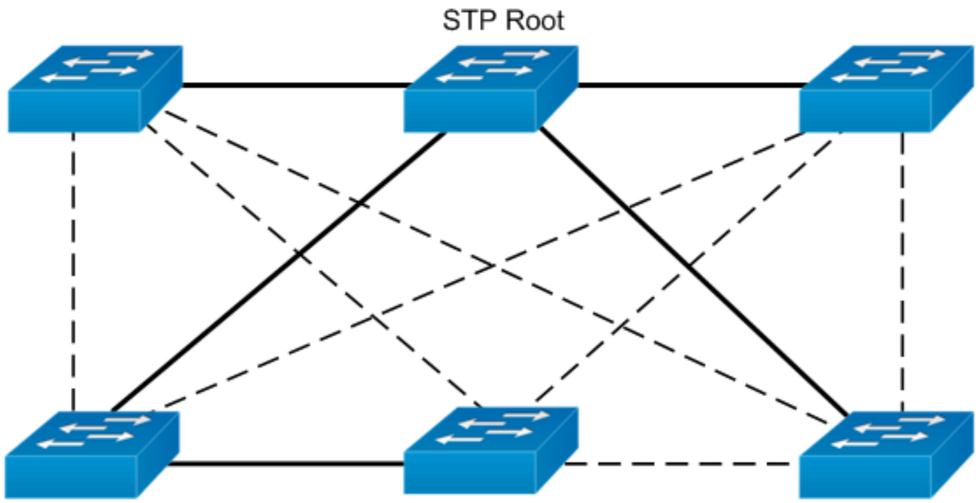
- SIX established in 1996 upon agreement of all major slovak ISPs
- Operations entrusted to Slovak University of Technology
  - Institution with long-term stability
  - Not a competitor to any ISP, telco, content provider, etc.
- Neutral and non-profit
  - Equal treatment for all SIX members
- 56 members, daily traffic peak ~70 Gbps
- Supports all kinds of interconnection:
  - Public IPv4 & IPv6 peering
  - Private peering
  - Ethernet, SDH, lambda, dark fibre, …

New SIX Platform

- Planning started in 2013
- Main goals:
  - Keep up with traffic demands
  - Provide enough available ports
  - Support new interfaces (40GE, 100GE)
  - Introduce state-of-the-art technology
  - Improve redundancy
  - Ensure easy upgradability
- Steps taken:
  - In-depth review of available technologies
  - Extensive lab testing of multiple devices & feedback to vendors
  - Selection of new core technology
  - Pilot project with academic network from Aug 5, 2014
  - Production from Sep 30, 2014

Rejected Technologies

- Technologies, which are unable to utilize all available links
- In principle all variants of spanning tree



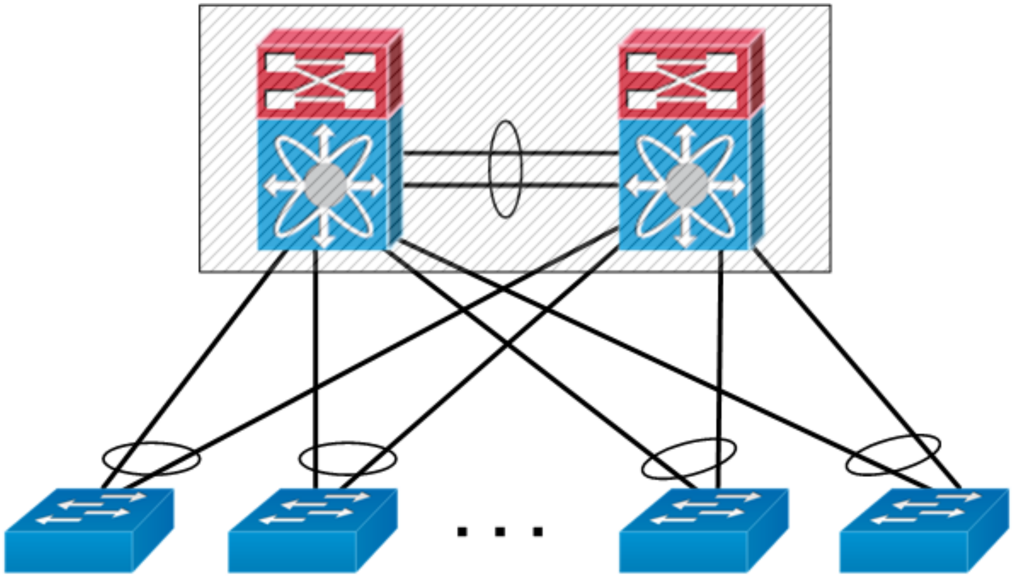
⛔ Blocking of redundant links is backwards

⛔ Huge waste of available bandwidth

⛔ Protocol failure leads to network meltdown

Rejected Technologies

- Technologies, which only work in very specific topology and/or proprietary to single vendor (or even single product)
- Typical example: MC-LAG / VSS / vPC / VLT / IRF



⛔ Complex synchronization of state between core switches
⛔ Doesn’t scale to more than 2 core units
⛔ No standardization in place

Evaluated Technologies

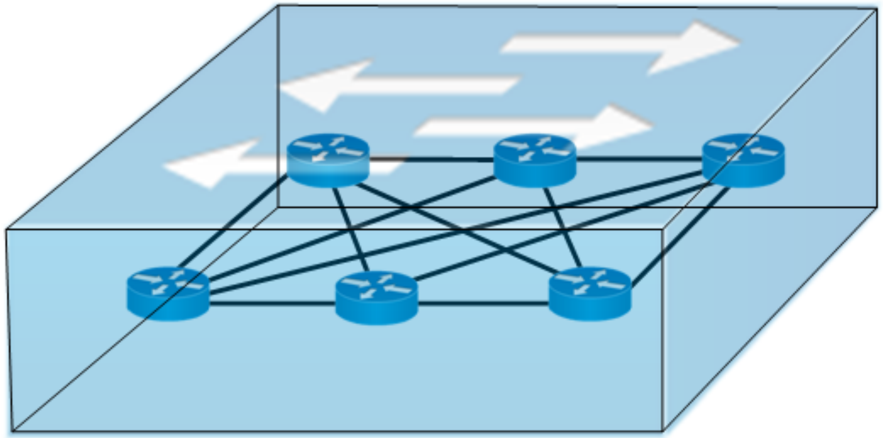
- List relatively short: VPLS, TRILL, SPB
- VPLS in production in large IXPs, so there's enough experience
  - Hands-on experience needed for new technologies
- TRILL equipment received for lab-testing from 3 vendors
  - We thoroughly checked the implementation
  - Very helpful for full understanding of TRILL operation
  - Found some limitations which we reported back to vendors
- Key differences:
  - VPLS: traffic flows over preconfigured tunnels
    number of LSPs grows fast (9000+ in large IXP)
  - TRILL: every switch makes independent routing decisions
    routing tables small and easy to check
- SPB not very useful for IXP
  - Needs spanning tree to work
  - Strange & suboptimal ECMP load balancing

The Decision: TRILL

- We strongly believe in KISS principle
  - Most systems work best if they're kept simple rather than made complicated
- IP routing is nice example
  - Key technology which enabled Internet in today's scale
  - Simple but very powerful and mature
  - No tunnels - each router independently decides about next hop
  - Not restricted to any predefined topology
- MPLS much more complex
  - Requires more expensive hardware
  - Configuration-intensive
  - Load balancing over parallel links can be tricky

TRILL Mechanics

- TRILL internally uses exactly the same principles as IP routing
  - Authors haven't tried to reinvent the wheel
  - TRILL headers are smaller, but have the same content
  - Builds on dynamic routing by field-proven IS-IS protocol
  - Natively makes use of all available links
  - Supports multiple paths (ECMP)
  - Utilizes IP safety belts like TTL check, RPF check
- External devices just see a huge ethernet switch



SIX Building Blocks

- Instead of installing one big switch, we went for distributed design similar to large clouds
- 4 Huawei CloudEngine 6850 switches connected by dual 40GE rings
- Switches are like building blocks of various sizes:

ASIC	Capacity	Ports (1RU)	Alt. Ports
Trident	0.64 Tbps	64 x 10GE	40GE
Trident +	0.64 Tbps	64 x 10GE	40GE
Trident 2	1.28 Tbps	32 x 40GE	10GE
Tomahawk	3.20 Tbps	32 x 100GE	10GE, 40GE

- When we need more capacity, we just add another switch
  - No need to upgrade/remodel existing switches
- TRILL supports arbitrary topology - when current rings reach their limits, we can easily change to full mesh, leaf & spine etc.

TRILL Configuration

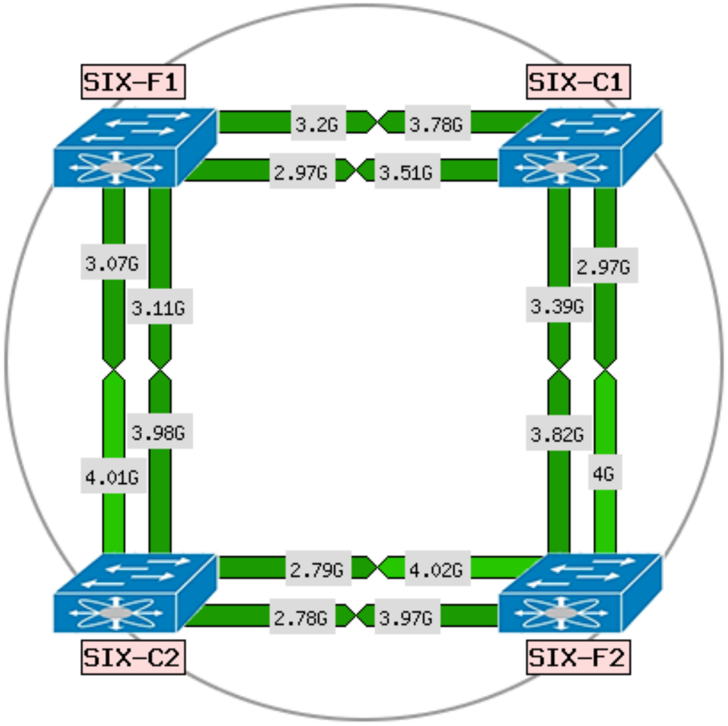
- TRILL requires minimal configuration to work
- IS-IS dynamically computes shortest paths over given topology
- TRILL enabled only on backbone ports
- Default link cost: 20000 / BW [Gbps]
- Link costs adjustable as needed

```
trill
 trill-name SIX-F1
 network-entity 00.0000.0000.0110.00
 nickname 110 root-priority 65200
 carrier-vlan 4000
 ce-vlan 666 700 to 720

interface range 40GE1/0/1 to 40GE1/0/4
 port link-type trunk
 trill enable
 trill cost 500
```

TRILL Load Balancing

- TRILL natively supports fine-grained per-flow ECMP load balancing
- No special provisions needed – just configure equal link costs



2x40GE TRILL RING

```
TRILL Unicast Routing Table
---------------------------------------------------
Flags: D-Download To Fib

Total Route(s): 3

Nickname        Cost  Flag  OutInterface     Hop
---------------------------------------------------
SIX-C1           500  D     40GE1/0/1         1
                            40GE1/0/2         1
SIX-C2           500  D     40GE1/0/3         1
                            40GE1/0/4         1
SIX-F2          1000  D     40GE1/0/1         2
                            40GE1/0/2         2
                            40GE1/0/3         2
                            40GE1/0/4         2
```

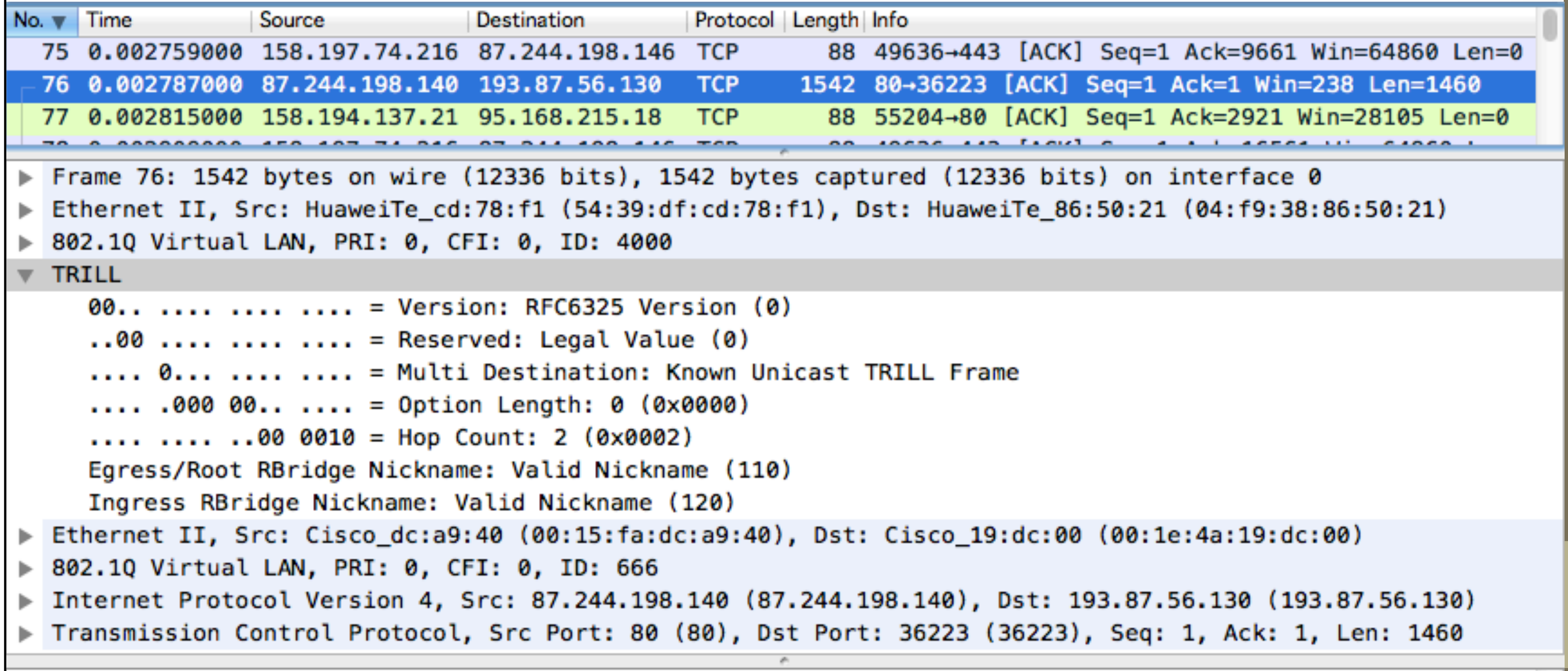
- Traffic between SIX-F1 and SIX-F2 uses all 4 available paths

Improved Maintenance

- TRILL allows reconfiguration of SIX core without single packet loss
- This is possible thanks to IS-IS protocol
- Well-known procedure from IP backbones:
  - Set cost of the link to maximum
  - Wait until all traffic gets rerouted
  - Disconnect the link
- We're able to change backbone topology, insert new switches or perform maintenance without any impact to SIX members
- Configuration done via commits
- Our switches also support hitless software patching
  - Security and bug fixes are applied to running system
  - No need to restart switches

TRILL Monitoring

- Port mirroring & sflow well supported on TRILL switches
- Major advantage over e.g. Cisco's Fabric Path
- We developed a few patches for Wireshark:



```
No.  Time        Source          Destination    Protocol Length Info
 75 0.002759000 158.197.74.216  87.244.198.146 TCP        88  49636→443 [ACK] Seq=1 Ack=9661 Win=64860 Len=0
 76 0.002787000 87.244.198.140  193.87.56.130  TCP      1542  80→36223 [ACK] Seq=1 Ack=1 Win=238 Len=1460
 77 0.002815000 158.194.137.21  95.168.215.18  TCP        88  55204→80 [ACK] Seq=1 Ack=2921 Win=28105 Len=0
```

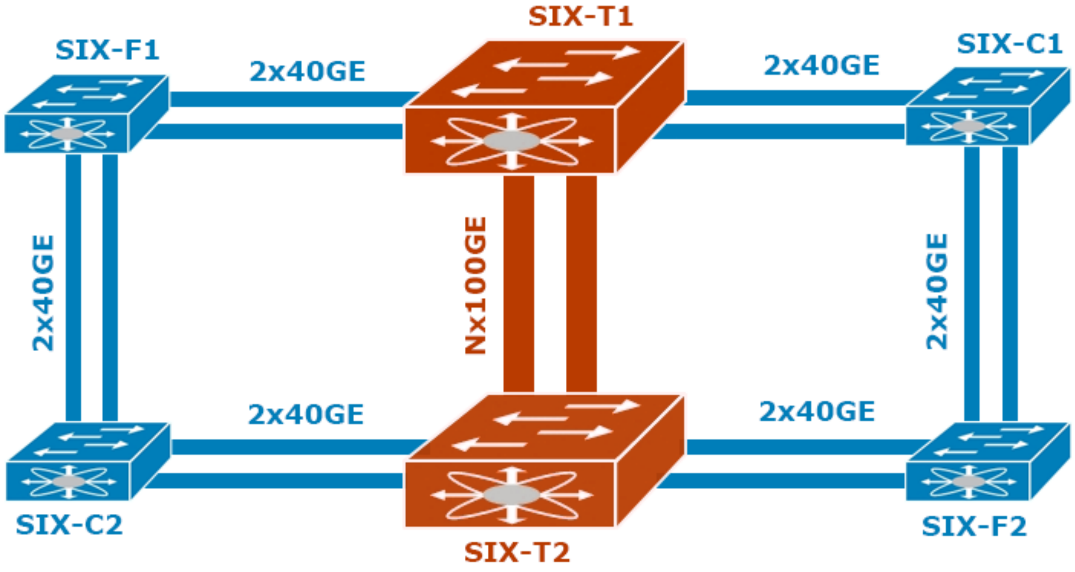
```
▶ Frame 76: 1542 bytes on wire (12336 bits), 1542 bytes captured (12336 bits) on interface 0
▶ Ethernet II, Src: HuaweiTe_cd:78:f1 (54:39:df:cd:78:f1), Dst: HuaweiTe_86:50:21 (04:f9:38:86:50:21)
▶ 802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 4000
▼ TRILL
    00.. .... .... .... = Version: RFC6325 Version (0)
    ..00 .... .... .... = Reserved: Legal Value (0)
    .... 0... .... .... = Multi Destination: Known Unicast TRILL Frame
    .... .000 00.. .... = Option Length: 0 (0x0000)
    .... .... ..00 0010 = Hop Count: 2 (0x0002)
    Egress/Root RBridge Nickname: Valid Nickname (110)
    Ingress RBridge Nickname: Valid Nickname (120)
▶ Ethernet II, Src: Cisco_dc:a9:40 (00:15:fa:dc:a9:40), Dst: Cisco_19:dc:00 (00:1e:4a:19:dc:00)
▶ 802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 666
▶ Internet Protocol Version 4, Src: 87.244.198.140 (87.244.198.140), Dst: 193.87.56.130 (193.87.56.130)
▶ Transmission Control Protocol, Src Port: 80 (80), Dst Port: 36223 (36223), Seq: 1, Ack: 1, Len: 1460
```

Experience with TRILL

- Initial software for lab testing didn't support per-flow load balancing
  - Major problem for IXP application
  - Supported in HW but needs non-default ASIC register settings
  - Implemented on our request in V1R3 software (Jul 30, 2014)
- During pilot with academic network we found a problem with ifHCInOctets/ifHCOutOctets SNMP counters
  - Fixed by a 24 kB patch applied before production
- Another minor SNMP issue discovered in Jan 2015 – ifHCInUcastPkts wrapping at 40-bit boundary
  - Patch applied to running system without any service impact
- TRILL implementation very robust and reliable
  - No problems found during 1 year of production

Near Future Plans

- CloudEngine 8860 switches currently in development
  - Based on Tomahawk ASIC (3.2 Tbps)
  - 2RU modular chassis with 4 slots
  - Subcards: 8 x 100GE, 16 x 40GE or 24 x 25/10GE + 2 x 100GE
- Install two CE8860s into existing TRILL ring
  - Provide 100GE access ports to SIX members



Conclusions

- TRILL met all our expectations about next-gen SIX infrastructure
- Distributed architecture consisting of fixed building blocks
- Currently available ports:
  - 96 x 10G/1G SFP
  - 96 x 10G/1G/100Base-T
  - Port grouping: 4 x 10G -> 40G
  - 100G and 40G (QSFP) ports coming soon
- SIX platform scalable upto 10s of Tbps as needed
- Solution based on industry standards
- Support for arbitrary topology
  - SIX core able to keep up with future demands
- Excellent support from Huawei
- TRILL planned as transport infrastructure for Slovak Academic Network