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
- A big step towards open networking software
- Shifts focus to build consistent and stable SDN applications

A big step towards open networking software.
77 members
48 Contributors
472 Commits
>60 meetings 2016
6 Releases
37 Proposals

Monthly Commits

<table>
<thead>
<tr>
<th>Month</th>
<th>Commits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar</td>
<td>10</td>
</tr>
<tr>
<td>April</td>
<td>5</td>
</tr>
<tr>
<td>May</td>
<td>20</td>
</tr>
<tr>
<td>June</td>
<td>15</td>
</tr>
<tr>
<td>July</td>
<td>15</td>
</tr>
<tr>
<td>August</td>
<td>20</td>
</tr>
<tr>
<td>Sept</td>
<td>25</td>
</tr>
<tr>
<td>Oct</td>
<td>30</td>
</tr>
<tr>
<td>Nov</td>
<td>30</td>
</tr>
<tr>
<td>Dec</td>
<td>20</td>
</tr>
<tr>
<td>Jan</td>
<td>40</td>
</tr>
<tr>
<td>Feb</td>
<td>15</td>
</tr>
</tbody>
</table>
SAI Releases

- Oct 14 V0.9.0
- Dec 14 V0.9.1
- Apr 15 V0.9.2
- Aug 15 V0.9.3
- Mar 16 V0.9.4
- Aug 16 V0.9.5
- Dec 16 SAI 1.0
- Mar 17 V1.1
- Jun 17 V1.2
- Sep 17 V1.3
- Dec 17 SAI 2.0

1/1/2014 to 1/1/2018
What Was Added to SAI 1.0
Enhanced ACL Model Proposal

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

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

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

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, Bridge

- Add flexibility – select the packet send/receive interface according to
  - \{packet type, port\}
  - \{packet type, Vlan\}
SAI Roadmap 2017

Monitoring
- TAM [Broadcom]
- Microburst [Marvell]
- Critical Resource Monitoring [MSFT]
- INT [Barefoot]

Protocol Support
- MPLS [Mellanox]
- 802.1BR [Dell]
- Segment Routing [Cavium]
- Open flow Extension [Cavium]

Reliability/QoS
- L3 Fast Reroute [Metaswitch]
- BFD [Dell]
- ECN [Dell]

Infrastructure
- SAI P4 Model [Mellanox]
- Multi-NPU [Dell]
- Capability Query [MSFT]
- SAI Ext API [Dell]
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

Speaker: Vitaly Vovnoboys
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

Speaker: Prem Jonnalagadda
Experimental Trak
In-band Network Telemetry (INT)

"Insert: switchID, time, matched rules, queue occupancy, switch metadata, ..., ..., ..."
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

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
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
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