OPEN
Compute Project

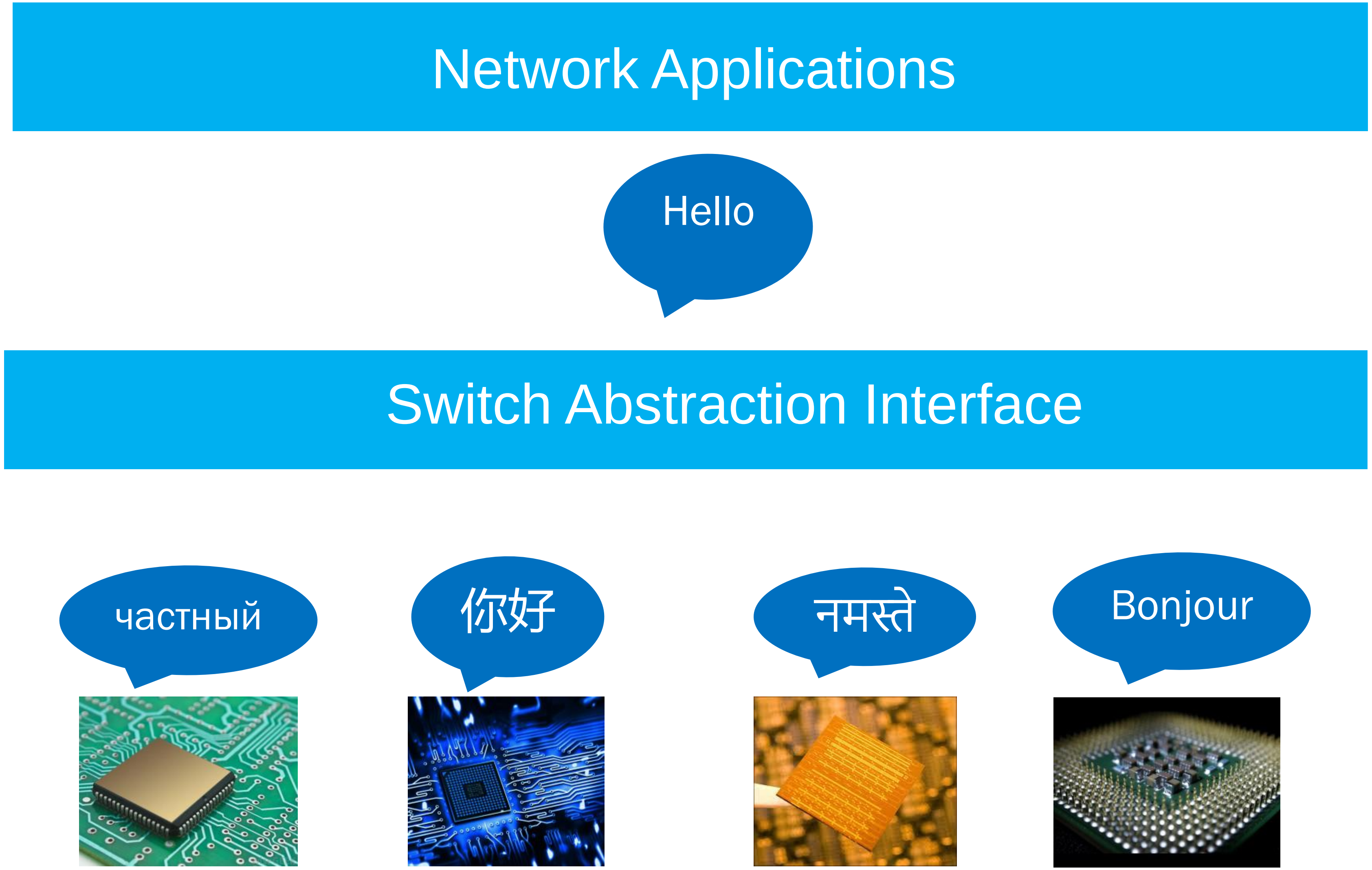
OCP U.S. SUMMIT 2017

Santa Clara, CA

SAI: Releasing the Potential of Switch ASIC

Xin Liu
Principal Product Manager
Microsoft

Switch Abstraction Interface (SAI)



Network Applications

Hello

Switch Abstraction Interface

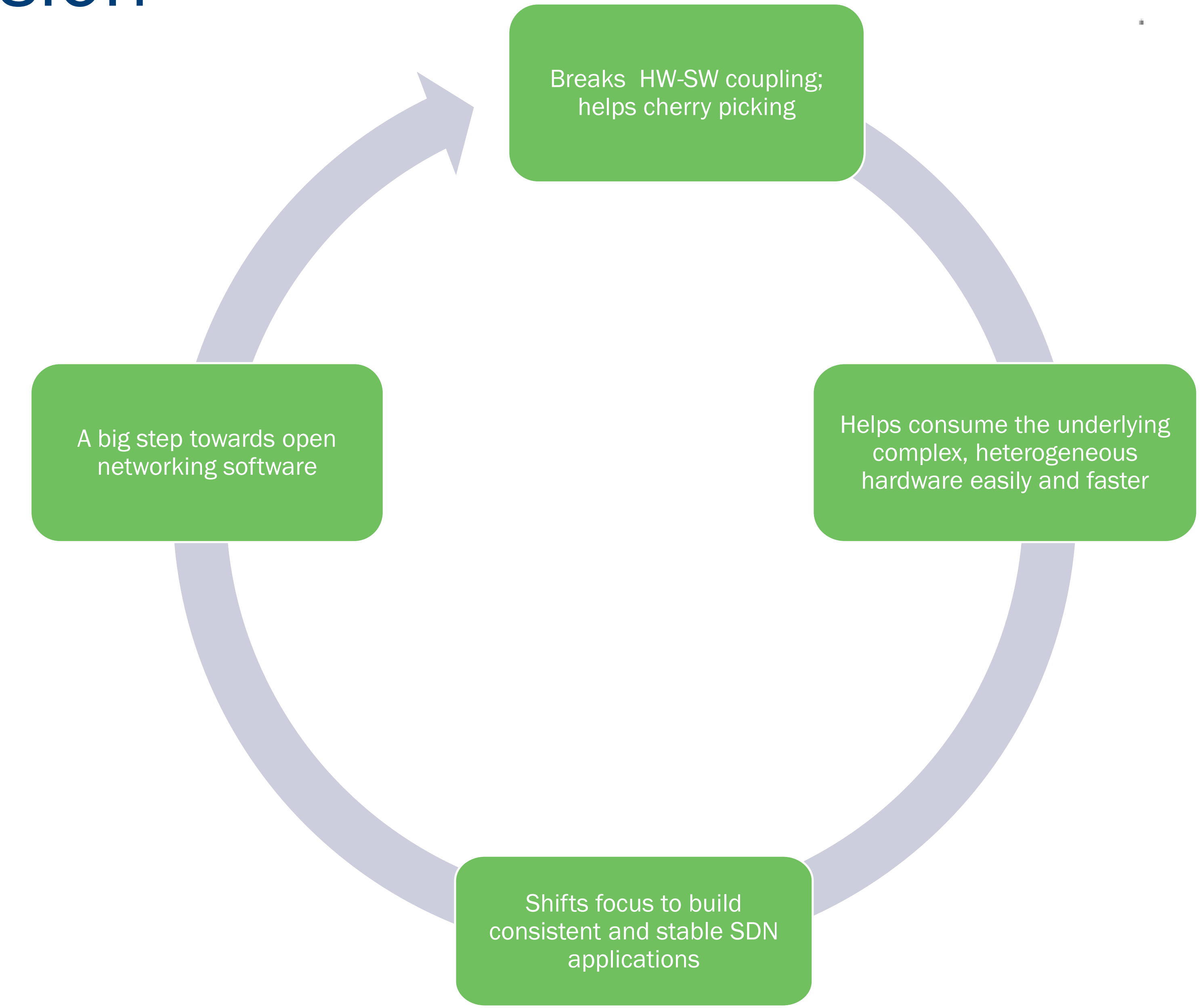
частный

你好

नमस्ते

Bonjour

SAI Mission



Breaks HW-SW coupling; helps cherry picking

Helps consume the underlying complex, heterogeneous hardware easily and faster

Shifts focus to build consistent and stable SDN applications

A big step towards open networking software

CNOS FlexSwitch Metaswitch OS10 OPX SONiC

DELL

metaswitch

Microsoft

Tofino, P4

BAREFOOT NETWORKS

Trident, Tomahawk

BROADCOM

XPliant

CAVIUM

Goldengate

centec networks

Prestera

MARVELL®

Spectrum

Mellanox® TECHNOLOGIES

Taurus

Nephos

77 members

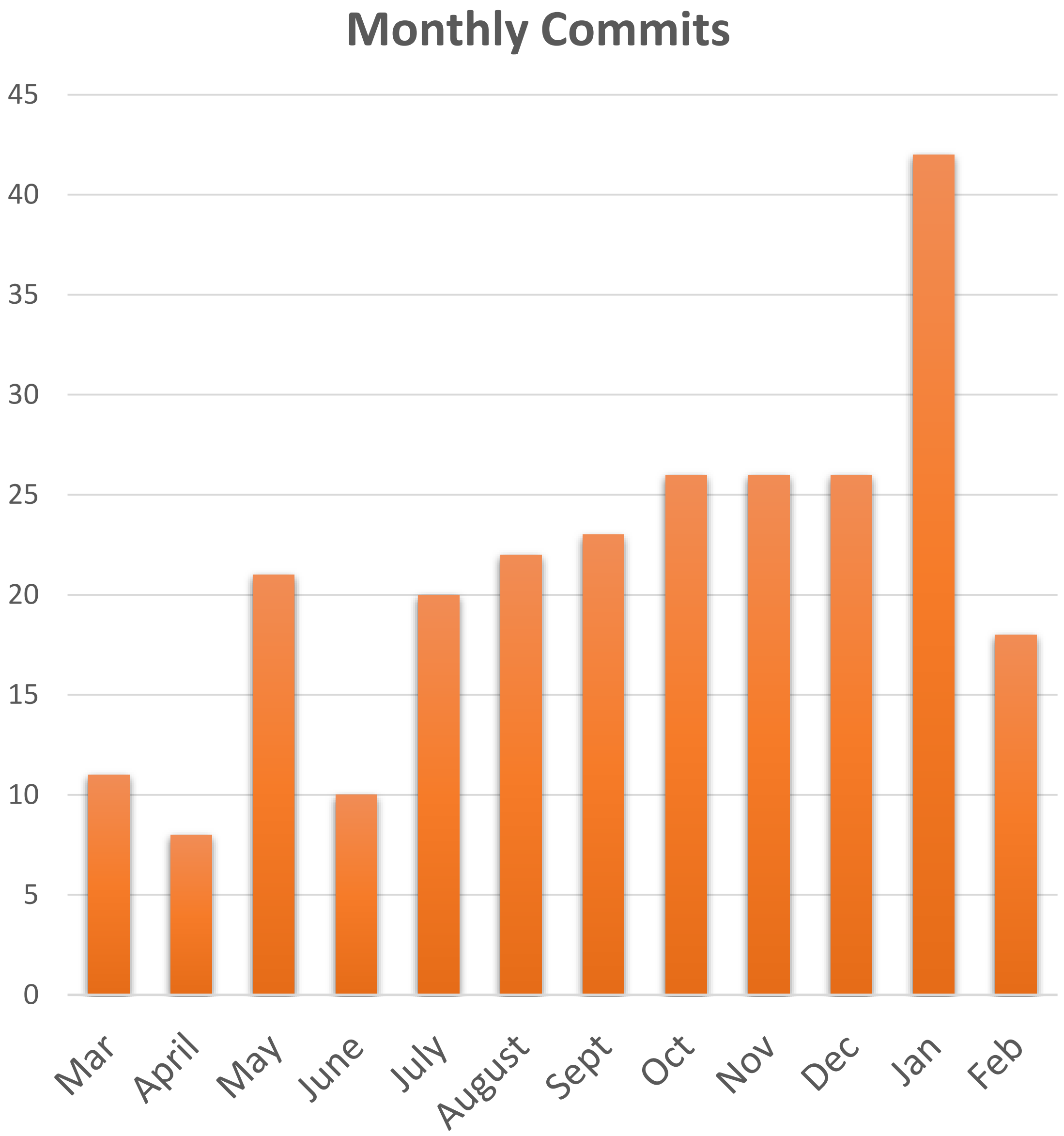
48 Contributors

472 Commits

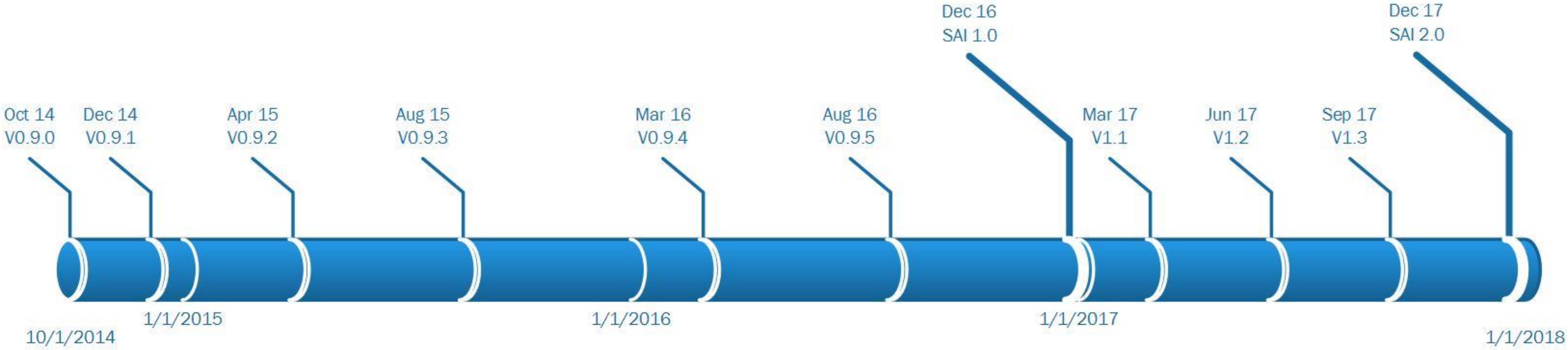
>60 meetings 2016

6 Releases

37 Proposals



SAI Releases



What Was Added to SAI 1.0

Enhanced ACL Model Proposal

CAVIUM

Speaker: Zubin Shah

Use Case and Motivation

- Use case
  - Universally deployed for N-tuple match and Security applications in Cloud, Enterprise, or WAN deployments

- Motivations
  - Operator centric, allows disaggregation of software from hardware
  - Simple configuration model through easy expression of filters, tables and rules as opposed to TCAMs
  - Better scaling and reusability of ACL table and hence achieving cloud-scale
  - ASIC agnostic, adopted by major silicon vendors

Proposal Details

- Introduced bind points
- Introduced ACL Groups and concluded a common abstract behavior
- Introduced behavioral model specs
  - Location of ACL tables and ACL groups in the model VLAN and Mirror cases : contributed and pending reviews and merged
  - Parallel versus Sequential lookups
  - Clarification of various fields, metadata, context available for ACL lookups

- Unit Test Cases
  - ACL case : 11+ UT cases , some merged and several available in PRs
  - VLAN and Mirror cases : contributed and pending reviews and merged

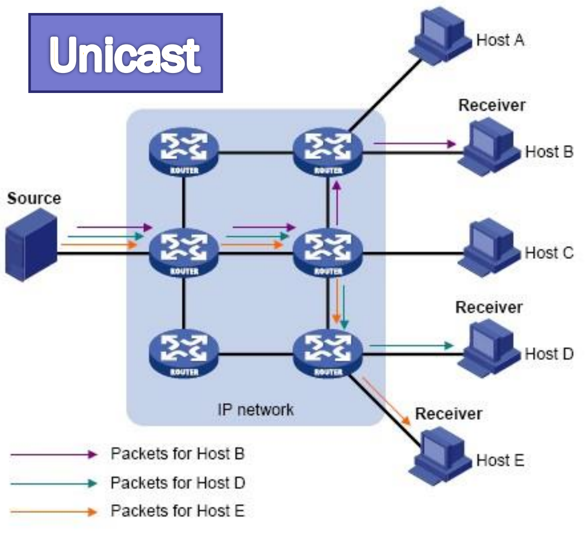
L2/L3 Multicast Proposal

centec
networks

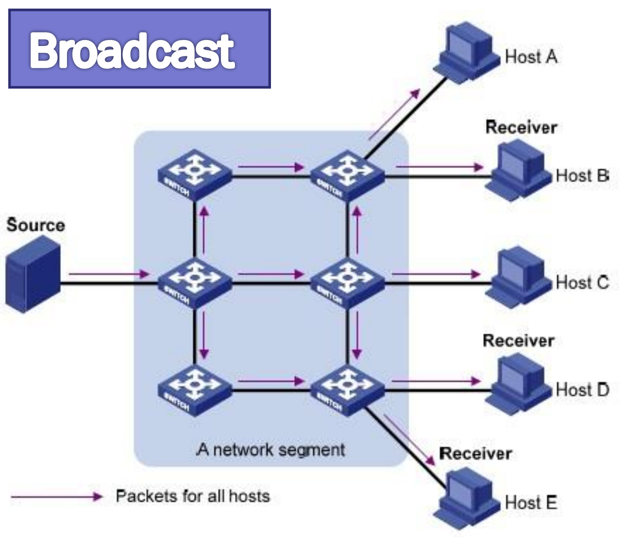
Speaker: Min Yao

Use Case and Motivation

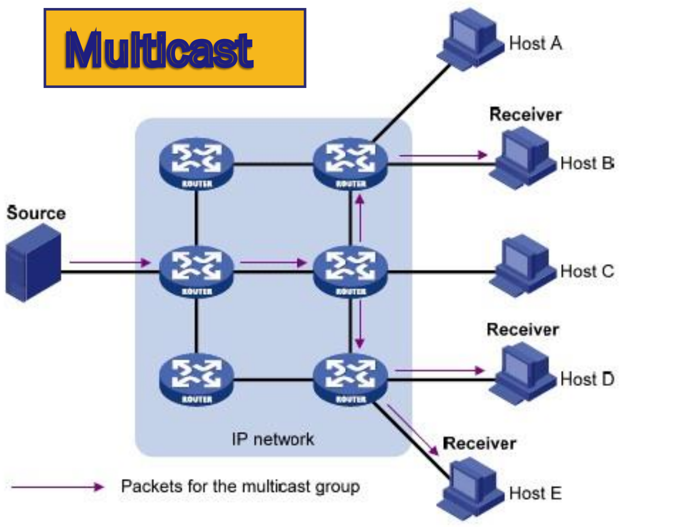
- Use case: multimedia distribution network, e.g. 2016 Rio Olympics
- Same copy of data, need to distribute to multiple nodes
- Multicast technology could save a lot of bandwidth, reduce the network traffic load



- Information transmitted is proportional to the receiver number



- The security of information can not be guaranteed, and bandwidth is wasted



- The packet will be forwarded to those hosts needed the information only.

Bridge Model for Multitenancy Proposal

Mellanox®
TECHNOLOGIES

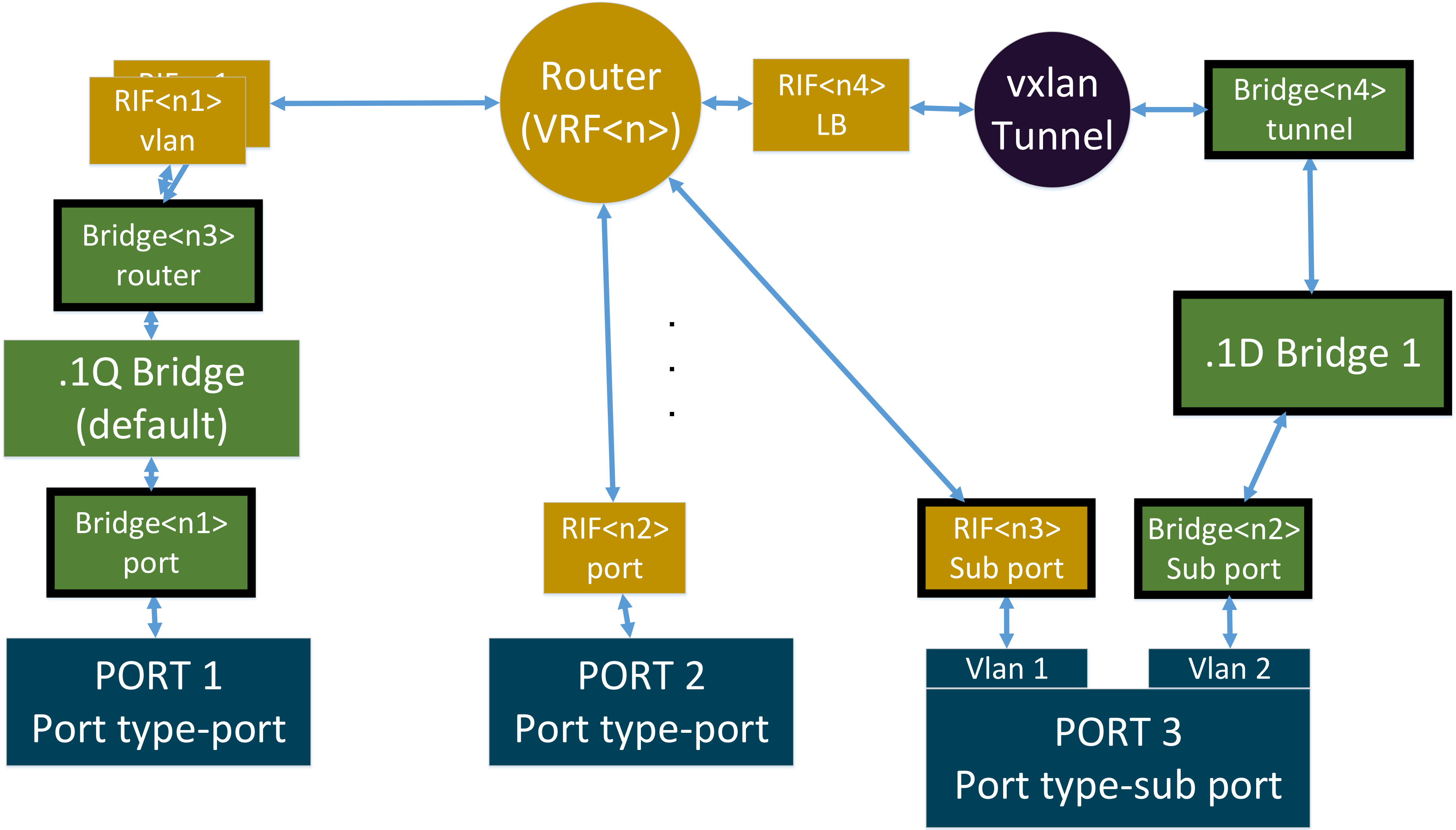
Speaker: Matty Kadosh

Use Case and Motivation

- Use case
  - To support multi-tenancy in the network

- Benefits
  - Enable user to create overlay networks
  - Increase the number of tenants and number of networks per tenant
    - by increasing SAI 4k Vlan broadcast domain
    - by adding ability to create interface base on {port, Vlan}

Proposal Details



Added a set of objects as Bridge Ports to build discrete pipeline

Flexible Host Interface for Network Management Proposal

Mellanox®
TECHNOLOGIES

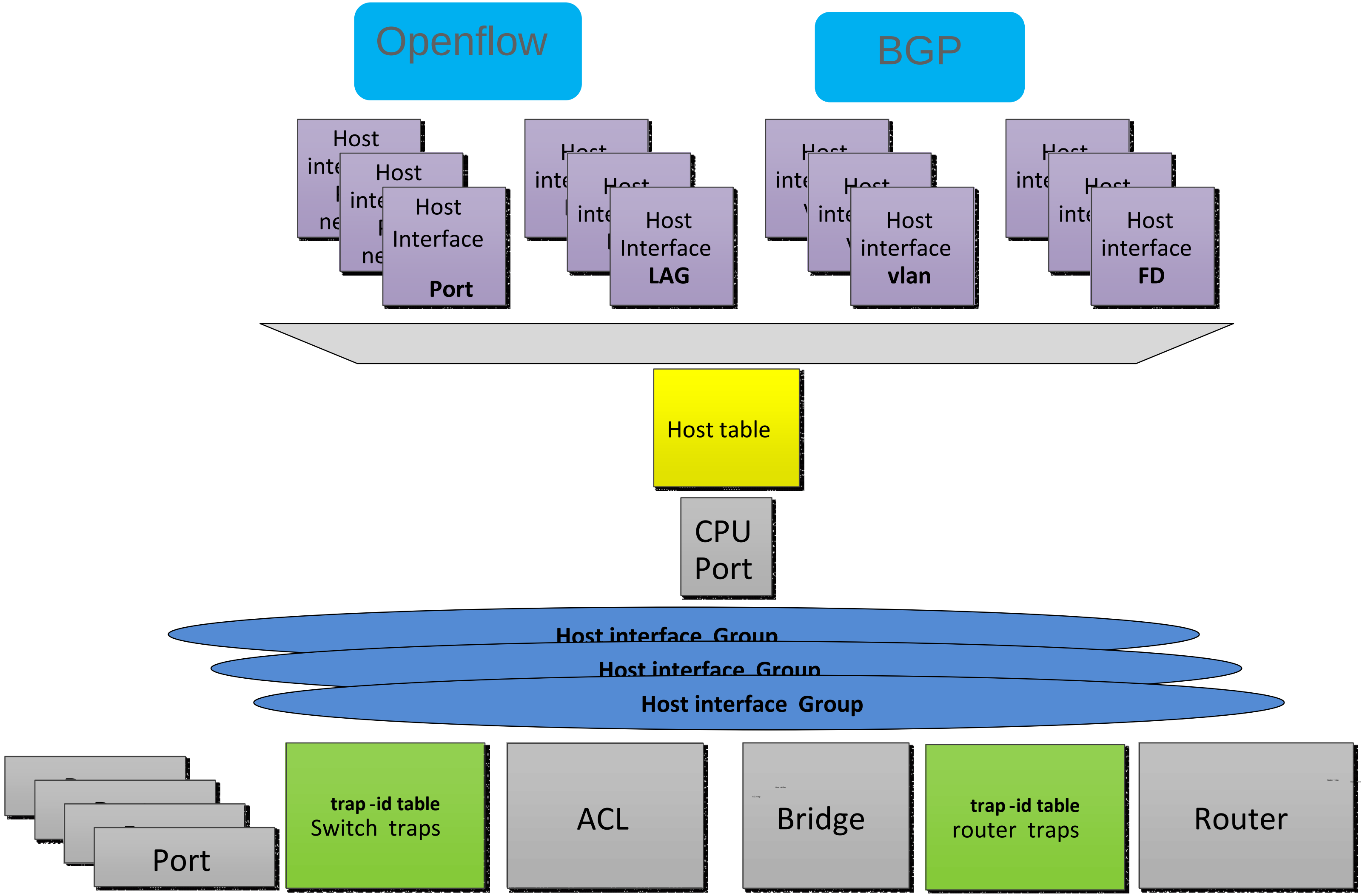
Speaker: Matty Kadosh

Use Case and Motivation

- Use case
  - For a network with multiple management mechanisms, e.g. overlay managed by Openflow, underlay managed by BGP, this greatly simplifies the flow

- Benefits
  - Enable engineers to be able to port SAI quickly
  - Better interface usability

Proposal Details



- Enhanced SAI application packet send /receive interface
  - Different type of Linux net devices
  - Port, LAG, Vlan, Brideg
- Add flexibility – select the packet send / receive interface according to
  - {packet type, port}
  - {packet type, Vlan}

SAI Roadmap 2017

Monitoring	Protocol Support	Reliability/QoS	Infrastructure
TAM [Broadcom]	MPLS [Mellanox]	L3 Fast Reroute [Metaswitch]	SAI P4 Model [Mellanox]
Microburst [Marvell]	802.1BR [Dell]	BFD [Dell]	Multi-NPU [Dell]
Critical Resource Monitoring [MSFT]	Segment Routing [Cavium]	ECN [Dell]	Capability Query [MSFT]
INT [Barefoot]	Open flow Extension [Cavium]		SAI Ext API [Dell]

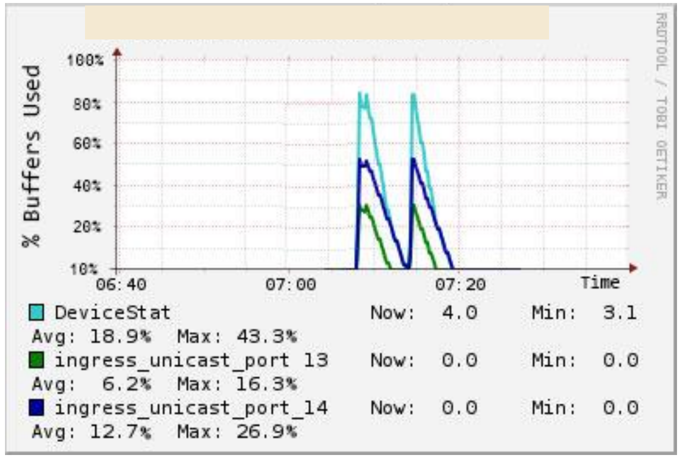
Telemetry and Monitoring (TAM) Proposal

BROADCOM

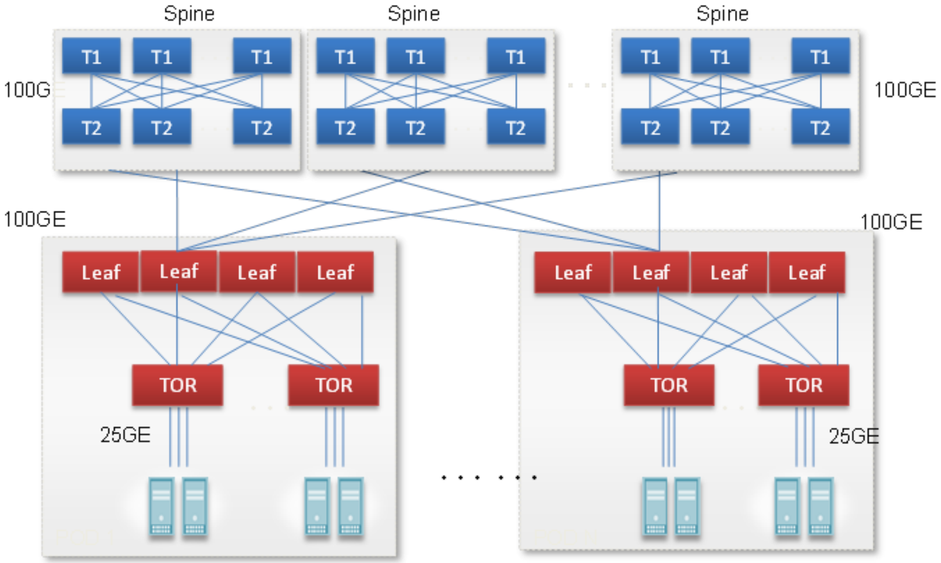
Speaker: Bhaskar Chinni

Scalable Monitoring of Data Center Networks

“How to monitor buffer occupancies in a large scale data center networks in a scalable way?”



Use case : Typically internet traffic flows from Spine to leaf and then to host. When multiple streams destined to servers connected through the same leaf/spine switch, they could create a congestion scenario.

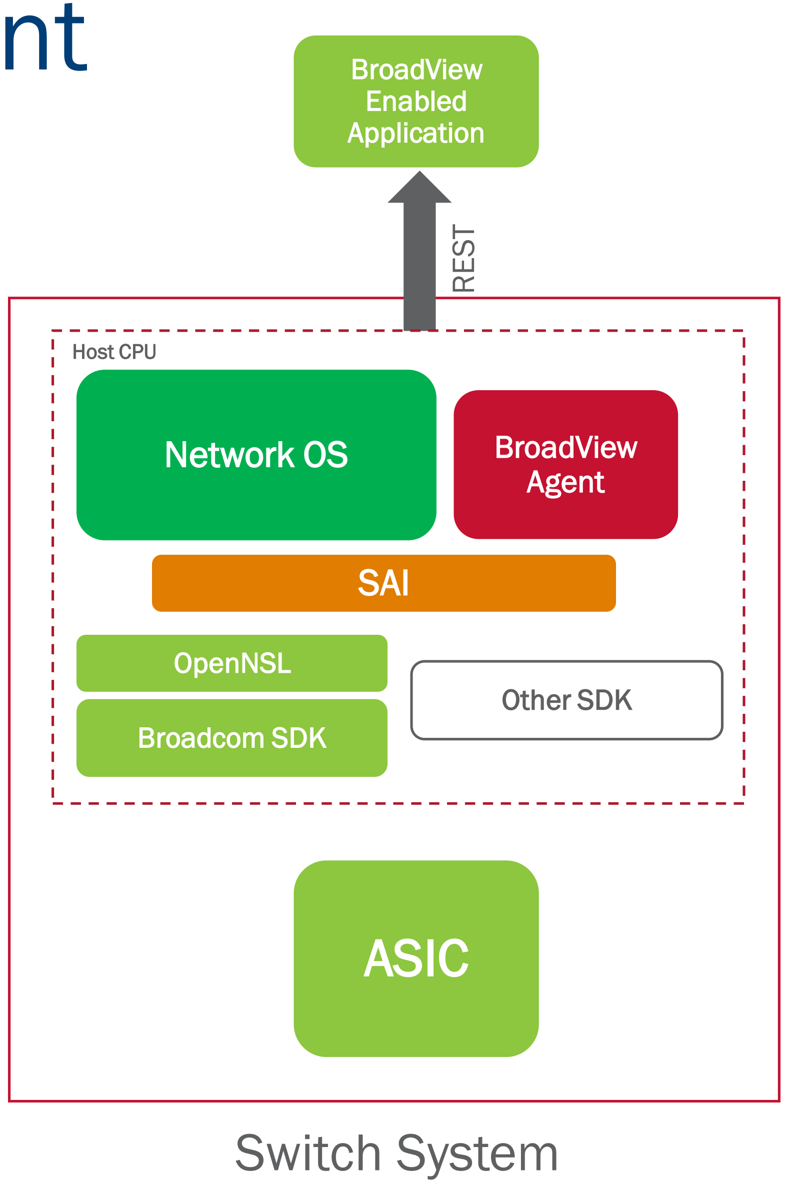


Proposal Details

- TAM is an API for monitoring and controlling buffer occupancies.
- TAM facilitates real-time microburst detection through watermark breach alerts
- TAM enables tracker objects to track multiple statistics
- TAM supports multiple snapshot objects for simultaneous capturing of different sets of statistics
- TAM uses transporter objects for delivering snapshots at a desired location
- TAM can be easily customized for underlying hardware

Broadview™ Instrumentation Agent

- Platform agnostic agent for advanced analytics
- Light weight with high scalability
- Working in progress to integrate into SONiC
- Pre-integrated with Open Ecosystem projects



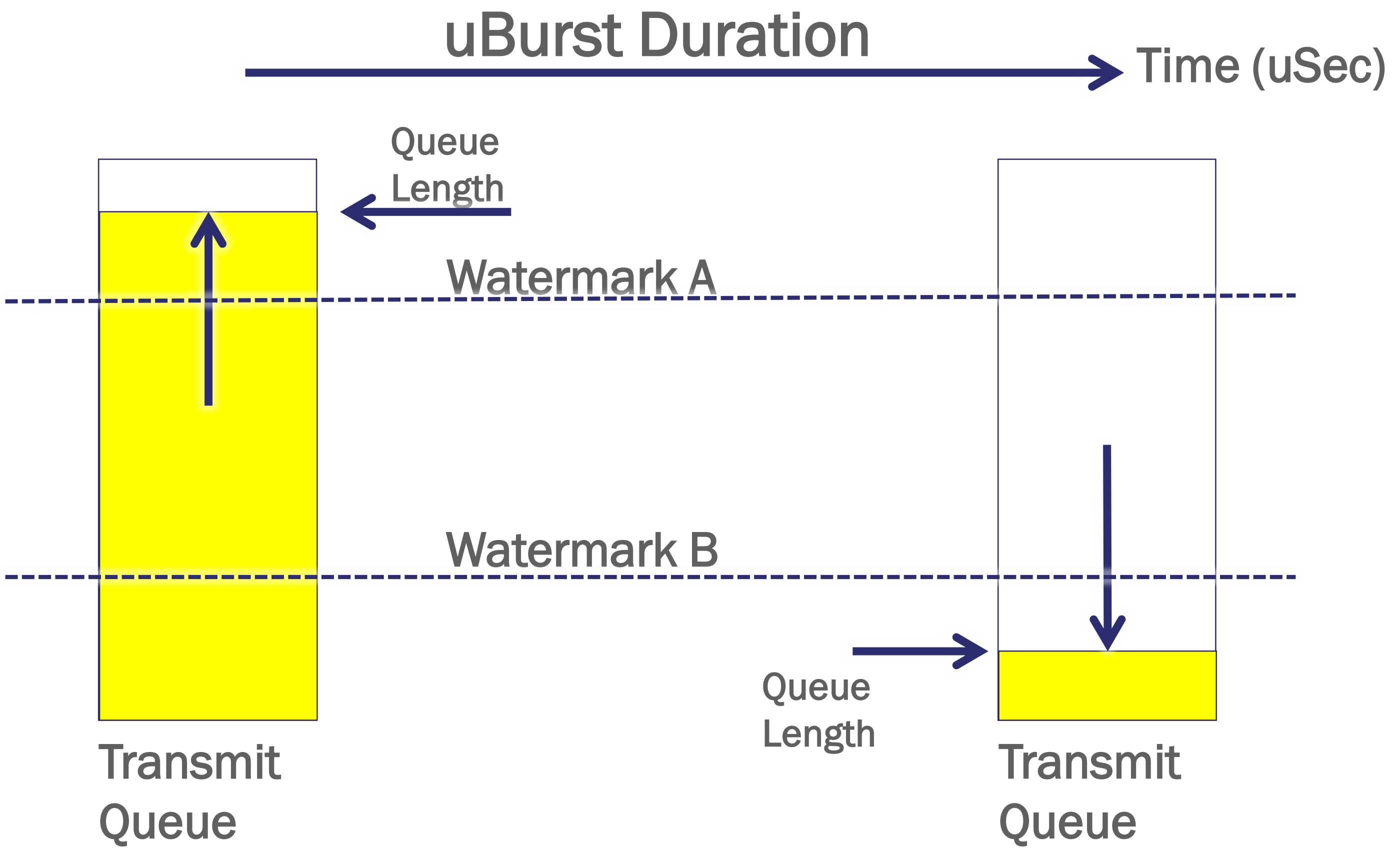
TAM Enhancements for Monitoring Microbursts

MARVELL®

Speaker: Vitaly Vovnoboy

Microburst Definition

- Microburst (uBurst) is an event in which a buffer-count (e.g., a queue length) crosses watermark A (from low to high) until it crosses watermark B (from high to low).



Benefits to Network Operators

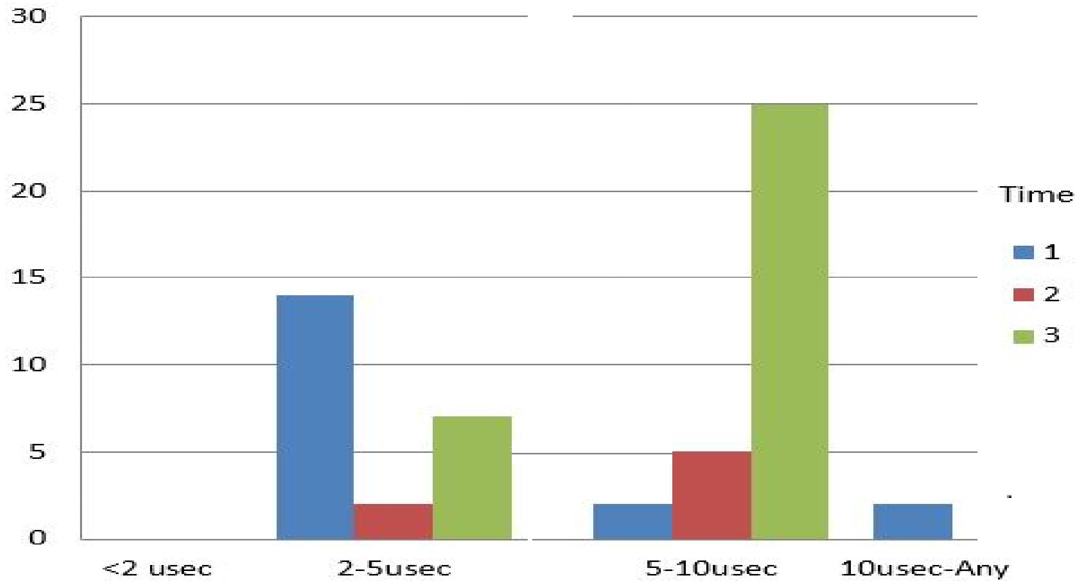
- Better characterize congestion events according to the different duration statistics.
- Correlate network congestion events with servers activities.
- Monitoring network health and identifying the severity of traffic events.
- Offload application CPU/controller from collecting huge number of events.

uBurst Duration Objects

- uBurst Duration Statistics:
  - Last uBurst duration
  - Longest duration (peak)
  - Shortest duration (min)
  - Average duration
  - Number of uBursts
  - Durations histogram

uBurst Durations Histogram

- Number of uBursts according to their durations in user-defined intervals
  - uBurst-duration-bin-a (from 0 to ‘a’ us)
  - uBurst-duration-bin-b (‘a’ to ‘b’)
  - uBurst-duration-bin-c (‘b’ to ‘c’)
  - uBurst-duration-bin-d (‘c’ to any)



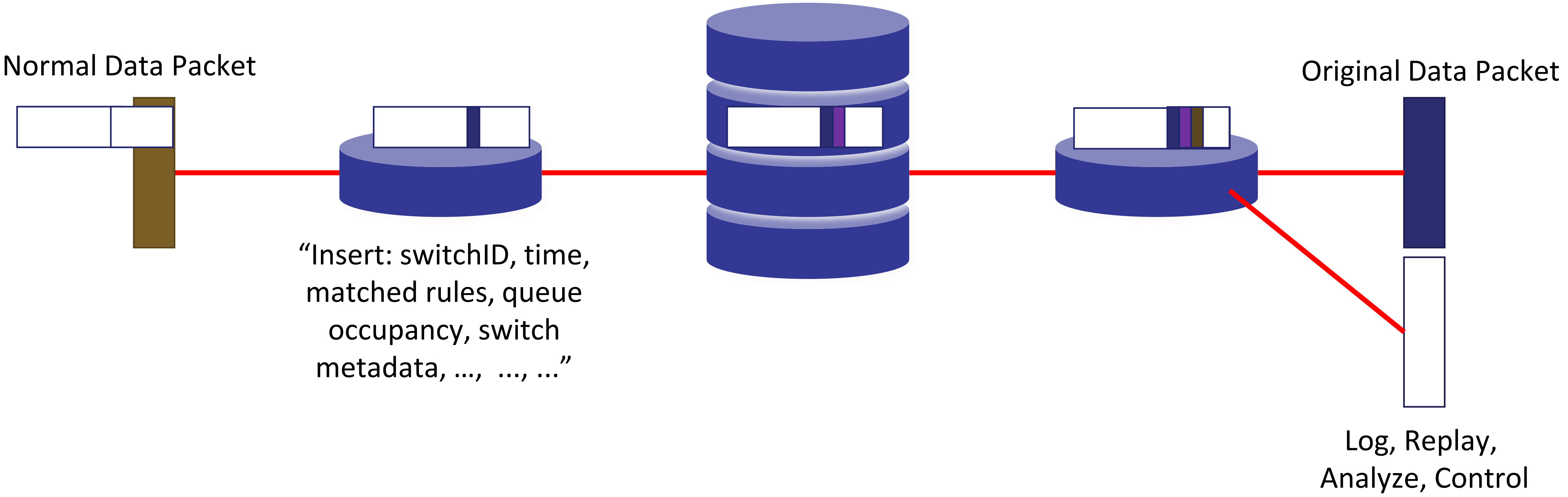
In-band Network Telemetry (INT) Proposal

BAREFOOT NETWORKS

Speaker: Prem Jonnalagadda

Experimental Trak

In-band Network Telemetry (INT)


Normal Data Packet
Original Data Packet
"Insert: switchID, time, matched rules, queue occupancy, switch metadata, …, …, …"
Log, Replay, Analyze, Control

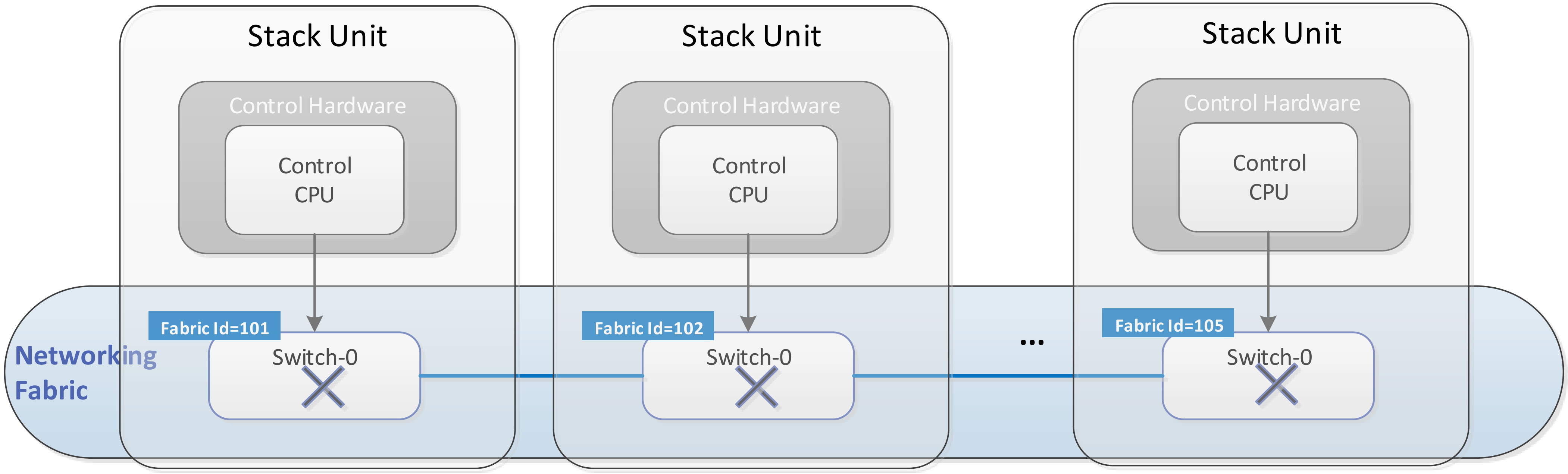
INT Proposal Details

- Description
  - APIs to enable INT Source/Transit/Sink functionality
  - Switches embed metadata in live packets
  - E.g., switch-id, port-id, hop-latency, queue-occupancy, tx-utilization, …
- Applications
  - Path Tracking
  - Latency Tracking
  - Congestion Tracking
  - …

Stacking using Multi-NPU/Networking Fabric

Speaker: Mihai Lazar

Use Case: stacking using Multi-NPU/Networking Fabric



- **Challenge:** provide a consistent API model for aggregating individual NPUs in a networking fabric
- **Benefit:** able to add new ports as needed to an existing network

Proposal Details

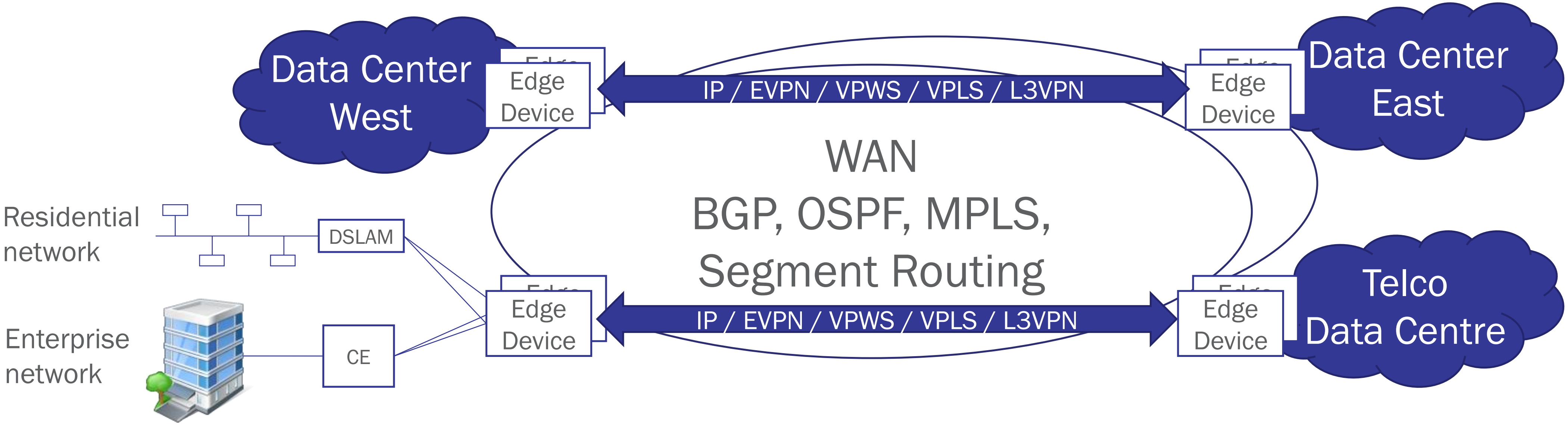
- Multi-NPU
  - Provides a means to aggregate multiple NPU’s into a Networking/Switching Fabric
  - The Switching Fabric behaves as a single NPU
- 802.1br
- BFD
- ECN at Port and Global level - queue level only in SAI 1.0
- SAI Vendor Extensions API

Reliable IP/MPLS Transport in the WAN with SAI Proposal

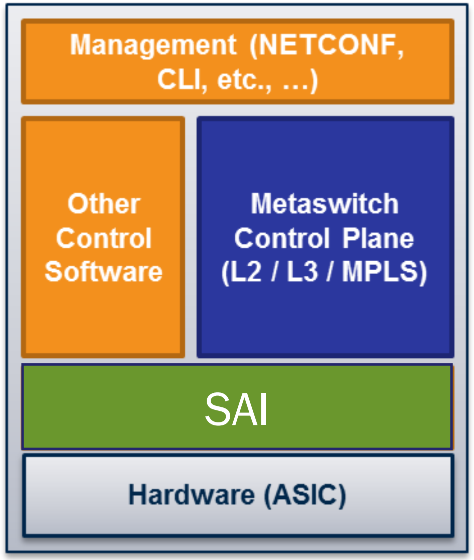
metaswitch

Speaker: Jonathan Hardwick

Use case: Disaggregation of WAN Edge Devices



- Who: Telcos and Hyperscale data centre operators
- What: WAN edge devices
- How: SAI to play a key role in disaggregating these complex, proprietary edge devices
  - Enabling cost reduction, innovation, SD-WAN

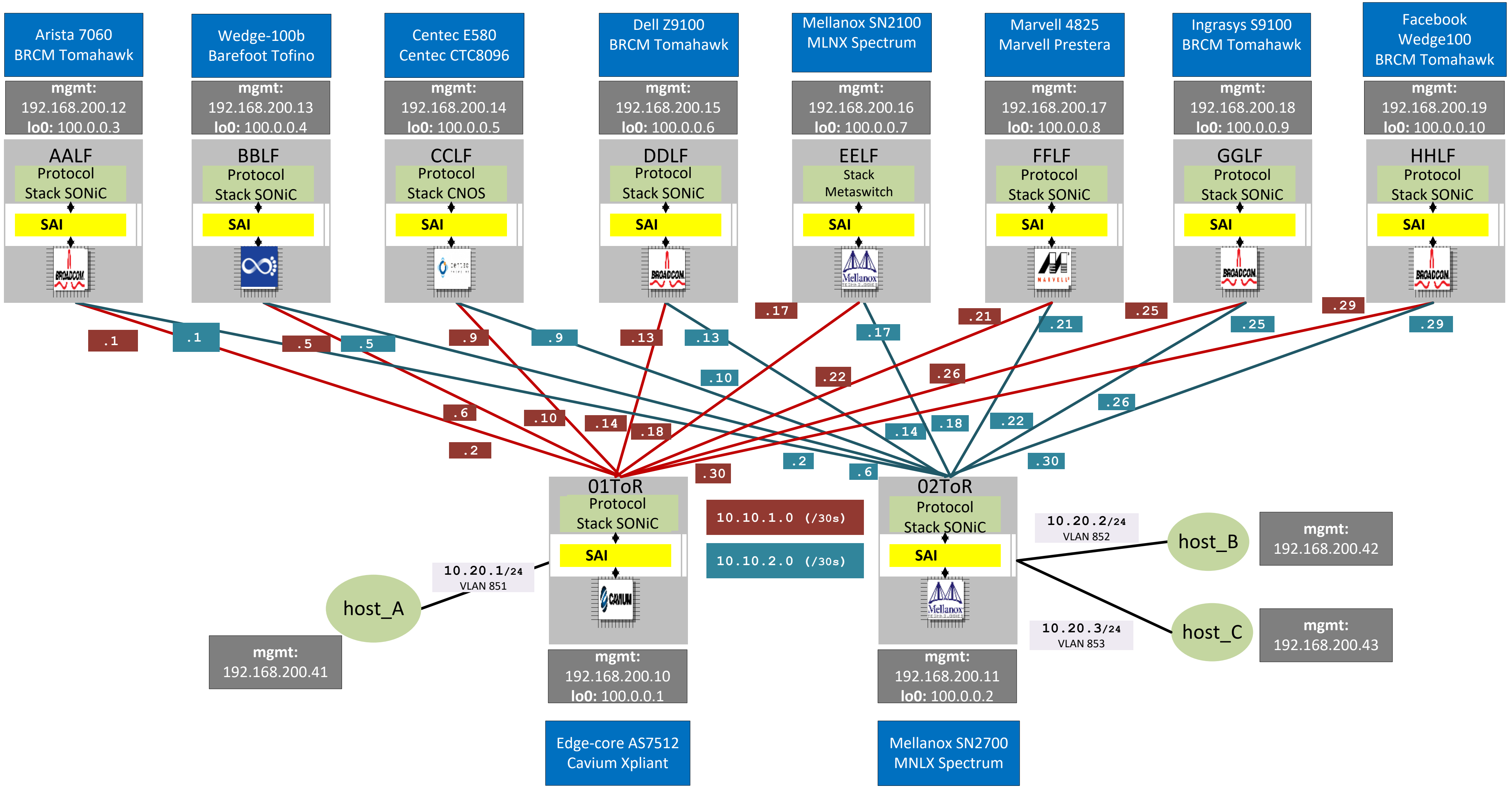

Management (NETCONF, CLI, etc., …)
Other Control Software
Metaswitch Control Plane (L2 / L3 / MPLS)
SAI
Hardware (ASIC)

Proposal Details

- New SAI features requested
  - IP and MPLS Fast Re-route:
    - SAI user’s responsibility to precompute the backup path and communicate it to the data plane
    - Enhance SAI with protection group semantics to enable rapid switchover
  - SAI Deep Integration with hardware-based BFD for fast fault detection

- Further SAI enhancements will also be required in future for VPN transport
  - (L2VPN) PWs, binding PWs to ACs, binding PWs to bridge domains, split horizon groups
  - (L3VPN) Labelled VRF routes
  - (EVPN) Labelled FDB entries

Demo Setup



Arista 7060
BRCM Tomahawk
mgmt: 192.168.200.12
lo0: 100.0.0.3
AALF
Protocol Stack SONiC
SAI

Wedge-100b
Barefoot Tofino
mgmt: 192.168.200.13
lo0: 100.0.0.4
BBLF
Protocol Stack SONiC
SAI

Centec E580
Centec CTC8096
mgmt: 192.168.200.14
lo0: 100.0.0.5
CCLF
Protocol Stack CNOS
SAI

Dell Z9100
BRCM Tomahawk
mgmt: 192.168.200.15
lo0: 100.0.0.6
DDLF
Protocol Stack SONiC
SAI

Mellanox SN2100
MLNX Spectrum
mgmt: 192.168.200.16
lo0: 100.0.0.7
EELF
Stack Metaswitch
SAI

Marvell 4825
Marvell Prestera
mgmt: 192.168.200.17
lo0: 100.0.0.8
FFLF
Protocol Stack SONiC
SAI

Ingrasys S9100
BRCM Tomahawk
mgmt: 192.168.200.18
lo0: 100.0.0.9
GGLF
Protocol Stack SONiC
SAI

Facebook Wedge100
BRCM Tomahawk
mgmt: 192.168.200.19
lo0: 100.0.0.10
HHLF
Protocol Stack SONiC
SAI

01ToR
Protocol Stack SONiC
SAI
mgmt: 192.168.200.10
lo0: 100.0.0.1
Edge-core AS7512
Cavium Xpliant

02ToR
Protocol Stack SONiC
SAI
mgmt: 192.168.200.11
lo0: 100.0.0.2
Mellanox SN2700
MNLX Spectrum

10.10.1.0 (/30s)
10.10.2.0 (/30s)

host_A
10.20.1/24
VLAN 851
mgmt: 192.168.200.41

host_B
10.20.2/24
VLAN 852
mgmt: 192.168.200.42

host_C
10.20.3/24
VLAN 853
mgmt: 192.168.200.43

Open Invitation

- Inviting contributions in all areas:
  - Bring up new proposals
  - Test and contribute test cases
  - Use it and report bugs

Github	https://github.com/opencomputeproject/SAI
Mailing list	opencompute-sai@lists.opencompute.org
Meeting	http://fuze.me/34034610
F2F Meeting	3/10 at Cavium Campus

OPEN
Compute Project