Collective Communication in a Layer 2 Clos Fat-tree

IEEE 802.1-24-0012-00-ICne

Roger Marks <roger@ethair.net>
EthAirNet Associates
+1 802 capable

14 March 2024

Introduction

Prior contributions to Nendica have argued for the importance of Nendica activity on Computing Networks, including AI Computing Networks

Contributions have indicated that collective communications are an important pattern in Computing Networks

Contribution considers Clos Fat-tree networks, common as Computing Networks

Contribution intends to demonstrate that collective multicast can be efficiently implemented in Clos Fat-tree networks at Layer 2

Related Contributions

- Data Center Collective Multicast using BARC-assigned Address Blocks

  https://ieee802.org/1/files/public/docs2024/cq-Marks-collective-multicast-0324-v00.pdf

- Implementation of Layer 2 Clos Fat-tree with Programmable Switches

  IEEE 802.1-24-0013

  https://mentor.ieee.org/802.1/documents?is_group=ICne&is_year=2024&is_dcn=0013

- Observations of a Layer 2 Clos Fat-tree

  IEEE 802.1-24-0014

  https://mentor.ieee.org/802.1/documents?is_group=ICne&is_year=2024&is_dcn=0014

Communication in Computing Network

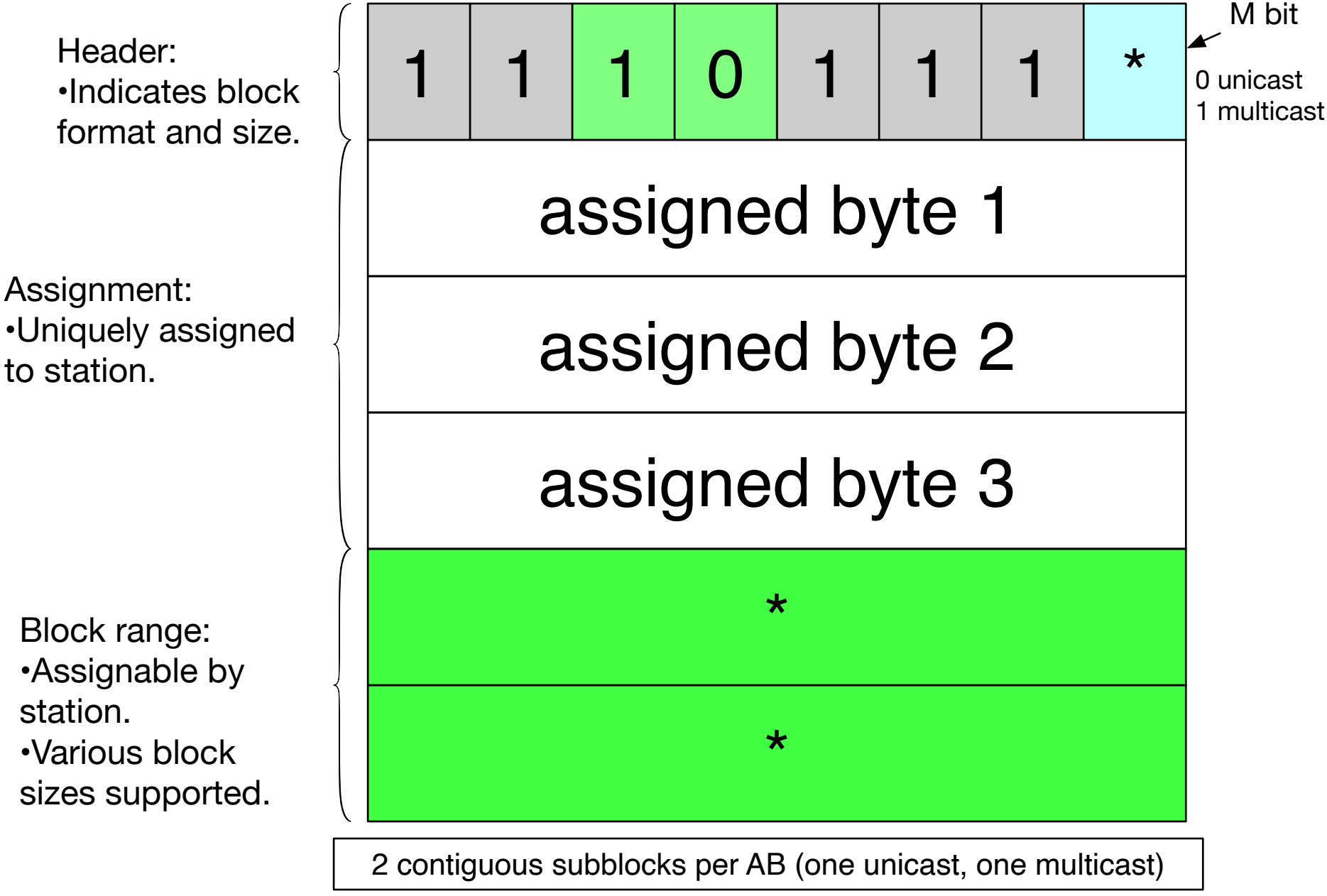
- High-performance computing and AI computing make use of distributed processing.

- A major communication pattern in computing-intensive networks, with processing distributed among various processors, is collective communication among a group of hosts.
  - noted in many Nendica contributions
  - e.g. AllReduce, which may use a collective multicast step

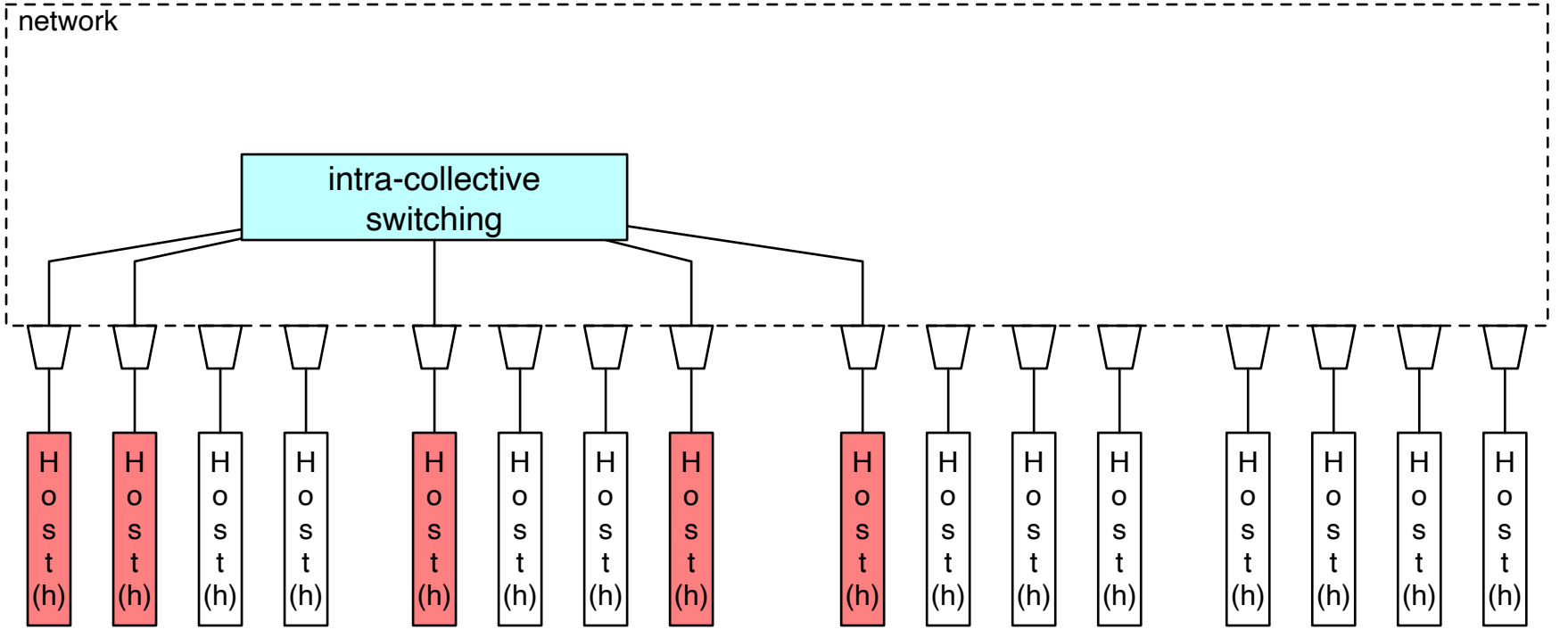
- Clos Fat-tree topology is frequently used in computing networks.

- Address Blocks, per the BARC protocol of P802.1CQ, provide the basis of simplified forwarding in a Clos Fat-tree, not only for unicast but also for collective multicast.

Example of BARC Address Block (AB) assigned to a station

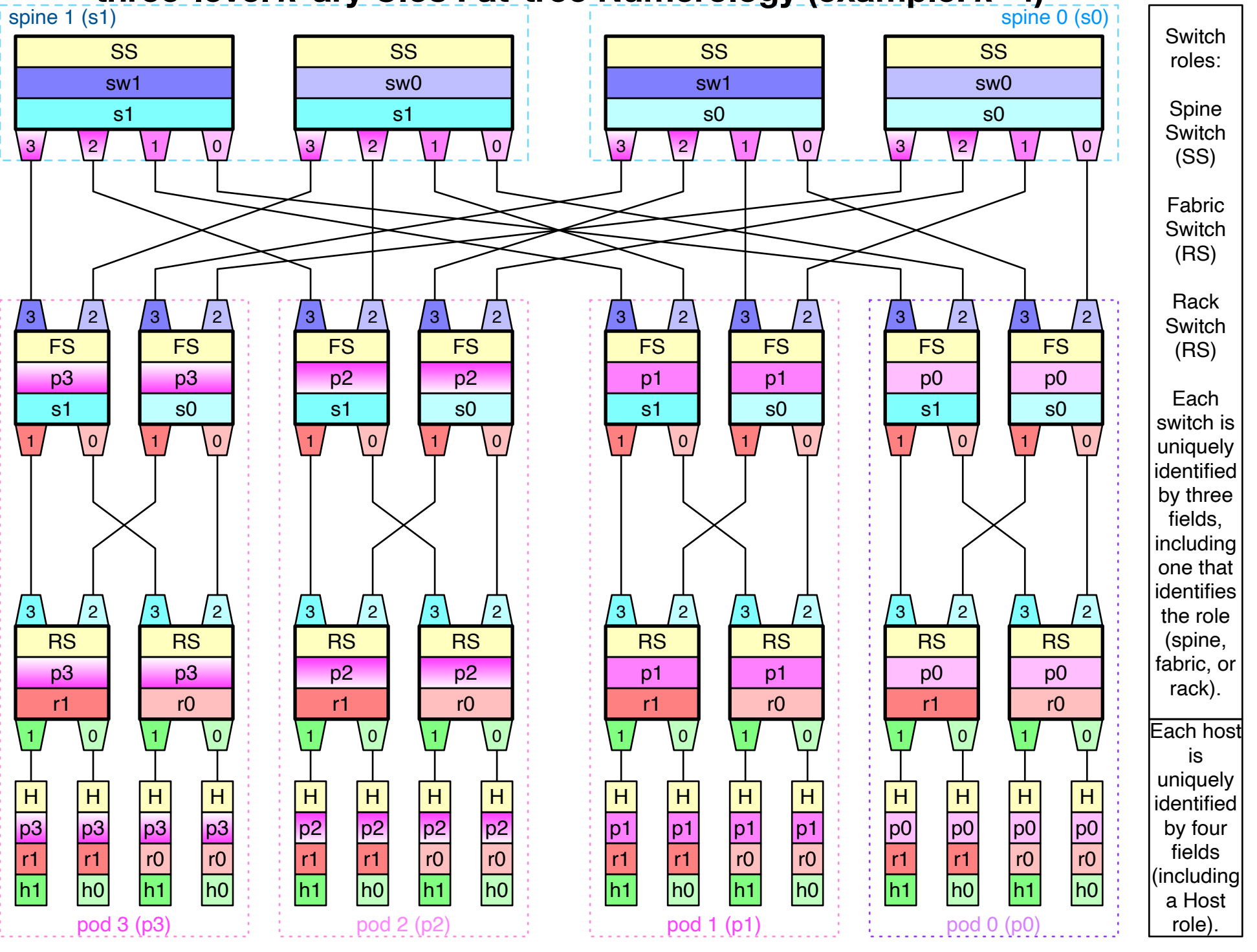


Typical Computing Network Communications: Collective



- Computation distributed among a collective of hosts
- Members of the collective communicate often among themselves
  - one useful operation is collective multicast
    - –any member sends a multicast message
    - –network delivers it to the others

three-level *k*-ary Clos Fat-tree Numerology (example: *k*=4)



Switch roles:

Spine Switch (SS)

Fabric Switch (RS)

Rack Switch (RS)

Each switch is uniquely identified by three fields, including one that identifies the role (spine, fabric, or rack).

Each host is uniquely identified by four fields (including a Host role).

Clos Fat-tree BARC Address Blocks (ABs)

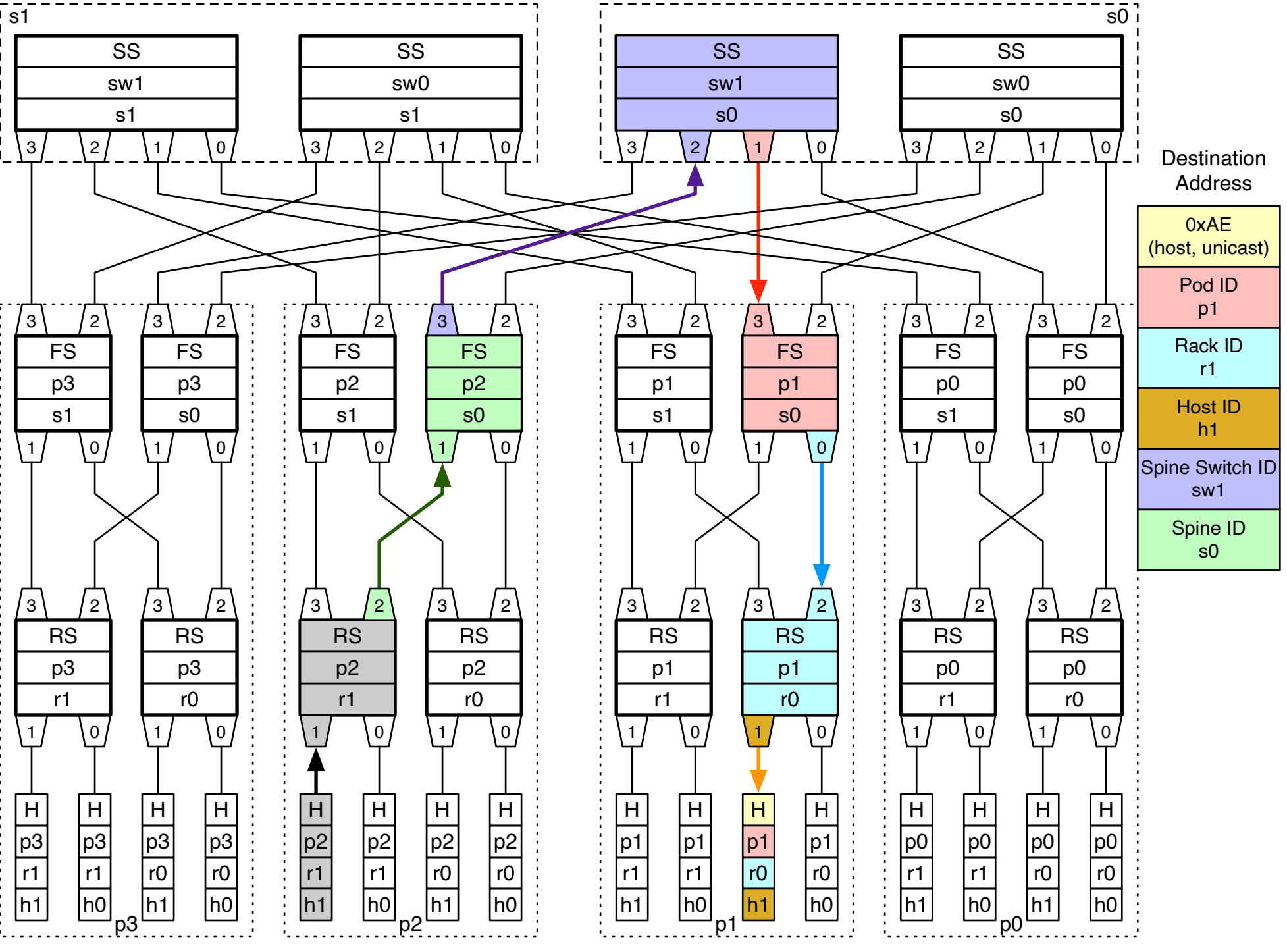
		spine switch (SS)	fabric switch (FS)	rack switch (RS)	host (H)
MSB	AB[0]	0xBE (unicast) 0xBF (multicast)	0xFE (unicast) 0xFF (multicast)	0xEE (unicast) 0xEF (multicast)	0xAE (unicast) 0xAF (multicast)
	AB[1]	Spine Switch ID (sw)	Pod ID (p)	Pod ID (p)	Pod ID (p)
	AB[2]	Spine ID (s)	Spine ID (s)	Rack ID (r)	Rack ID (r)
	AB[3]	*	*	*	Host ID (h)
	AB[4]	*	*	*	*
LSB	AB[5]	*	*	*	*

Using the specified numerology, and Address Blocks aligned with the numerology, any unicast frame can be forwarded directly toward its destination address in any of these address blocks, by any switch, without a forwarding database. The egress port can be read directly from the destination address.

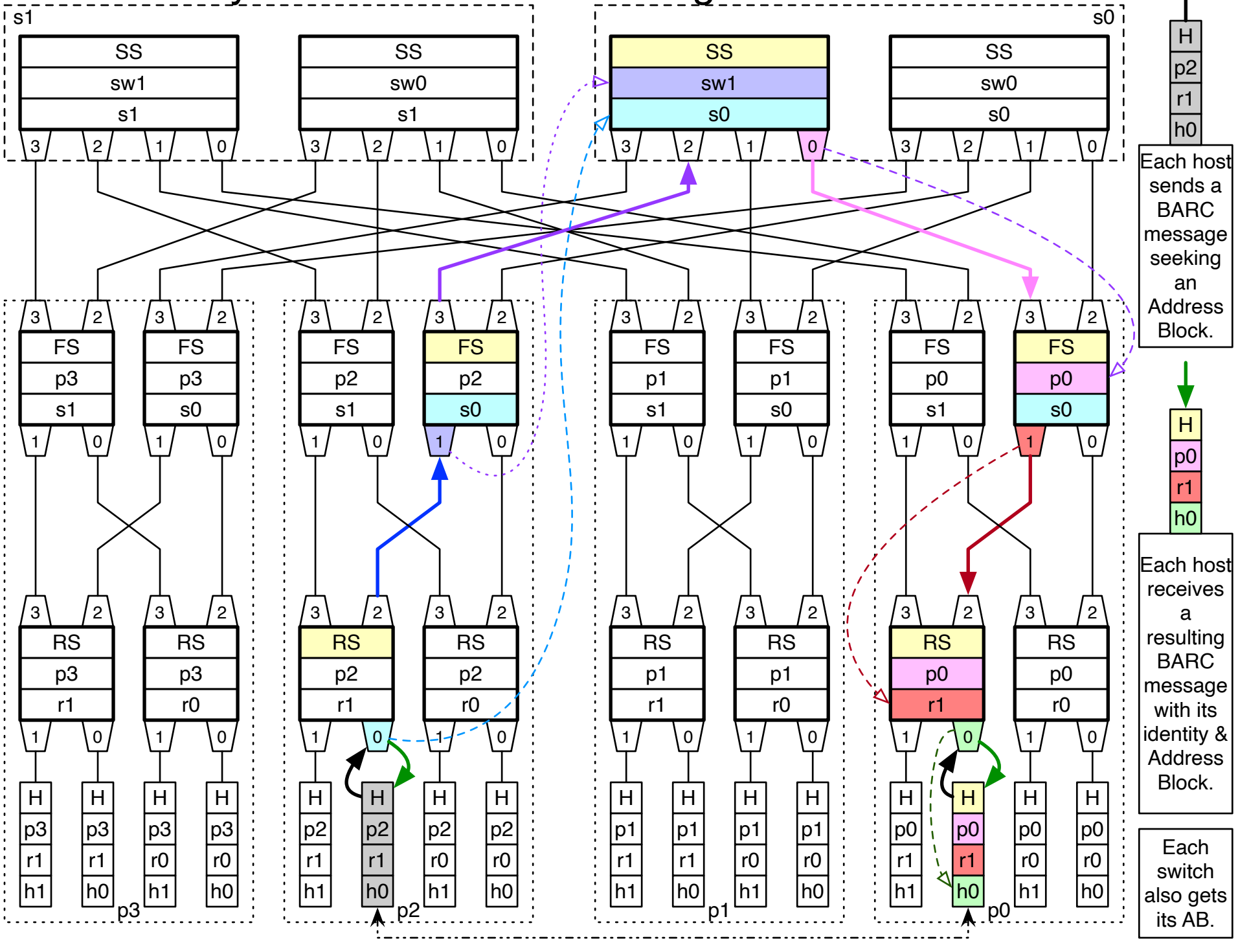
See, for example, *Address Assignment for Stateless Flow-Zone Switching in the Data Center* https://mentor.ieee.org/omniran/dcn/18/omniran-18-0059-00-CQ00-address-assignment-for-stateless-flow-zone-switching-in-the-data-center.pdf and *Stateless Flow-Zone Switching Using Software-Defined Addressing* https://ieeexplore.ieee.org/document/9424558

This dimensioning scales to over 16 Mi hosts; *k*=256 has only 4 Mi host.

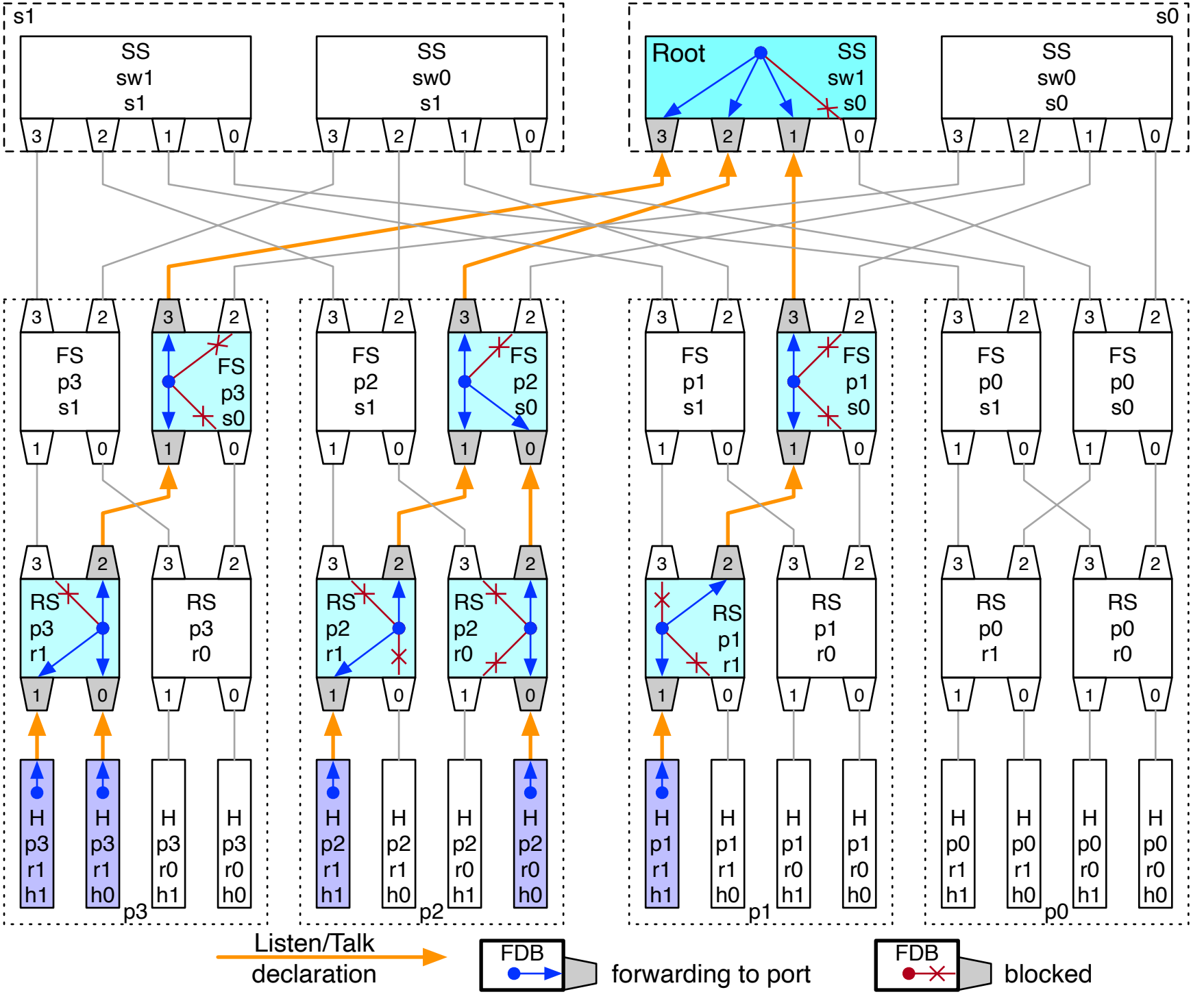
Stateless Unicast forwarding (fully source-specified)



Identity and Address Block Assignment with BARC

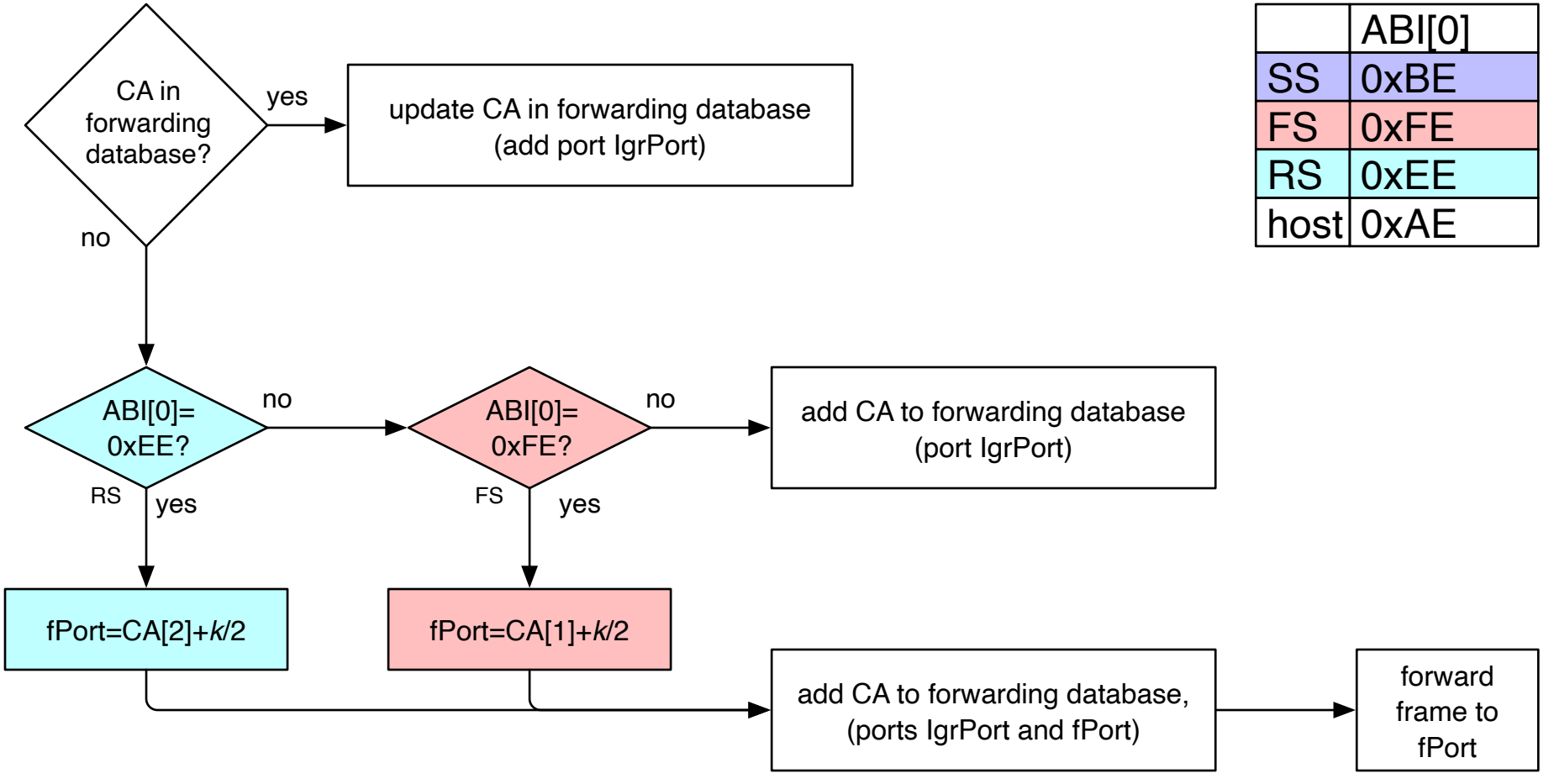


Collective Multicast in Clos Fat-tree: host-driven

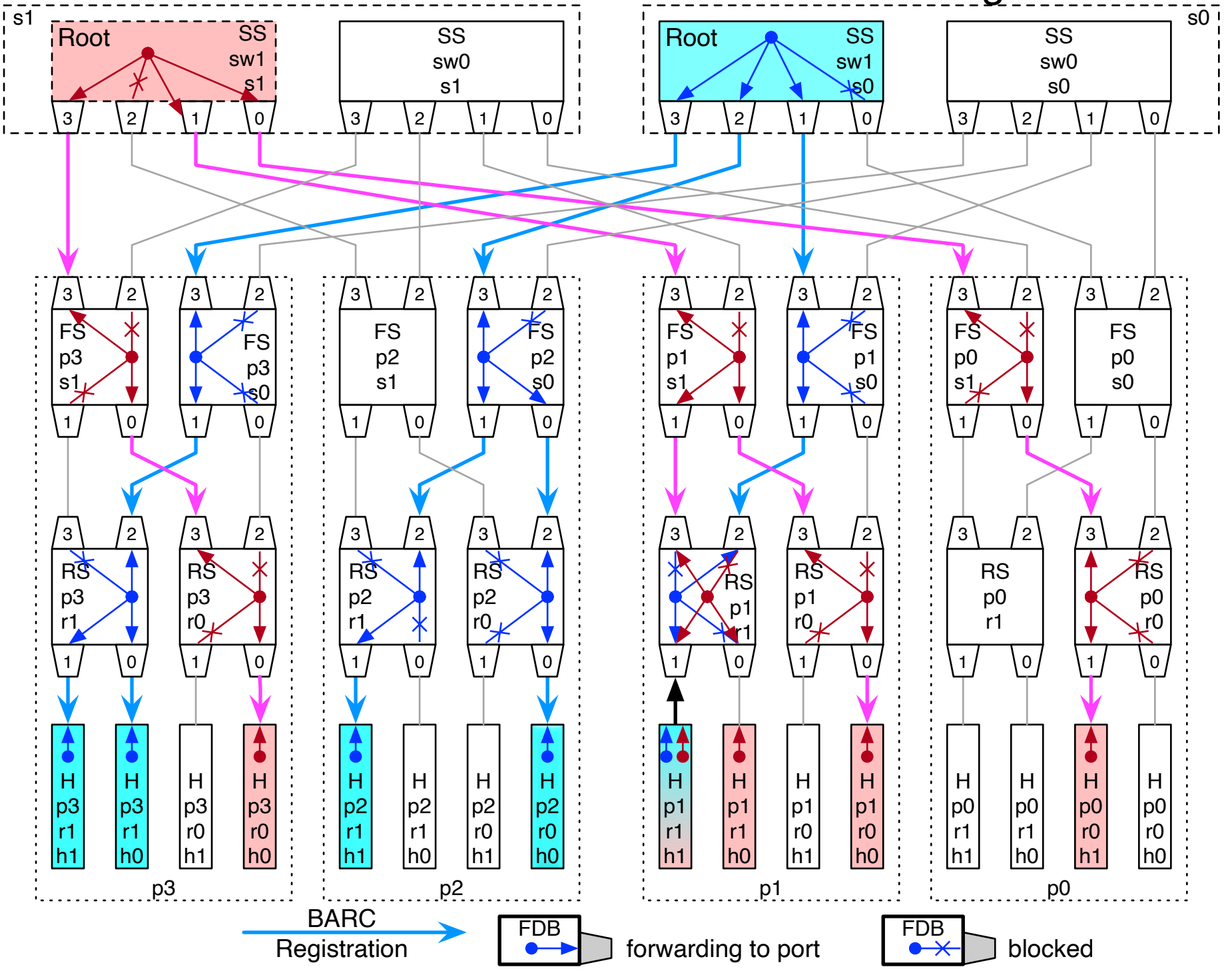


Collective Multicast Join

- DA: 01-80-C2-00-00-00 (non-forwarding Nearest Customer Bridge)
- Join frame flagged by specific EtherType and subtype
- Collective Address (CA) carried in payload
- ”IngPort” is ingress port



	ABI[0]
SS	0xBE
FS	0xFE
RS	0xEE
host	0xAE


Collective Multicast in Clos Fat-tree: root-driven registration
s1
Root
SS
sw1
s1
SS
sw0
s1
Root
SS
sw1
s0
SS
sw0
s0
s0
FS
p3
s1
FS
p3
s0
FS
p2
s1
FS
p2
s0
FS
p1
s1
FS
p1
s0
FS
p0
s1
FS
p0
s0
RS
p3
r1
RS
p3
r0
RS
p2
r1
RS
p2
r0
RS
p1
r1
RS
p1
r0
RS
p0
r1
RS
p0
r0
H
p3
r1
h1
H
p3
r1
h0
H
p3
r0
h1
H
p3
r0
h0
H
p2
r1
h1
H
p2
r1
h0
H
p2
r0
h1
H
p2
r0
h0
H
p1
r1
h1
H
p1
r1
h0
H
p1
r0
h1
H
p1
r0
h0
H
p0
r1
h1
H
p0
r1
h0
H
p0
r0
h1
H
p0
r0
h0
p3
p2
p1
p0
BARC
Registration
FDB
forwarding to port
FDB
blocked

Conclusion

- Further Nendica work on Computing Networks is warranted.

- Collective communications in Computing Networks deserves further study.

- Collective communications can be efficiently constructed for a Clos Fat-tree.
  - within Layer 2

- Further studies on this topic may be useful to Nendica progress on Computing Networks.

Nendica: IEEE 802 Network Enhancements for the Next Decade Industry Connections Activity
  - enhancing IEEE 802 Networks
  - for the next decade