

|  |
| --- |
|  |

Switch Abstraction Interface

Change Proposal

|  |  |
| --- | --- |
| **Title** | **IP multicast** |
| **Authors** | **CENTEC** |
| **Status** | **In Review** |
| **Type** | **Standards Track** |
| **Created** | **17/10/2015** |
| **SAI-Version** | **0.9.3** |

**Contents**

[List of Changes i](#_Toc457504291)

[1 Overview 1](#_Toc457504292)

[2 Specification 1](#_Toc457504293)

[2.1 Changes to sai.h 1](#_Toc457504294)

[2.2 Change to sairouterinterface.h 1](#_Toc457504295)

[2.3 Change to sainexthopgroup.h 1](#_Toc457504296)

[2.4 New definitions in saiipmc.h 1](#_Toc457504297)

[2.5 Change to saihostif.h 4](#_Toc457504298)

[3 Examples 4](#_Toc457504299)

[3.1 Create nexthop group 4](#_Toc457504300)

[3.2 Create IPMC entry 4](#_Toc457504301)

[4 Appendix (review meeting questions) 5](#_Toc457504302)

# List of Changes

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Changes | Name | Date |
| 0.9.3 | Base version |  | 17/10/2015 |
| 0.9.3 | Redesign | Min Yao | 17/10/2015 |

License

© 2014 Microsoft Corporation, Dell Inc., Facebook, Inc, Broadcom Corporation, Intel Corporation, Mellanox Technologies Ltd.

As of September 9, 2014, the following persons or entities have made this Specification available under the Open Web Foundation Final Specification Agreement (OWFa 1.0), which is available at <http://www.openwebfoundation.org/legal/the-owf-1-0-agreements/owfa-1-0>

Microsoft Corporation, Dell Inc., Facebook, Inc, Intel Corporation, Mellanox Technologies Ltd.

You can review the signed copies of the Open Web Foundation Agreement Version 1.0 for this Specification at <http://opencompute.org/licensing/>, which may also include additional parties to those listed above.

Your use of this Specification may be subject to other third party rights. THIS SPECIFICATION IS PROVIDED "AS IS." The contributors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, noninfringement, fitness for a particular purpose, or title, related to the Specification. The entire risk as to implementing or otherwise using the Specification is assumed by the Specification implementer and user. IN NO EVENT WILL ANY PARTY BE LIABLE TO ANY OTHER PARTY FOR LOST PROFITS OR ANY FORM OF INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION OR ITS GOVERNING AGREEMENT, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND WHETHER OR NOT THE OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

THE FOLLOWING IS A LIST OF MERELY REFERENCED TECHNOLOGY: Microprocessor technology, semiconductor manufacturing technology, operating system technology (including without limitation networking operating system technology), emulation technology, graphics technology, video technology, integrated circuit packaging technology and the like, compiler technologies, object oriented technology, optical/RF communications technology including chip I/O and driver technology, bus technology, memory chip technology (including, without limitation, NAND memory, NOR memory, resistive RAM (RRAM), seek scan probe (SSP) memory, nonvolatile memory (including without limitation, memory based on chalcogenide materials, phase change memory (PCM), one or more stacked layers of memory cells, embedded PCM memories, non-volatile cache memory, solid state drives, SRAM, embedded DRAM, ferro-electric memory, and polymer memory)) and/or health-related and medical technology. IMPLEMENTATION OF THESE TECHNOLOGIES MAY BE SUBJECT TO THEIR OWN LEGAL TERMS.

# Overview

IP Multicast provides a solution to eliminate traffic redundancy and reduce server and CPU loads. This document covers details of IP multicast object –IP multicast group creating, IP multicast group removing, IP multicast group attributes setting and getting, RPF interface adding to IP multicast group, and RPF interface removing from IP multicast group.

# Specification

[This section describes the details of the proposed interface/API]

## Changes to sai.h

/\*

typedef enum \_sai\_api\_t {

SAI\_API\_IPMC= 45, /\* sai\_ipmc\_api\_t \*/

} sai\_api\_t;

## Change to sairouterinterface.h

/\*\*

\* @brief Routing interface attribute IDs

\*/

typedef enum \_sai\_router\_interface\_attr\_t

{

/\*\* READ-WRITE \*/

**/\*\* V4 Mcast enable [bool] (CREATE\_AND\_SET) (default to TRUE) \*/**

**SAI\_ROUTER\_INTERFACE\_ATTR\_V4\_MCAST\_ENABLE,**

**/\*\* V6 Mcast enable [bool] (CREATE\_AND\_SET) (default to TRUE) \*/**

**SAI\_ROUTER\_INTERFACE\_ATTR\_V6\_MCAST\_ENABLE,**

} sai\_router\_interface\_attr\_t;

## Change to sainexthopgroup.h

/\*

\* Next hop group type

\*/

typedef enum \_sai\_next\_hop\_group\_type\_t

{

/\*\* Next hop group is ECMP \*/

SAI\_NEXT\_HOP\_GROUP\_ECMP,

**/\*\* Next hop group is IP multicast, group member could be routeinterface/l3tunnel \*/**

**SAI\_NEXT\_HOP\_GROUP\_L3MC,**

/\* Other types of next hop group to be defined in the future, e.g., WCMP \*/

} sai\_next\_hop\_group\_type\_t;

## New definitions in saiipmc.h

/\*\*

\* @brief Enum defining ipmc entry types.

\*/

typedef enum \_sai\_ipmc\_entry\_type\_t

{

/\*\* IPMC entry with type (S,G) \*/

SAI\_IPMC\_TYPE\_SG = 0x00000001,

/\*\* IPMC entry with type (\*,G) \*/

SAI\_IPMC\_TYPE\_XG = 0x00000002,

} sai\_ipmc\_entry\_type\_t;

/\*

\* IP multicast entry key

\*/

typedef struct \_sai\_ipmc\_entry\_t

{

sai\_object\_id\_t vrf\_id;

sai\_ipmc\_entry\_type\_t type;

sai\_ip\_address\_t destination;

sai\_ip\_address\_t source;

} sai\_ipmc\_entry\_t;

/\*

\* Attribute Id for IP multicast entry

\*/

typedef enum \_sai\_ipmc\_entry\_attr\_t

{

SAI\_IPMC\_ENTRY\_ATTR\_START,

/\*\* READ-ONLY \*/

/\*\* READ-WRITE \*/

/\*\* Packet action [sai\_packet\_action\_t]

(default to SAI\_PACKET\_ACTION\_FORWARD) \*/

SAI\_IPMC\_ATTR\_PACKET\_ACTION = SAI\_IPMC\_ENTRY\_ATTR\_START,

/\*\* Packet priority for trap/log actions [uint8\_t]

(default to 0) \*/

SAI\_IPMC\_ATTR\_TRAP\_PRIORITY,

/\*\* Next hop group id for the packet [sai\_object\_id\_t]

\* The next hop group type should be SAI\_NEXT\_HOP\_GROUP\_IPMC

\* (MANDATORY\_ON\_CREATE when SAI\_IPMC\_ATTR\_PACKET\_ACTION == SAI\_PACKET\_ACTION\_FORWARD)

\* (CREATE\_AND\_SET) \*/

SAI\_IPMC\_ATTR\_NEXT\_HOP\_GROUP\_ID,

/\* Custom range base value \*/

SAI\_IPMC\_ATTR\_CUSTOM\_RANGE\_BASE = 0x10000000

} sai\_ipmc\_entry\_attr\_t;

/\*

\* Routine Description:

\* Create IP multicast entry

\*

\* Arguments:

\* [in] ipmc\_entry - IP multicast entry

\* [in] attr\_count - number of attributes

\* [in] attr\_list - array of attributes

\*

\* Return Values:

\* SAI\_STATUS\_SUCCESS on success

\* Failure status code on error

\*/

typedef sai\_status\_t (\*sai\_create\_ipmc\_entry\_fn)(

\_In\_ const sai\_ipmc\_entry\_t\* ipmc\_entry,

\_In\_ uint32\_t attr\_count,

\_In\_ const sai\_attribute\_t \*attr\_list

);

/\*

\* Routine Description:

\* Remove IP multicast entry

\*

\* Arguments:

\* [in] ipmc\_entry - IP multicast entry

\*

\* Return Values:

\* SAI\_STATUS\_SUCCESS on success

\* Failure status code on error

\*/

typedef sai\_status\_t (\*sai\_remove\_ipmc\_entry\_fn)(

\_In\_ const sai\_ipmc\_entry\_t\* ipmc\_entry

);

/\*

\* Routine Description:

\* Set IP multicast entry attribute value

\*

\* Arguments:

\* [in] IP multicast - IP multicast entry

\* [in] attr - attribute

\*

\* Return Values:

\* SAI\_STATUS\_SUCCESS on success

\* Failure status code on error

\*/

typedef sai\_status\_t (\*sai\_set\_ipmc\_entry\_attribute\_fn)(

\_In\_ const sai\_ipmc\_entry\_t\* ipmc\_entry,

\_In\_ const sai\_attribute\_t \*attr

);

/\*

\* Routine Description:

\* Get IP multicast entry attribute value

\*

\* Arguments:

\* [in] ipmc\_entry - IP multicast entry

\* [in] attr\_count - number of attributes

\* [inout] attr\_list - array of attributes

\*

\* Return Values:

\* SAI\_STATUS\_SUCCESS on success

\* Failure status code on error

\*/

typedef sai\_status\_t (\*sai\_get\_ipmc\_entry\_attribute\_fn)(

\_In\_ const sai\_ipmc\_entry\_t\* ipmc\_entry,

\_In\_ uint32\_t attr\_count,

\_Inout\_ sai\_attribute\_t \*attr\_list

);

/\*

\* IP multicast method table retrieved with sai\_api\_query()

\*/

typedef struct \_sai\_ipmc\_api\_t

{

sai\_create\_ipmc\_entry\_fn create\_ipmc\_entry;

sai\_remove\_ipmc\_entry\_fn remove\_ipmc\_entry;

sai\_set\_ipmc\_entry\_attribute\_fn set\_ipmc\_entry\_attribute;

sai\_get\_ipmc\_entry\_attribute\_fn get\_ipmc\_entry\_attribute;

} sai\_ipmc\_api\_t;

## Change to sairouter.h

typedef enum \_sai\_virtual\_router\_attr\_t

{

SAI\_VIRTUAL\_ROUTER\_ATTR\_START,

/\*\* READ-WRITE \*/

**/\*\* Action for Unknown L3 multicast packets [sai\_packet\_action\_t]**

**\* (default to SAI\_PACKET\_ACTION\_DROP) \*/**

**SAI\_VIRTUAL\_ROUTER\_ATTR\_UNKNOWN\_L3\_MULTICAST\_PACKET\_ACTION,**

} sai\_virtual\_router\_attr\_t;

## Change to saihostif.h

typedef enum \_sai\_hostif\_trap\_id\_t

{

**/\*\* unknown l3 multicast packets**

**\* (default packet action is drop) \*/**

**SAI\_HOSTIF\_TRAP\_ID\_UNKNOWN\_L3\_MULTICAST,**

SAI\_HOSTIF\_TRAP\_ID\_CUSTOM\_EXCEPTION\_RANGE\_BASE = 0x00005000,

} sai\_hostif\_trap\_id\_t;

# Examples

Router interface should be created first, for example, rif1\_object, rif2\_object and rpf\_object have been created.

## Create nexthop group

sai\_api\_query(SAI\_API\_NEXT\_HOP\_GROUP, &sai\_nexthop\_group\_api);

sai\_object\_id\_t nexthop\_group\_object;

sai\_object\_id\_t nexthop\_list\_obj[2] = {0};

sai\_attribute\_t attr[2] = {0};

attr[0].id = SAI\_NEXT\_HOP\_GROUP\_ATTR\_TYPE;

attr[0].value.oid = SAI\_NEXT\_HOP\_GROUP\_L3MC;

attr[1].id = SAI\_NEXT\_HOP\_GROUP\_ATTR\_NEXT\_HOP\_LIST;

nexthop\_list\_obj [0] = rif1\_object;

nexthop\_list\_obj [1] = rif2\_object;

attr[1].value.objlist.count = 2;

attr[1].value.objlist.list = nexthop\_list\_obj;

sai\_nexthop\_group\_api->create\_next\_hop\_group(&nexthop\_group\_object, 2, attr);

## Create IPMC entry

sai\_api\_query(SAI\_API\_IPMC, &sai\_ipmc\_api);

sai\_ipmc\_entry\_t ipmc\_entry;

sai\_attribute\_t attr[1] = {0};

ipmc\_entry.vrf\_id = vr\_id\_obj;

ipmc\_entry.type = SAI\_IPMC\_TYPE\_XG;

ipmc\_entry.destination.addr\_family = SAI\_IP\_ADDR\_FAMILY\_IPV4;

ipmc\_entry.destination.addr.ip4 = 0xE0010101;

ipmc\_entry.destination.mask.ip4 = 0xFFFFFFFF;

attr[0].id = SAI\_IPMC\_ATTR\_PACKET\_ACTION;

attr[0].value.s32 = SAI\_PACKET\_ACTION\_FORWARD;

attr[1].id = SAI\_IPMC\_ATTR\_NEXT\_HOP\_GROUP\_ID;

attr[1].value.oid = nexthop\_group\_object;

sai\_ipmc\_api->create\_ipmc\_entry(&ipmc\_entry, 2, attr);

# Appendix (review meeting questions)

1. Suggest the multicast group entry to be pointers to a group structure instead of port itself. That group structure could be shared by L2 and L3 multicast, containing the port bitmap for L2 and maybe VLAN info for L3.

Answer: About IPMC application with IGMP Snooping, the logical is similar to IPUC on VLAN interface. In this case, there are no neighbor table lookup and unicast FDB table lookup logical. But it will involve multicast FDB table lookup instead.

**Update[7/23]Use nexthop group to manage both L2/L3 multicast members.**

1. What would be a IPMC event as defined? Any example?

Answer: IPMC event will notify a IPMC aging event to software. I will add an example to proposal.

**Update[7/23]In some scenarios, control stack wants to age the multicast group entry when there is no packet hitting the mcast entry in a period of time. That’s why we add the aging event. Control stack could remove the IPMC entry after receiving the aging event.**

1. How to determine if a packet would go to multicast?

Answer: the condition is that the packet is from router interface and the packets is a legal IP multicast packet. in this case, the packet should go to multicast process.

**Update[7/23]1.packet comes into “Router” in pipeline 2.Ipda is a multicast address.**

1. Please clarify the type of IPMC nexthop.

Answer: First, we add a new type nexthop group called SAI\_NEXT\_HOP\_GROUP\_ECMP. And second we define a new type of nexthop called SAI\_NEXT\_HOP\_IPMC. One IPMC nexthop group is composed by multiple ipmc nexthop.

**Update[7/23]no need to create IPMC nexthop, use routeinterface object in nexthop group directly.**

1. Why it is defined as prefix for the IPMC address? Should it be specific 32 bits?

Answer: the original definition is used for (\*,G). now prefix is changed to sai\_ip\_address\_t.and we add a new type to differentiate (S,G) and (\*,G).

1. Please detail the possible attributes of nexthop group

Answer: please refer to sai\_ipmc\_entry\_attr\_t.currently SAI\_IPMC\_ATTR\_NEXT\_HOP\_GROUP\_ID is MANDATORY\_ON\_CREATE attribute. And another attribute is rpf\_interface\_list.

1. Please describe the logical pipeline of L2 and L3 multicast. Better to separate these two to make it clearer.

Answer: yes, will add L2 and L3 pipeline into <<pipleline.vsd>> on git.