Data Center Collective Multicast using BARC-assigned Address Blocks

Roger Marks
(EthAirNet Associates)
roger@ethair.net
+1 802 capable

13 March 2024

Introduction

Contribution intends to:

- Review P802.1CQ BARC Block Address assignments
  unicast/multicast Address Blocks

- Distinguish the multicast environment in computing networks from models considered in 802.1Q.

- Describe applicability of BARC Address Blocks to collective communication patterns in computing networks.

- Encourage this BARC use case as an additional reason to progress P802.1CQ.

Related Contributions

- Collective Communication in a Layer 2 Clos Fat Tree
  IEEE 802.1-24-0012
  https://mentor.ieee.org/802.1/documents?is_group=ICne&is_year=2024&is_dcn=0012

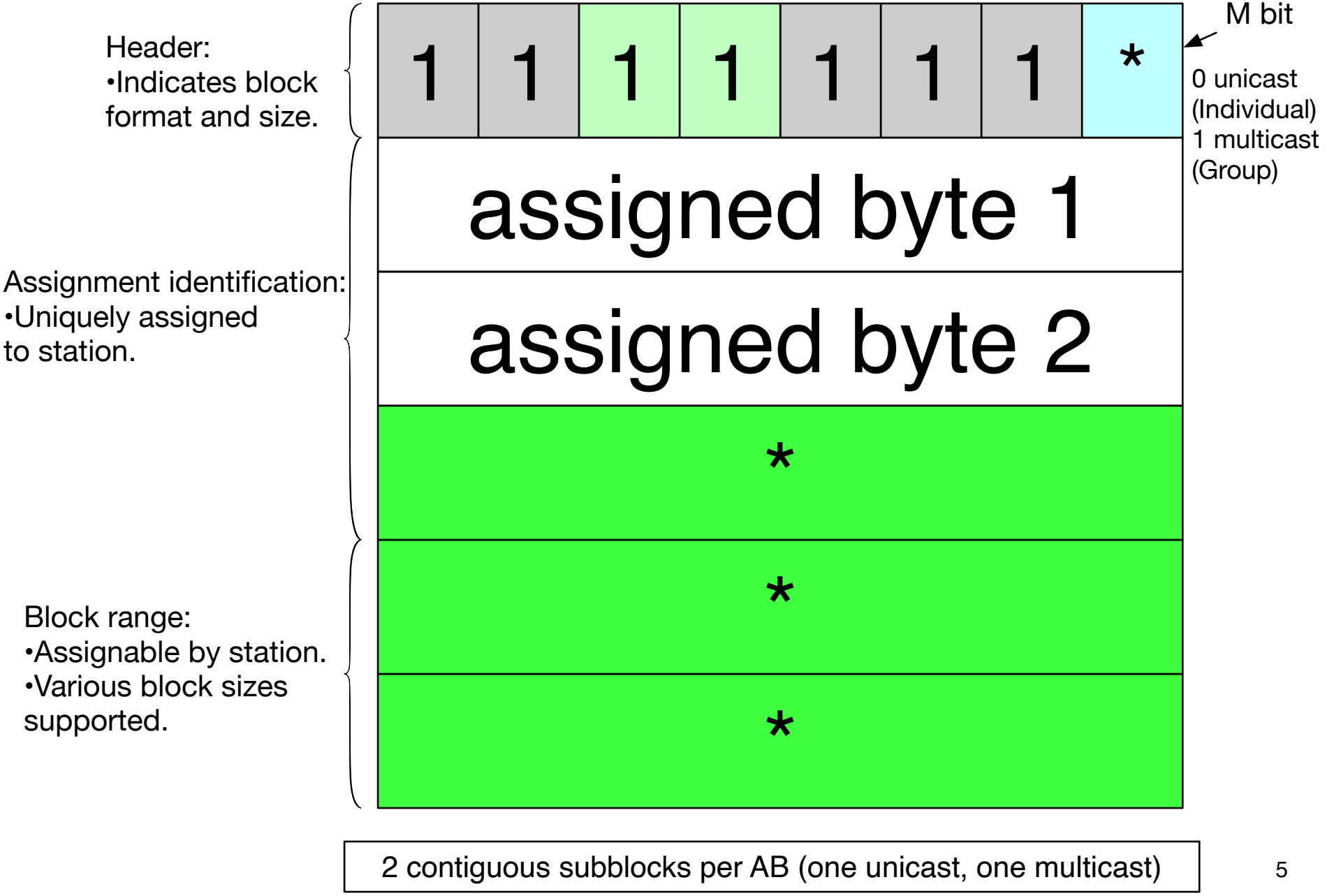
- Implementation of Layer 2 Clos Fat Tree with Programmable Switches
  IEEE 802.1-24-0013
  https://mentor.ieee.org/802.1/documents?is_group=ICne&is_year=2024&is_dcn=0013

- Observations of Configuration, Unicast, and Collective Multicast in 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

BARC Background

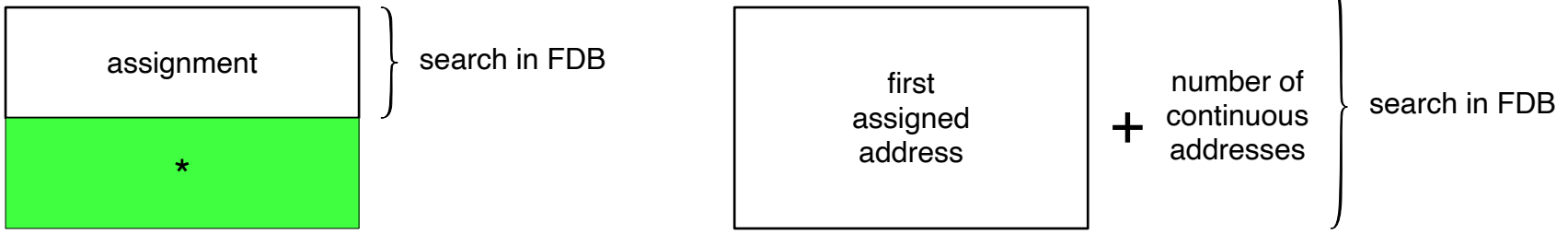
- P802.1CQ ("Multicast and Local Address Assignment") :
  - *specifies protocols, procedures, and management objects for locally-unique assignment of 48-bit and 64-bit addresses in IEEE 802 networks. Peer-to-peer address claiming and address server capabilities are specified.*
- specifically uses the **Block Address Registration and Claiming** (BARC) protocol to assign blocks of unicast and multicast addresses
- P802.1CQ/D0.8 TSN Task Group Ballot comments requested information on use cases
- A multicast use case was presented
  - *Multicast Applications of P802.1CQ BARC Address Blocks*
  - See that document for more explanation of BARC.
- This is a different multicast use case.

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



BARC Q&A

- Why assign addresses in blocks?
  - sometimes more than one address is needed
  - block sizes are variable
  - it's a compact operation
    - an address conveys the block size and the addresses within
    - no separate "number of addresses" value is necessary
  - forwarding can be based on aggregate addressing
    - e.g., forward based on the common, aggregated part of the address
      - ignore the remainder
    - may be easier than looking up a start address and a count



- Why assign Individual and Group addresses together?
  - sometimes both are needed
  - it's a compact operation
  - the network knows the owner of a group address

- Why assign multicast addresses to a station?
  - a station may need multicast address for its own transmissions
  - a station may a coordinate multicast addresses for frames sent by others

Communication in Computing Network

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

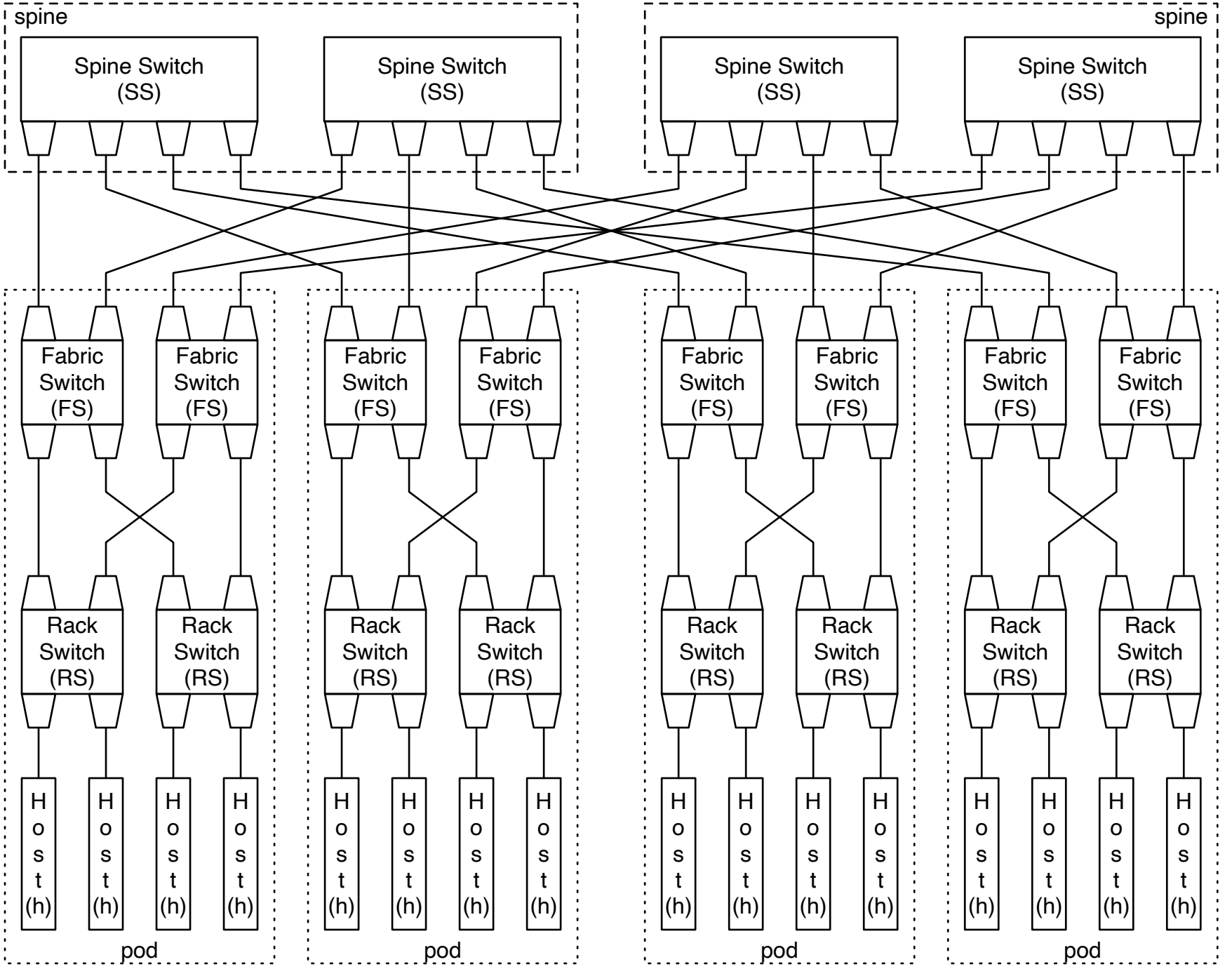
- A major communication pattern in computing-intensive networks, with parallel processing, is collective communication among a group of hosts.

- An example is collective multicast among the group.

- Data center networks are strongly multipathed, not trees.

- Data center networks could have a vast quantity of hosts, way more than the size of a group, so registration to the entire network can be inefficient.

Typical Computing Network: Clos Fat-tree
It's not a Tree!



This example of a *k*-ary Clos Fat-tree uses *k*=4.

Each switches has *k* ports.

$k^3/4$ hosts

k=64: 65 536 hosts

k=128: 524 288 hosts

Clos Fat-tree BARC Address Blocks (ABs)

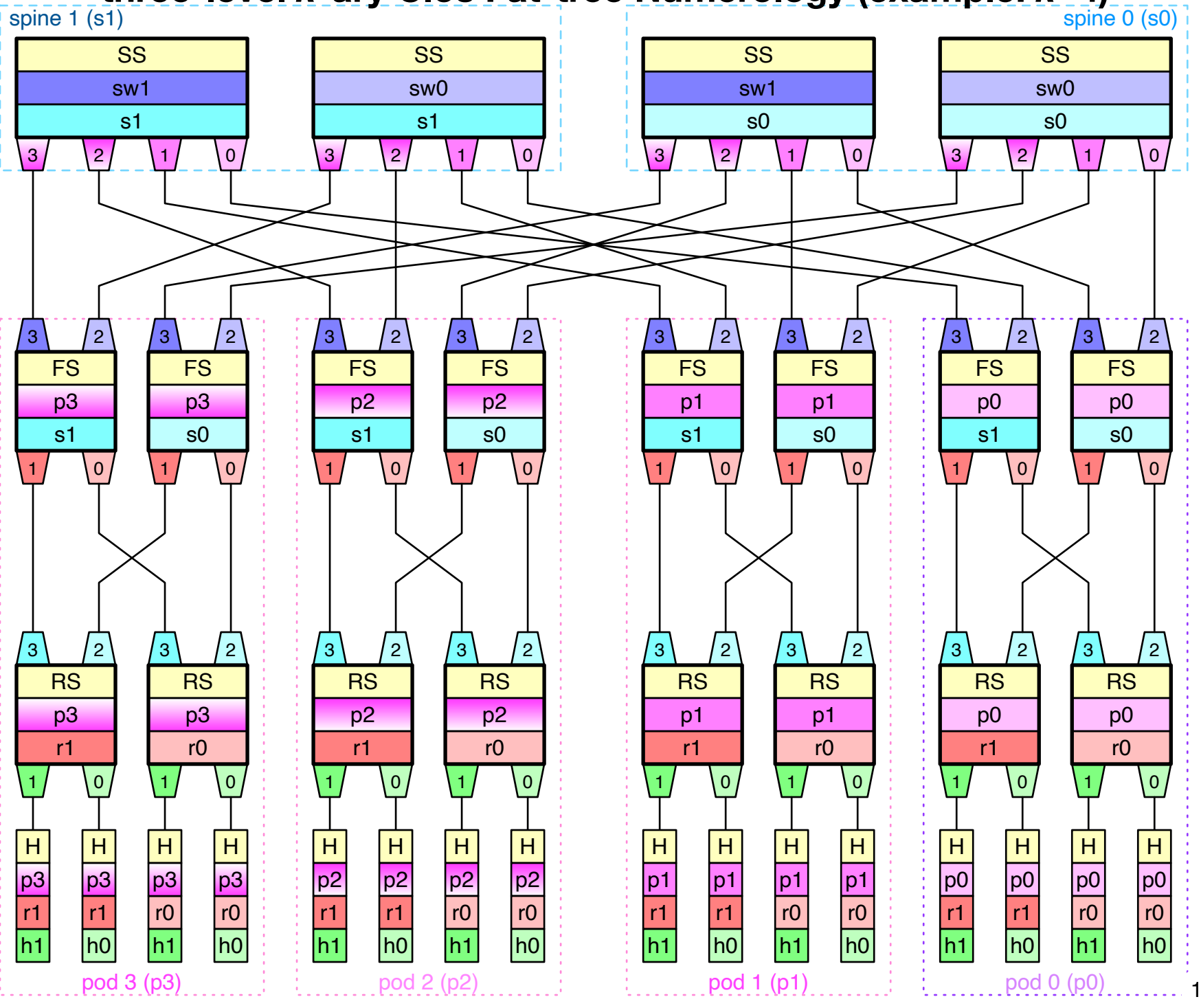
		spine switch (SS)	fabric switch (FS)	rack switch (RS)	host (H)
MSB	AB[0]	0xBE (unicast) 0xBF (multicast)	0xFE (unicast) 0xFF (multicast)	0xEE (unicast) 0xEF (multicast)	0xAE (unicast) 0xAF (multicast)
	AB[1]	Spine Switch ID (sw)	Pod ID (p)	Pod ID (p)	Pod ID (p)
	AB[2]	Spine ID (s)	Spine ID (s)	Rack ID (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 the use 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 over 4 Mi host.

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



Switch roles:

Spine Switch (SS)

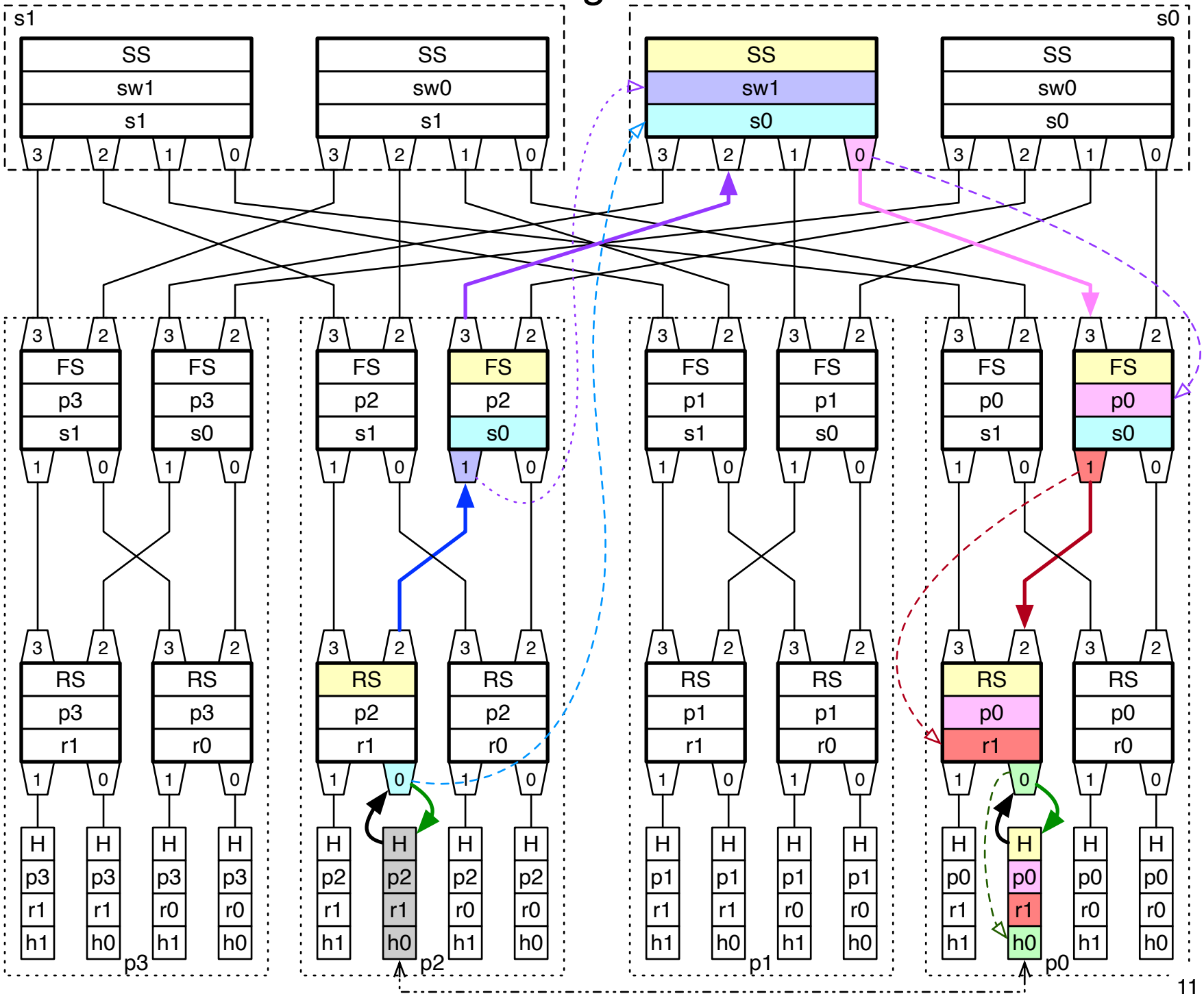
Fabric Switch (RS)

Rack Switch (RS)

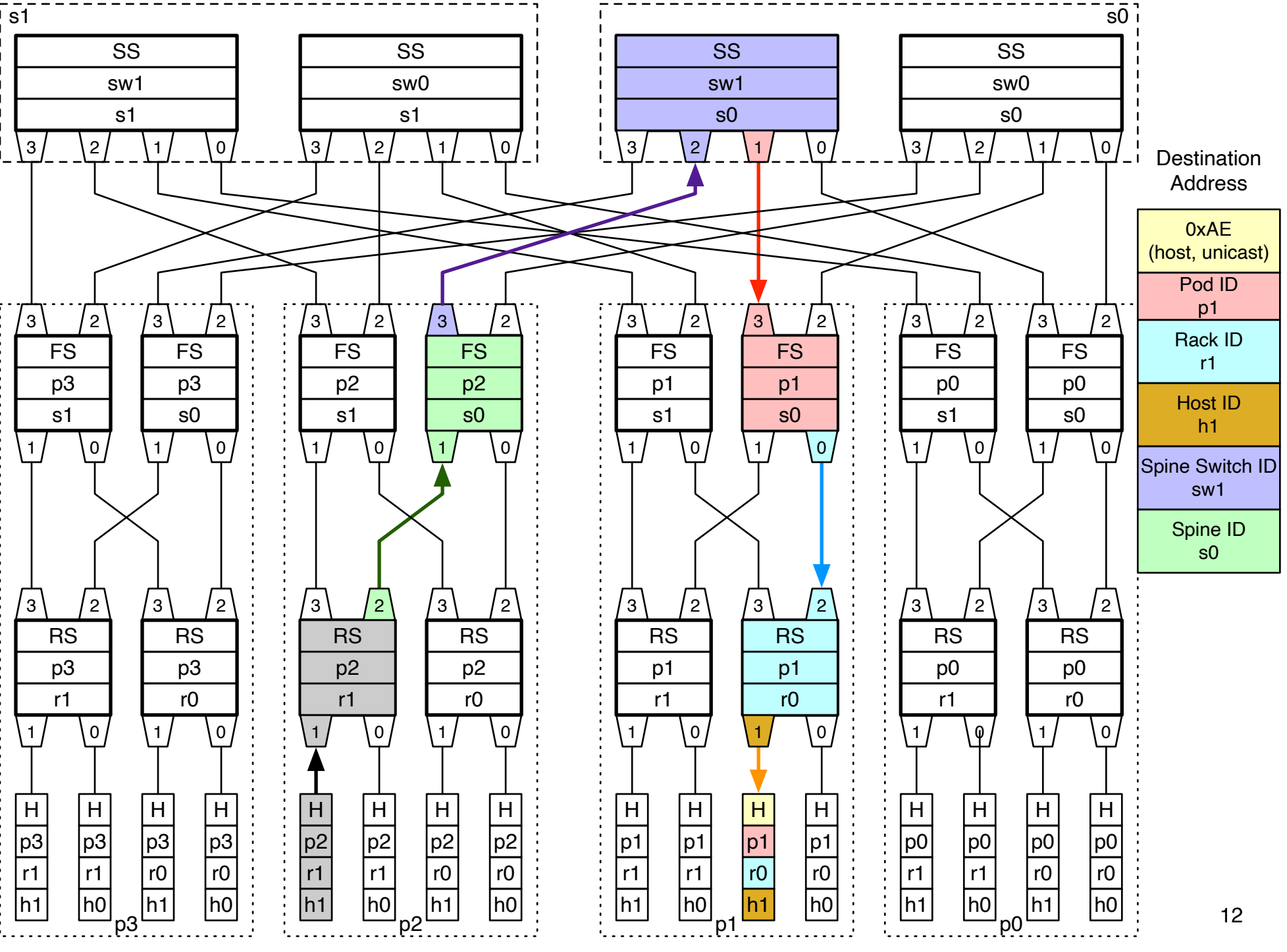
Each switch is uniquely identified by three fields, including one that identifies the role (spine, fabric, or rack).

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

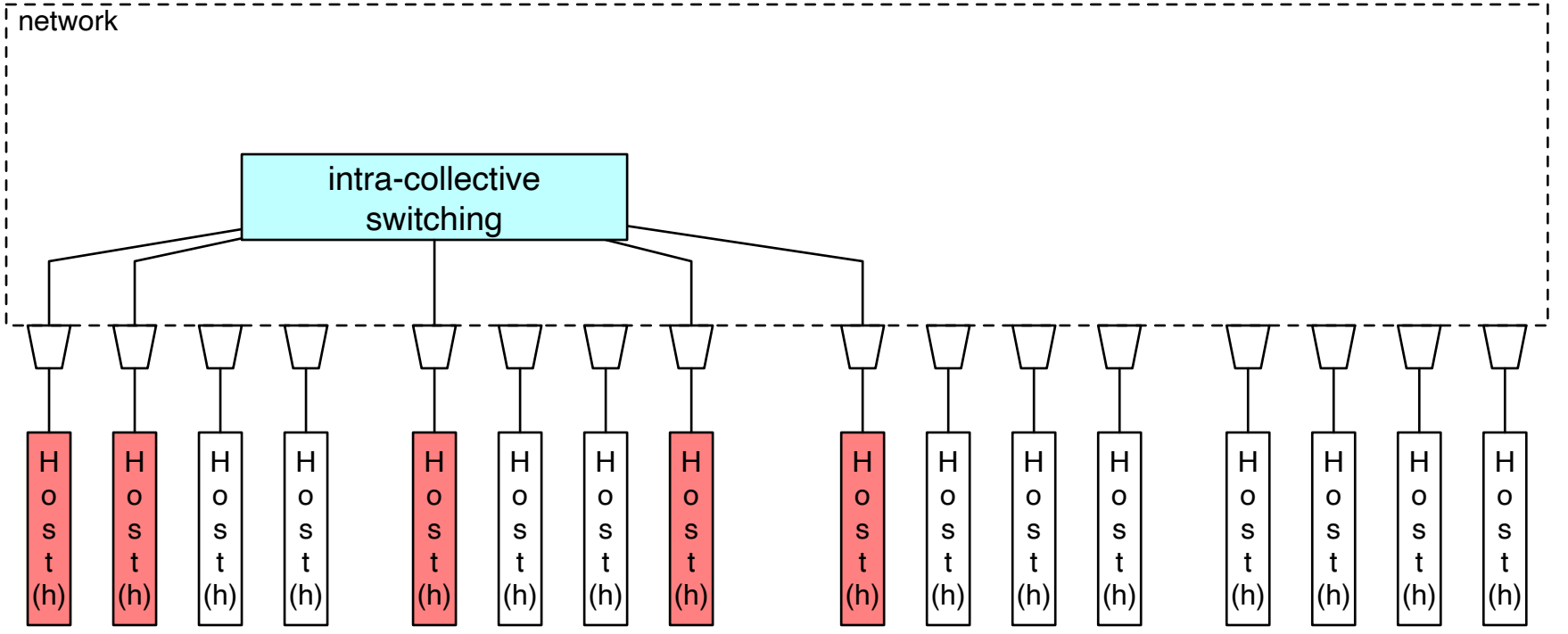
Address Block Assignment with BARC



Stateless Unicast forwarding (fully source-specified)

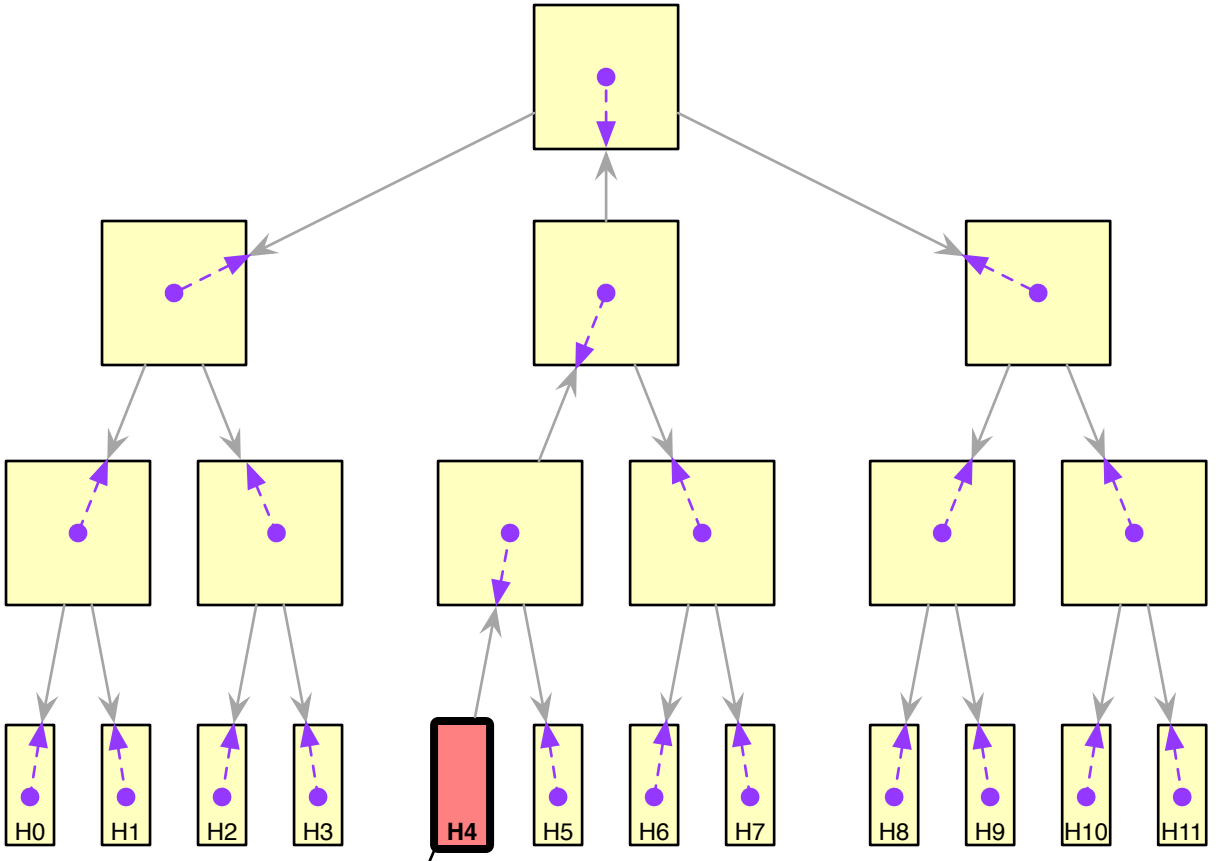


Typical Computing Network Communications: Collective



- 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

Tree network:
Listener MMRP declaration, per 802.1Q



Open Host Group

Per 802.1Q
10.10 MMRP Model of operation:

*By receiving frames from all Ports … Bridges facilitate Group distribution mechanisms based on the concept of an **Open Host Group**…Any MAC Service user that wishes to send frames to a particular Group can do so from any point of attachment to the Bridged Network.*

Listener/
declarer

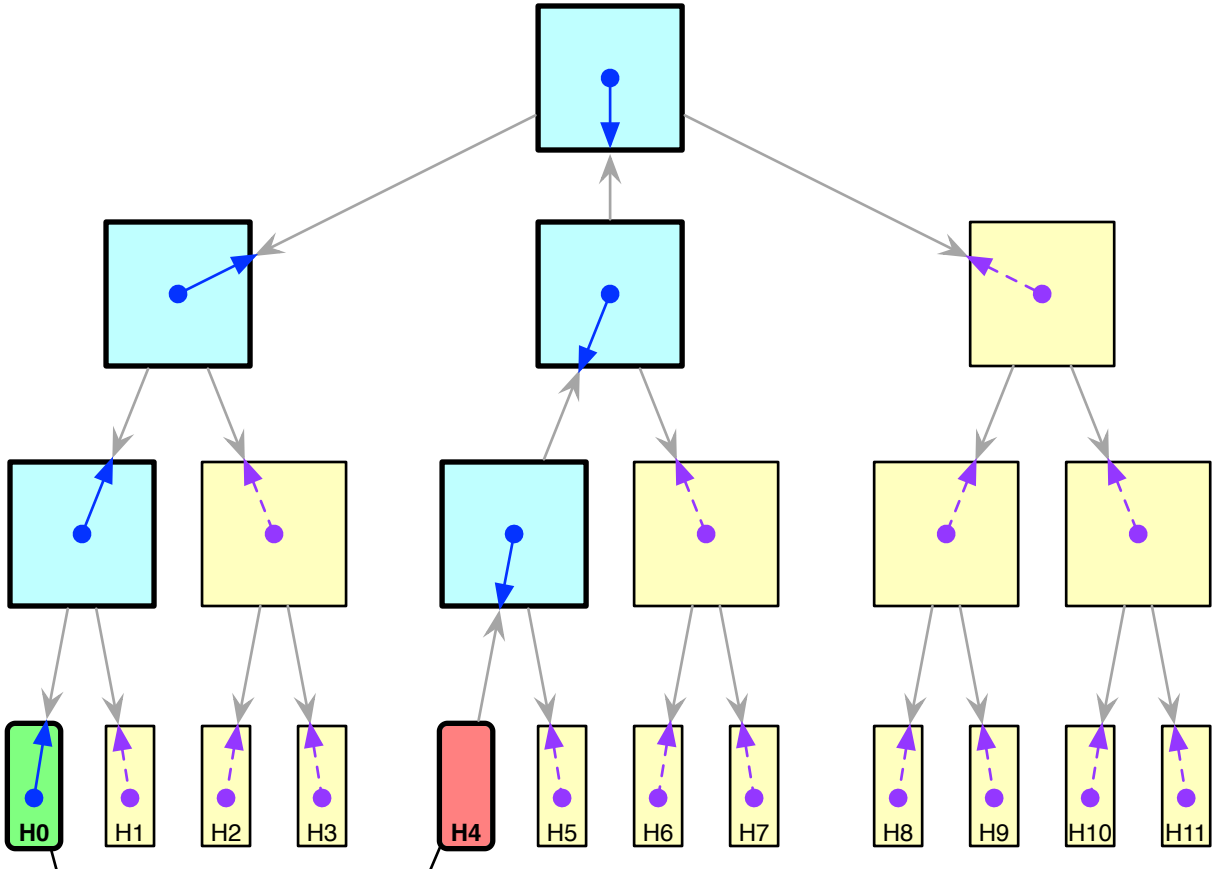
Listen
declaration

declaration of interest in receiving frames whose DA is a specified value (typically multicast)

FDB

forwarding to

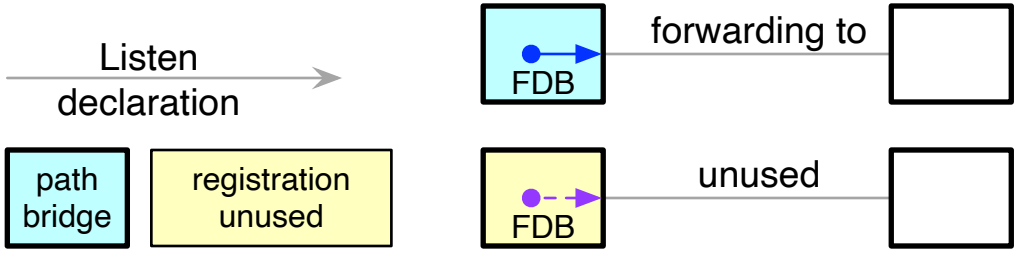
Tree network:
Listener MMRP declaration, per 802.1Q



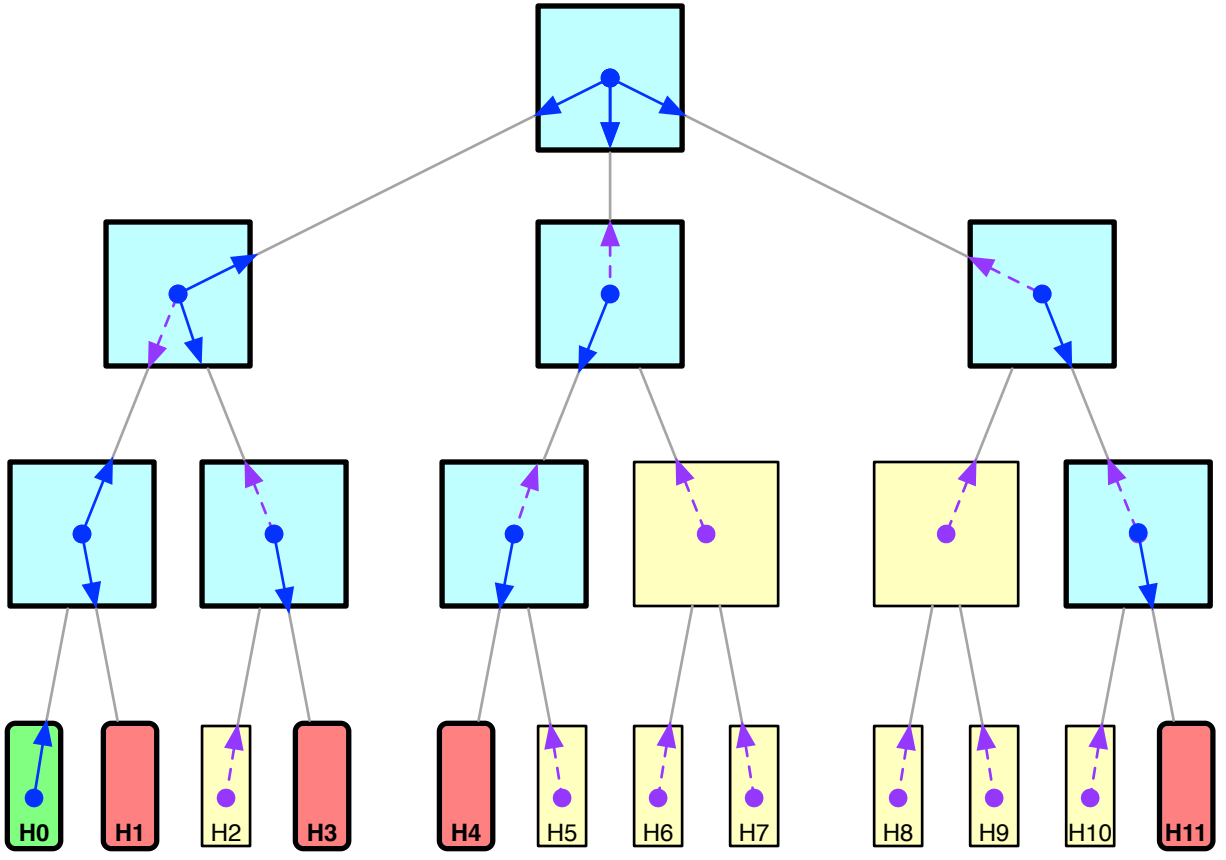
Open Host Group

Per 802.1Q
10.10 MMRP Model of operation:

*By receiving frames from all Ports … Bridges facilitate Group distribution mechanisms based on the concept of an **Open Host Group**…Any MAC Service user that wishes to send frames to a particular Group can do so from any point of attachment to the Bridged Network.*



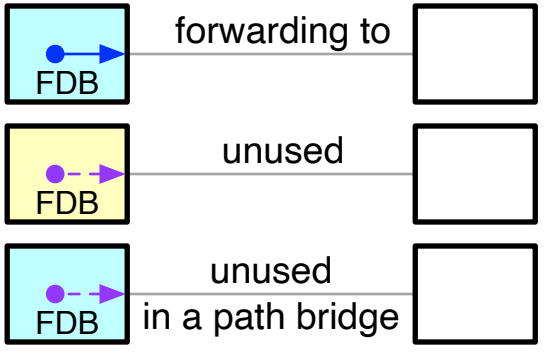
Tree network:
Listener MMRP declaration, per 802.1Q



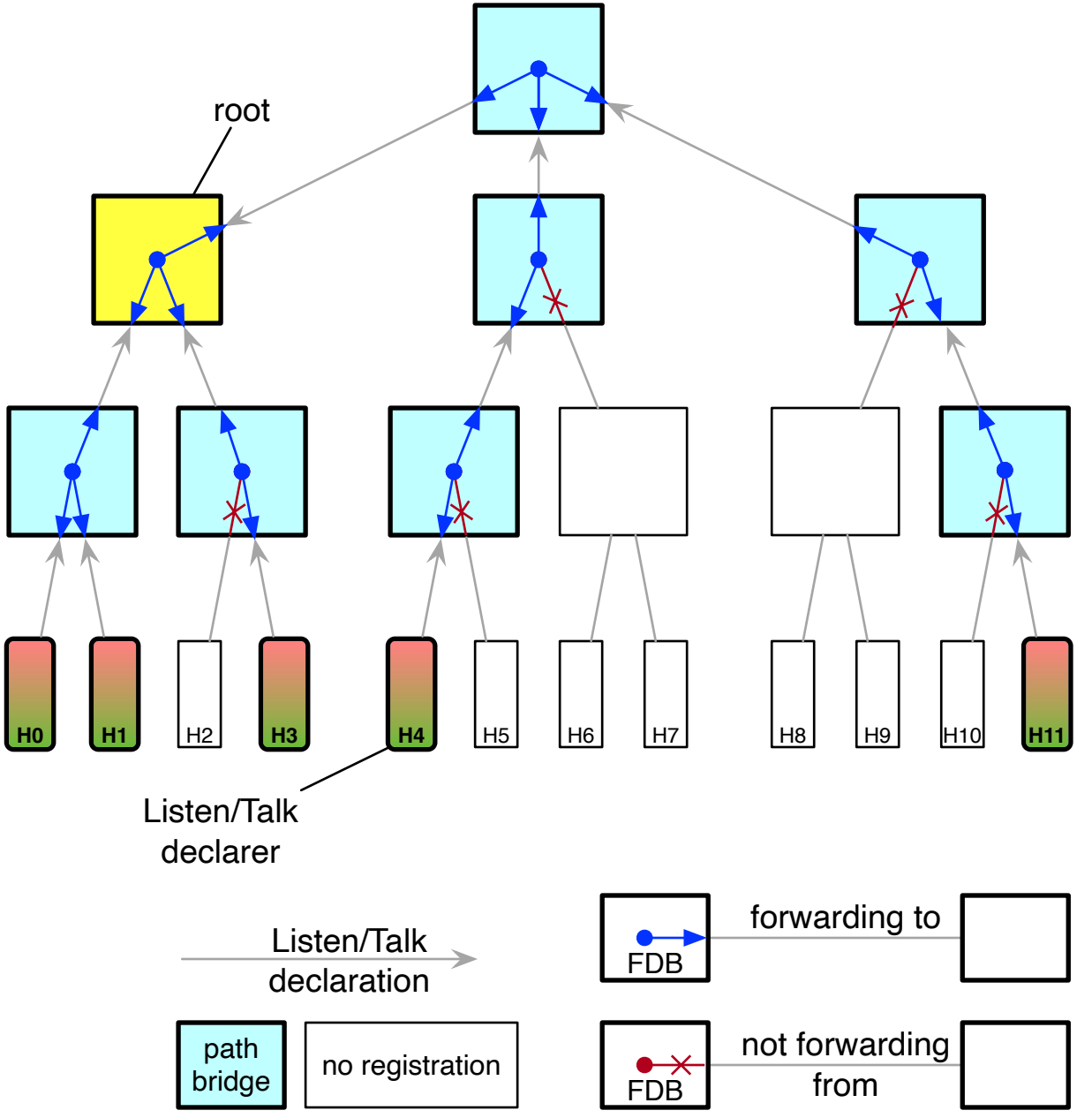
Open Host Group

Per 802.1Q
10.10 MMRP Model of operation:

*By receiving frames from all Ports … Bridges facilitate Group distribution mechanisms based on the concept of an **Open Host Group**…Any MAC Service user that wishes to send frames to a particular Group can do so from any point of attachment to the Bridged Network.*



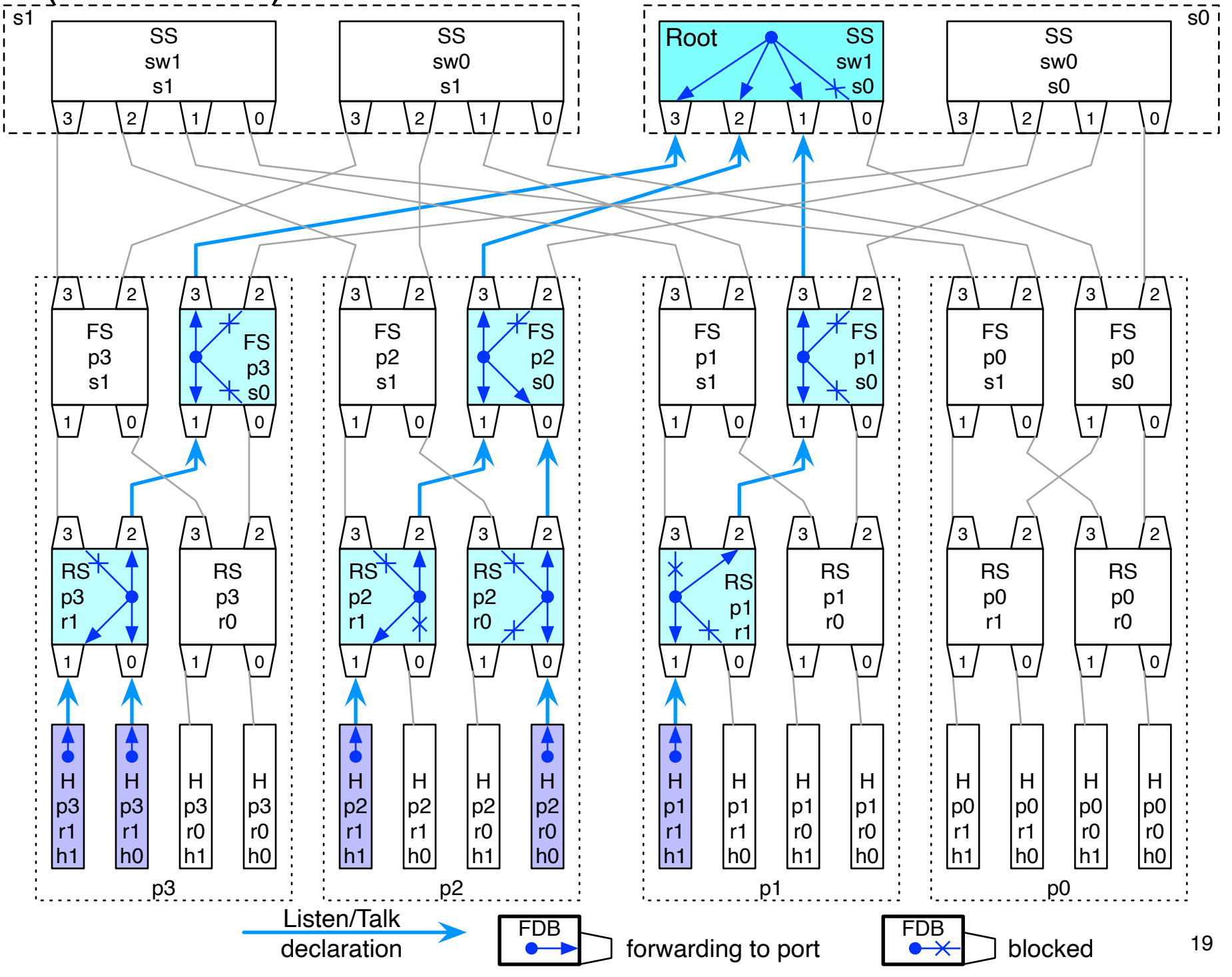
Tree network:
Closed Host Group Listen/Talk Declaration



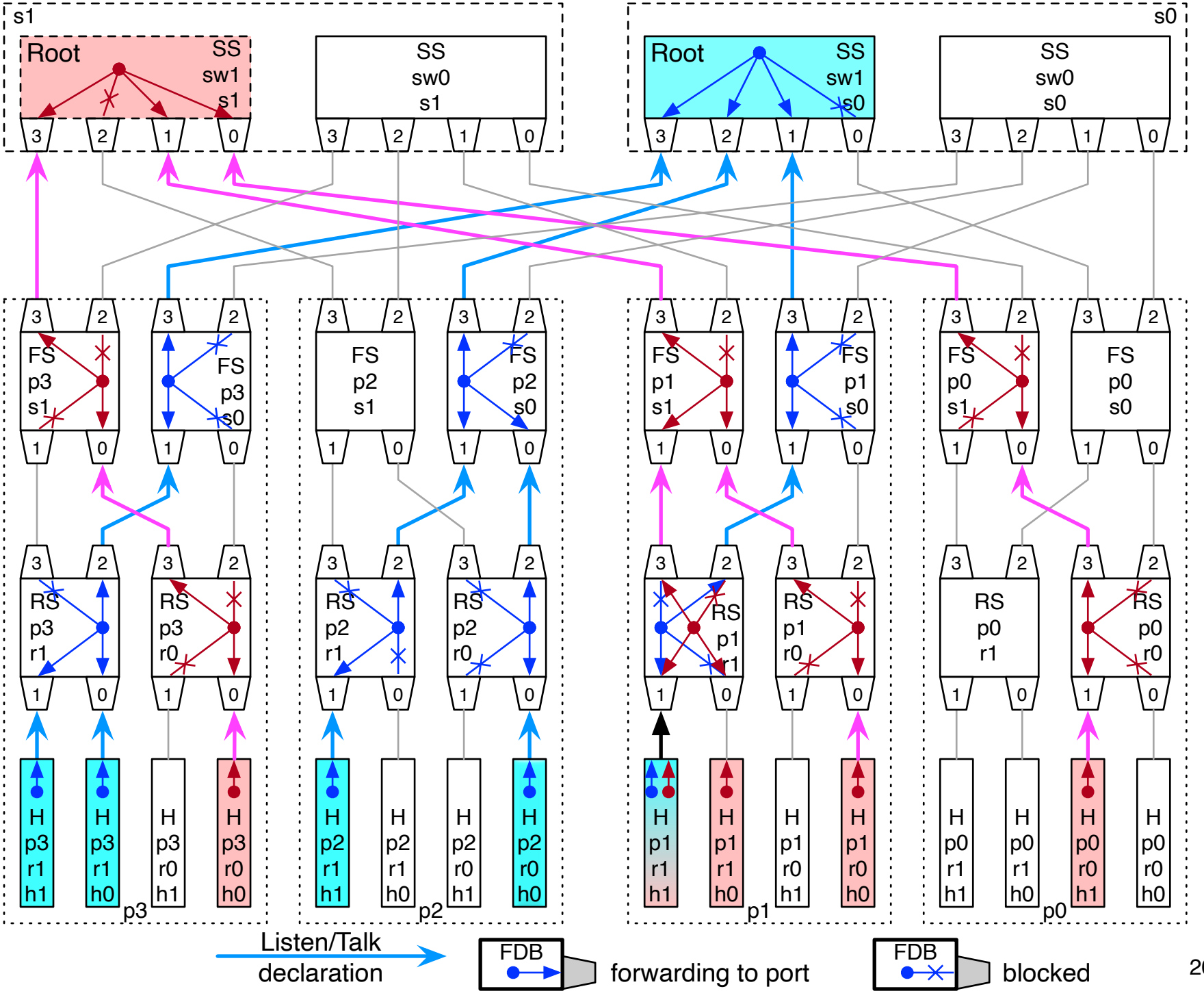
Comparison

MMRP	Clos Fat-Tree Collective Multicast
Tree	not a Tree
arbitrary tree topology	structured topology
Open Host	Closed Host: only a specific set of hosts participates
listener declaration propagates throughout	may be wasteful to propagate declaration throughout; could be many hosts and far fewer collective members
Talkers; Listeners; Listener/Talkers	Listener/Talkers (could add Talk-only & Listen-only)
	multiple roots, each managing their own collective addresses

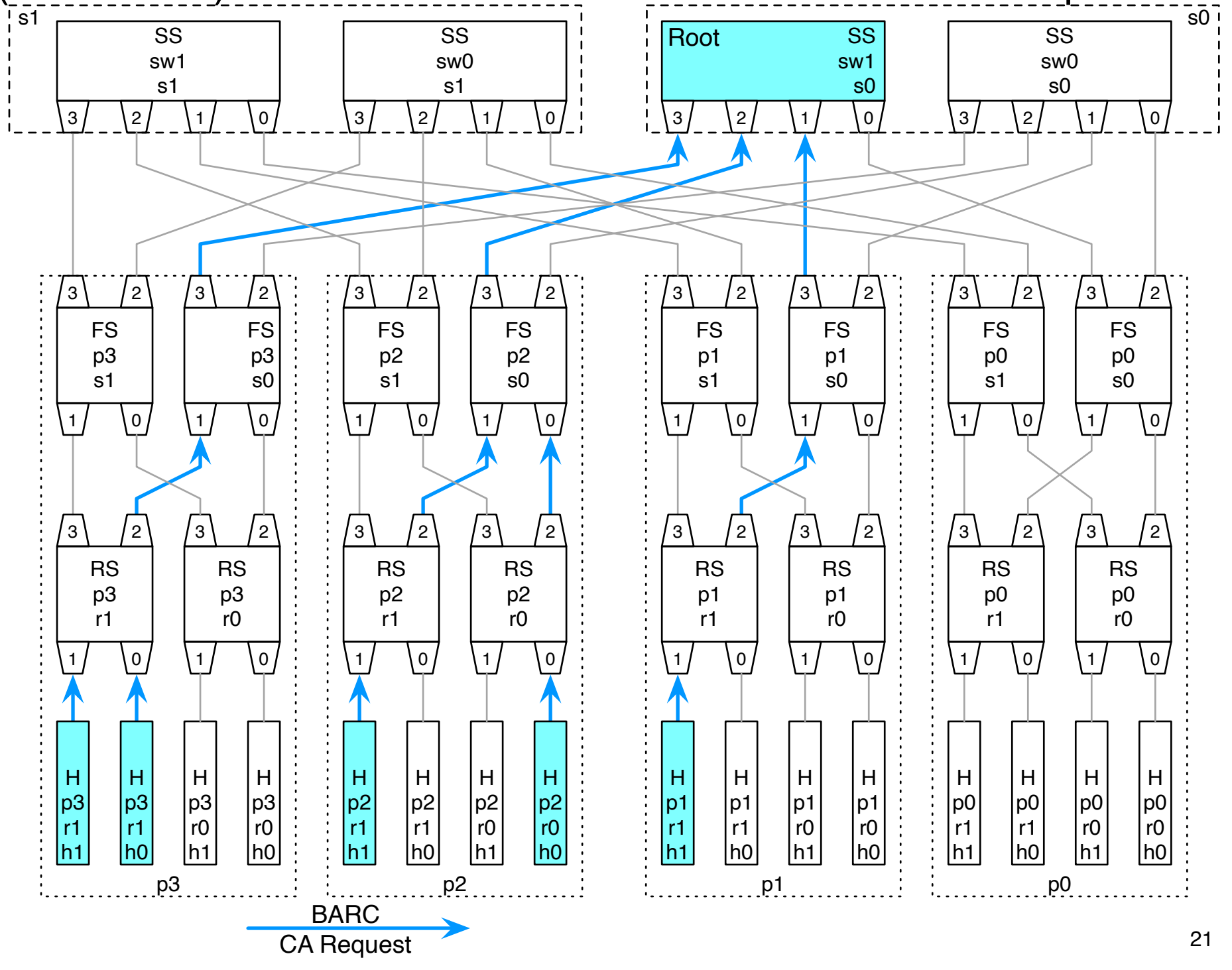
(host-driven) Collective Multicast in Clos Fat-tree: declaration



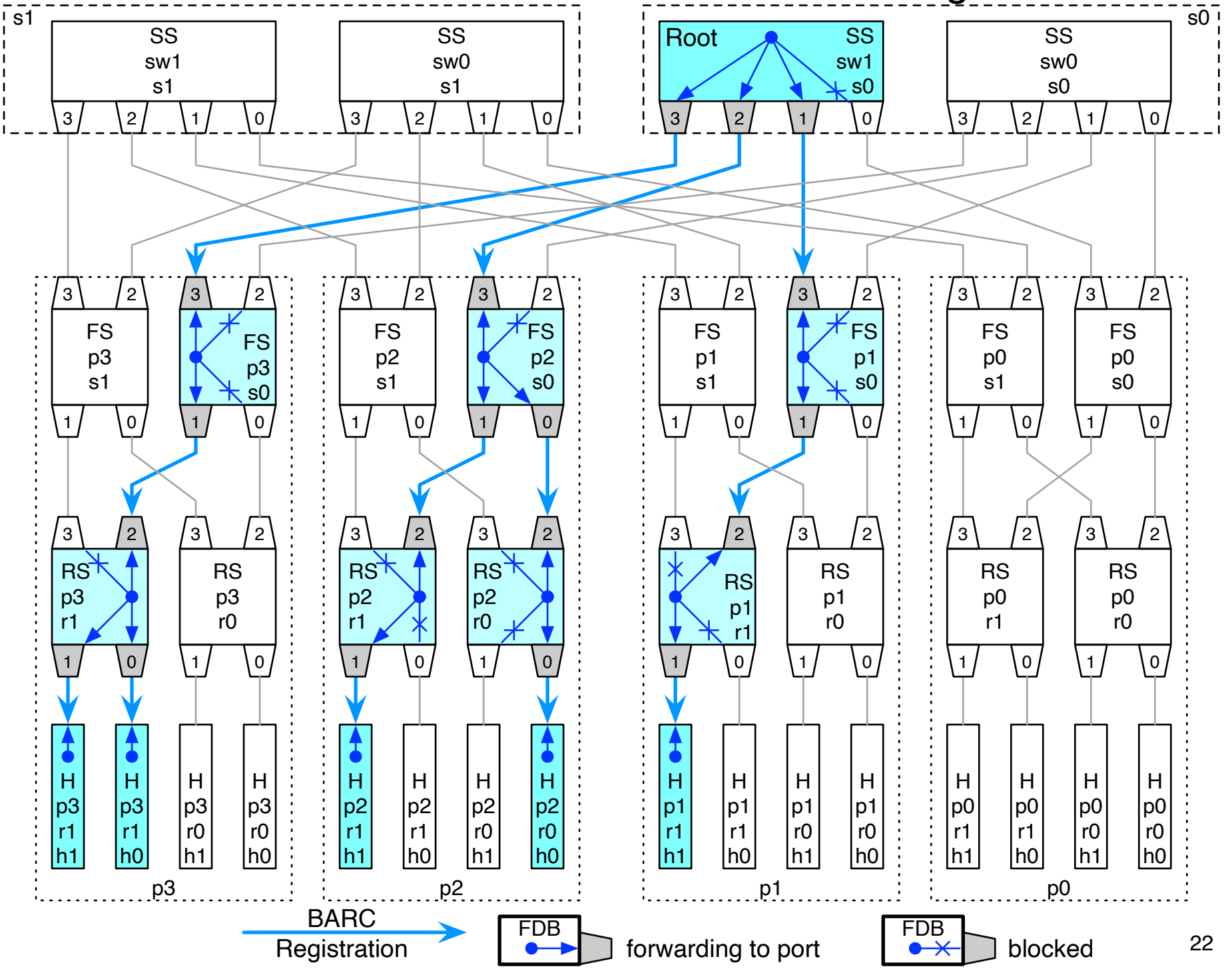
Collective Multicast in Clos Fat-tree: declaration



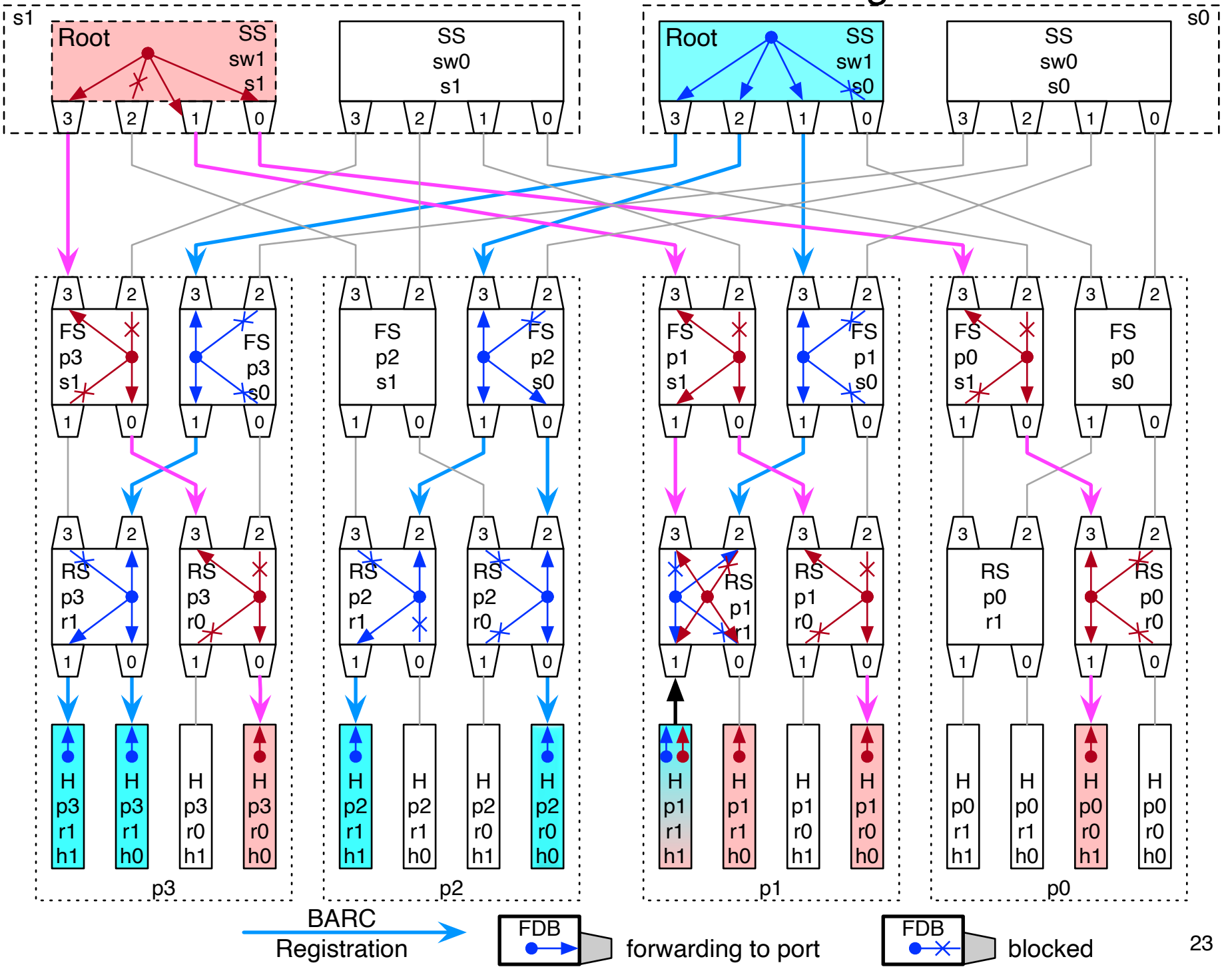
(root-driven) Collective Multicast in Clos Fat-tree: CA Request



Collective Multicast in Clos Fat-tree: registration



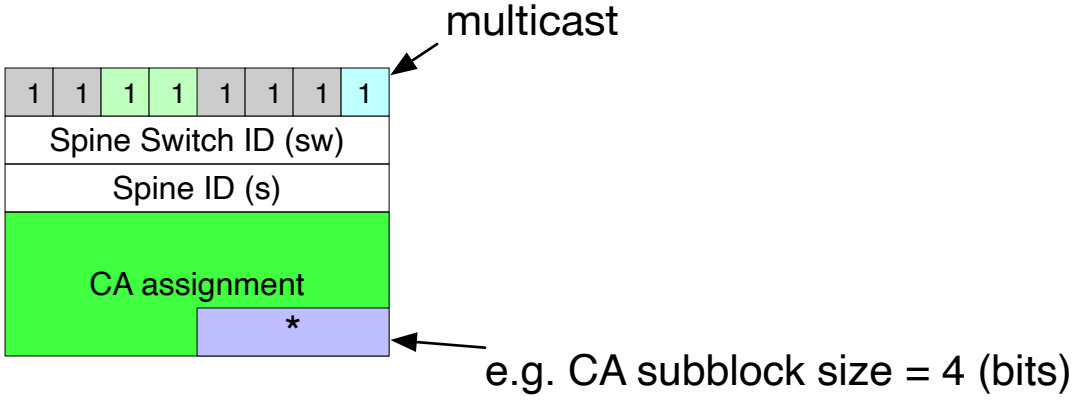
Collective Multicast in Clos Fat-tree: registration



possible content of Collective Multicast registration frame

		bytes	
BARC PDU fields	DA	6	01-80-C2-00-00-00
	State	0.5	CA assignment indication
	BI	6	Collective Address assignment
	Info	6	CA subblock size

example Collective Address assignment with multiple CAs to one group



Summary

- Collective multicast benefits from consideration as a Closed Host Group model, rather than the Open Host Group model of MRP/ MMRP.

- Collective communications for computing networks, particularly in structured networks such as a Clos Fat-tree, provide another use case for BARC Address Blocks.

Note: Further details to be discussed in Nendica meeting of 2024-03-14.