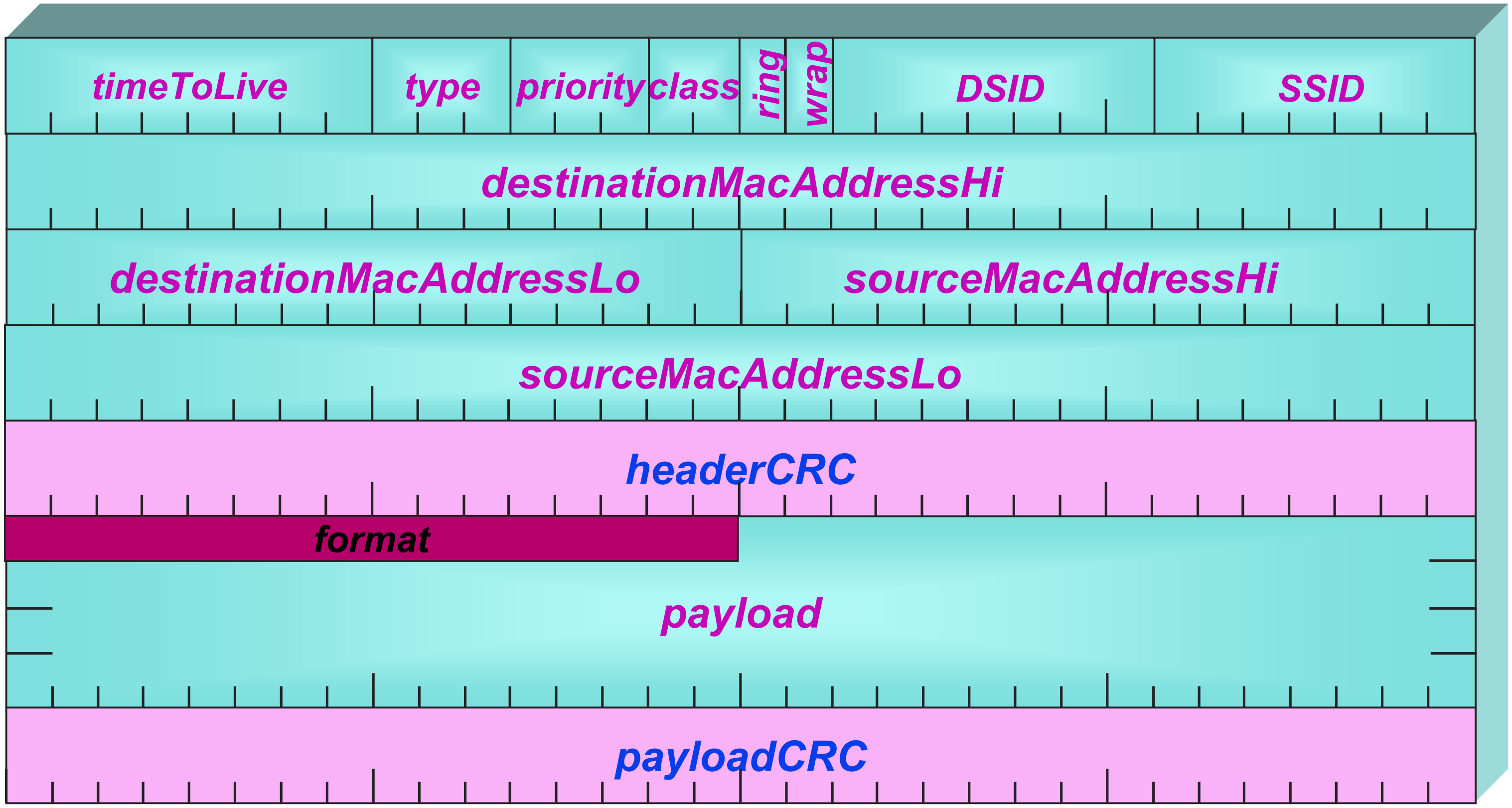
802.17 presentation

- Prepared for 802.17, November 2001
- Dr. David V. James
  Chief Architect
  Network Processing Solutions
  Data Communications Division
  110 Nortech Parkway
  San Jose, CA 95134-2307
  Tel: +1.408.942.2010
  Fax: +1.408.942.2099
  Base: dvj@alum.mit.edu
  Work: djz@cypress.com

Frame formats

RPR Frame Format



timeToLive | type | priority | class | ring | wrap | DSID | SSID

destinationMacAddressHi

destinationMacAddressLo | sourceMacAddressHi

sourceMacAddressLo

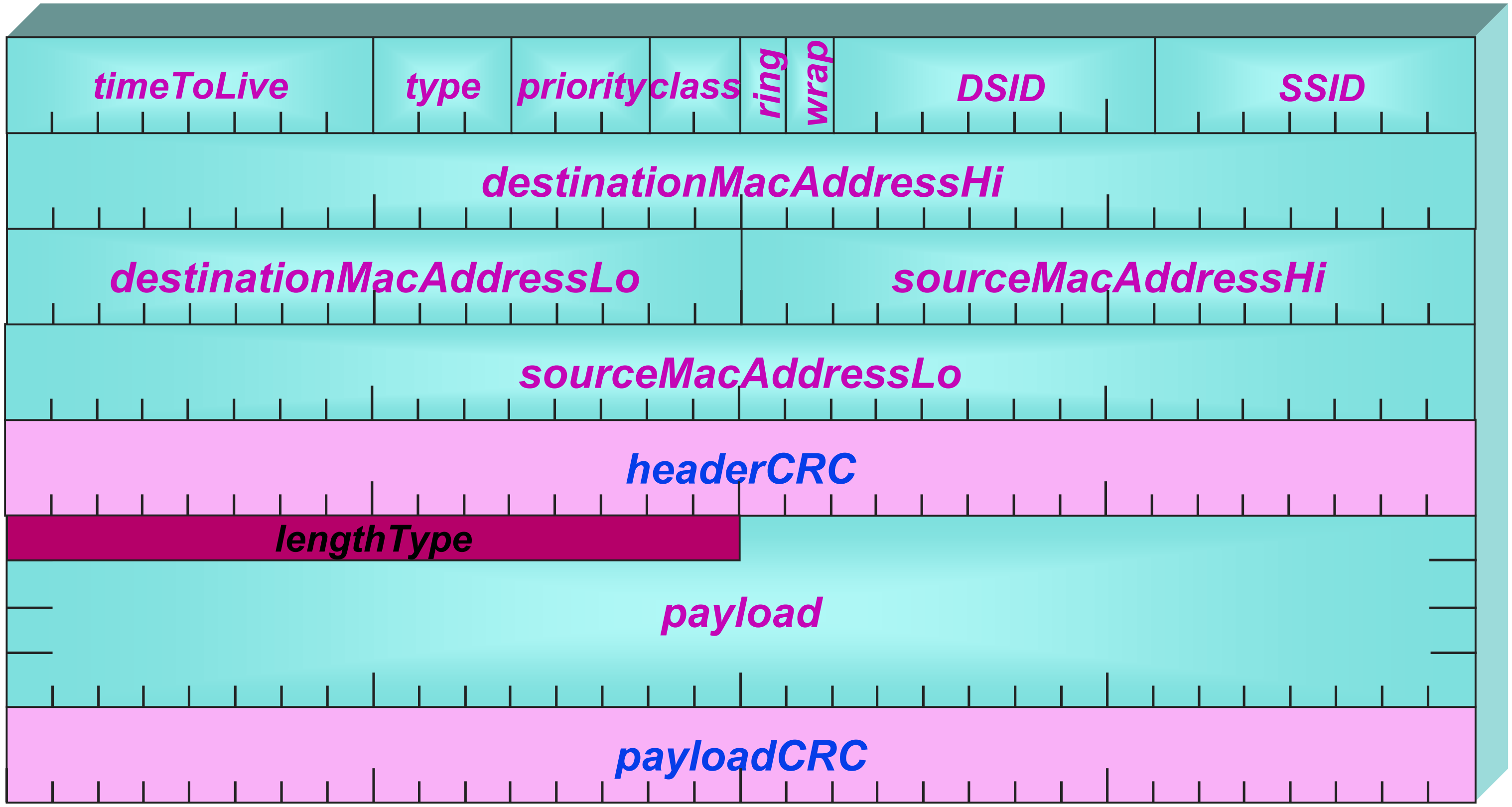
headerCRC

format

payload

payloadCRC

Ethernet Frame



timeToLive | type | priority | class | ring | wrap | DSID | SSID

destinationMacAddressHi

destinationMacAddressLo | sourceMacAddressHi

sourceMacAddressLo

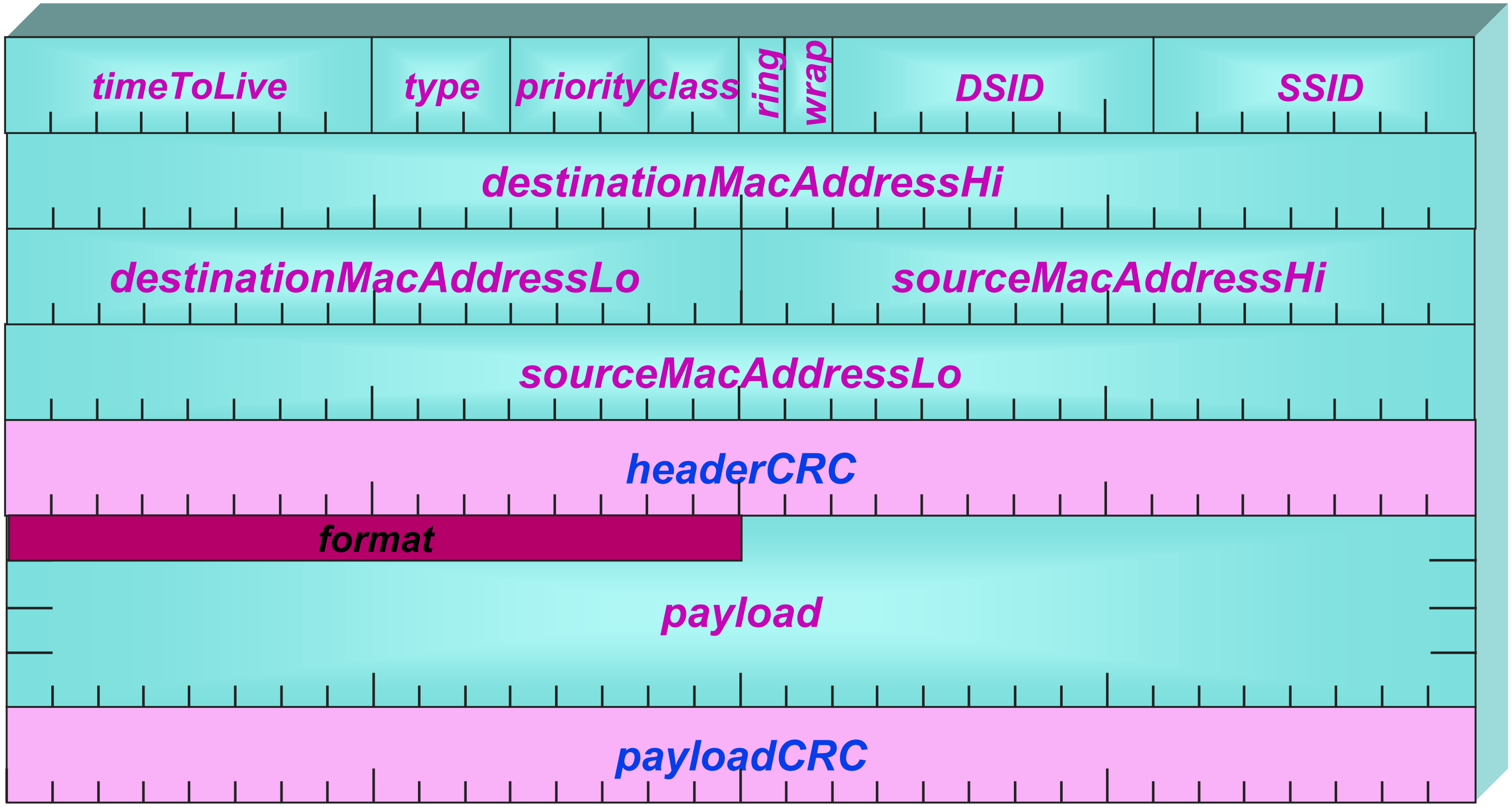
headerCRC

lengthType

payload

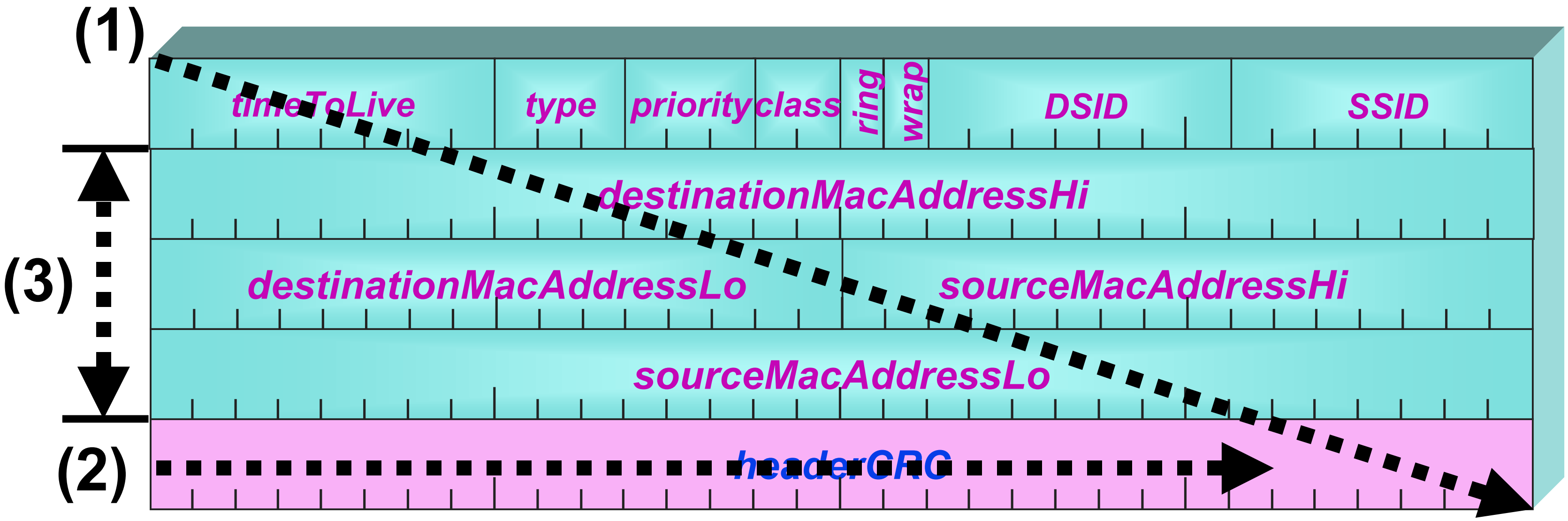
payloadCRC

Control Frame



Control Field Functionality

RPR Frame Format



timeToLive | type | priority | class | ring | wrap | DSID | SSID

destinationMacAddressHi

destinationMacAddressLo | sourceMacAddressHi

sourceMacAddressLo

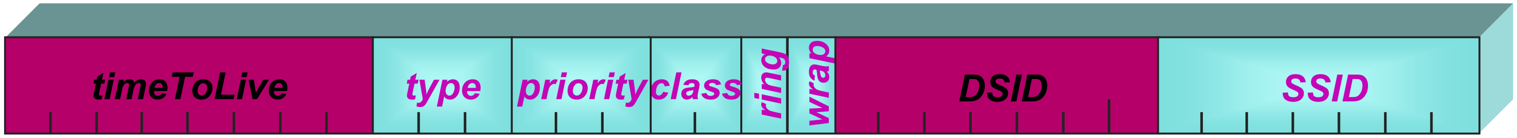
headerCRC

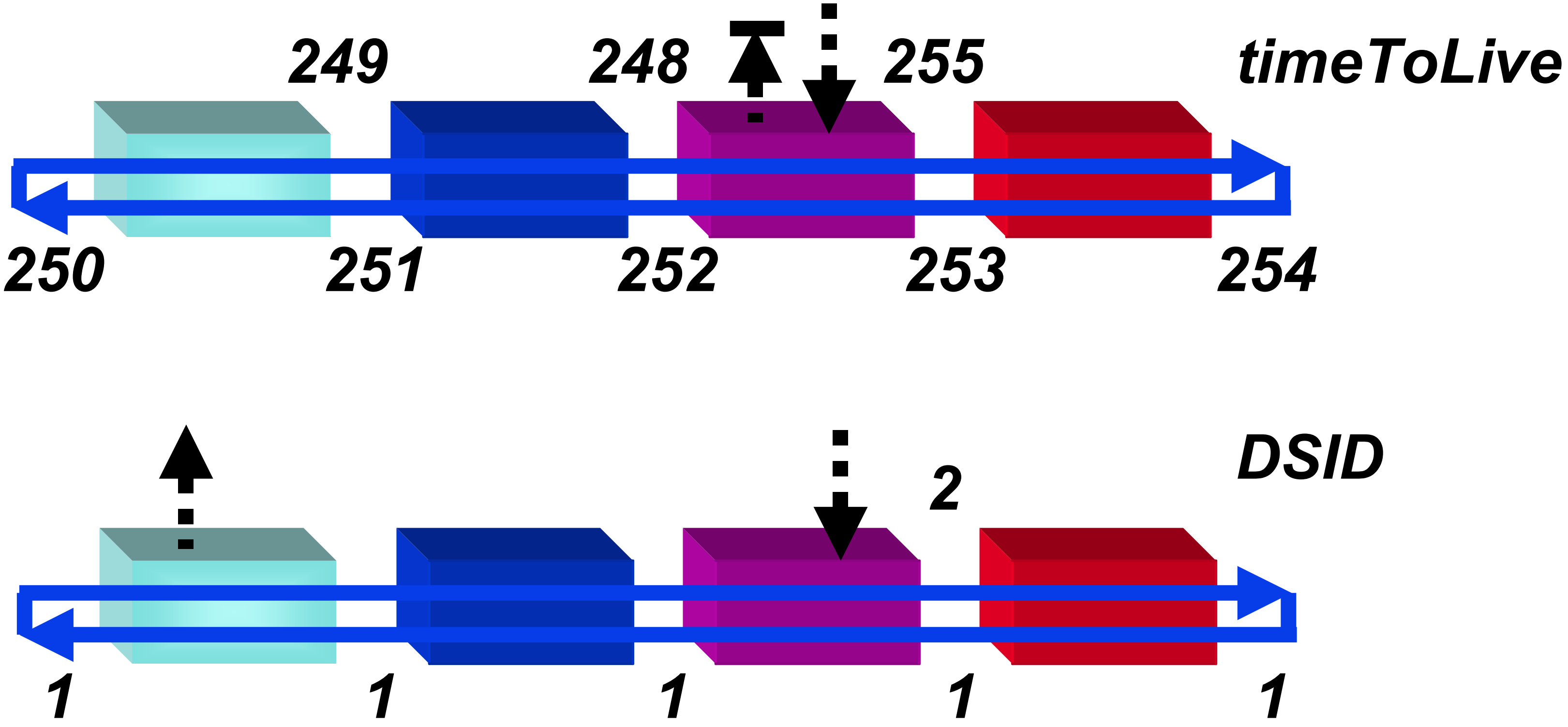
1) 32-bit aligned

2) 32-bit checksum

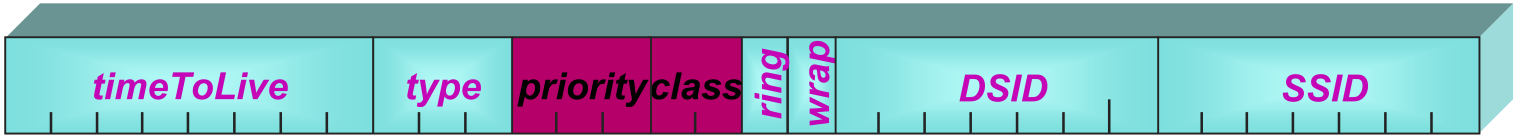
3) Global MAC addresses (not local)

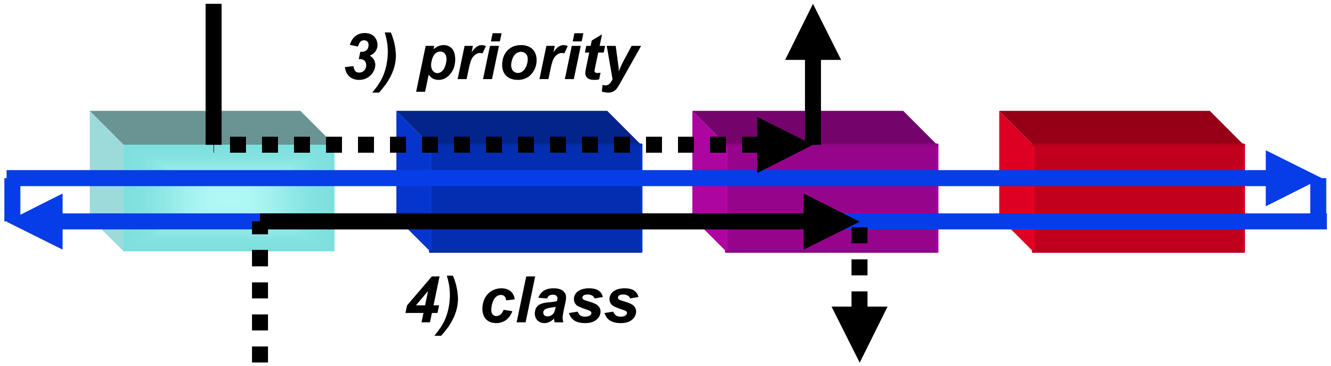
Robust TTL accounting



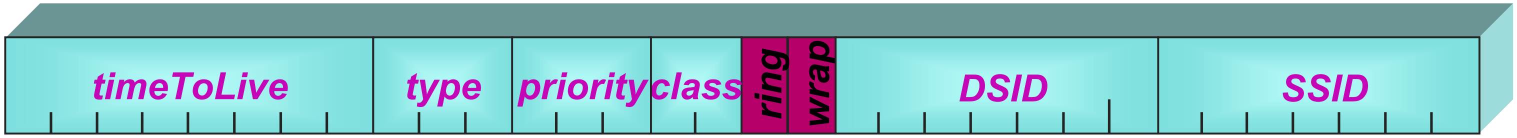


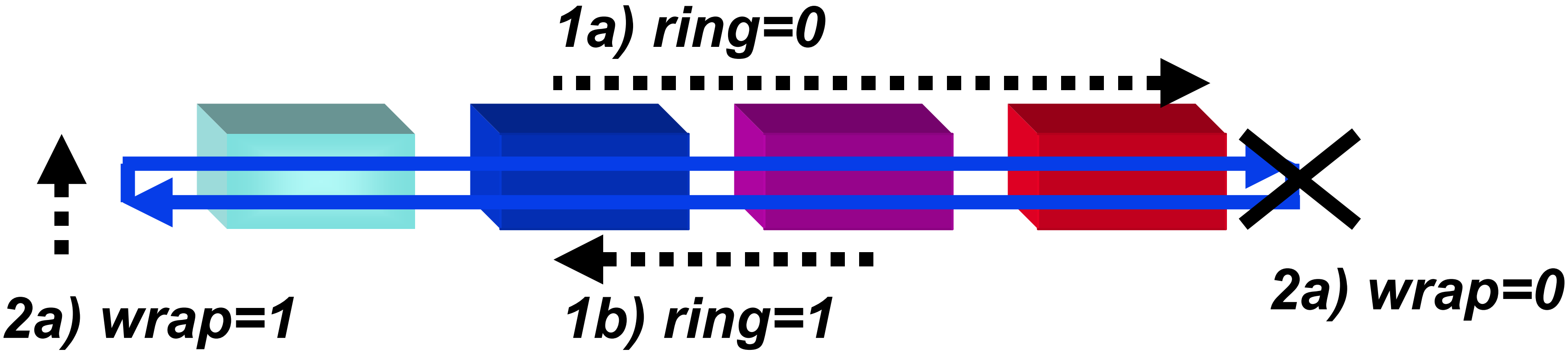
Global and local priorities



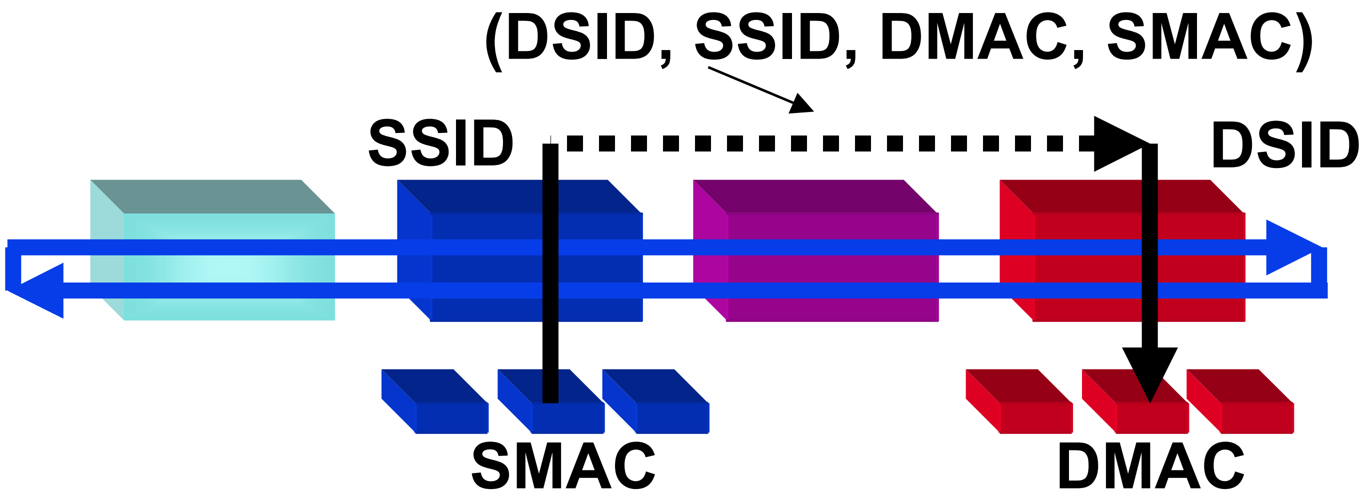


Ring&wrap flags





Source/Destination Coding



Fixed

TTL:8
DMAC:48
SMAC:48
DSID:48
SSID:48

+12 bytes

Stable

TTL:8
DSID:8
SSID:8
DMAC:48
SMAC:48

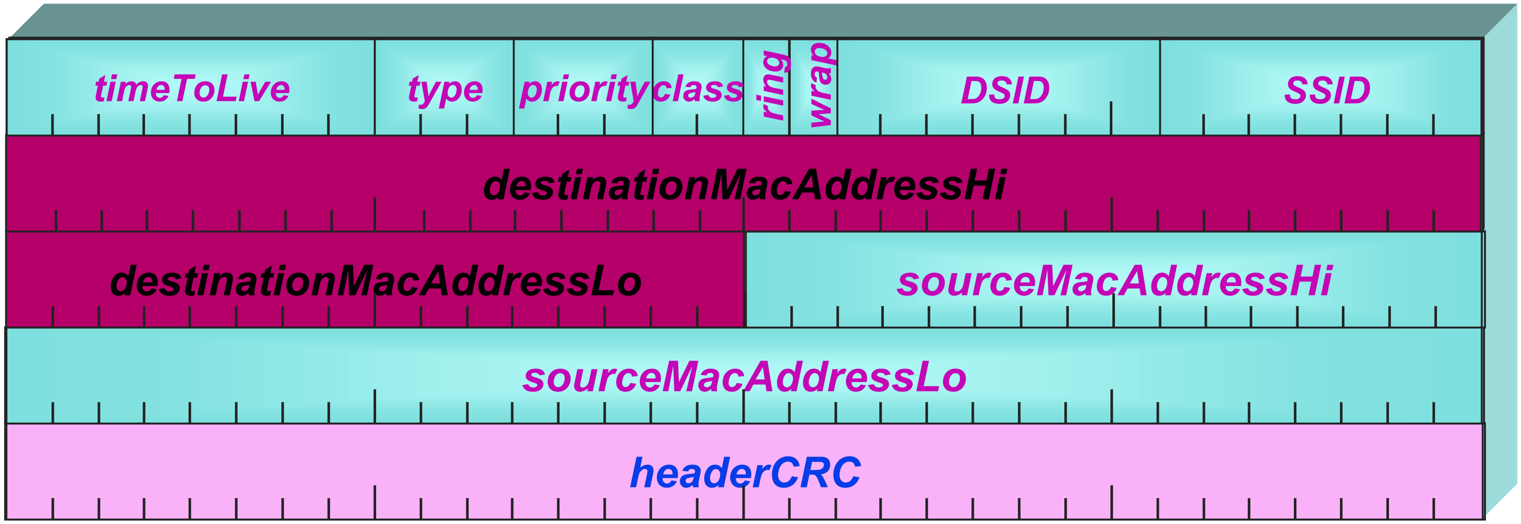
+2 bytes

Relative

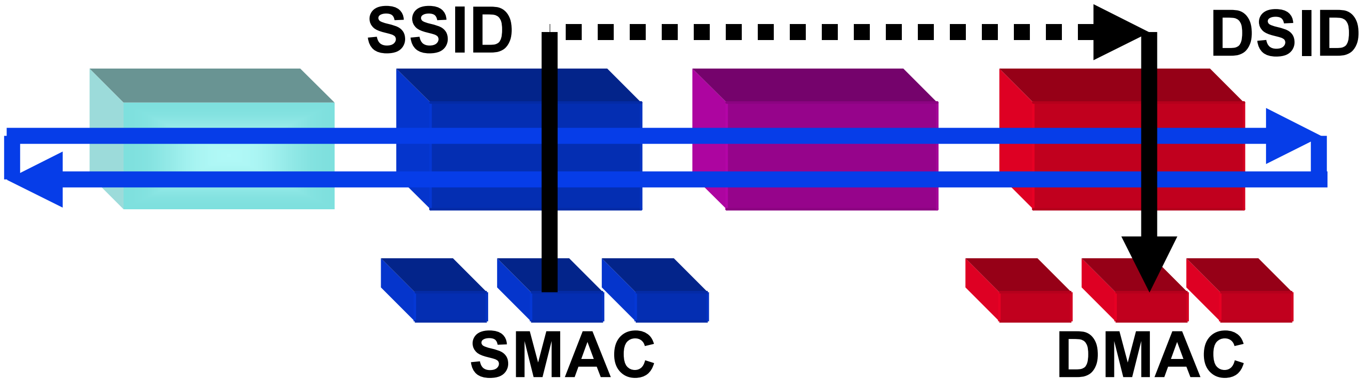
DSID:8
SSID:8
DMAC:48
SMAC:48

(+1 byte)

Ethernet Bridging

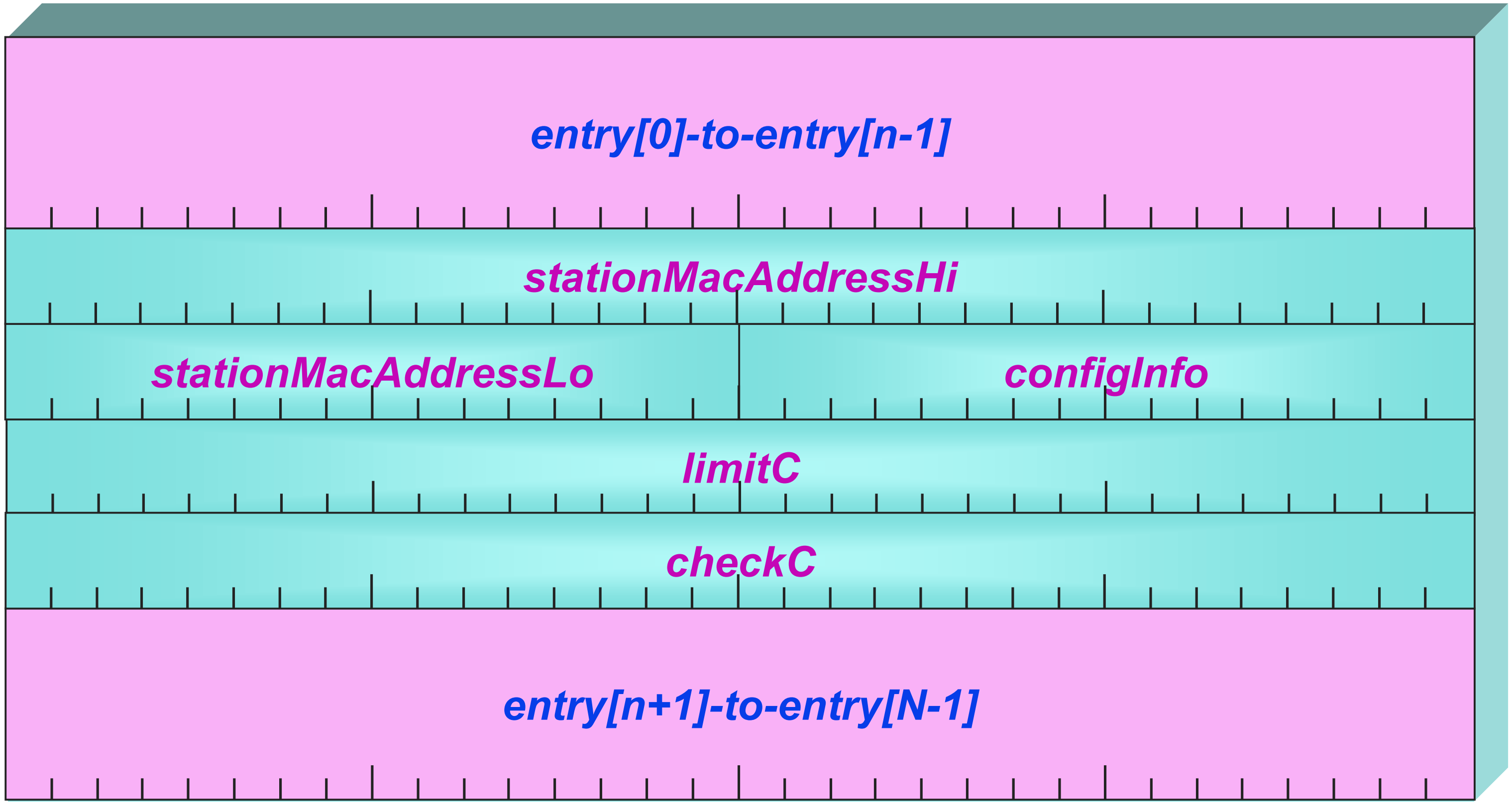


timeToLive
type
priority
class
ring
wrap
DSID
SSID
destinationMacAddressHi
destinationMacAddressLo
sourceMacAddressHi
sourceMacAddressLo
headerCRC



Control Frame Formats

Discovery Frame Format



entry[0]-to-entry[n-1]

stationMacAddressHi

stationMacAddressLo

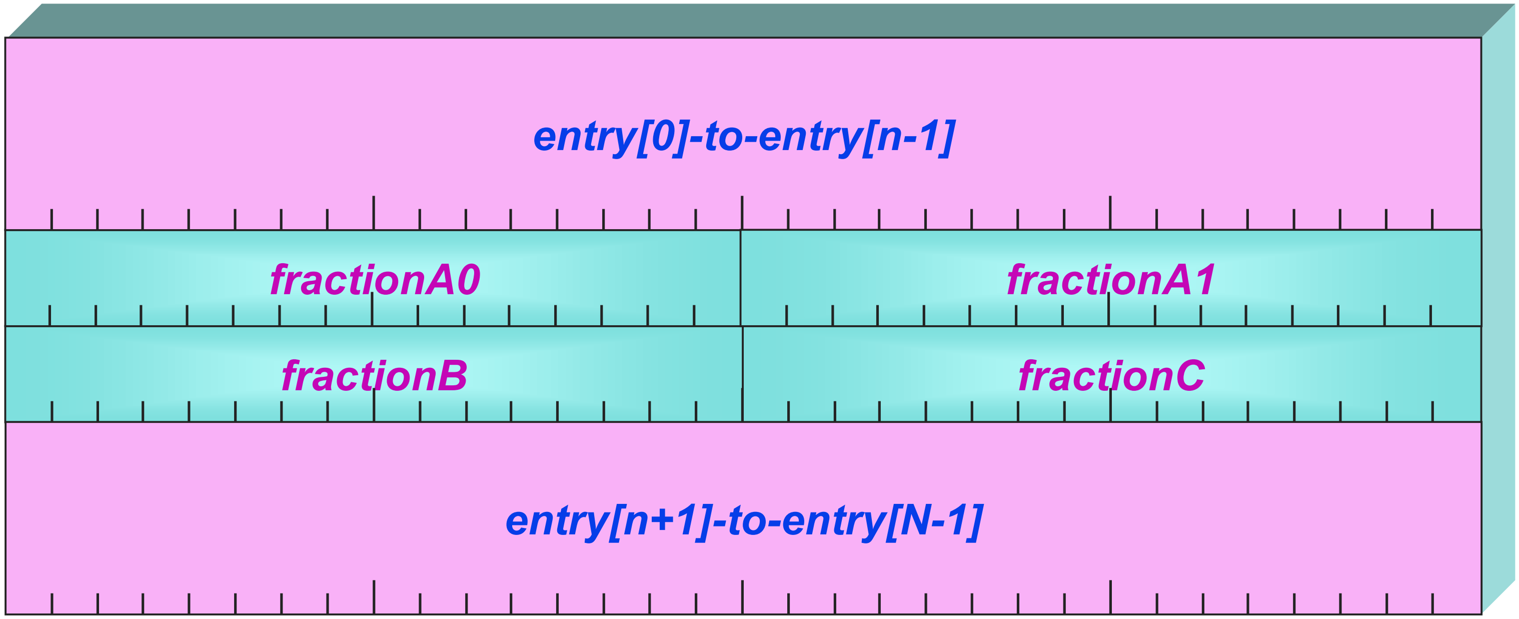
configInfo

limitC

checkC

entry[n+1]-to-entry[N-1]

Survey Frame Format

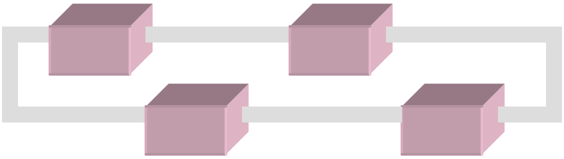


Format Issues

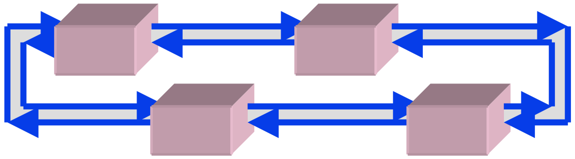
- Wrap: static versus dynamic
- Structural differences:
    - ñ Alignment: 32-bit versus *16-bit
    - ñ CRC coverage: 32-bit versus *16-bit
- Ethernet-type: payload vs *header
- Priority and class: distinct vs *merged
- Local addressing:
    - ñ SSID= TTL, destination= DSID
    - ñ *DSID= TTL, SSID= ????
- Class-A flow-control: embedded vs distinct

Discovery Sequencing

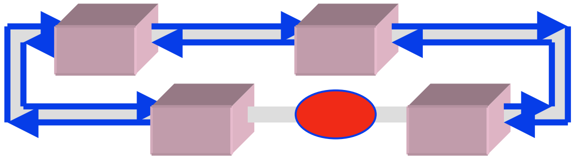
Supported topologies



- A physical ring

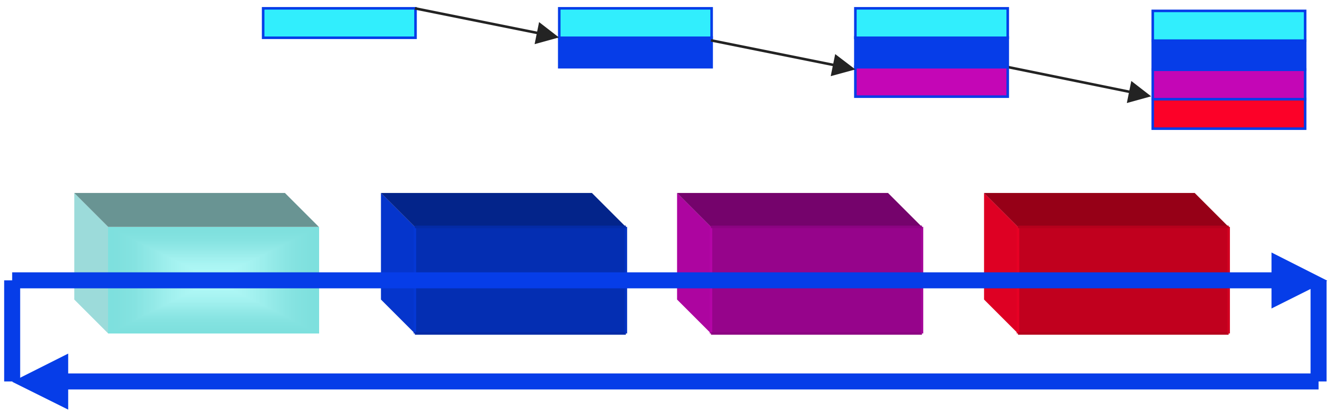


- Dual ringlets



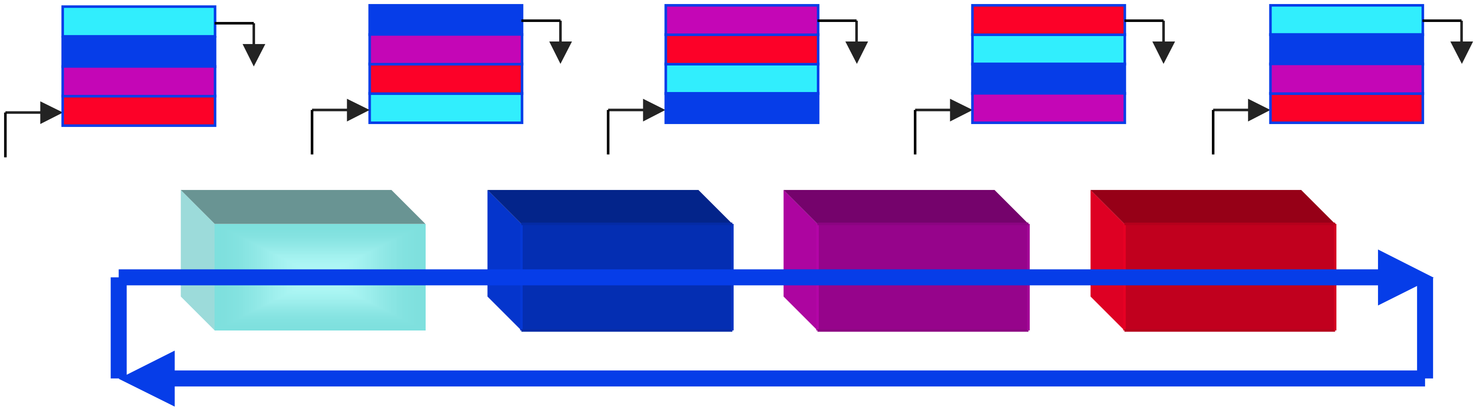
- Duplex ringlet

Topology collection



ï Append your macAddress & info (no duplicate copies presentÖ)

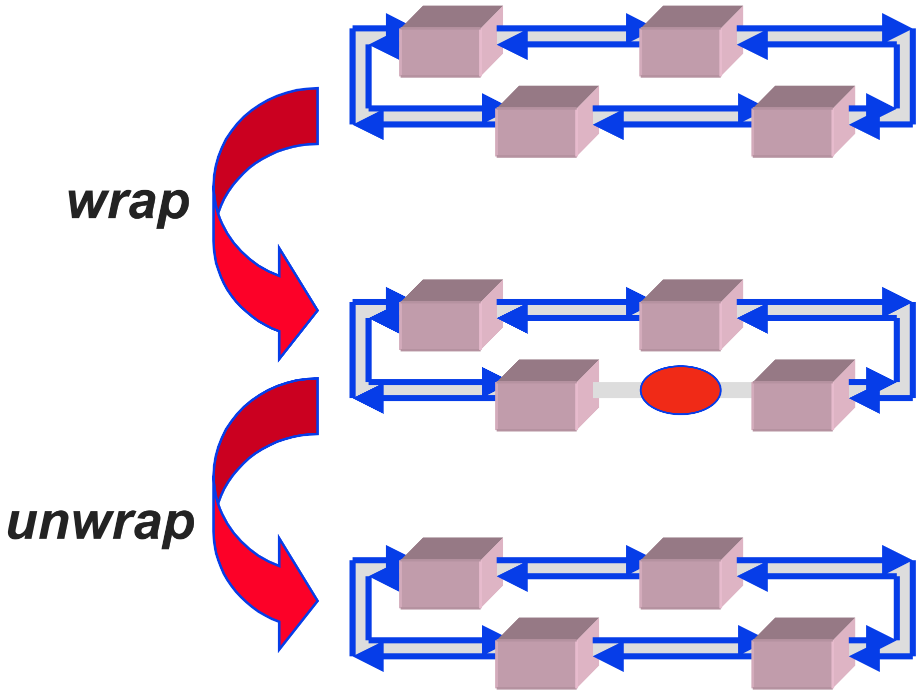
Topology Discovery



ï Strip up-to existing macAddress (inclusive)

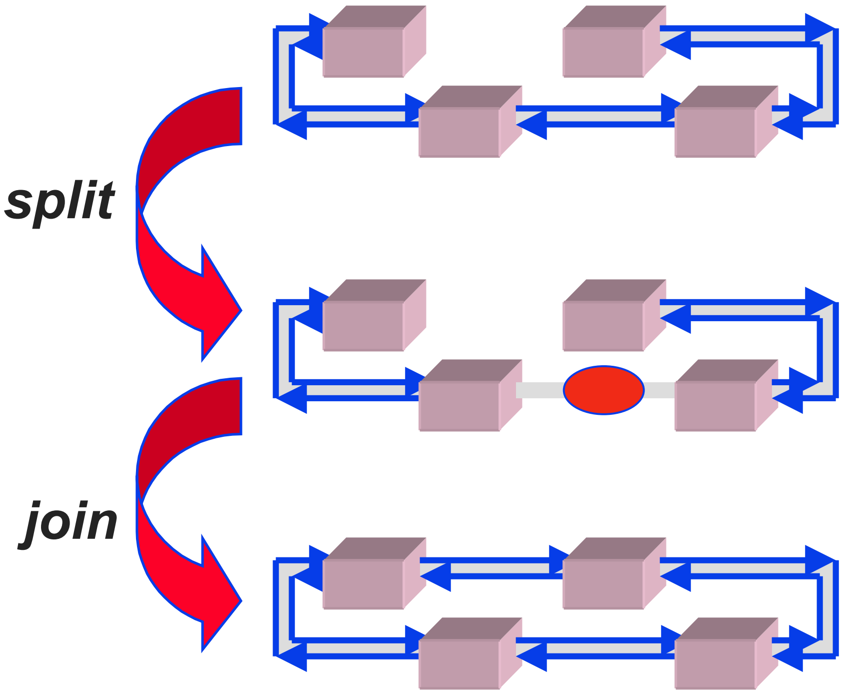
ï Postpend your macAddress & information

Link failures: wrap & unwrap



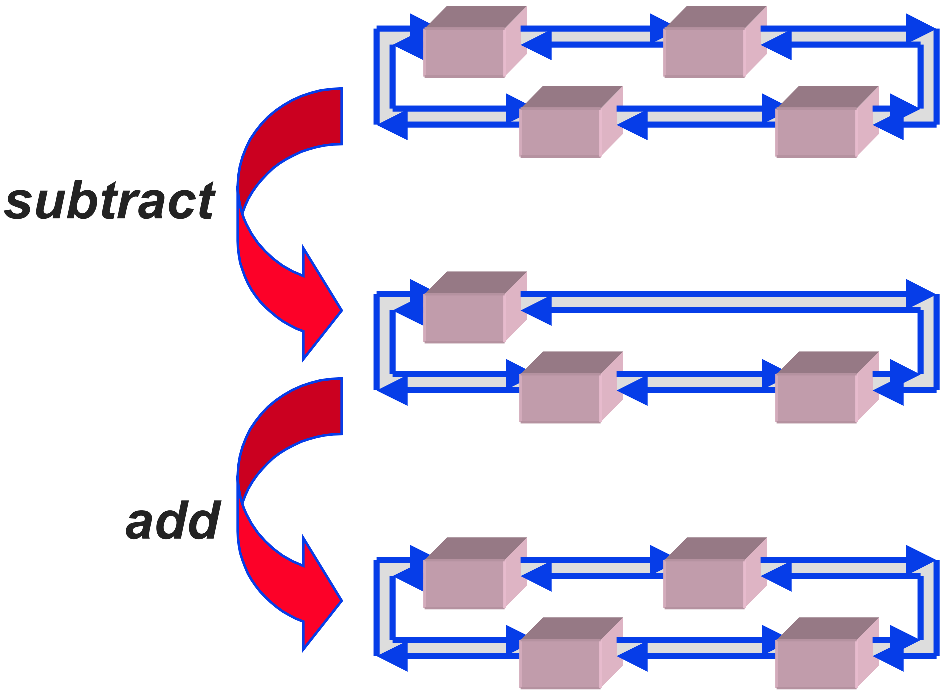
- Dual ringlets
- Duplex ringlet
- Dual ringlets

Link failures: split&join



- One segment
- Two segments
- One segment

Link failures: subtract & add



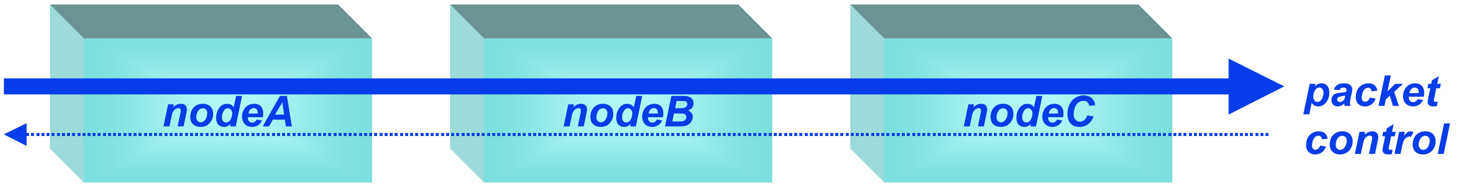
- Basic ringlet
- Reduced ringlet
- Basic ringlet

Discovery properties

- During topology changes, chaos is inevitable
  - ñ Cannot distinguish link failure or topology change
  - ñ Periodicity with event-invoked trigger
- Periodic transmission to neighbor :
  - ñ broadcast relies on DSID, which is unknown
  - ñ broadcast implies ì ownerî , which is unknown
  - ñ cumulative transmission is efficient & robust
- Common features, sent every ~millisecond:
  - ñ Heartbeat
  - ñ Discovery
  - ñ Flow control

Flow control

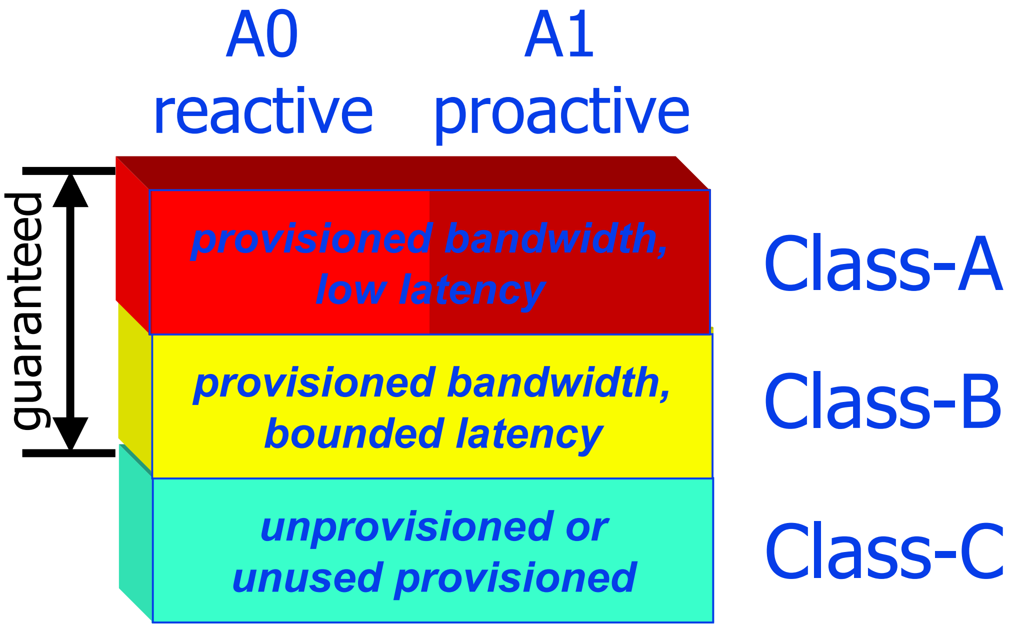
Opposing arbitration



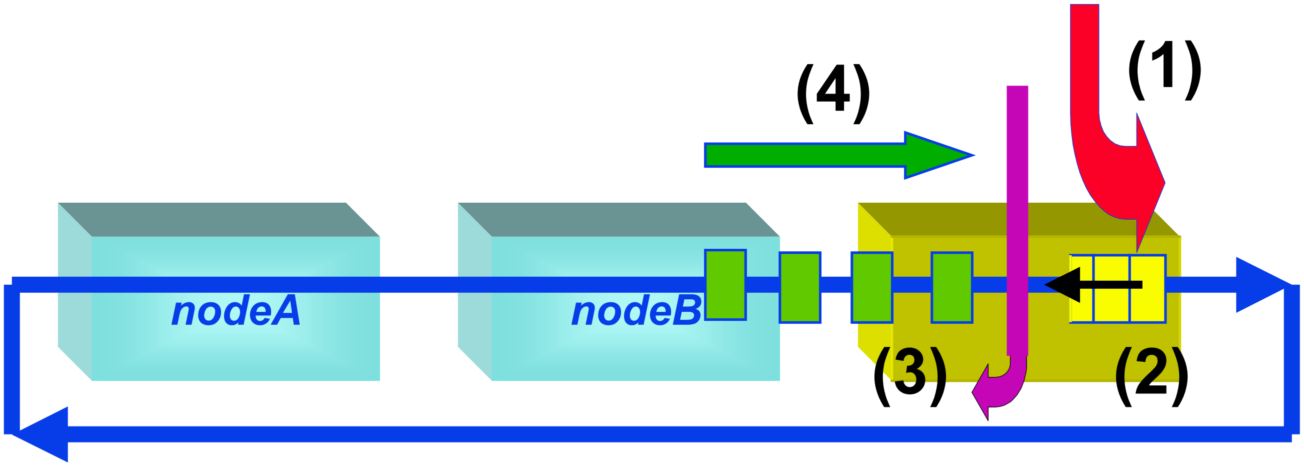
ï Data packets flow in one direction

ï Arbitration control flows in the other*

Arbitration classes

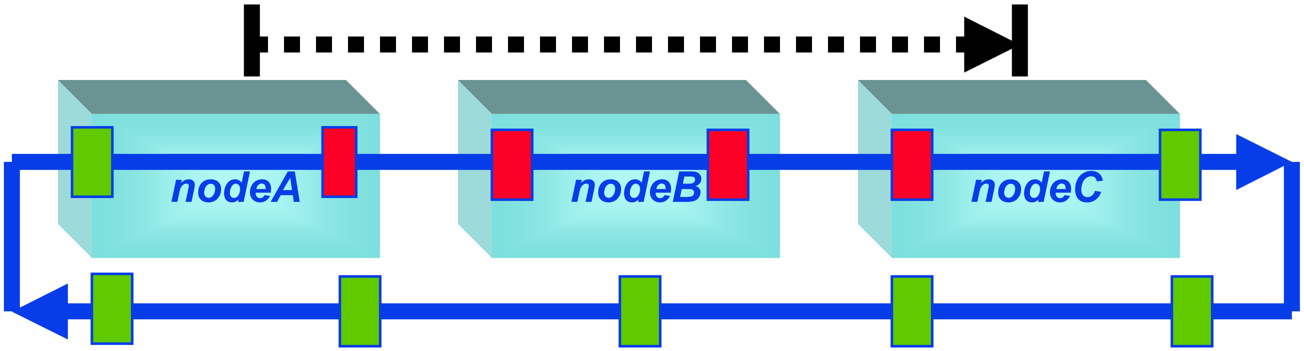


Reactive class-A0 control



- ï Transmission of packets causes
- ï Backup of passBC FIFO that
- ï Returns flow-control information that
- ï Provides consumable idle packets

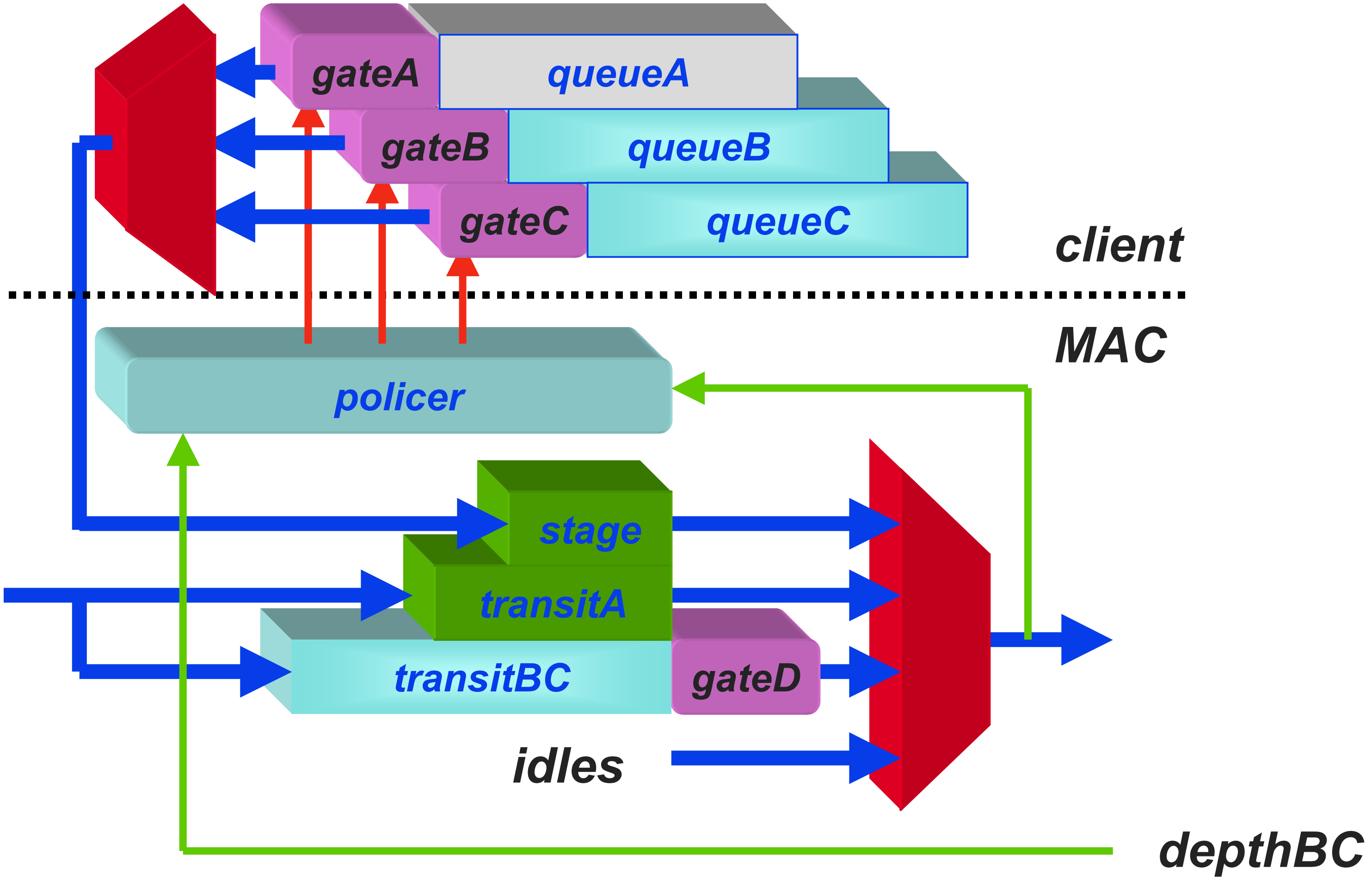
Proactive class-A1 partitions



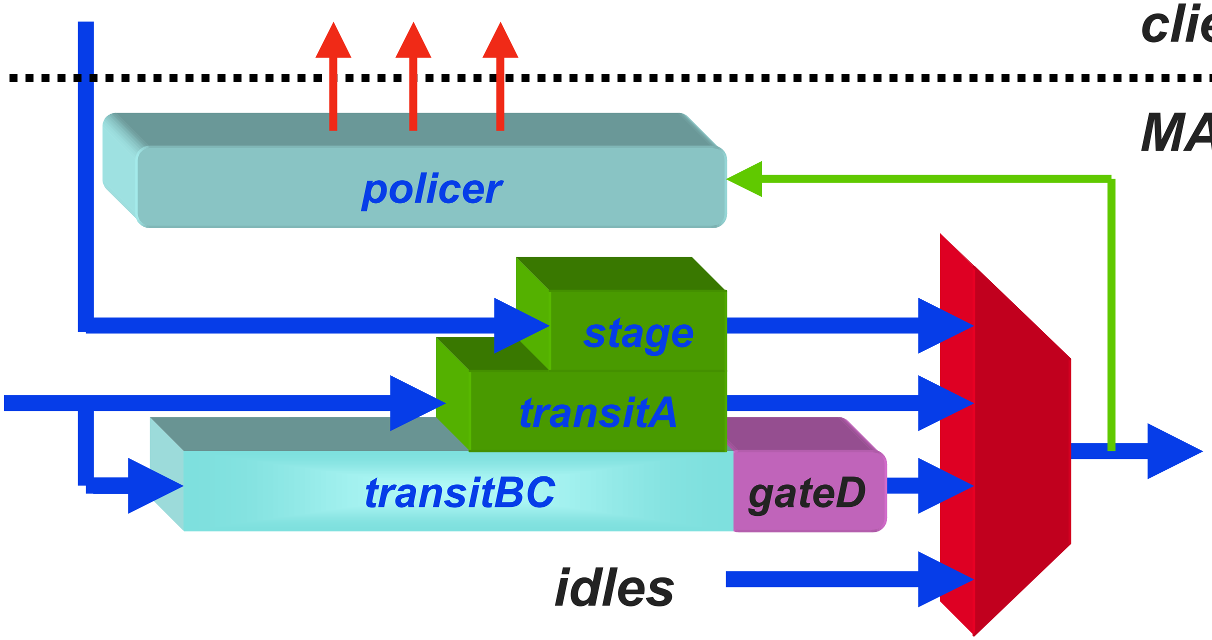
ï Data packets go source-to-destination

ï Residue returns destination-to-source
to provide subsistence for transmissions

Arbitration components



Small-to-large transmitBC

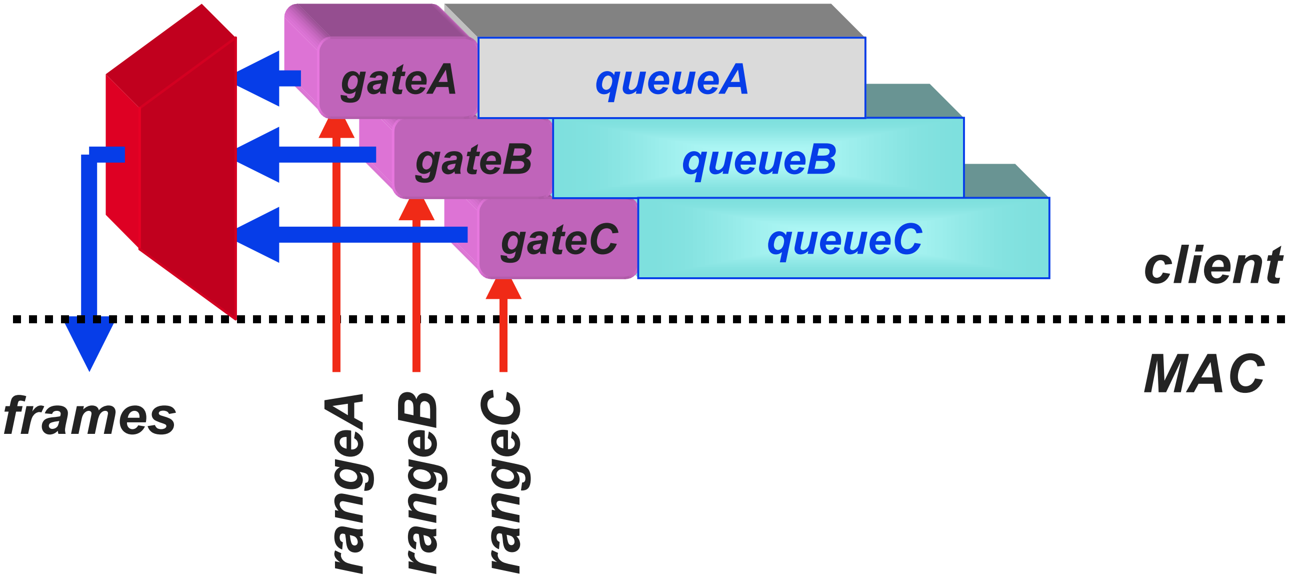


1) Small => proactive classA1

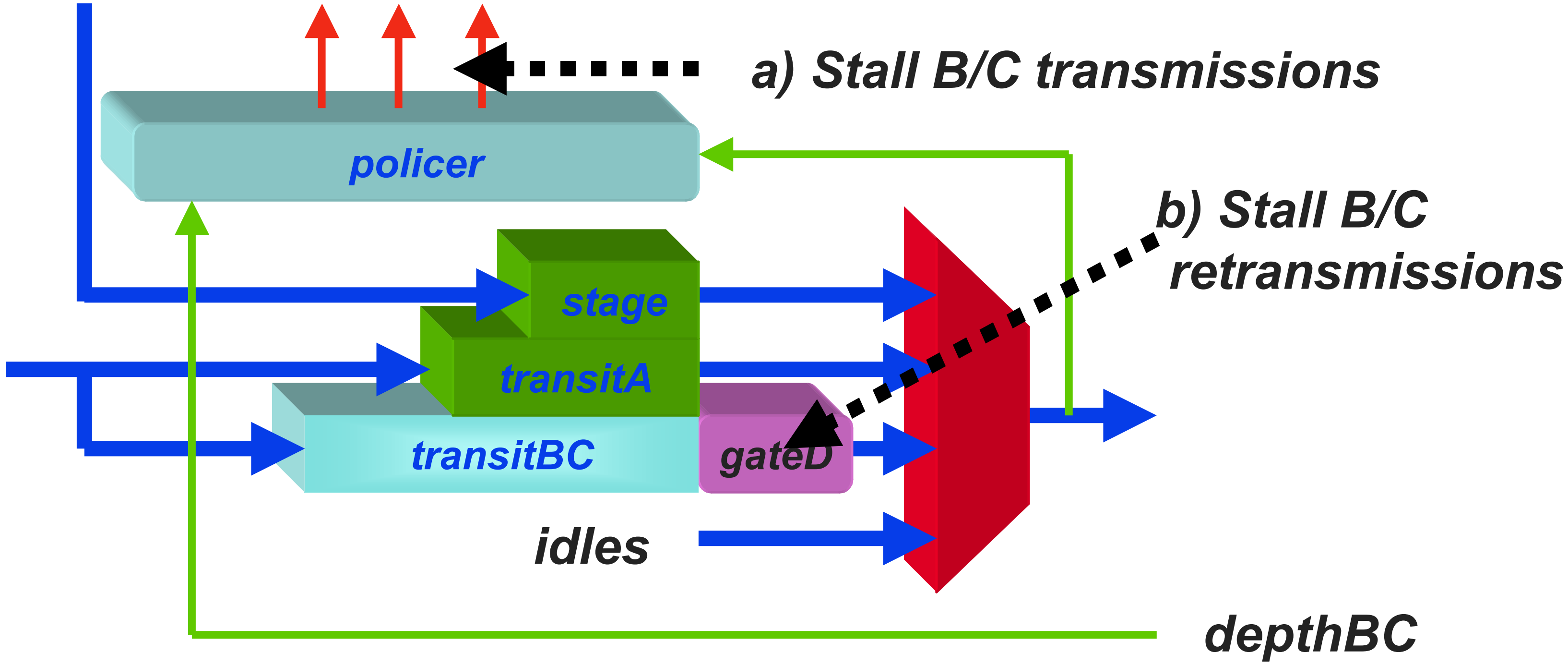
2) Medium => mixed classA0/classA1

3) Large => reactive classA0

MAC-Client interface signals

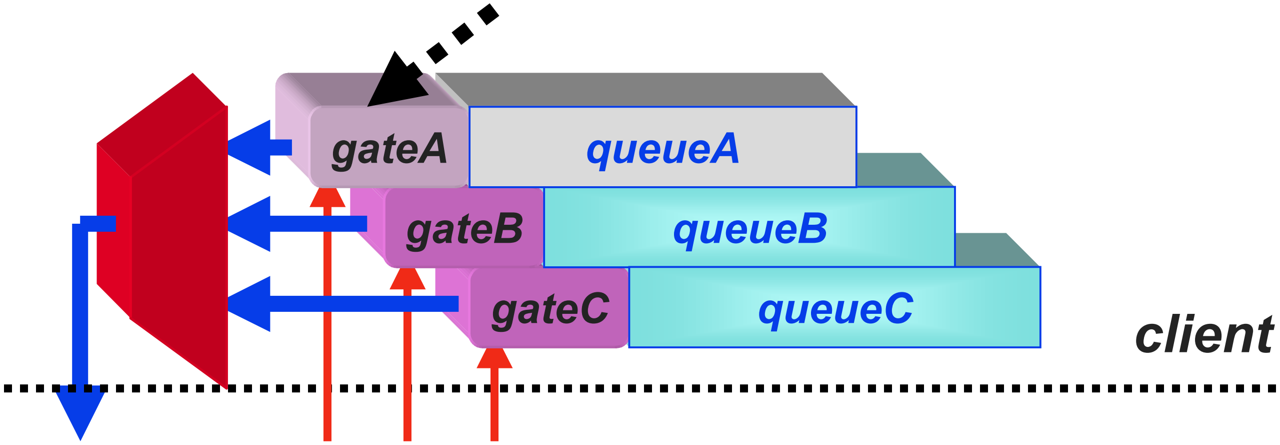


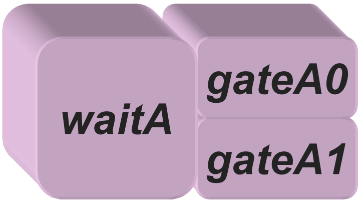
Class-A precedence



```
If (congested(depthBC0, depthBC1))
  rate < ratedA0+ratedA1
else
  rate < rateA0
```

Class-A send-queue gating



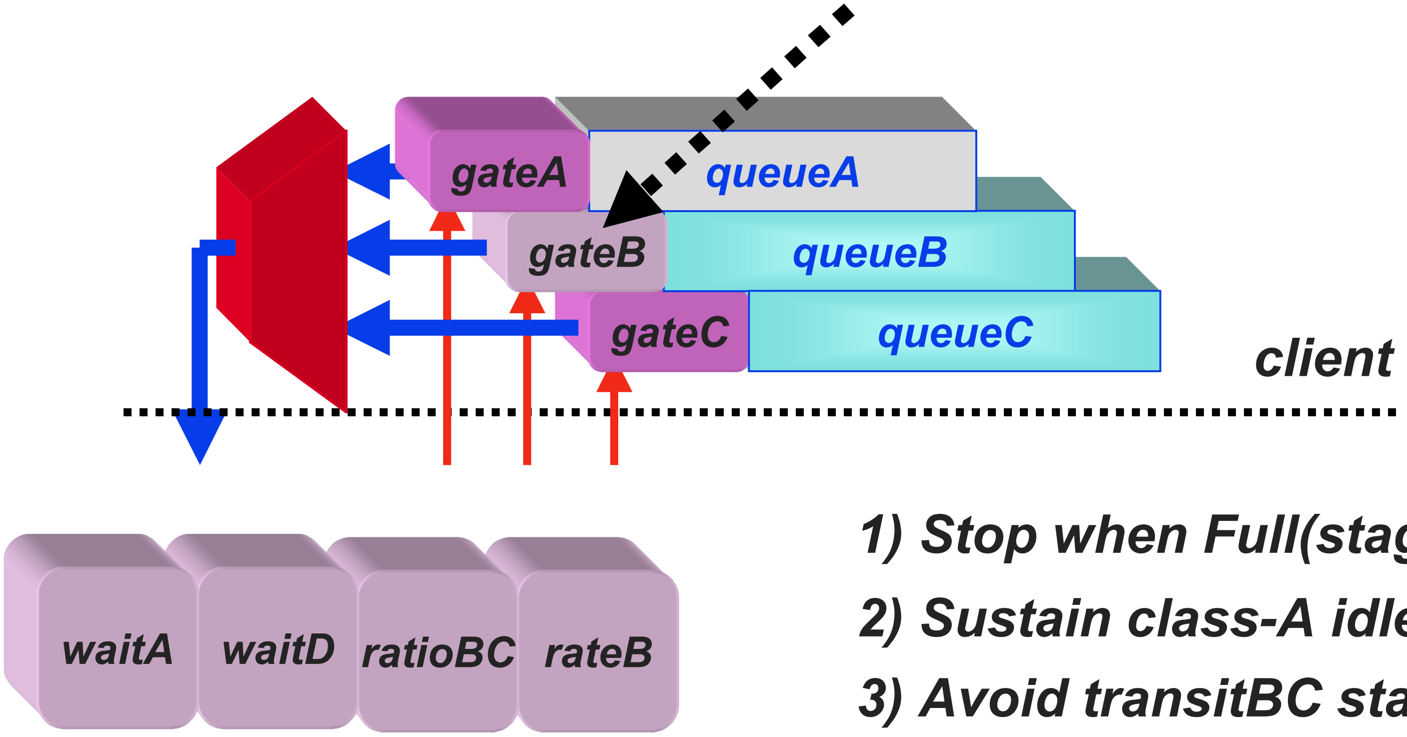


1) Rate limit on class-A0

2) Rate limit on class-A1

3) Stop when Full(stage)

Class-B send-queue gating



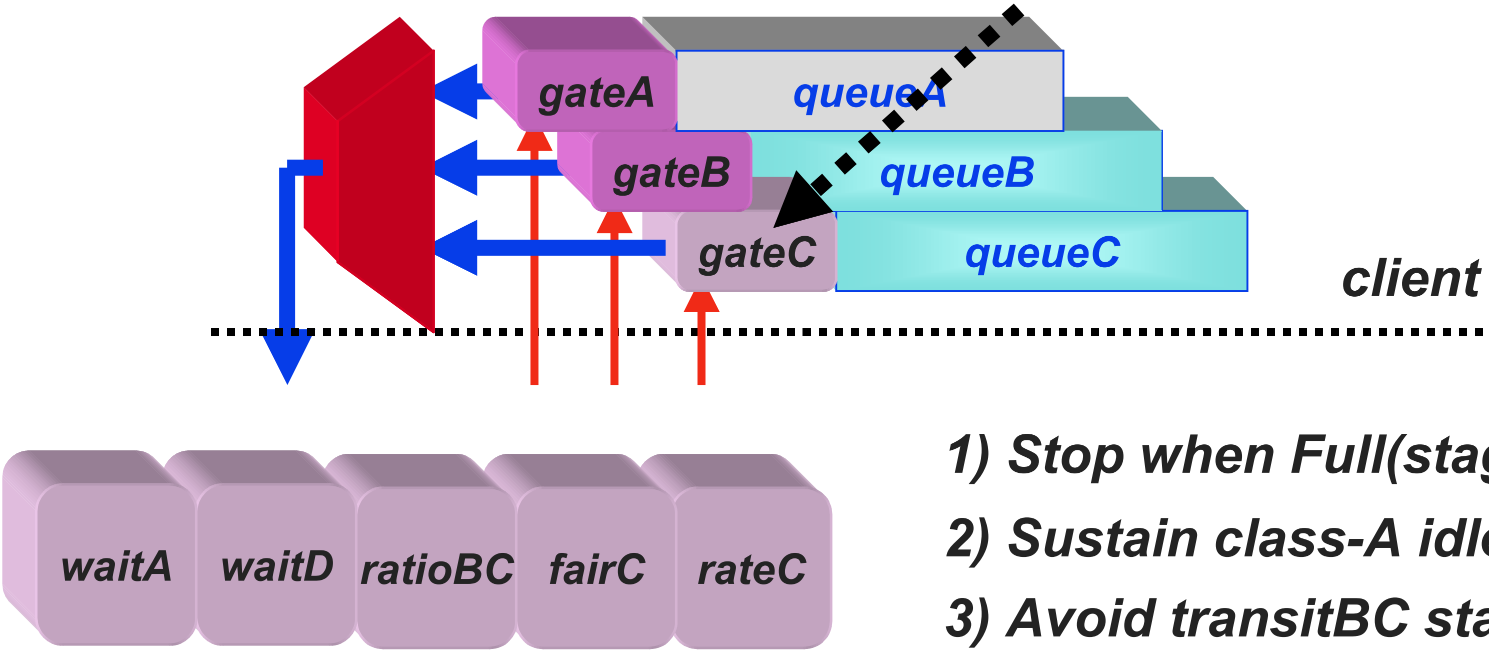
1) Stop when Full(stage)

2) Sustain class-A idles

3) Avoid transitBC starvation

4) Provisioned class-B rate

Class-C send-queue gating



1) Stop when Full(stage)

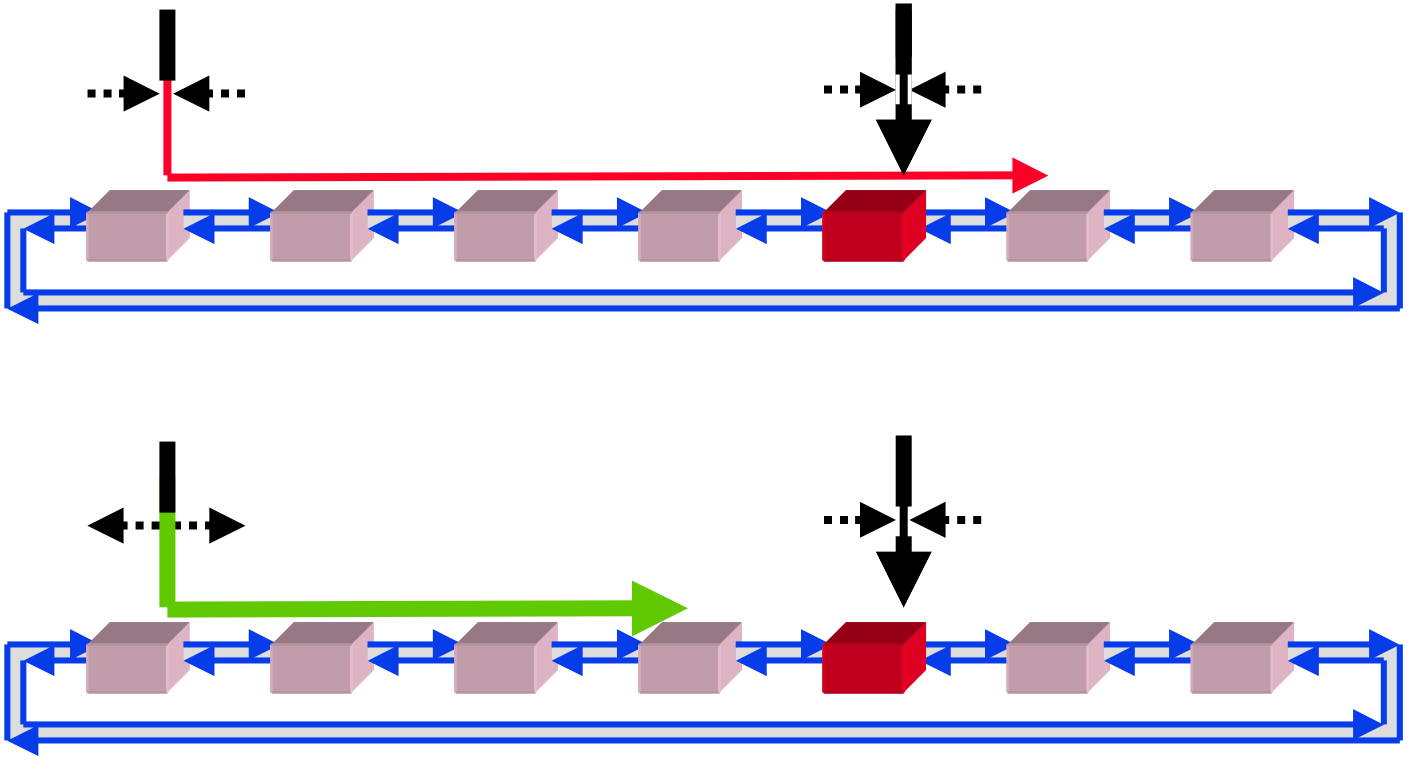
2) Sustain class-A idles

3) Avoid transitBC starvation

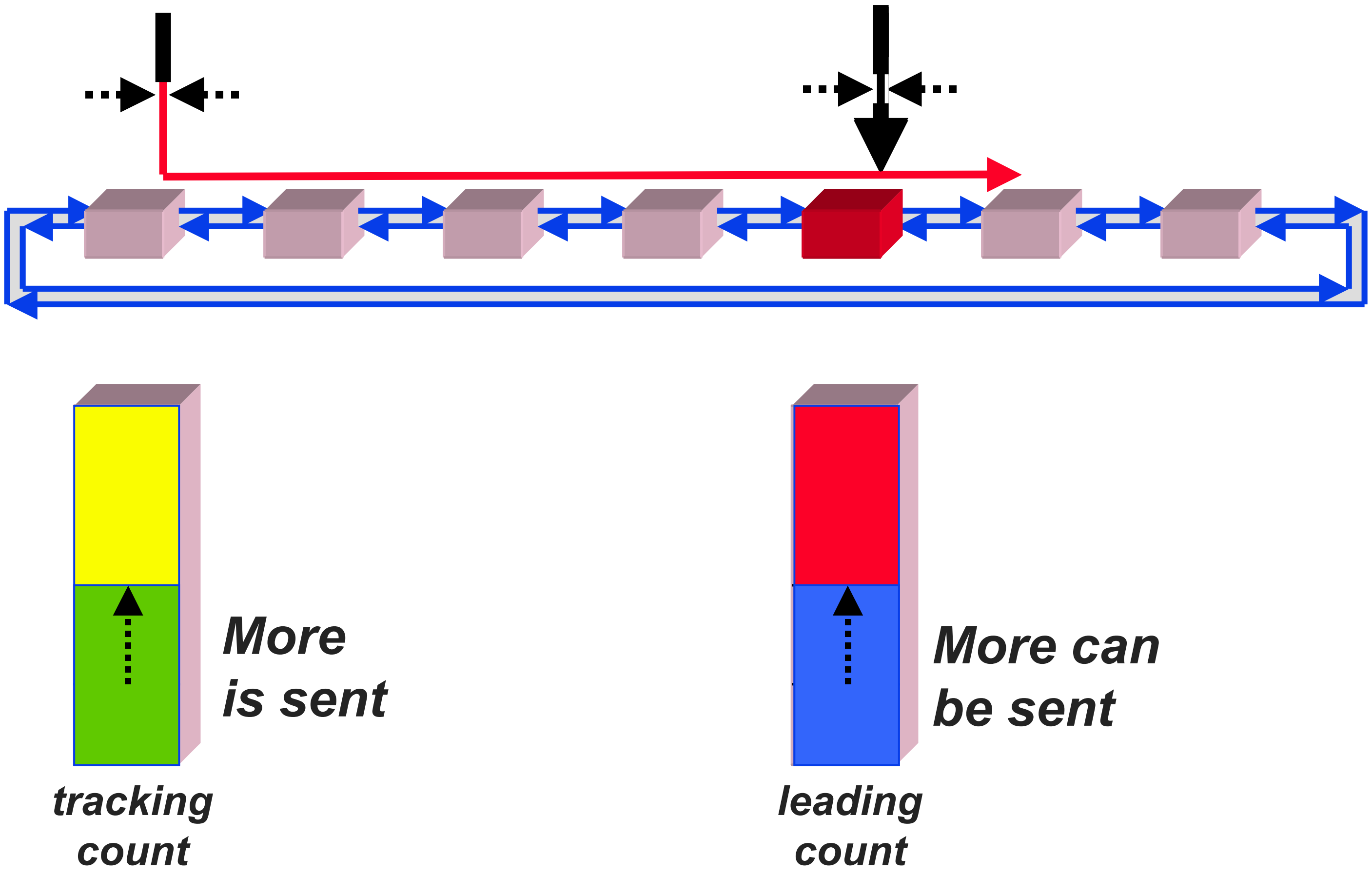
4) Weighted class-C fairness

5) Bounded class-C rate

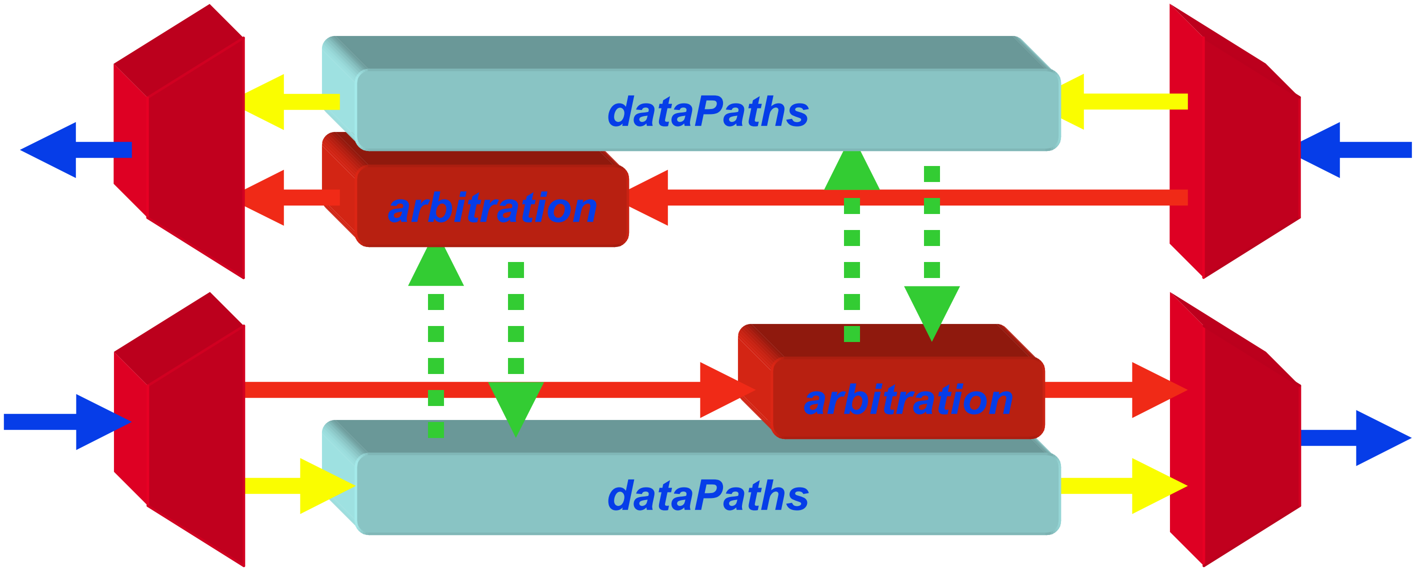
Class-C principles



Class-C fairness counts



Internal MAC arbitration signals



ï Arbitration affects opposing run

ï My congestion affects upstream node

ï Downstream congestion affects me

Class-A flow control (proactive and reactive)

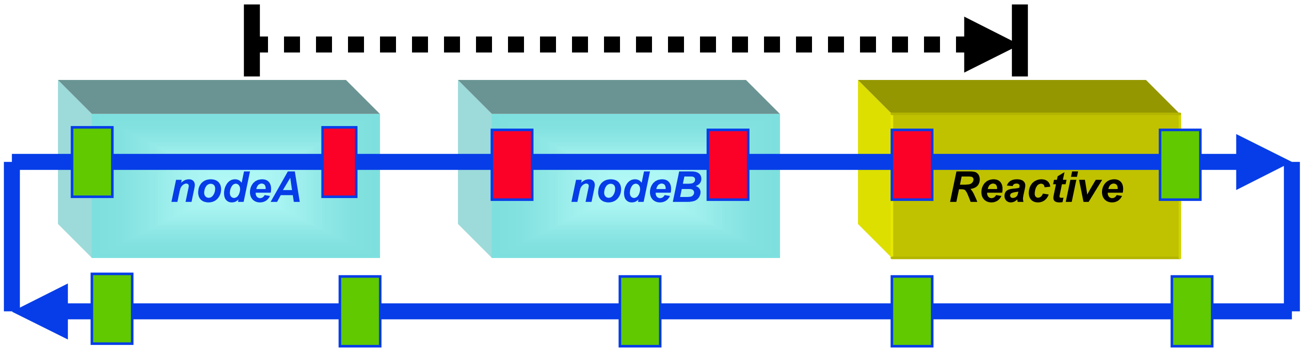
Class-A flow control

- Proactive
  Minimal (nonexistent?) passBC transit buffer
  Less available bandwidth
  Each station maintains constant classAp traffic

- Reactive
  Significant passBC transit buffer
  Full bandwidth utilization
  Each station responds/regenerates throttle messages

- Interoperable?
  This is a bandwidth vs memory $$ tradeoff

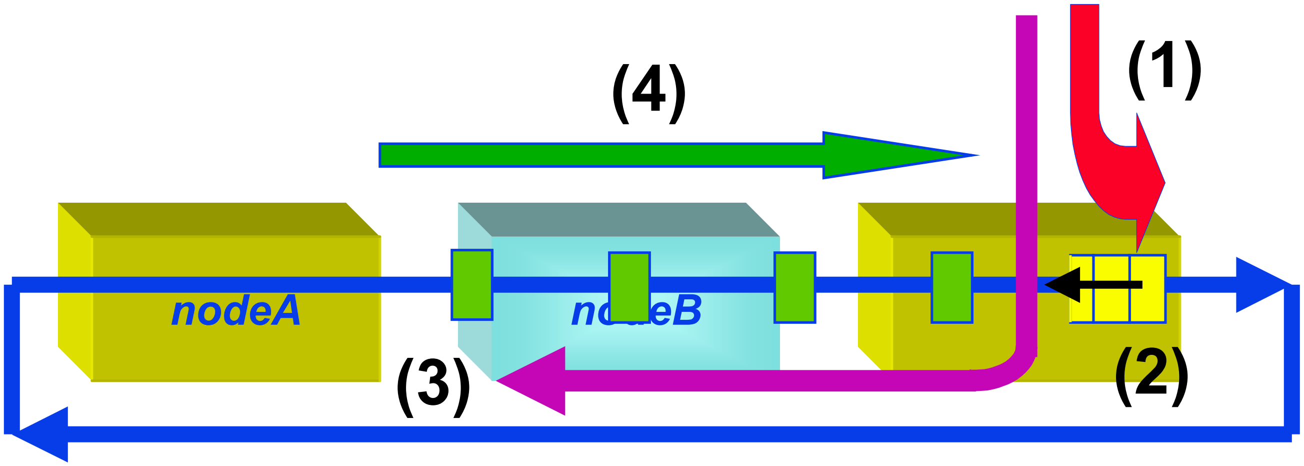
Proactive class-A compatibility options



ï Reactive node trickles class-A bandwidth

ï Reactive node recycles class-A bandwidth class-A => class-Aí, thus preserving BW

Reactive class-A compatibility



ï Flow control passes upstream

ï Proactive stations pass these indications

Topology discovery

Frame interchanges

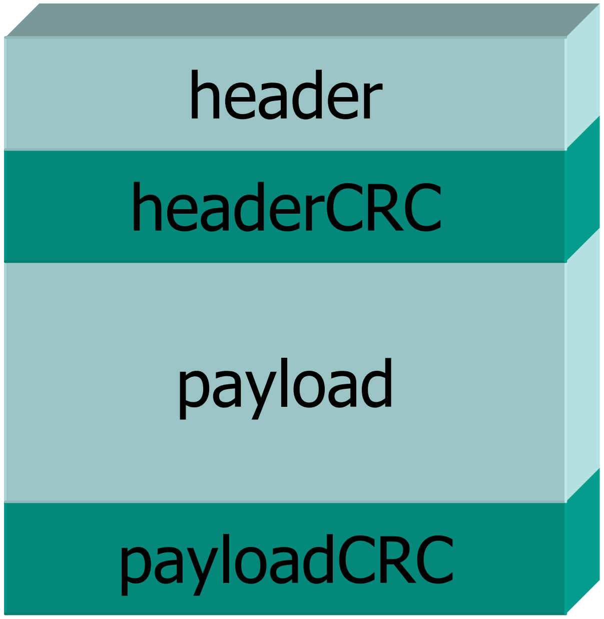
- Triggered on state change
- Triggered on state change
- Also sent periodically
  - Automatic fault recovery
  - Piggyback on heartbeat
- Also distributes stationID addresses
  - Previous: derived from topology and EUI-48 info
  - Bit map supportive ì reclaimingî precedence
- Robust!
  - Context-less behavior (update rate only)
  - No addressing or timeouts required

CRC processing

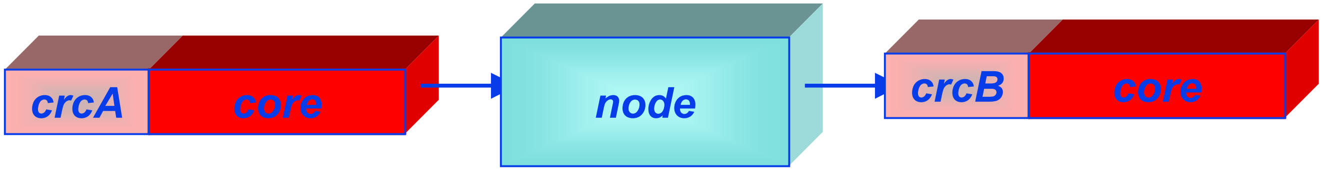
CRC processing

- Store&forward/Cut-through agnostic
- Invalid data is effectively discarded
  - ñ store-and-forward discards
  - ñ cut-through stomps the CRC
- Maximize error-logging accuracy
  - ñ Separate header&data CRCs
  - ñ ì mostî corruptions hit the data

Separate header and data CRCs



Cut-through CRCs

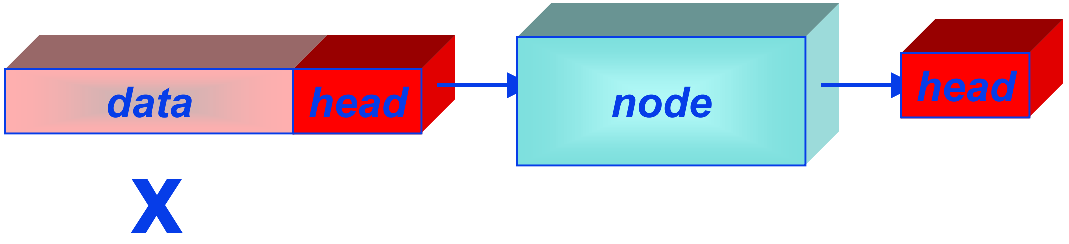


ï Corrupted packet remains corrupted

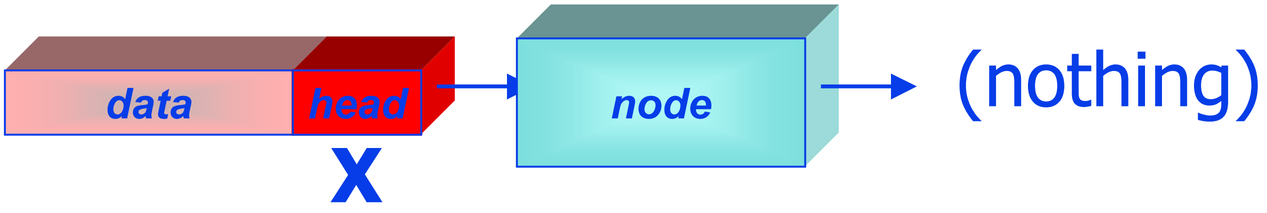
ï Error logged when first detected

```
ï if (crcA!=crc) {
        errorCount+= (crcA!=crc^STOMP);
        crcB= crc^STOMP;
  }
```

Distinct CRCs reduces discards

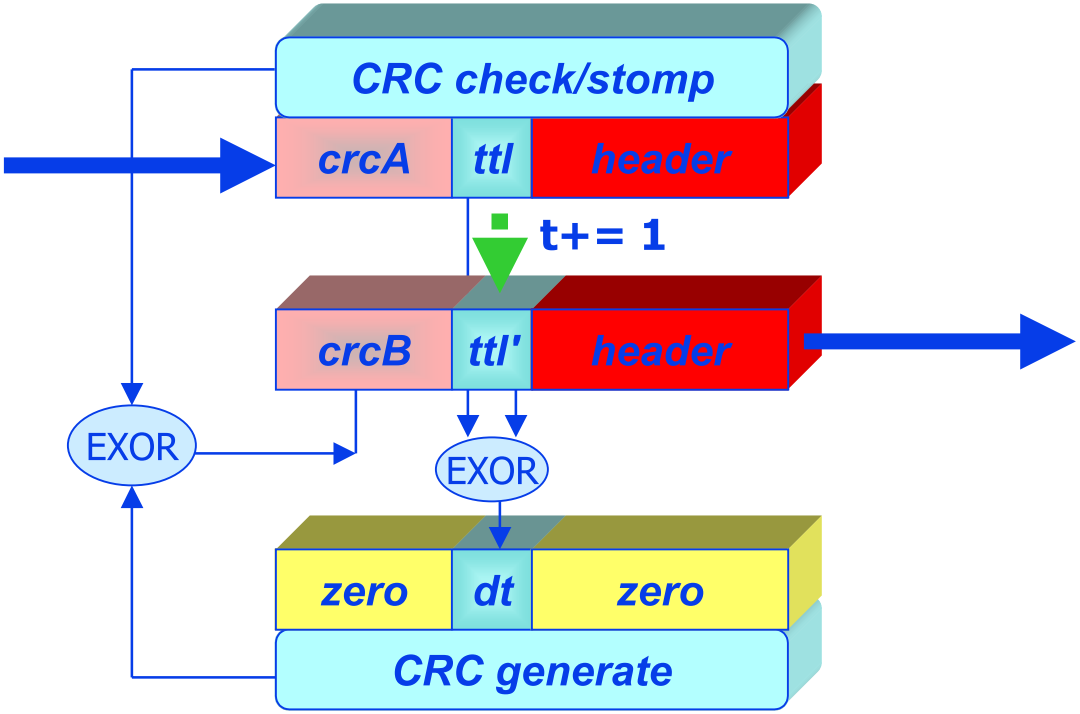


ï Discard the corrupted data



ï Discard the corrupted packet

End-to-end CRC protected TTL

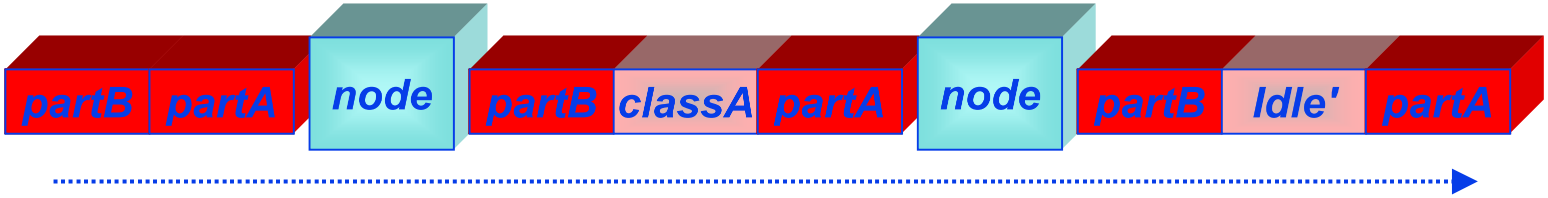


Pre-emption
(a physical layer decision)

Pre-emption

- **Suspend class-B/C for class-A packet**
- **Only one level is sufficient**
  - ñ **class-A is the latency critical traffic**
  - ñ **more levels complicate hardware**
- **Physical layer dependent**
  - ñ **marginal for high BW & small packets**
  - ñ **distinctive ì suspendî symbol required**

Pre-emption fragments



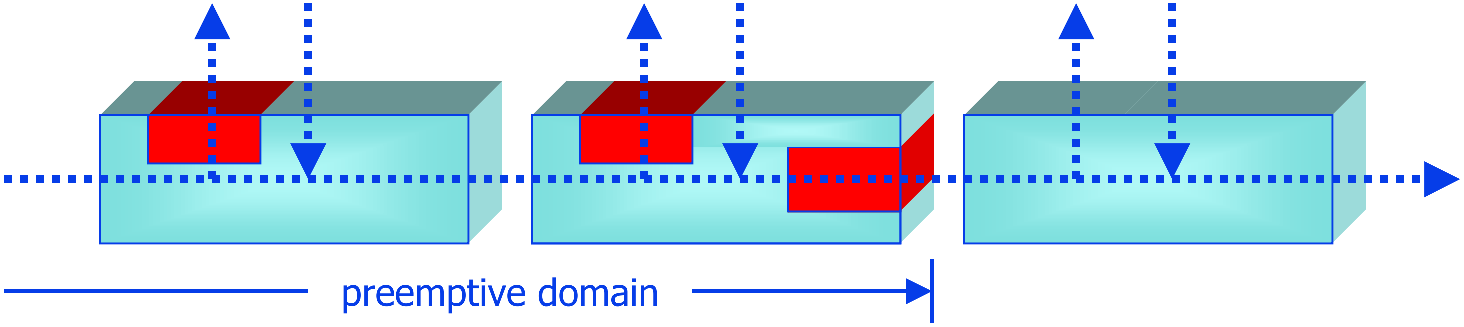
ï Packets can be suspended

ï The class-A packet can be stripped

  ï egress queues are store&forward

  ï distinctive idle markers needed

Pre-emption compatibility



ï Pre-emption mandates egress S&F

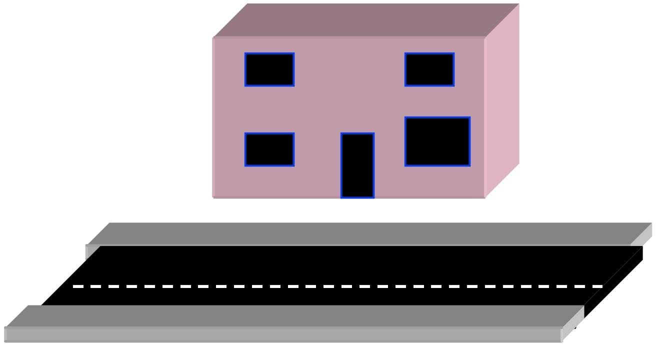
ï Simplistic node has no such S&F

ï Interoperability burden on elegant

  ï boundary node has S&F bypass

  ï cut-through in preemptive domain

Limits of scalability



- Geosynchronous
  - Terrestrial
    - The metro area
      - To the curb
        - To the home

Lessons of the pastÖ

- Flow control mandates 2-out-of-3
  - Low latency transmissions
  - Fair bandwidth allocation
  - High bandwidth utilization
- Feedback control systems
  - Low latency signaling
  - Control can pass class-B/C packets
  - Separate class-A queue is utilized
- Other observations
  - Local control => global perversions
  - Fairness is inherently ì approximateî
  - Strange beating sequences DO OCCUR

Allowed transmissions

	warnings		transmissions		
	LO	HI	none	LO	HI
≥3/4	send	send	A,F	A,F	A,F
≥1/2	send	pass	A,F	A,F	A
≥1/4	pass	--	A,B,C_b,F	A,B	
≥0	--	--	A,B,C_b,C_c,F		

Arbitration summary

- Dual levels
  - Class-A, pre-emptive low latency
  - Class-B, less latency sensitive
- Jumbo frames
  - Affect asynchronous latencies
  - NO IMPACT on synchronous latency
- Cut-through vs store-and-forward
  - Either should be allowed
  - Light-load latency DOES matter

Common features

Common features

- +Separate header and payload CRCs
- +Virtual output queues for efficient spatial reuse
- +Proactive&reactive class-A traffic options
- +Weighted fairness
- +Three fairness classes but distinct naming
  high/medium/low vs A/B/C
- +Node count: >=63, with a desire for 256
  (TTL w/wrap is much simpler if <=127)
- +Wrap and steering supported

Similar themes

- +Duplex queues: Gandolf & DVJ
- +Cumulative discovery: Gandolf & DVJ
- +Steering/wrapping specified on per-packet basis
- #DVJ: Client-to-MAC physical interface
- #DVJ: Clock differences (elasticity buffer mgmnt)
- #DVJ: Time-of-day (stratum check)
- #DVJ: Brandwidth reservation management (for consistent provisioning)
- #DVJ: CRC-32 formats (MAC assumes only one?)

Contending mechanisms

- -More than duplex (x2) ringlets
  DVJ&Gandolf: x2 duplex ONLY
  Alladin: xN if not found to be ì overlyî complex
- -Flow control (B and C)
- -Frame format fields
  - ñ Presence or absence of stationID fields
  - ñ "Questionable" value fields
  - ñ header vs payload, for type & CID
- Discovery
  DVJ&Gandolf: Cumulative discovery
  Alladin: Multistep