N
E
W

C
H
A
P
T
E
R

Part II
« The present »

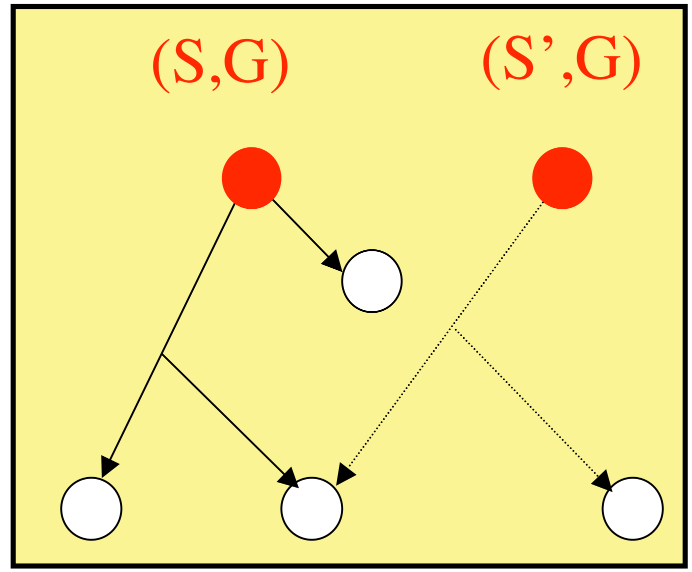
- Advanced group management
- Advanced routing
- Advanced reliability features
- Multicast congestion control
- IETF standards

IGMP v3, RFC 3376

- IGMP v1&2 follow the any-source model
  - Any receiver joins to all the sources in a given group: noted as (*,G)
  - Can lead to an overwhelming overhead at the routing level
- IGMP v3 introduce the specific source model
  - A receiver can join to a specific source in a gicen group: noted as (S,G)

Single-Source Multicast (SSM)

- Current infrastructure uses Any-Source Multicast (ASM)
  - any source can send to any group at any time
- Source-specific channel (S,G)
  - only S can send to G
  - another source S' must use a separate channel (S',G)
  - hosts join channels, so a member joining only (S,G) will NOT receive traffic from S'



Source Shivkumar Kalyanaraman

Why SSM?

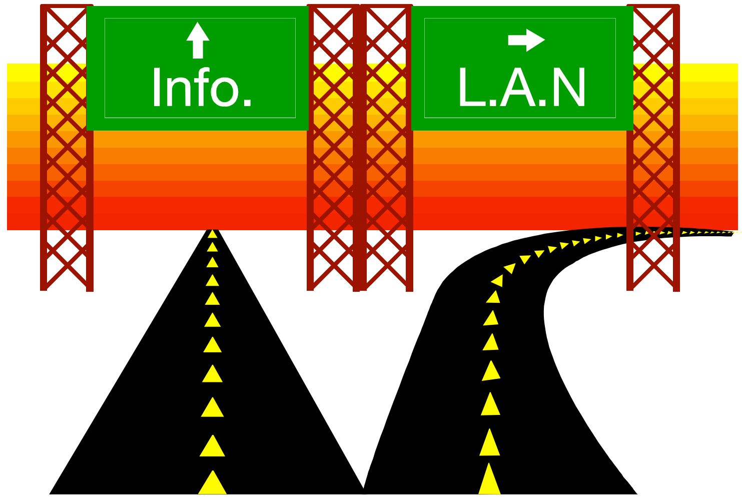
- Network Operator
  - trivial address allocation (16 million addresses per host)
  - no network-layer source discovery (PIM RP and/or MSDP moved to the application layer)
  - overcomes two significant obstacles to deployment

- Content Provider
  - exclusive access to multicast groups (no interruptions)
  - permanent multicast groups (easy to advertise)
  - provides better service

SSM Advantages

- All joins are (S,G), so no need for Class D address allocation
- More security
- Receivers find out about sources through out-of-band means (such as a web site)
- Works with limited modifications of current protocols
  - use IGMPv3 in hosts and 1st hop routers
  - use a modified (simpler) version of PIM-SM
    - No RP, No Bootstrap RP Election
    - No Register state machine
    - No need to keep (*,G), (S,G,rpt) and (*,*,RP) state
    - No (*,G) Assert State

Part II
« The present »



Advanced group management

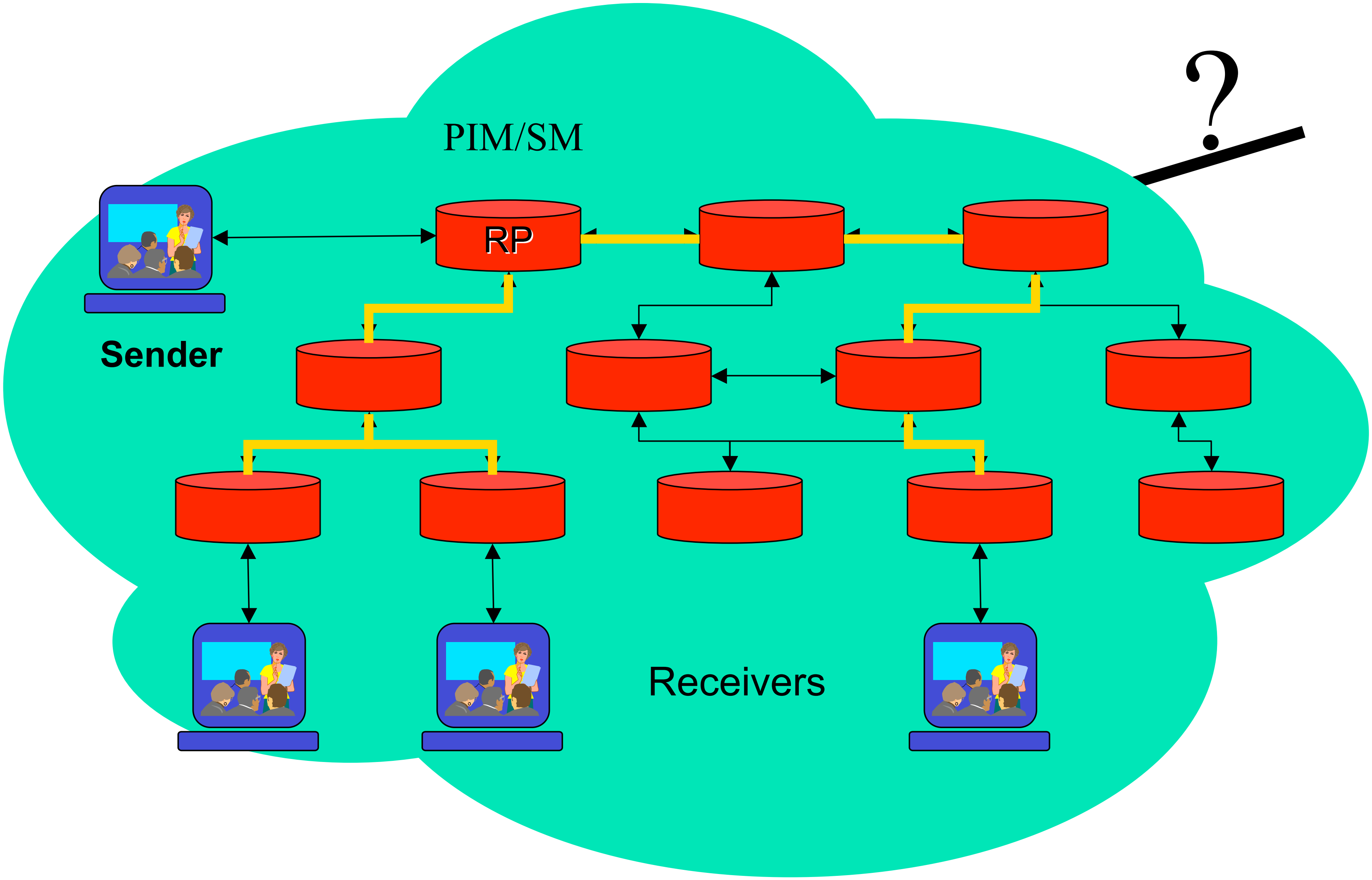
Advanced routing

Advanced reliability features

Multicast congestion control

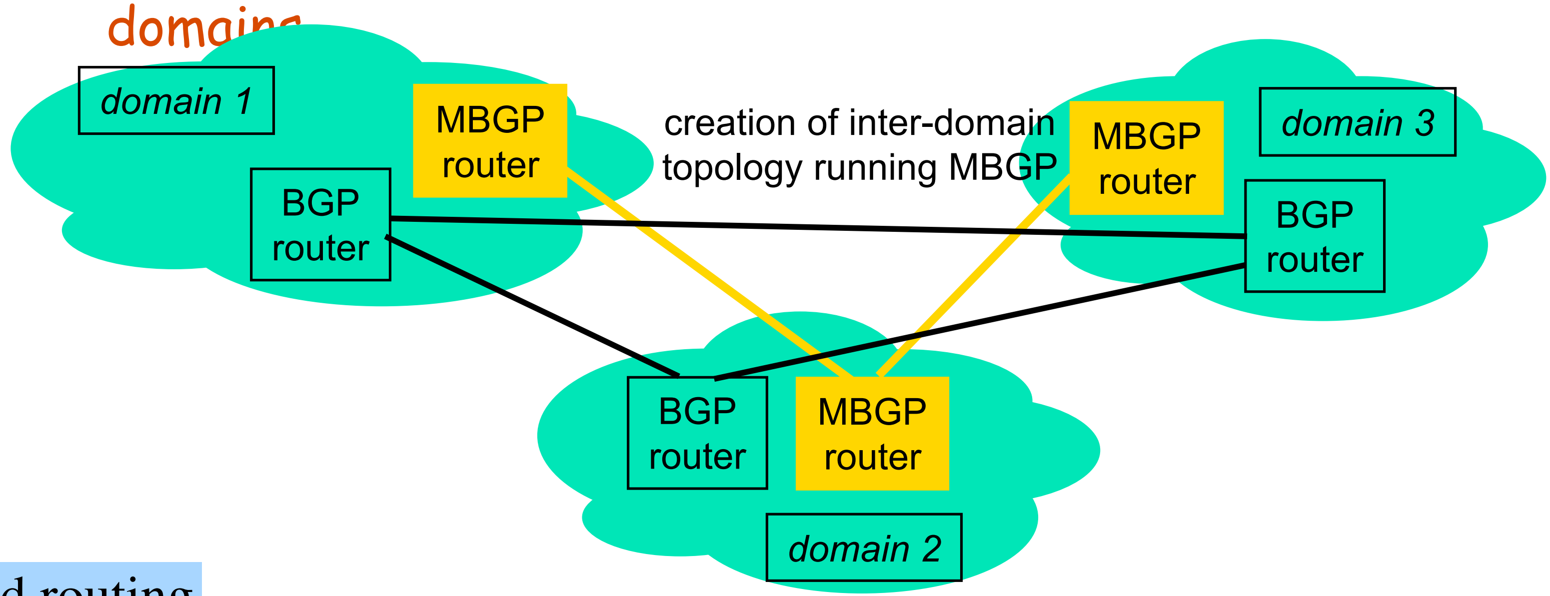
IETF standards

Ok, now I have a tree, so what?

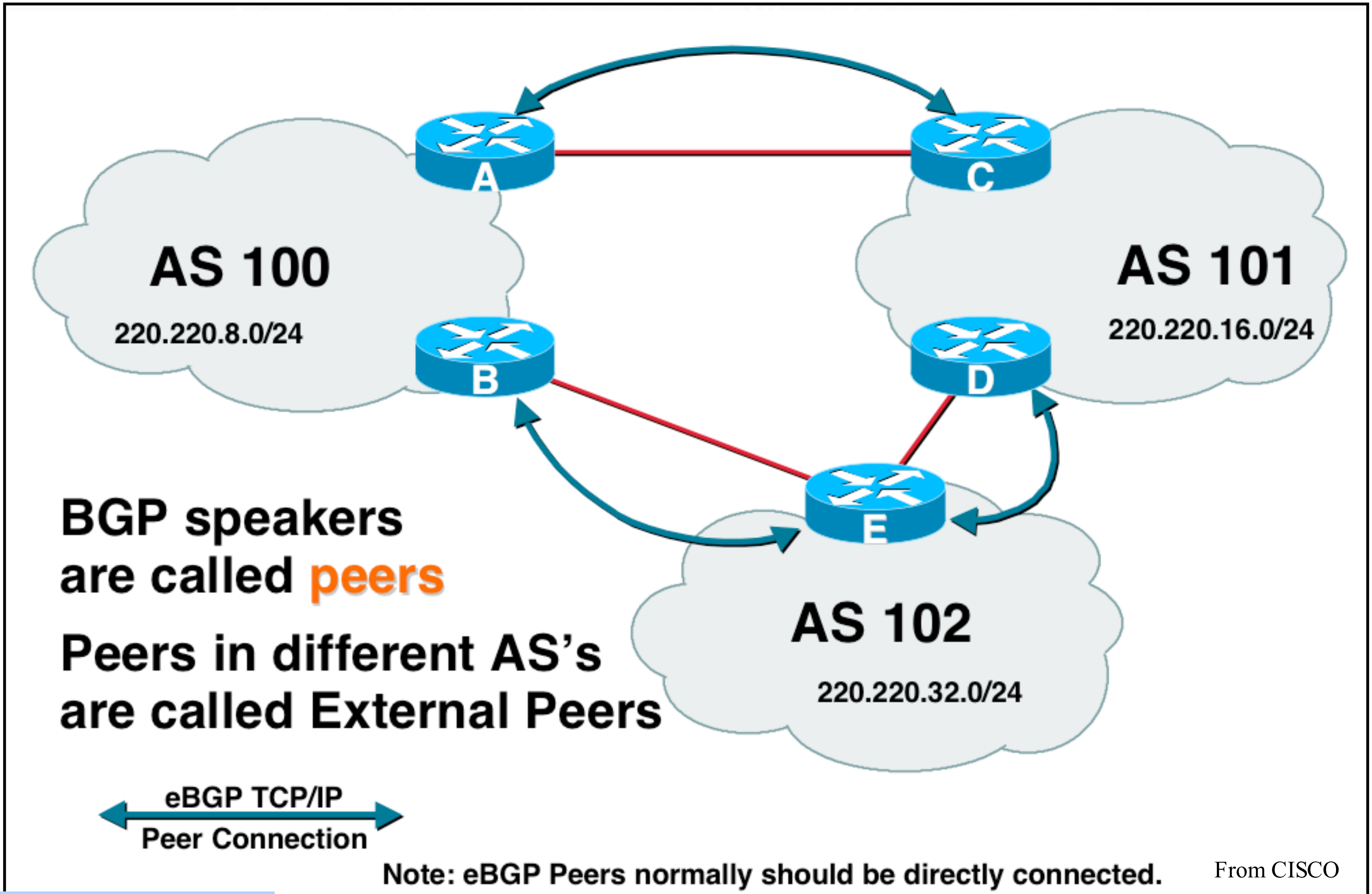


MBGP for inter-domain connectivity

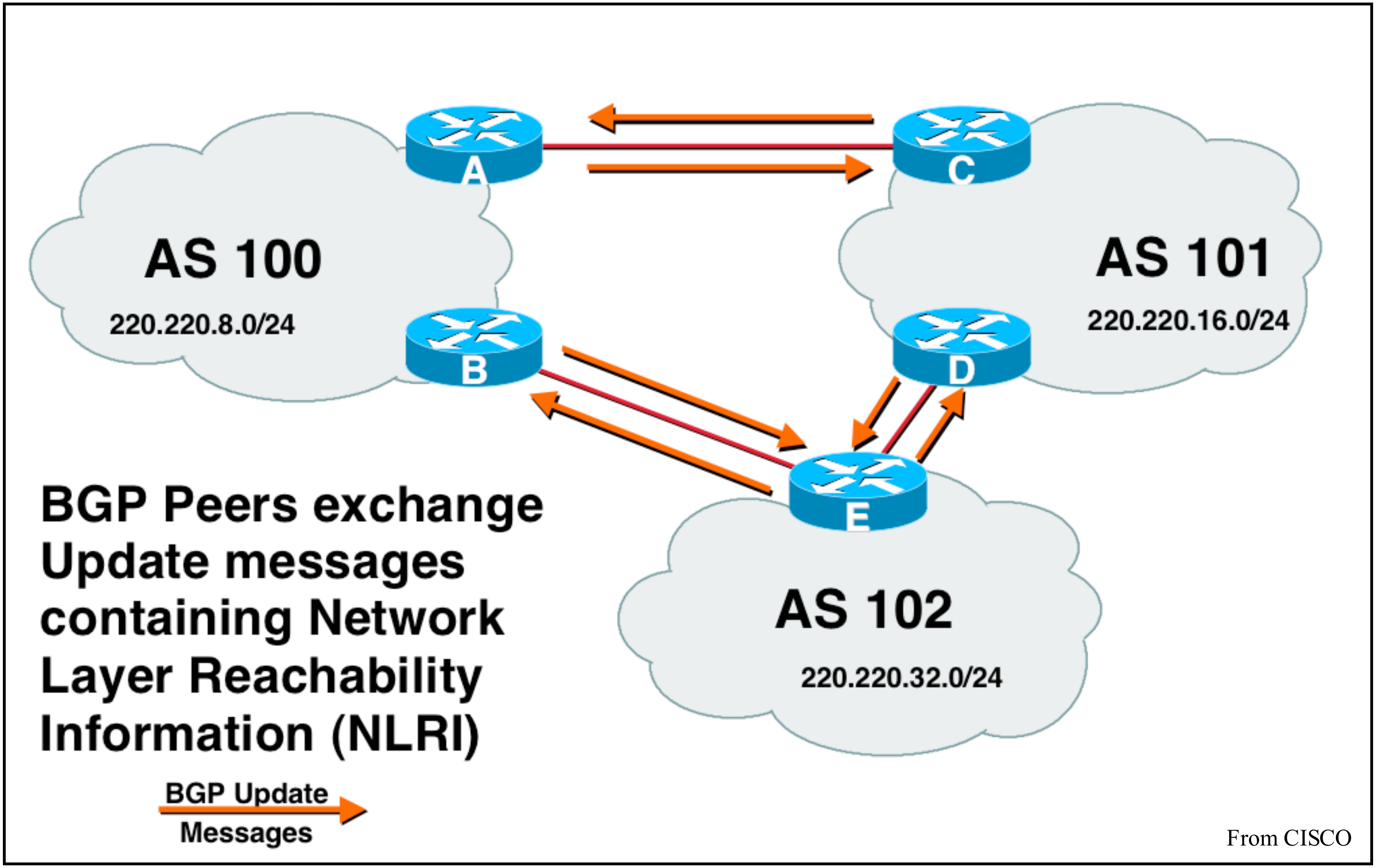
- MBGP (MultiProtocol BGP, RFC 2283) is an extension to BGP4 to carry more than IPv4 route prefix (MP_REACH_NLRI)
- Maintained a separate M(ulticast)-RIB in order to perform RPF between AS
- The internal domain's topology is only known to the local MBGP router
- Each MBGP router only knows how to reach other multicast domains



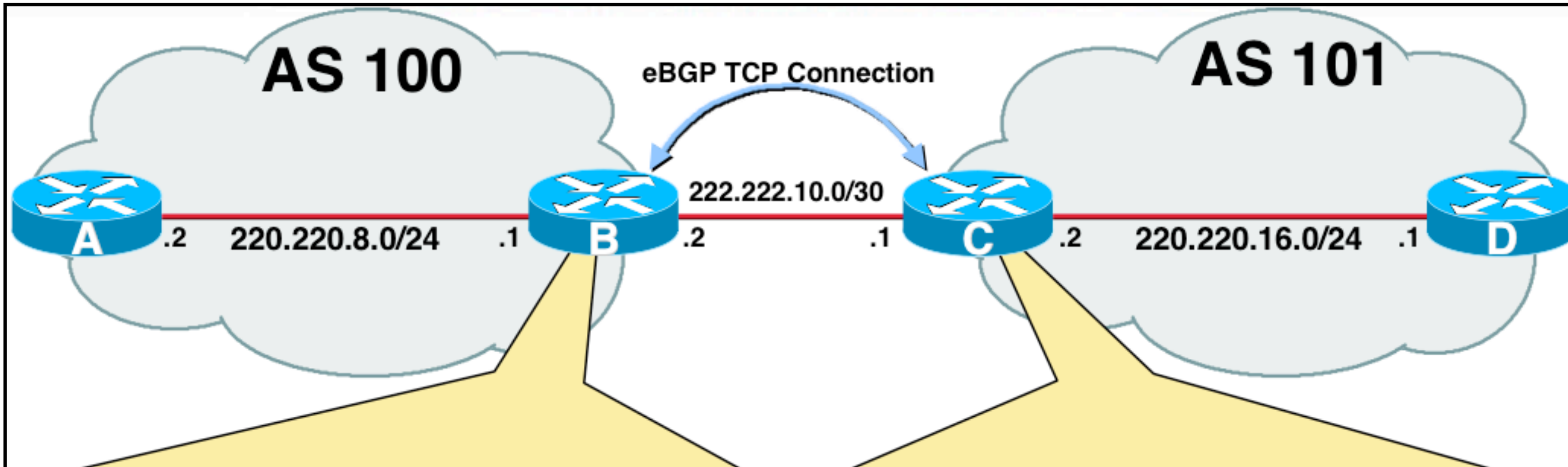
BGP background (1)



BGP background (2)



BGP background (3)




```
interface Serial 0
ip address 222.222.10.2 255.255.255.252

router bgp 100
 network 220.220.8.0 mask 255.255.255.0
 neighbor 222.222.10.1 remote-as 101
```

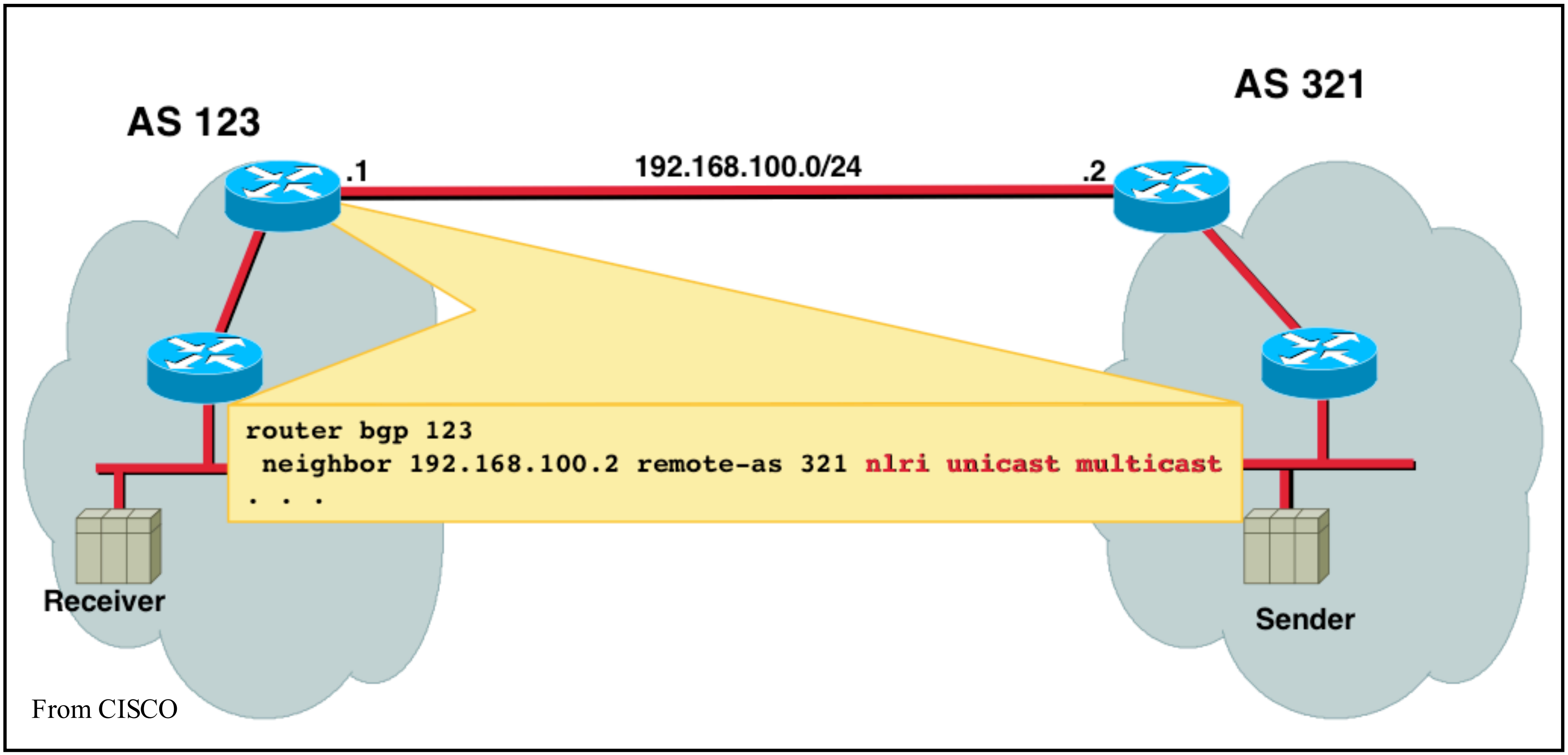
```
interface Serial 0
ip address 222.222.10.1 255.255.255.252

router bgp 101
 network 220.220.16.0 mask 255.255.255.0
 neighbor 222.222.10.2 remote-as 100
```

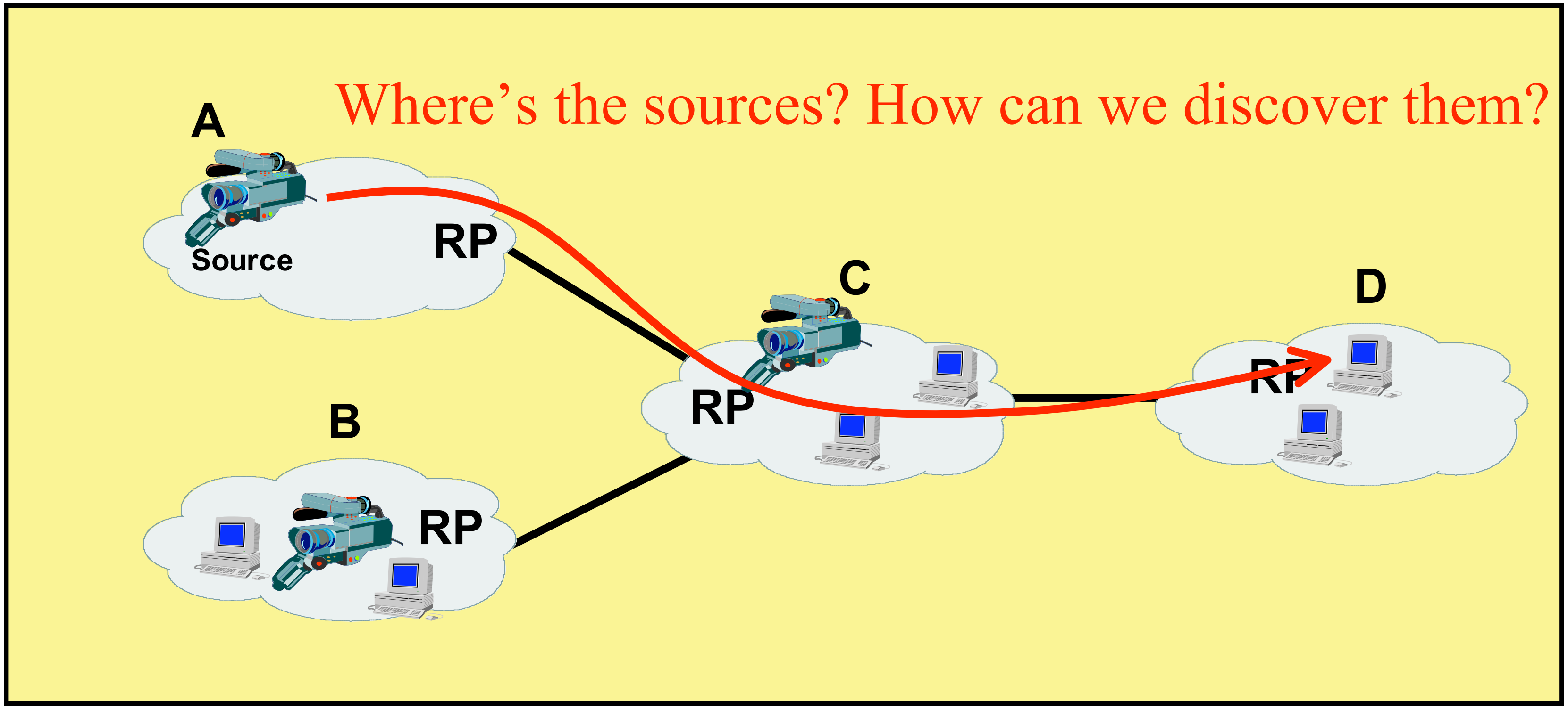
- **BGP Peering sessions are established using the BGP “neighbor” configuration command**
  - External (eBGP) is configured when AS numbers are different

From CISCO

Multiprotocol BGP

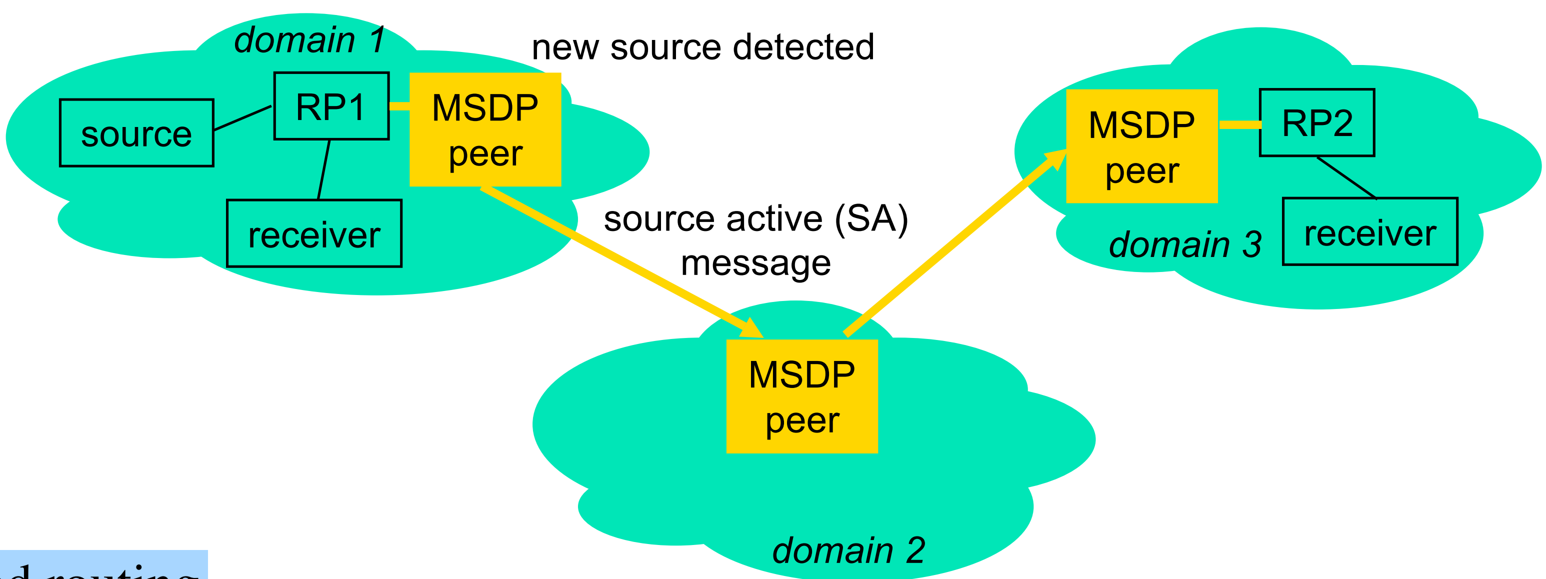


Ok, now I have inter-domain routing, so what?

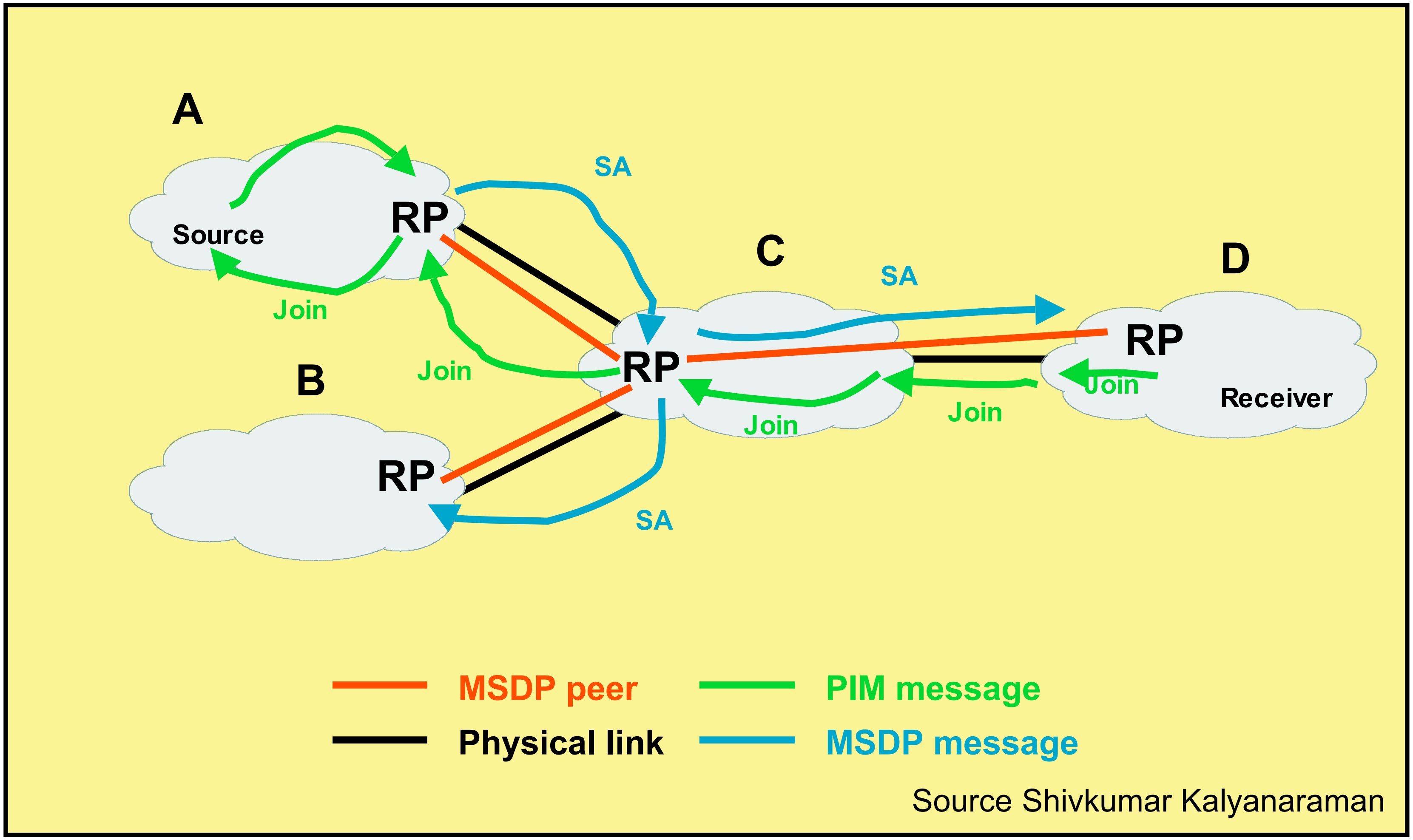


MSDP for inter-domain src discov.

- each domain runs PIM-SM with its own local RP to avoid third-party dependency
- problem: how can a receiver in a domain be informed of a source located in another domain... with MSDP!

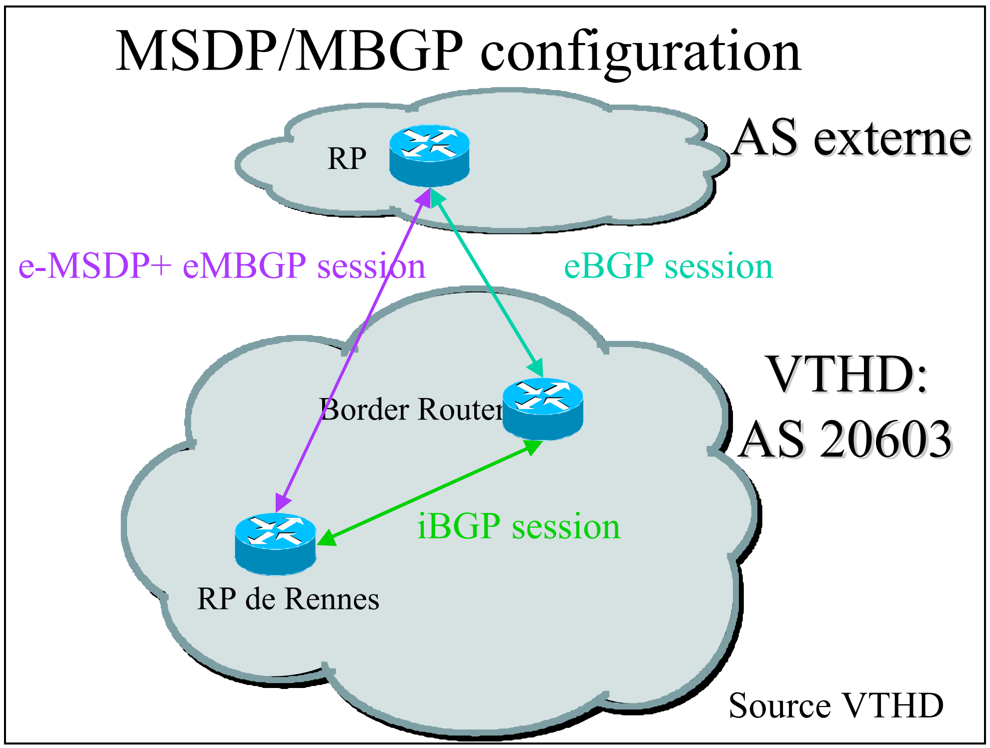


How MSDP works with PIM-SM



Example: MBGP/MSDP on VTHD

- RP's address is announced with MBGP
- External active sources are discovered with MSDP



MSDP… (cont')

- problem with some applications
  - reducing the join latency requires using a cache in each peer of active sources
  - follows a soft-state model, where entries must be periodically refreshed
  - does not work with low frequency bursty applications
    - soft-state is lost each time a packet sent… receivers never get any packet
- limited scalability in terms of nb groups
  - each peer informs every other peer of local sources, and everybody knows everything !

Conclusions PIM-SM/MBGP/MSDP

- works, currently operational
  - deployed in VTHD (http://www.vthd.org)
  - deployed in the GEANT European network

    http://www.dante.net/nep/GEANT-MULTICAST/
- but this is not the long term solution...
  - high signaling load for dynamic groups
  - problems with low frequency bursty applications
  - limited scalability with the number of groups
- long term solution may be quite different...

Part II
« The present »

Advanced group management

Advanced routing

Advanced reliability features

Multicast congestion control

IETF standards

Advanced reliability features

FEC-based solutions

Router-assisted solutions

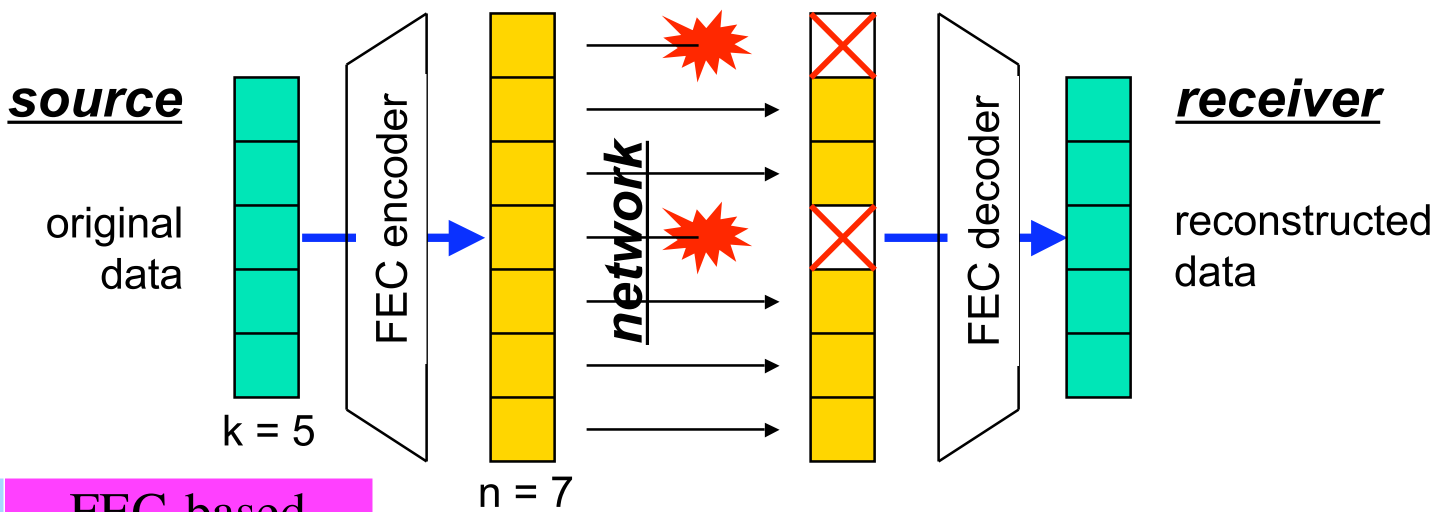
Slides from V. Roca
INRIA Planète

FEC (Forward Error Correction)

- Add some redundancy to the data flow
- A single FEC packet can recover different losses at different receivers $\Rightarrow$ improves scalability
- We only consider packet-based erasure channels (like the Internet)
  - packets are either perfectly received or lost
  - mimics the effects of congested routers
  - FEC operates on a packet basis

MDS property

- Maximum Distance Separable FEC code
  - sender: FEC (k, n)
    - for k original data symbols, add n-k FEC symbols
    - ⇒ total of n symbols (or packets) sent
  - receiver:
    - as soon as it receives any k symbols out of n, a receiver can reconstruct the original k symbols
    - a FEC code with this property is called "MDS"



FEC classification

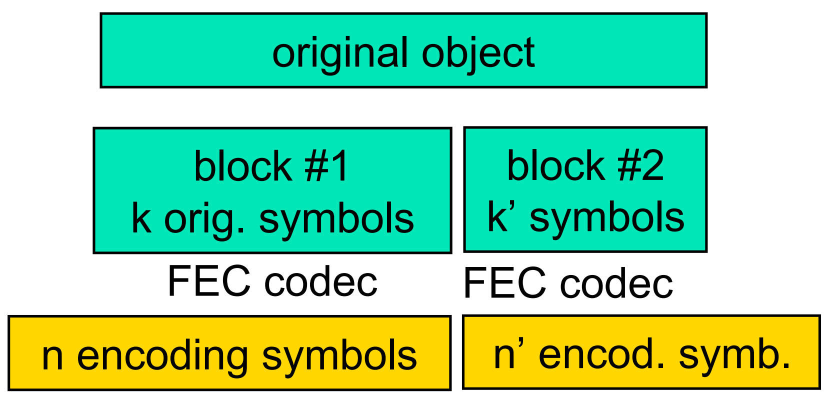
- Classification based on the (k, n) parameters
  - small block FEC codes (small k)
    - Reed-Solomon (based on Vandermonde matrices, or Cauchy matrices), Reed-Muller…
  - large block FEC codes (large k)
    - LDPC, Tornado
    - belong to the "codes on graph" category
  - expandable FEC codes (large k and n)
    - LT

FEC classification... (cont')

- other codes exist but are
  - either lossy codes (ok for video/audio transmission)
  - or dedicated to bit stream transmissions over noisy channels
  - not for us!

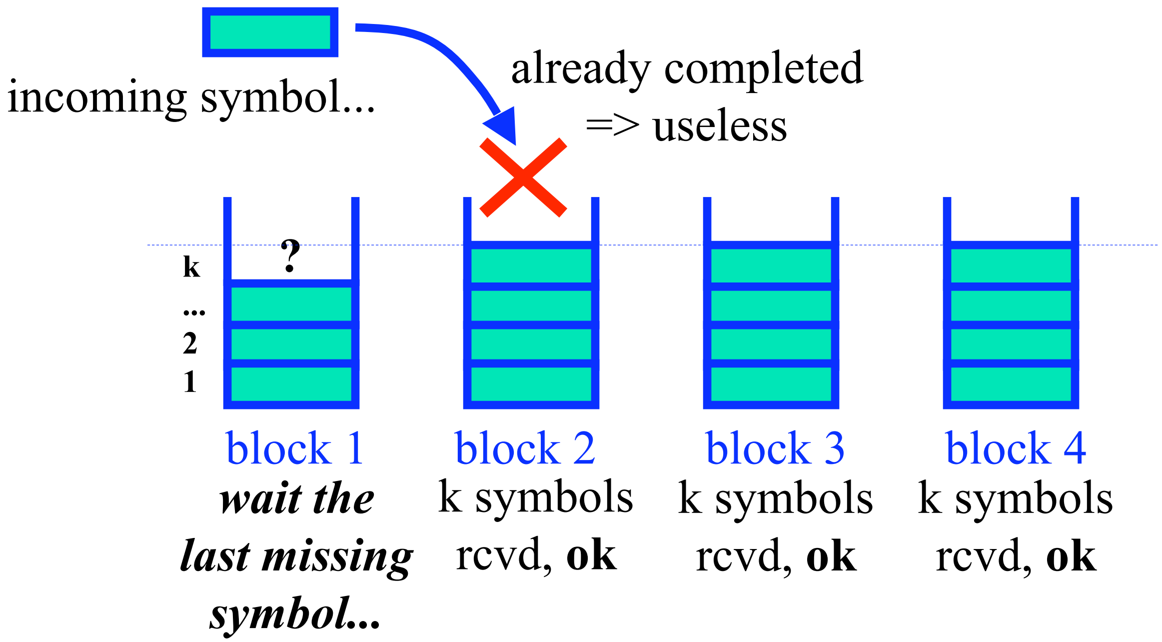
Small block FEC codes

- e.g. Reed-Solomon codes [Rizzo97]
- this is an "MDS code"
  - any k out of n is sufficient to build original pkts
- the k parameter is < a few tens for computational reasons
  - split large data objects into several blocks
  - limits correction capability of a FEC symbol
  - limits the global efficiency



Small block FEC codes... (cont')

- an example of problem generated by a small k



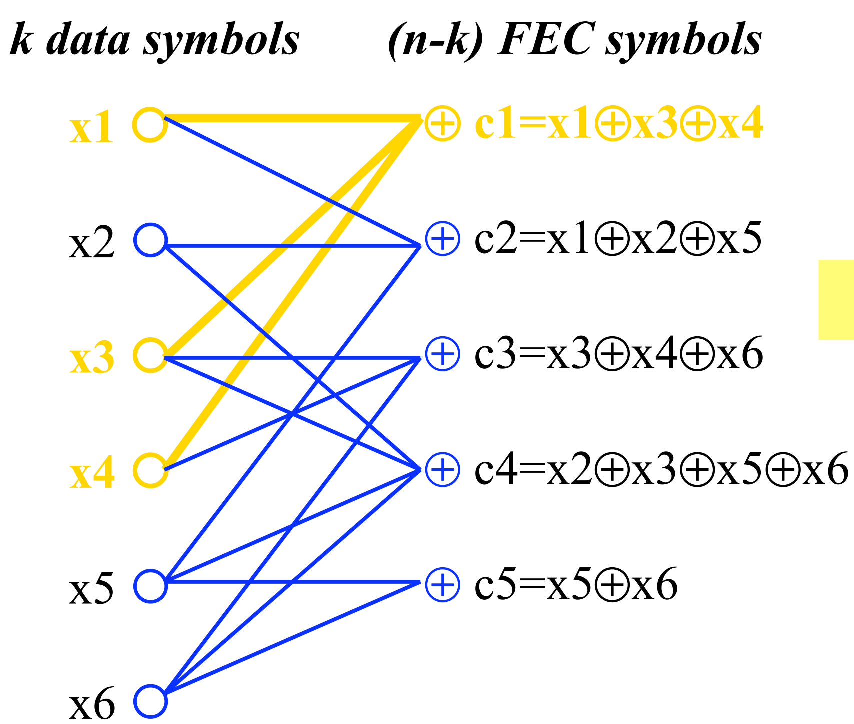
- limited number of n-k FEC symbols created
  - ⇒ can lead to packet duplications
- high quality open-source implementation available

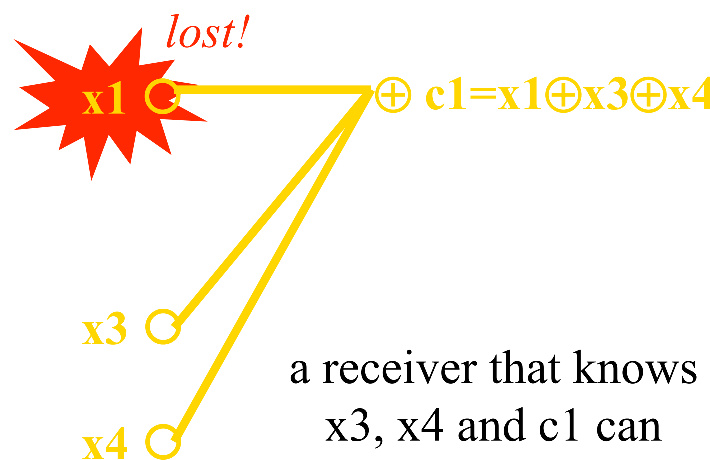
Large block FEC codes

- e.g. LDPC and Tornado codes
- (k,n) with a **very large k**
- but n is limited in practice (e.g. n = 2k)
- decoding requires (1+ε)k, i.e. a bit more than k symbols
  - ε is around %10 (for the best codes) to 40%
  - this is not an MDS code
- high-speed encoding/decoding

Large block FEC codes... (cont')

- an example: LDPC code
  - based on XOR operations (⊕)
  - uses bipartite graphs between source and FEC symbols
  - iterative decoding



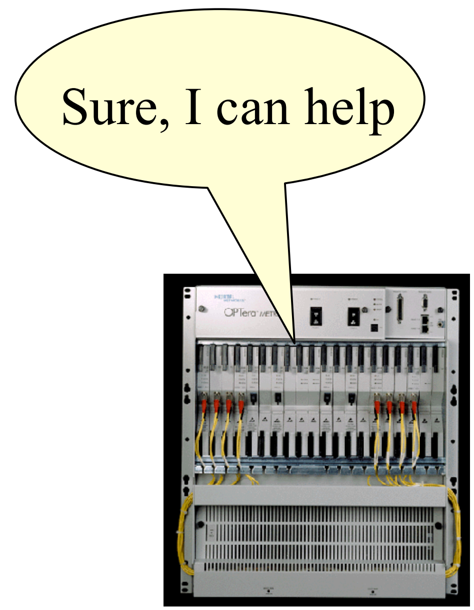


a receiver that knows x3, x4 and c1 can recover x1:
x1=c1+x3+x4

Additional functions in routers

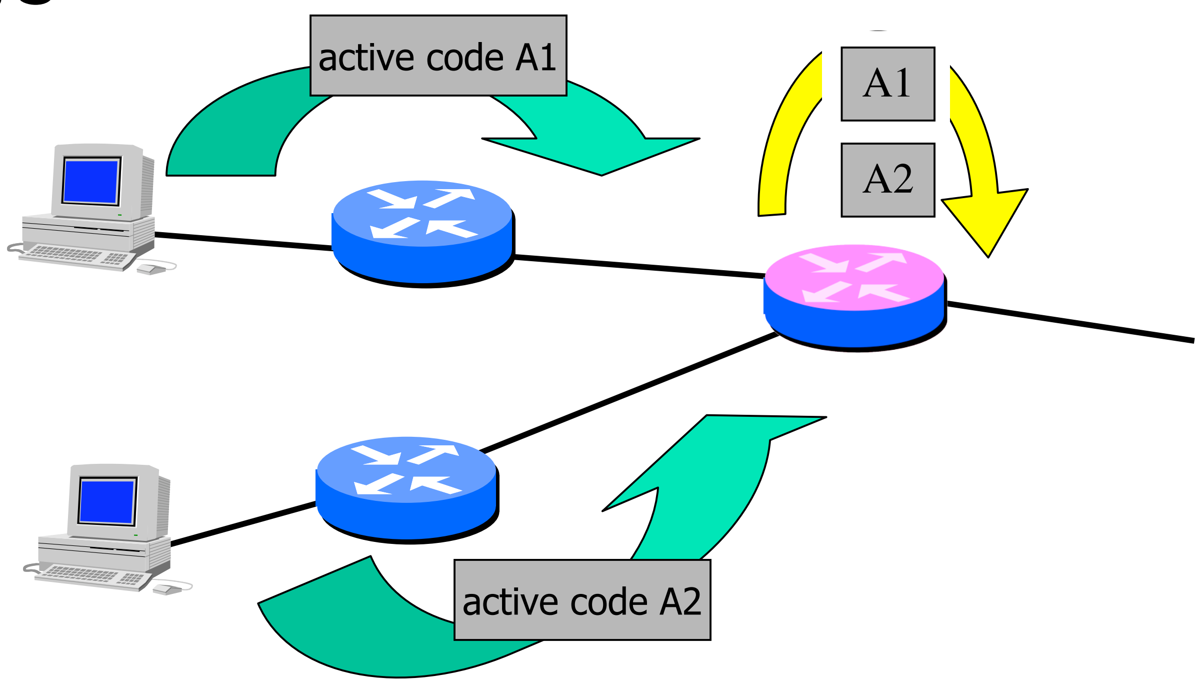
- Traditional approaches
  - end-to-end retransmission schemes
  - scoped retransmission with the TTL fields
  - receiver-based local NACK suppression

- Router-assisted contributions
  - feedback aggregation
  - cache of data to allow local recoveries
  - subcast
  - early lost packet detection
  - ...



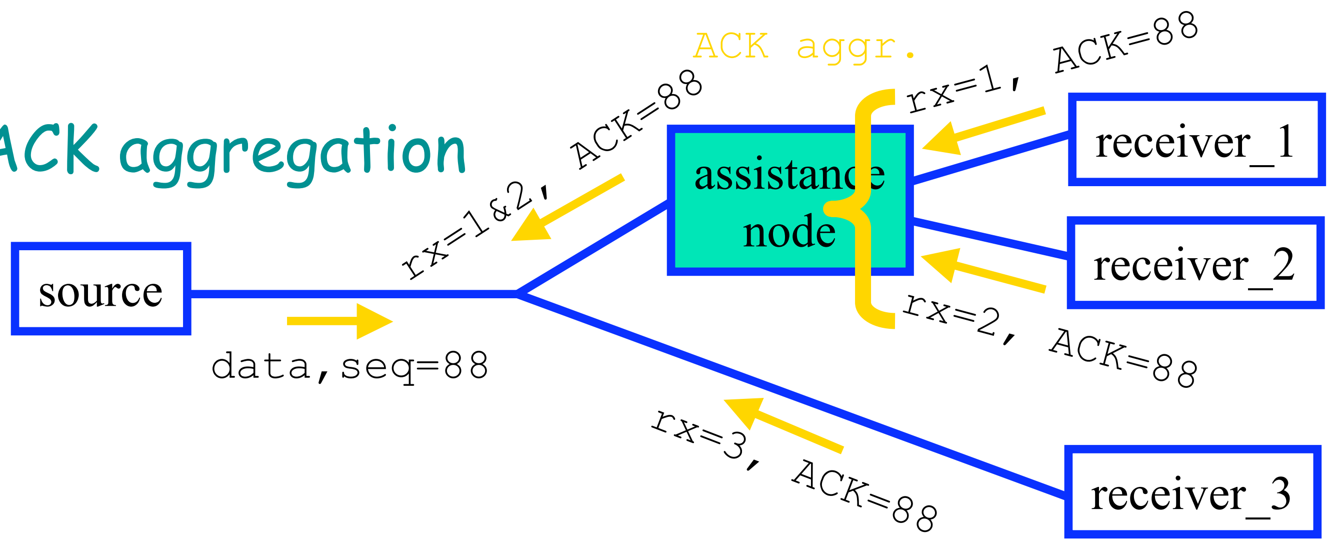
The active network approach

- An execution environment, acting like an OS, can perform dedicated task (specified by the end-user) on incoming packets

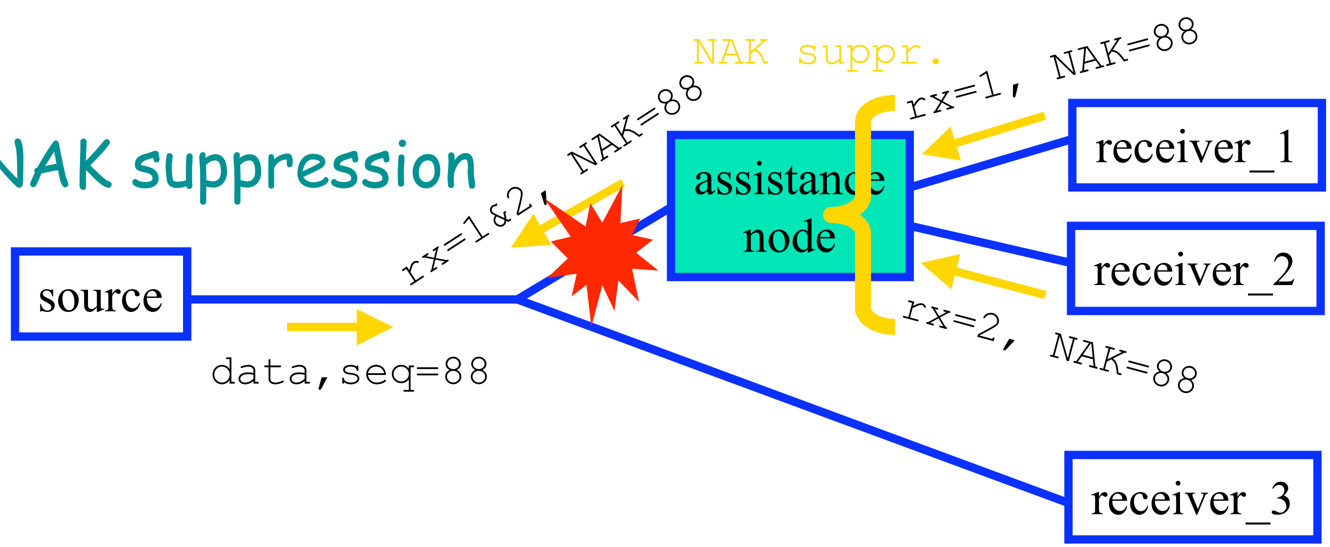


Feedback aggregation example

- ACK aggregation

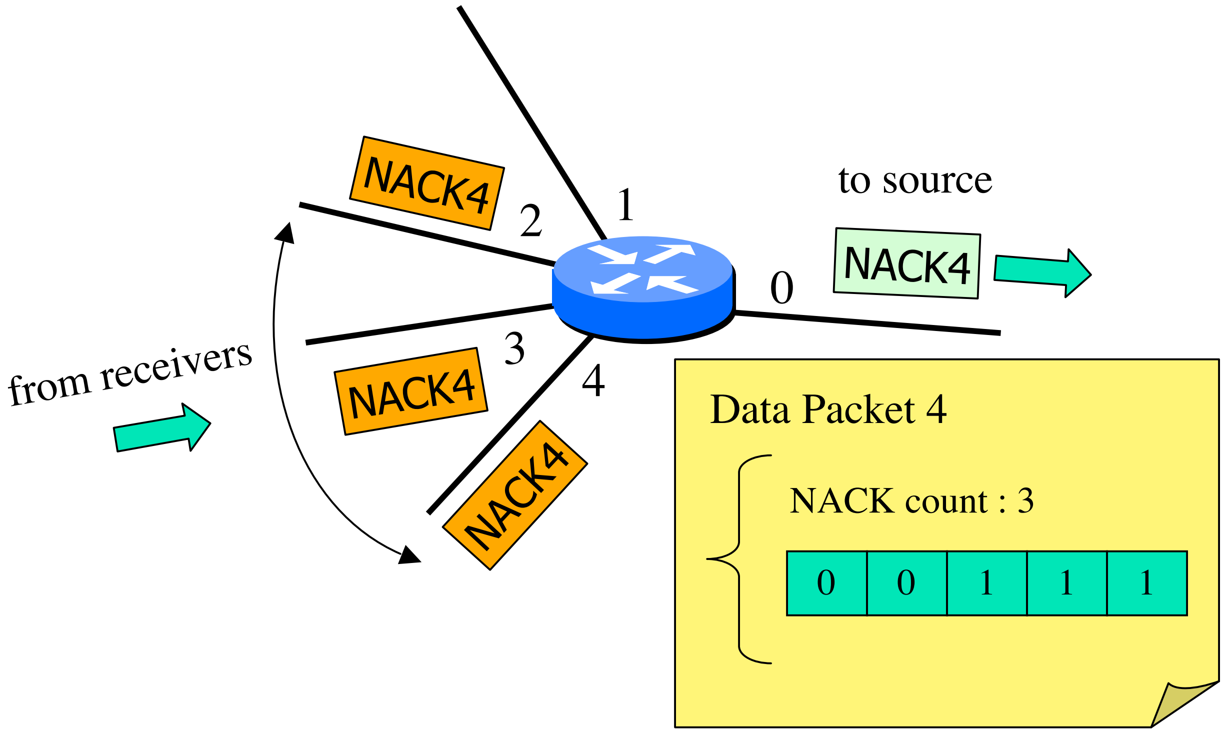


- NAK suppression

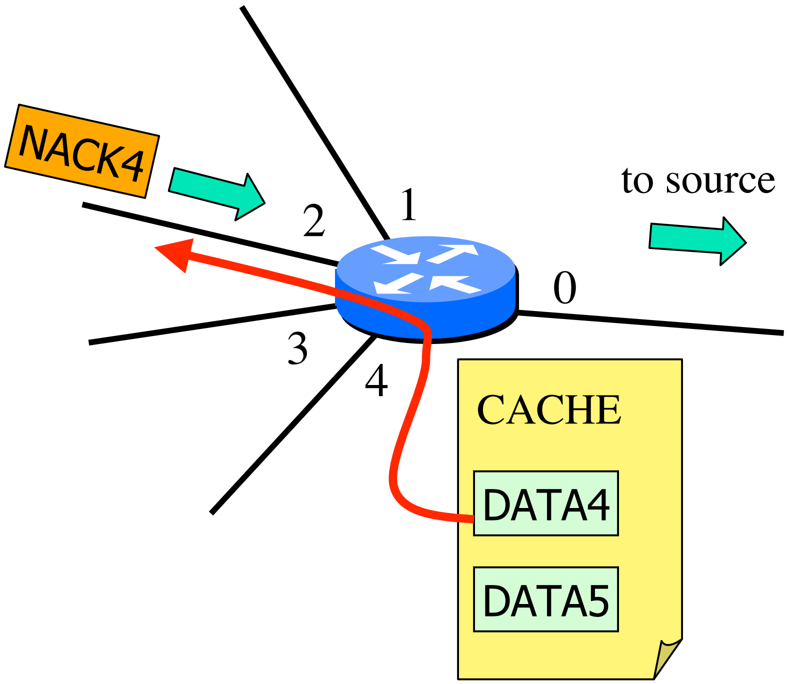


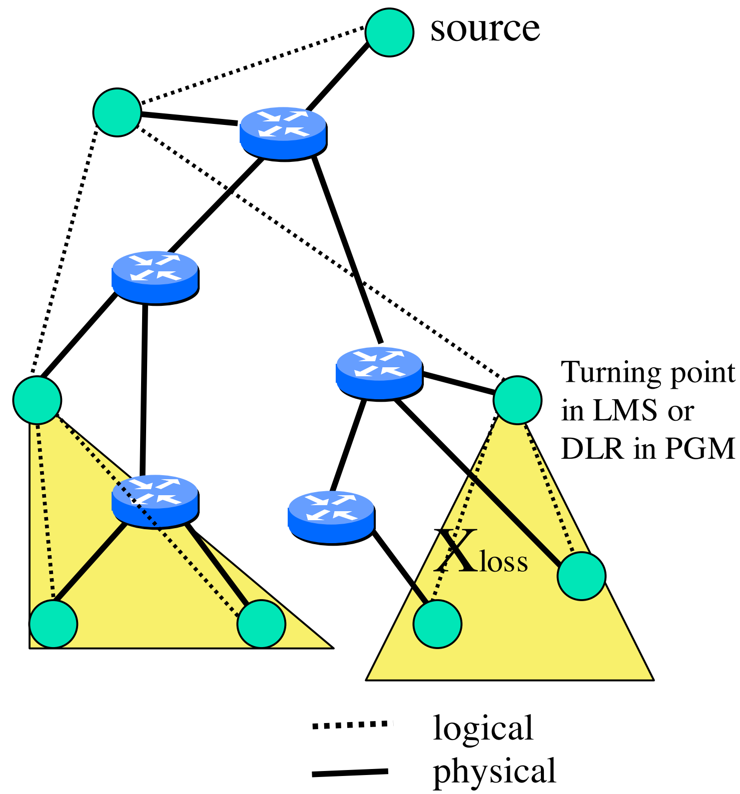
Adv. reliability | Router-assisted

Implementing NACK aggregation



Advanced functionnalities

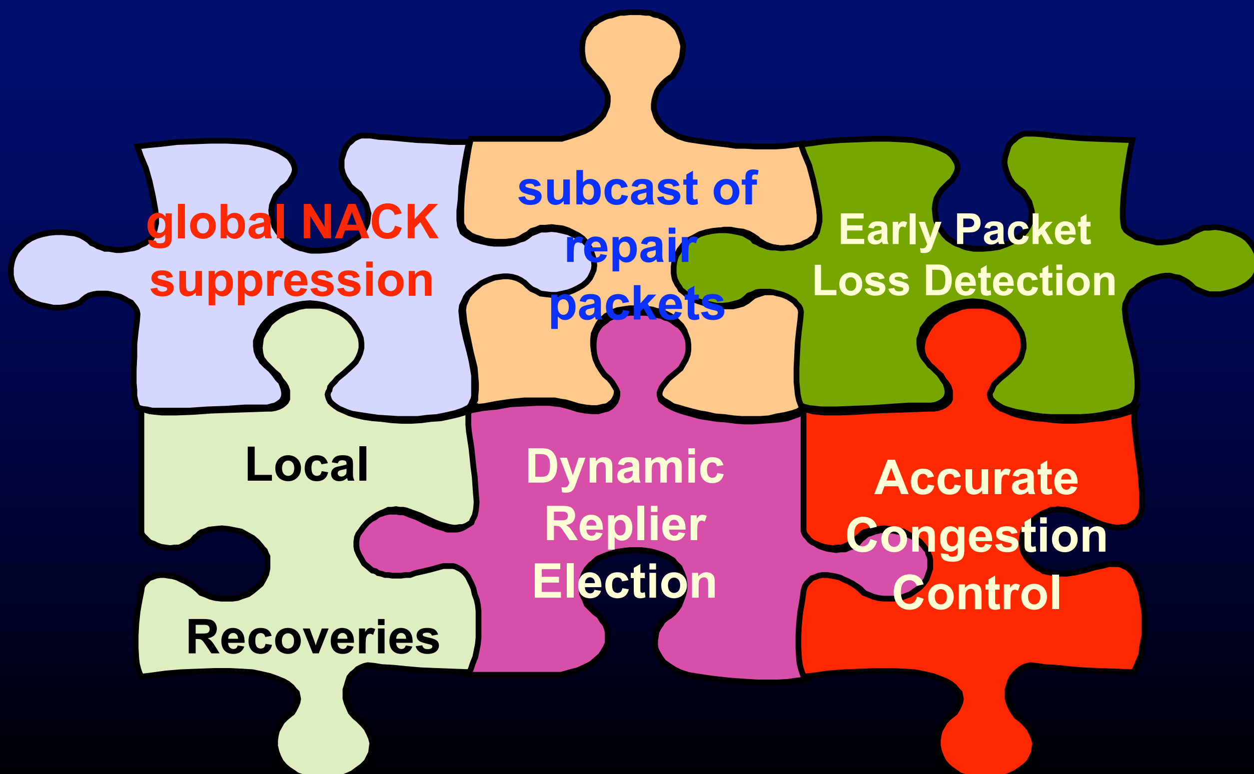




Data packet cache **Representative election**

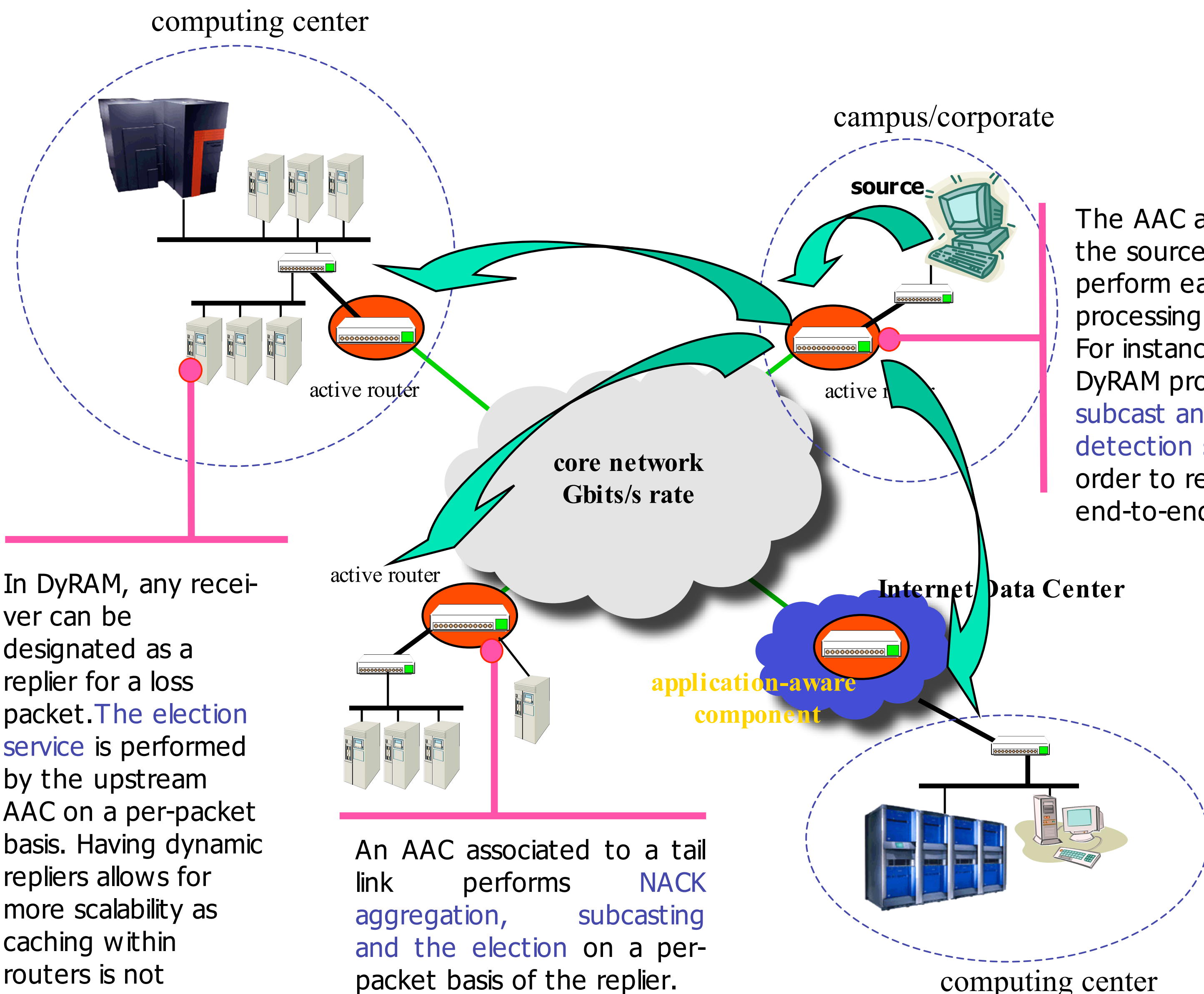
DyRAM (Maimour & Pham, 2001)

Protocol with modular services for achieving reliability, scalability and low latencies



Adv. reliability | Router-assisted

DyRAM on a grid infrastructure



The AAC associated to the source can perform early processing on packets. For instance the DyRAM protocol uses subcast and loss detection services in order to reduce the end-to-end latency.

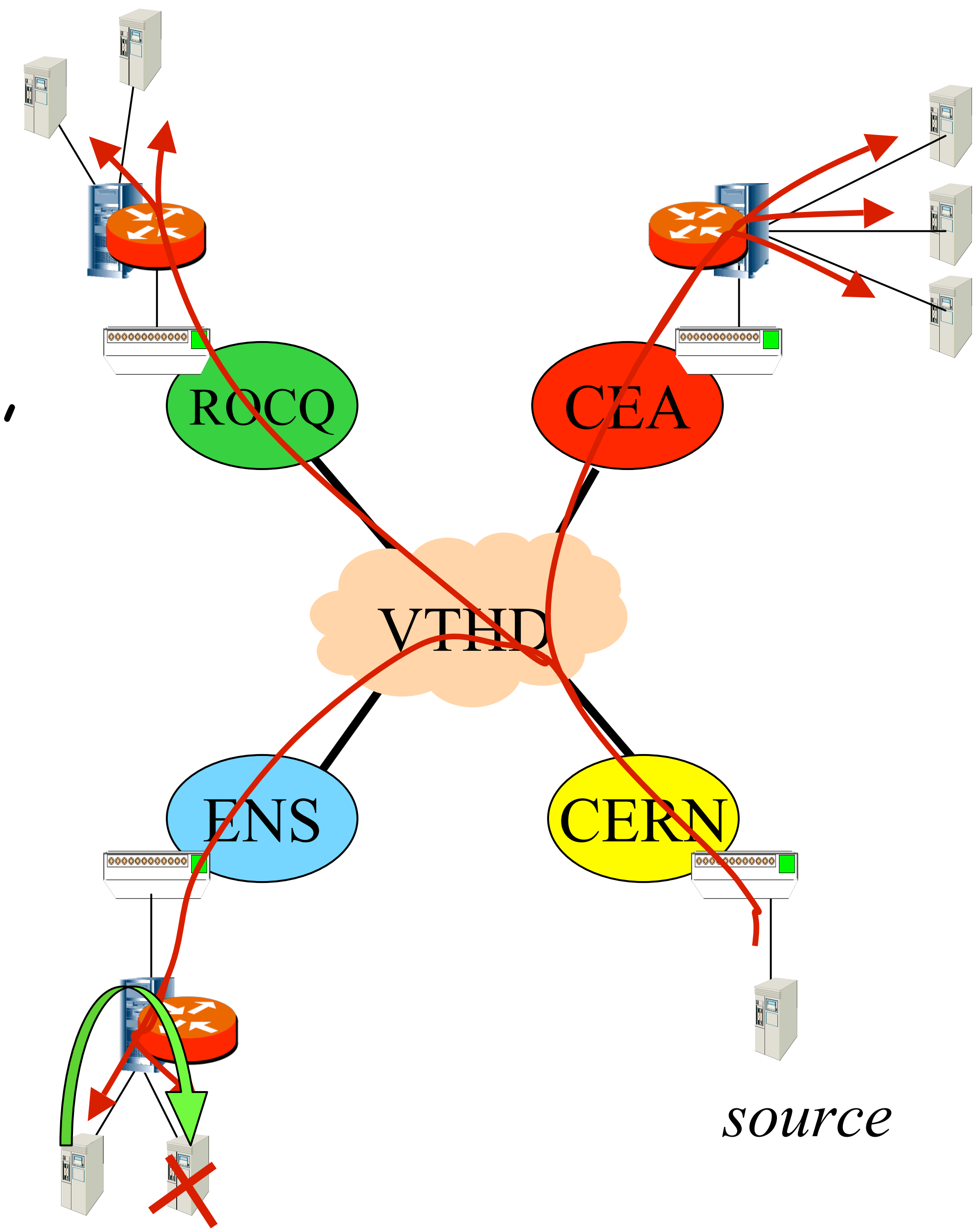
In DyRAM, any receiver can be designated as a replier for a loss packet. The election service is performed by the upstream AAC on a per-packet basis. Having dynamic repliers allows for more scalability as caching within routers is not required.

An AAC associated to a tail link performs NACK aggregation, subcasting and the election on a per-packet basis of the replier.

Adv. reliability | Router-assisted

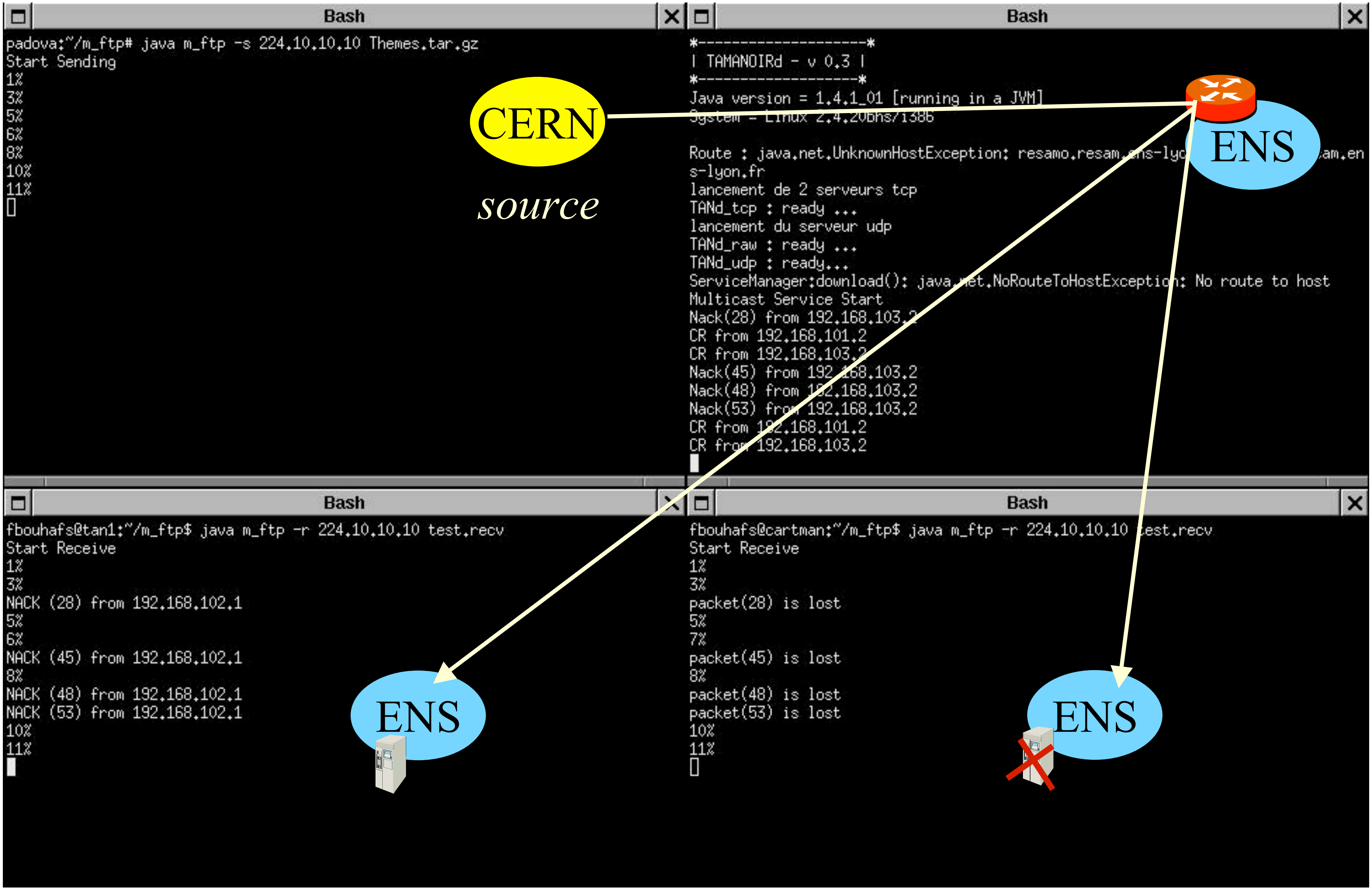
Multicast on E-Toile (RNTL)

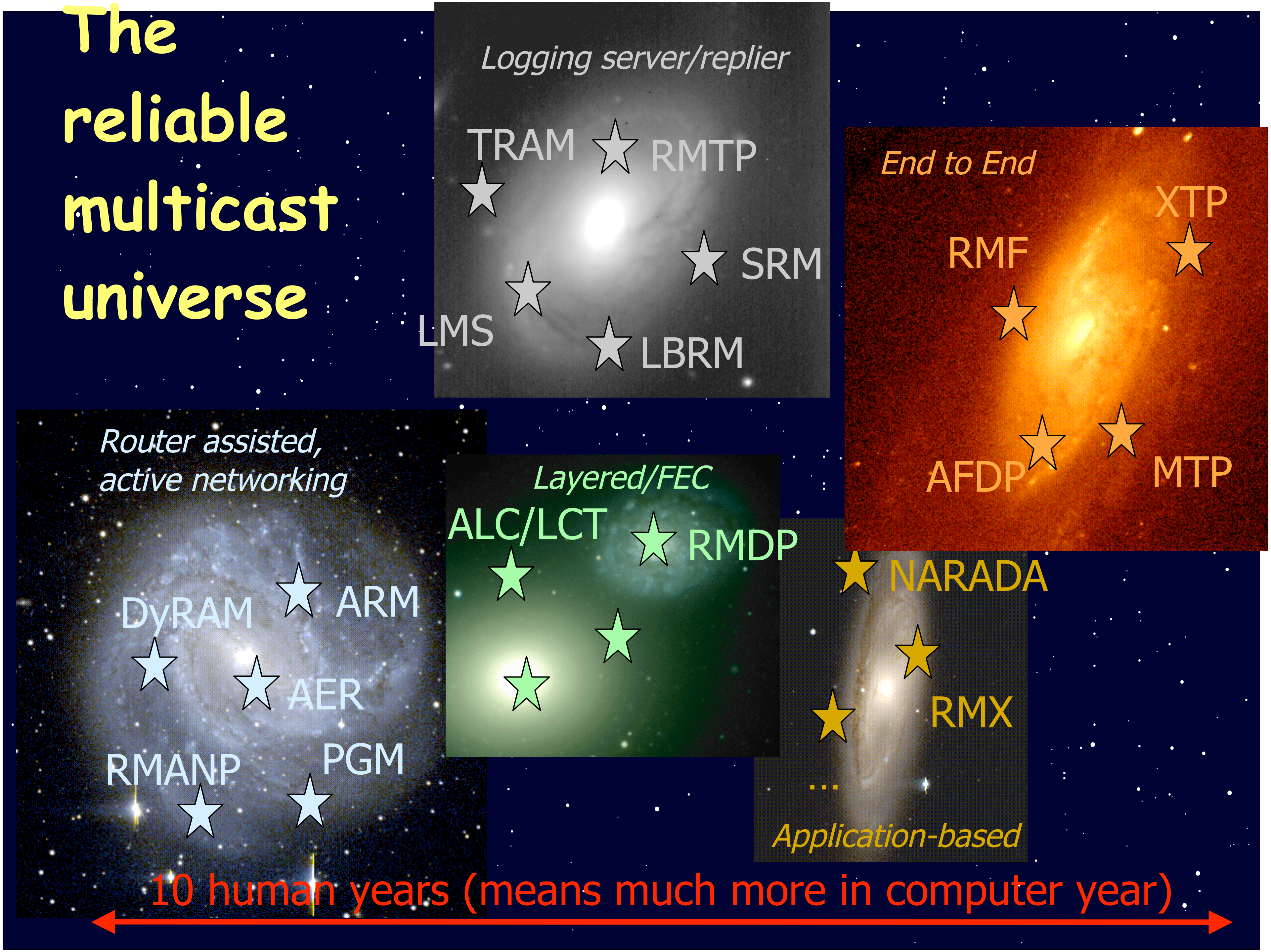
- Demo June 5th, 2003 showing active reliable multicast on computational grids



Adv. reliability | Router-assisted

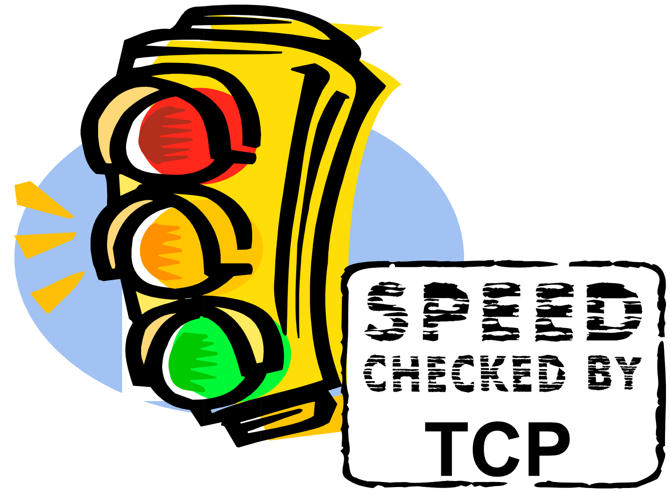
Demo was successfull!





The reliable multicast universe
Logging server/replier
TRAM
RMTP
SRM
LMS
LBRM
End to End
XTP
RMF
AFDP
MTP
Router assisted, active networking
DyRAM
ARM
AER
RMANP
PGM
Layered/FEC
ALC/LCT
RMDP
NARADA
RMX
...
Application-based
10 human years (means much more in computer year)

Part II
« The present »



Advanced group management

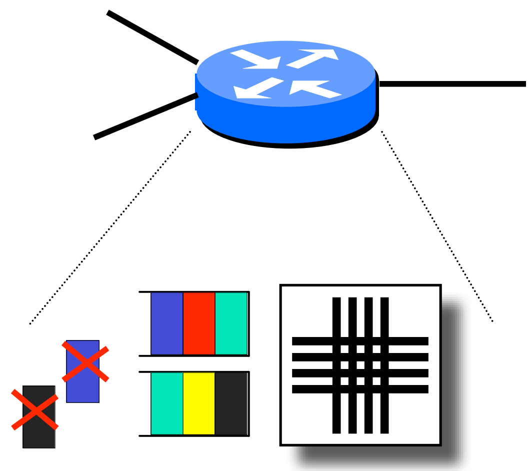
Advanced routing

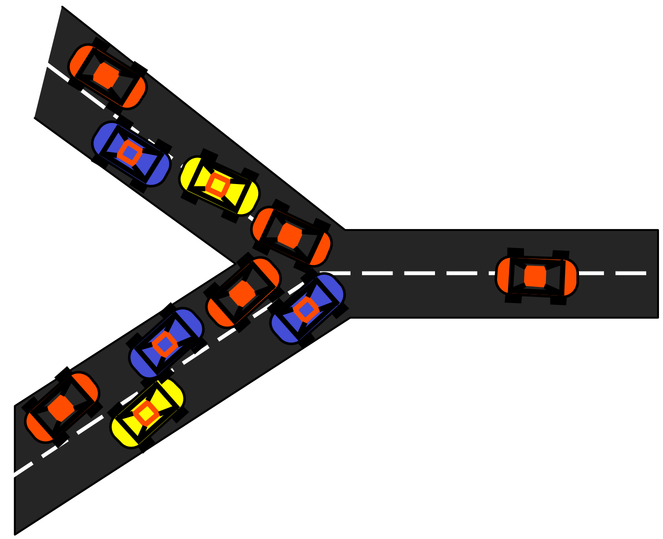
Advanced reliability features

Multicast congestion control

IETF standards

What is congestion?



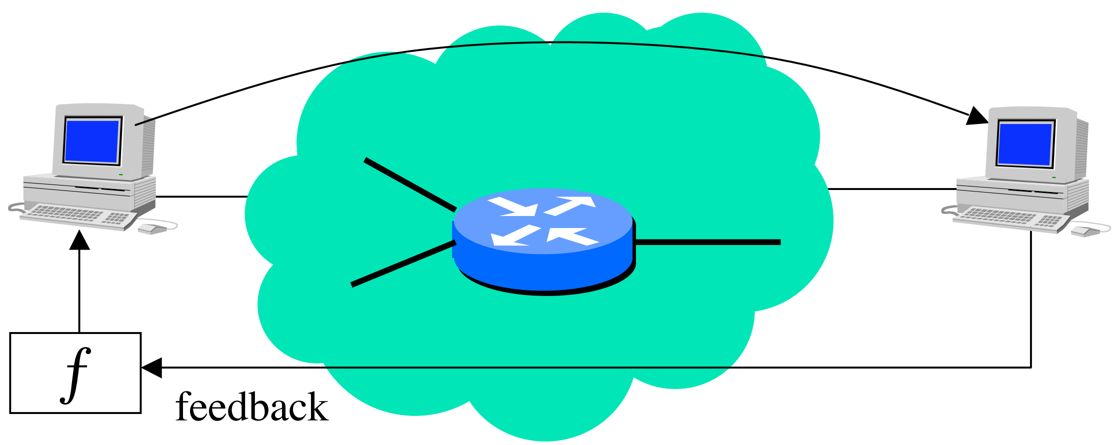


- Congestion appears when too many packets are injected in a network with limited resources
- Main consequences: packet losses

Congestion Control

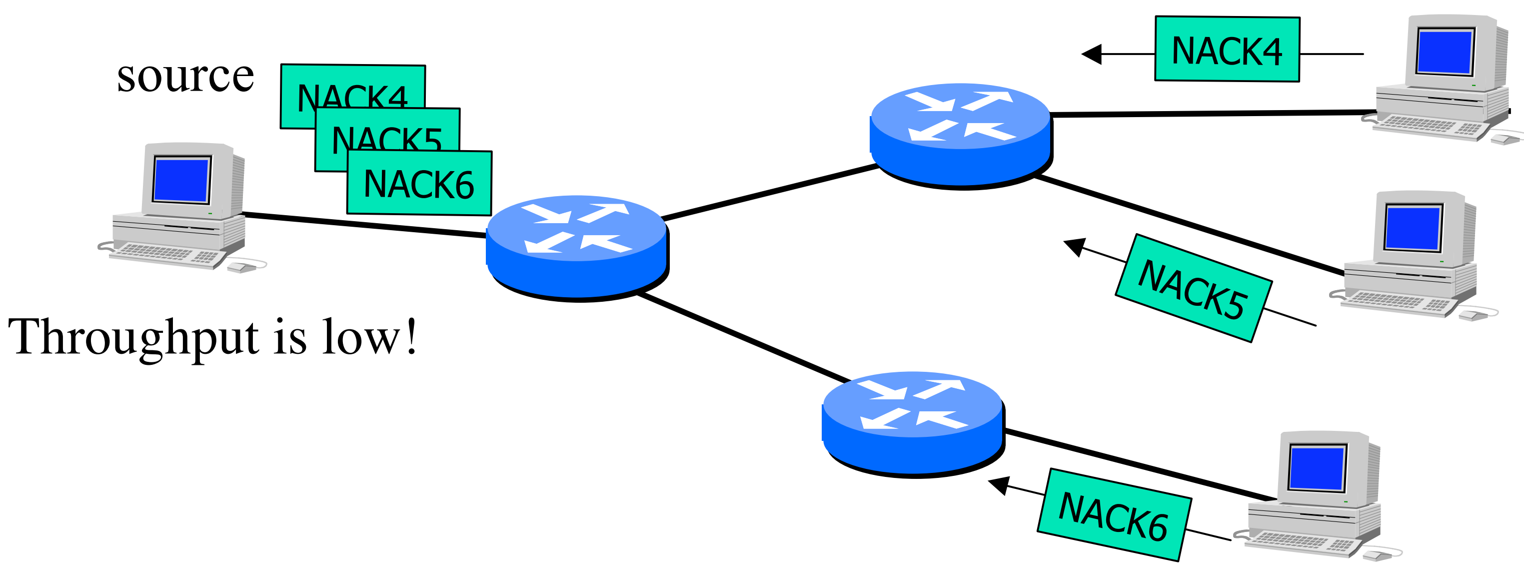
- general goals of CC
  - be ***fair*** with other data flows (be "TCP friendly")
    - no single definition
    - be responsive to network conditions
  - be ***stable***, i.e. avoid oscillations
  - utilize network resources ***efficiently***
    - if only one flow, then use all the available bandwidth

Typical closed-loop control



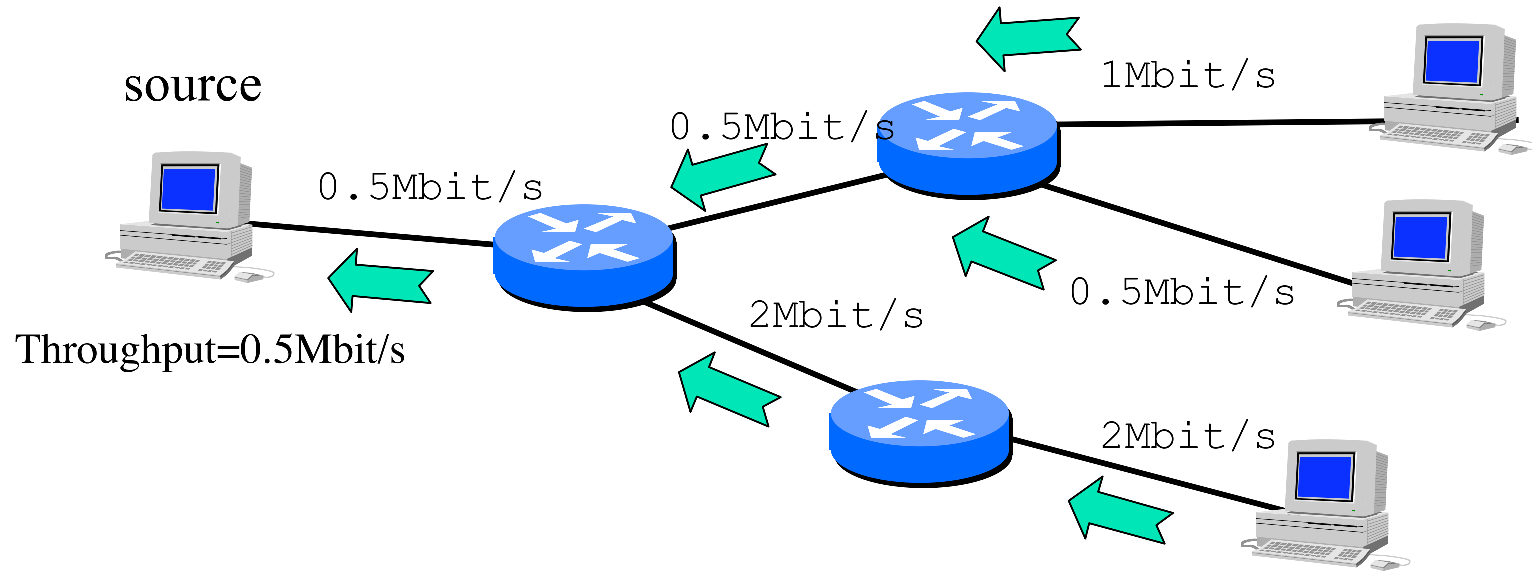
Multicast congestion control (1)

- Multiple receivers, multiple notifications
  - Source implosion problem (similar to the reliability problem)
  - Drop-to-zero syndrom: uncorrelated packet losses are seen as correlated!



Multicast congestion control (2)

- Representativity: who should I follow?



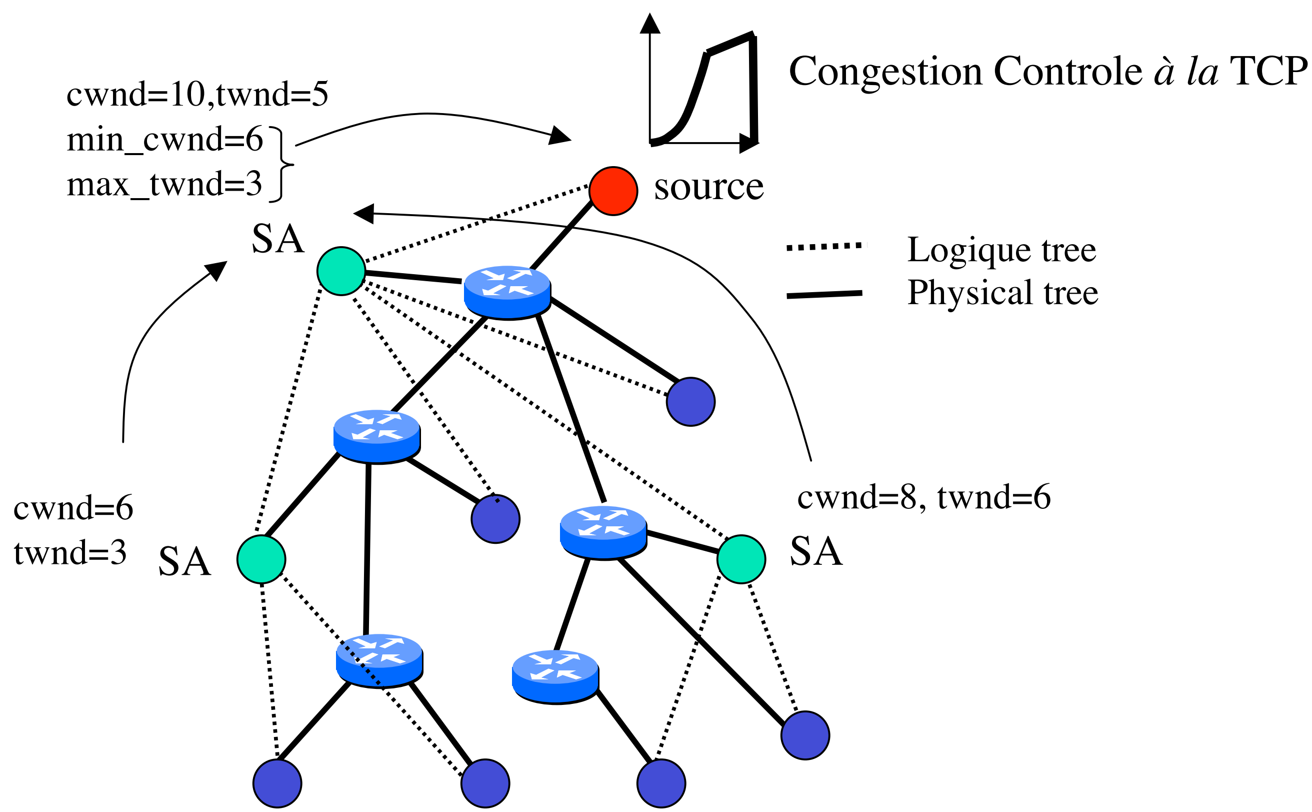
- Single-rate: pace of the slowest
- Multi-rate

Multicast Congestion Control

- Regulation could be
  - Sender-initiated
    - Most approaches are single-rate
    - Uses window or throughput as the regulation parameter
  - Receiver-initiated
    - Most approaches are multi-rate
    - Most approaches use throughput as the regulation parameter
- Congestion notifications could be
  - Losses, delay, queue size…

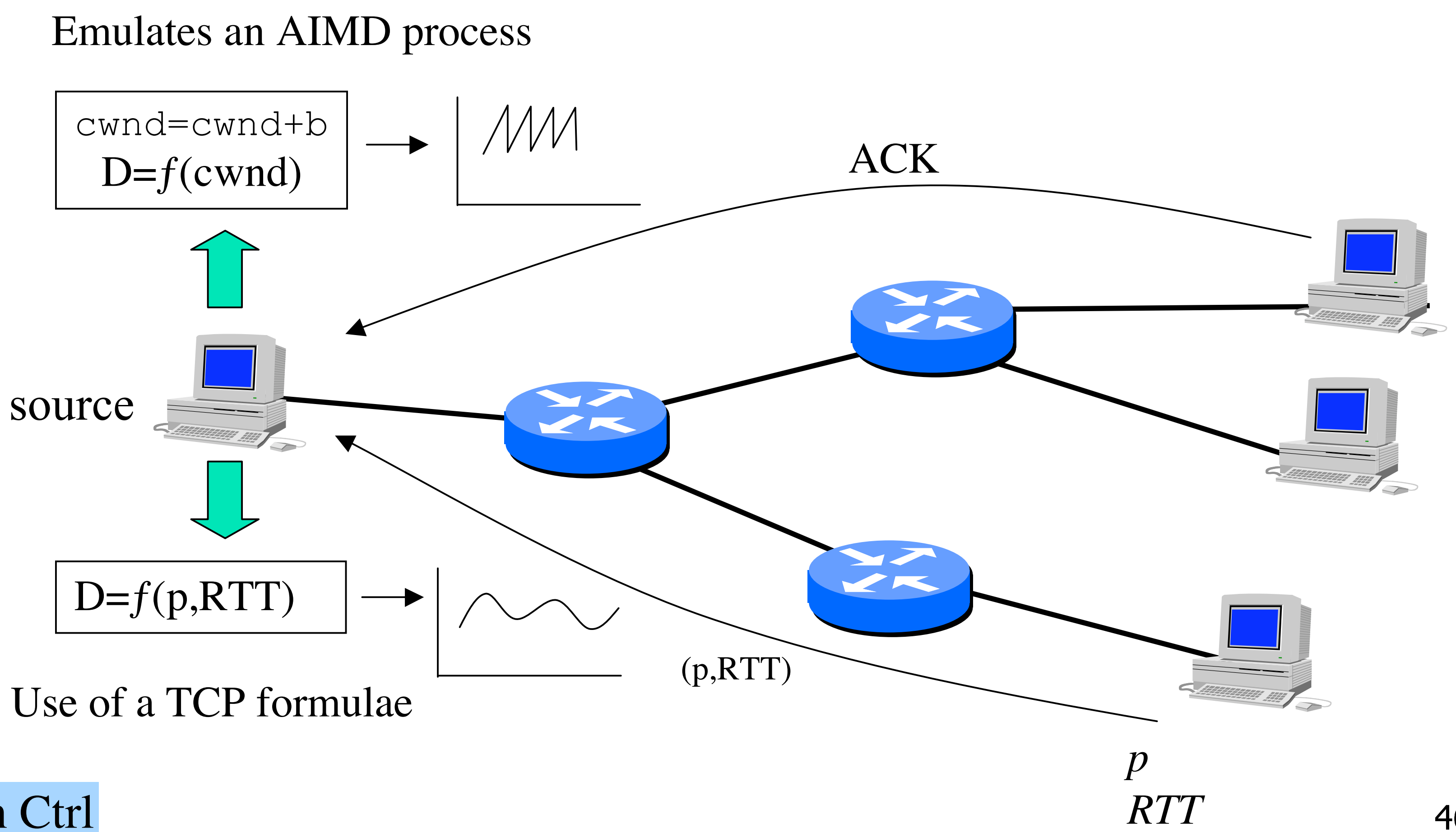
CC: single-rate, window-based

- MTCP: Multicast Transfert Control Protocol



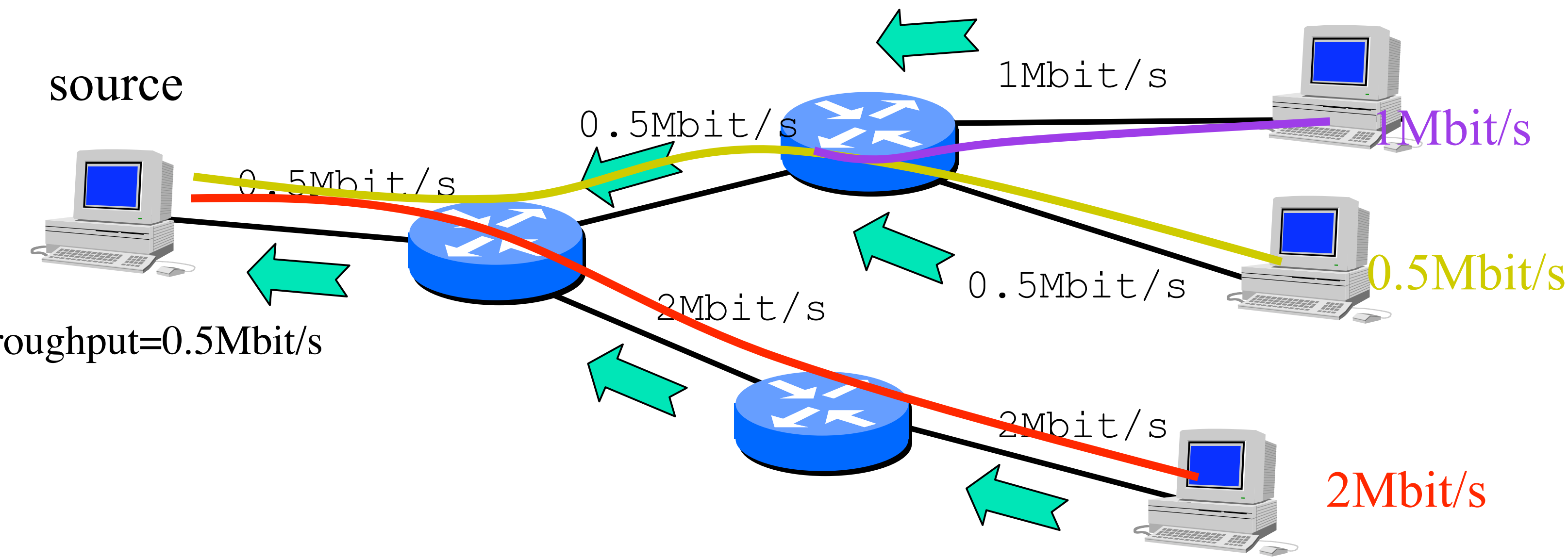
CC: single-rate, formulae-based

- TFMCC: TCP-Friendly Multicast Congestion Control



Multi-rate congestion control

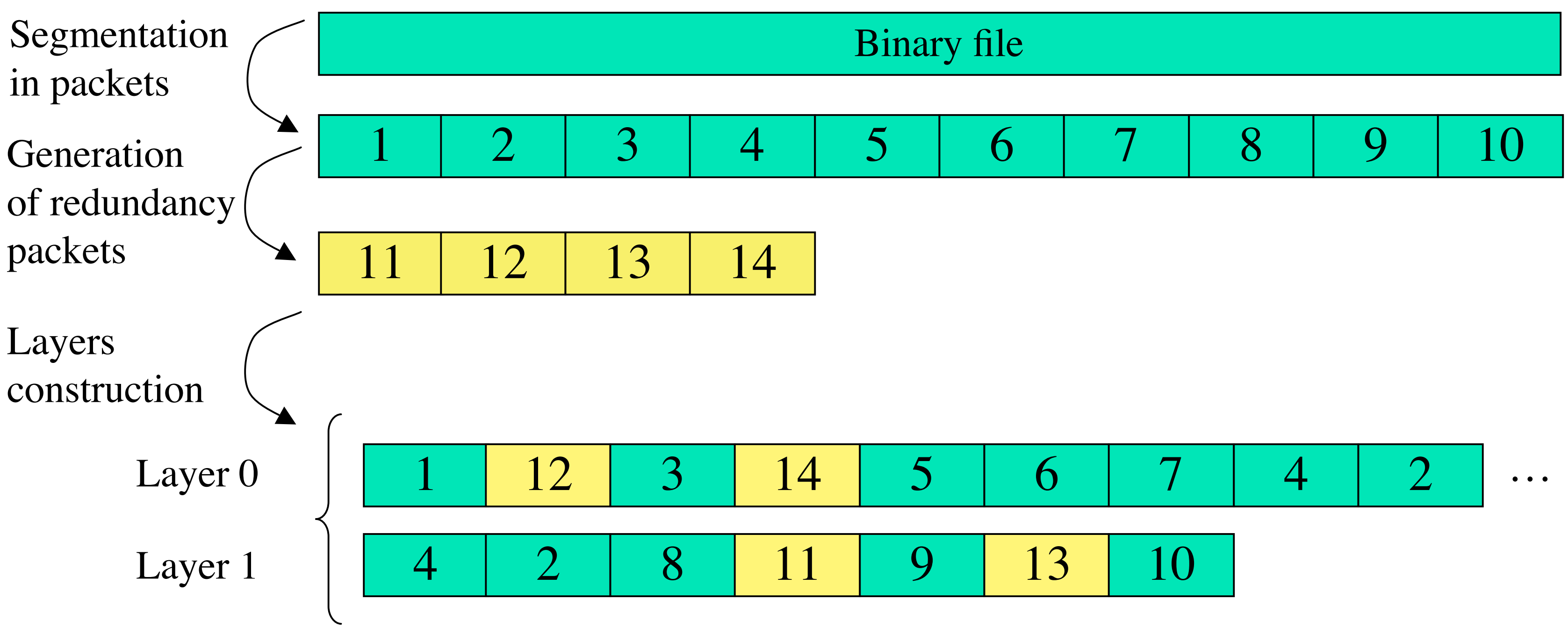
- Obviously more efficient: no need to keep with the slowest receiver



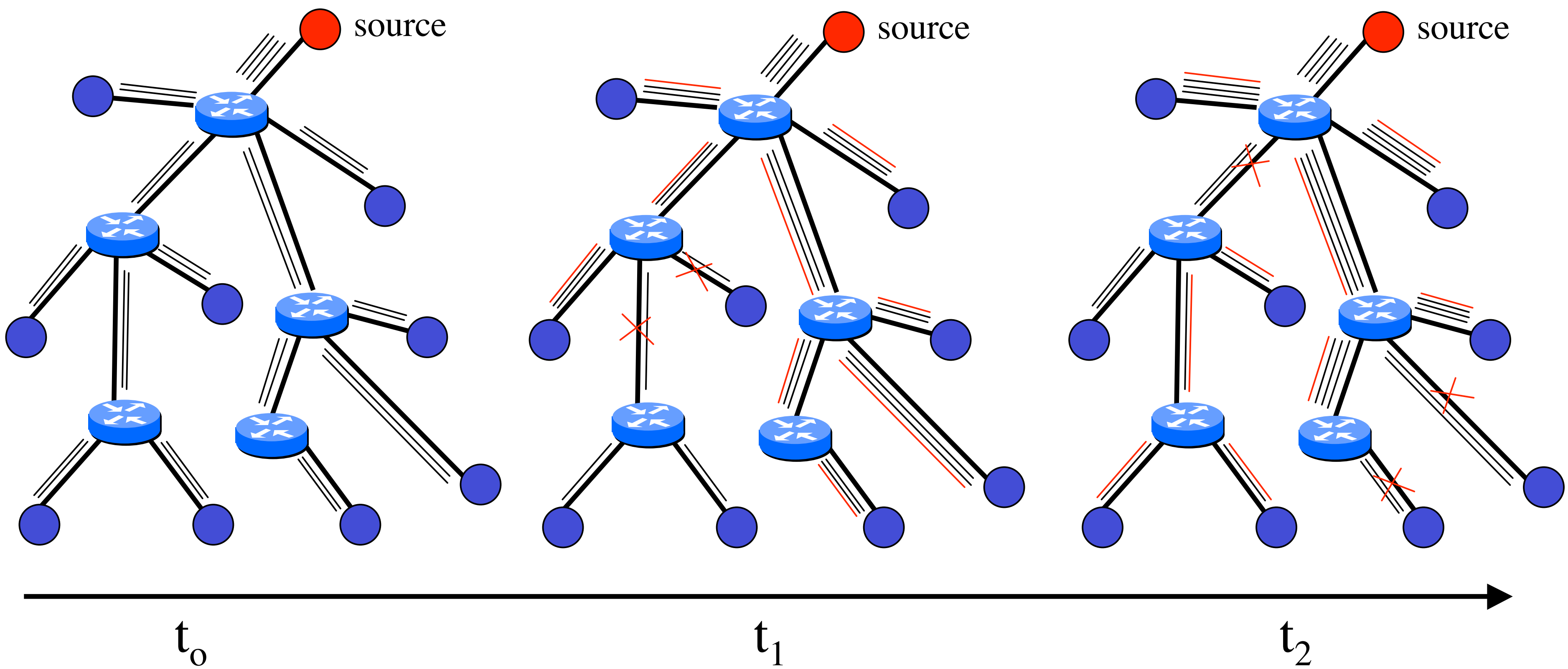
- Usually needs a layered encoding scheme

Principles of multi-layering

- 1 multicast group is assigned to 1 layer
- Throughput on each layer could be identical or increasing
- Subscription to a layer means subscription to a new group



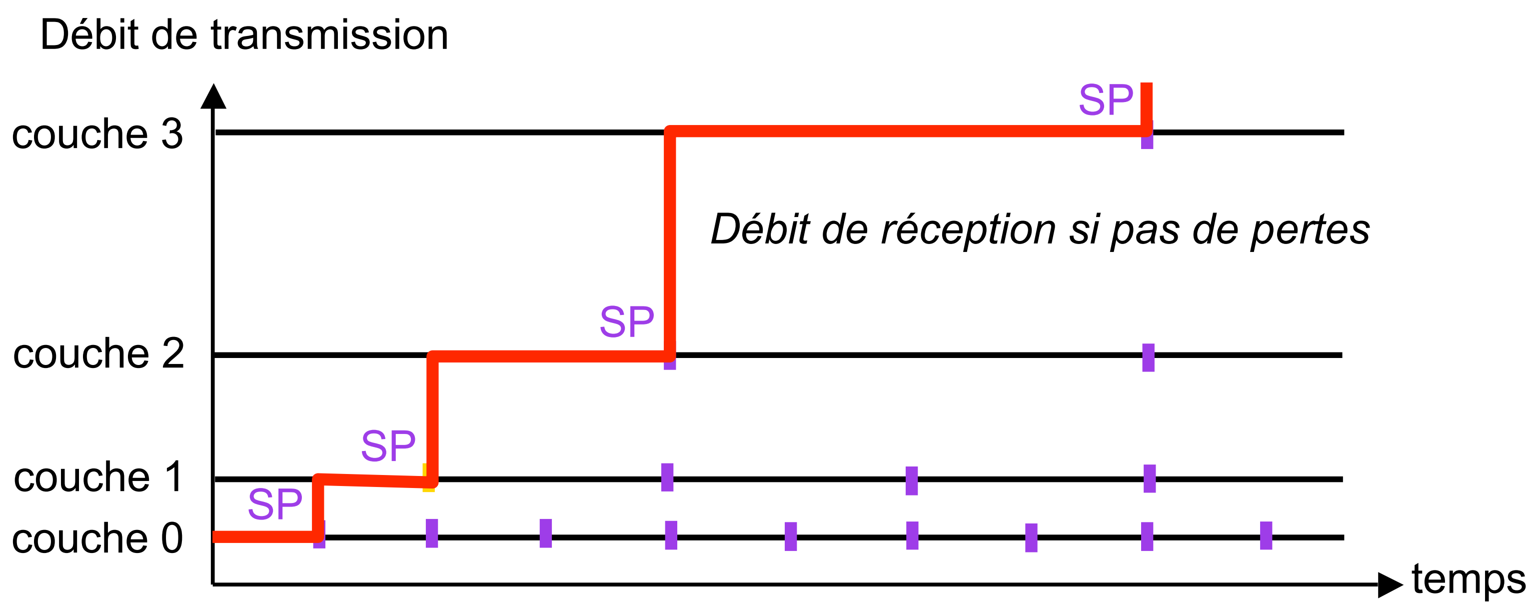
Example of layer operations



- Assuming that
  - Throughput in each layer is the same
  - There are a maximum of 4 layers

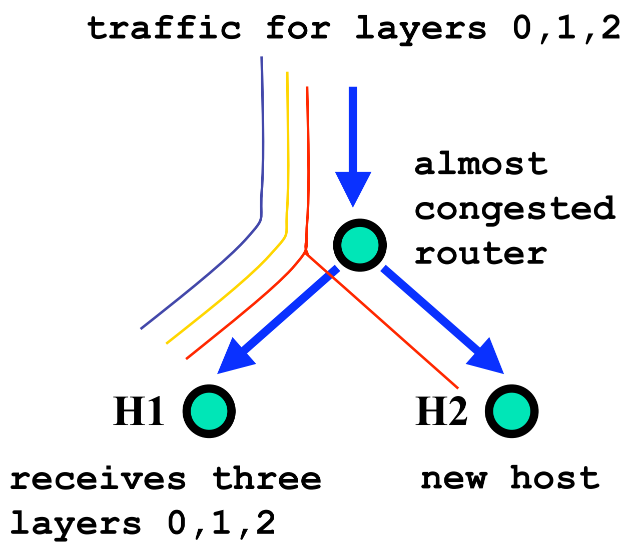
Synchronizing joins and leaves

- Layered approaches rely on fast joins and leaves from receivers
- More efficient if joins/leaves operations are synchronized



Example with RLC

- H1 adds L3 and H2 adds L1 (SP both on L0 and L2)
- router becomes congested ➔ losses
- H1 drops L3 and H2 drops L1
- no more losses
- H2 adds L1 (SP on L0)
- H2 adds L2 (SP on L1)
- H1 adds L3 and H2 adds L1 (SP on L2)



Part II
« The present »

Advanced group management

Advanced routing

Advanced reliability features

Multicast congestion control

IETF standards

ALC: Asynchronous Layered Coding

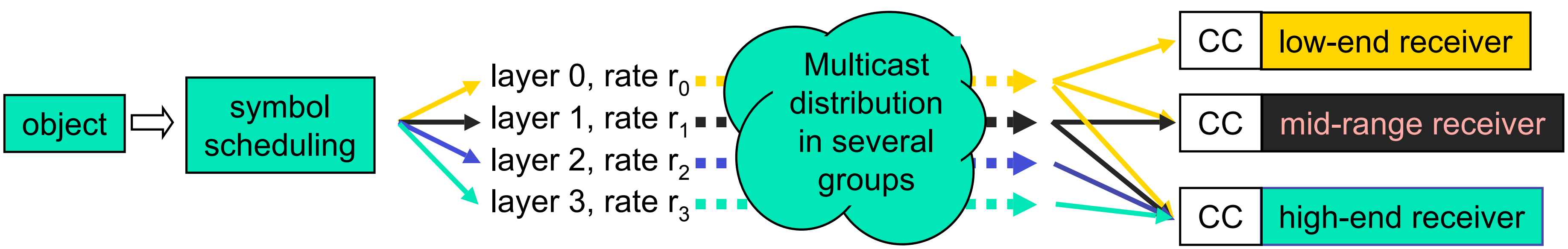
- ALC/LCT standard
  - one the two reliable multicast protocols being standardized at the RMT IETF working group
  - ***RFC 3450 up to RFC 3453***
  - offers unlimited scalability (no feedback)
  - supports receiver heterogeneity
  - supports ``push", ``on-demand" and ``streaming" delivery modes
  - suited to the distribution of popular content

ALC… (cont')

- Building blocks required by ALC
  - LCT (glue + header definition)
  - FEC (any FEC code)
  - layered congestion control (e.g. RLC)
  - security (e.g. TESLA authentication)

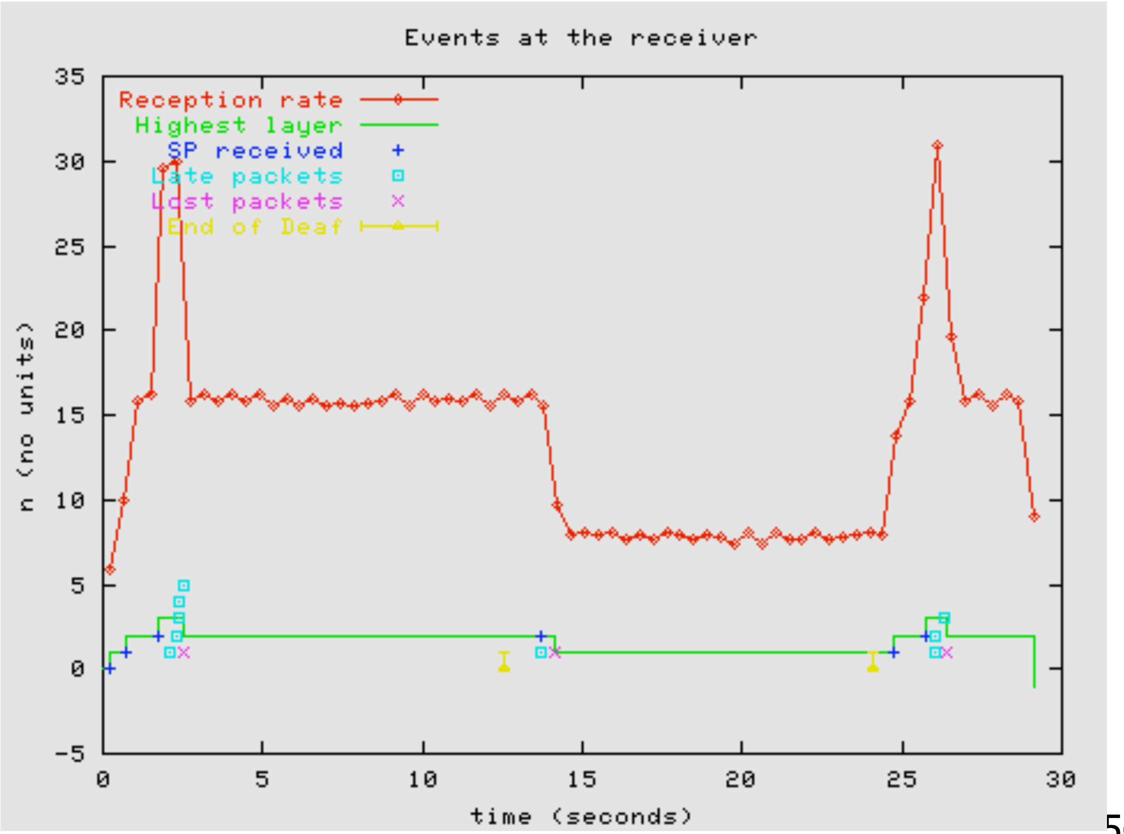
ALC… (cont')

- How does it work?
  - multi-rate transmissions, over several multicast groups, one per layer
  - the congestion control BB (e.g. RLC) tells a receiver when to add or drop a layer



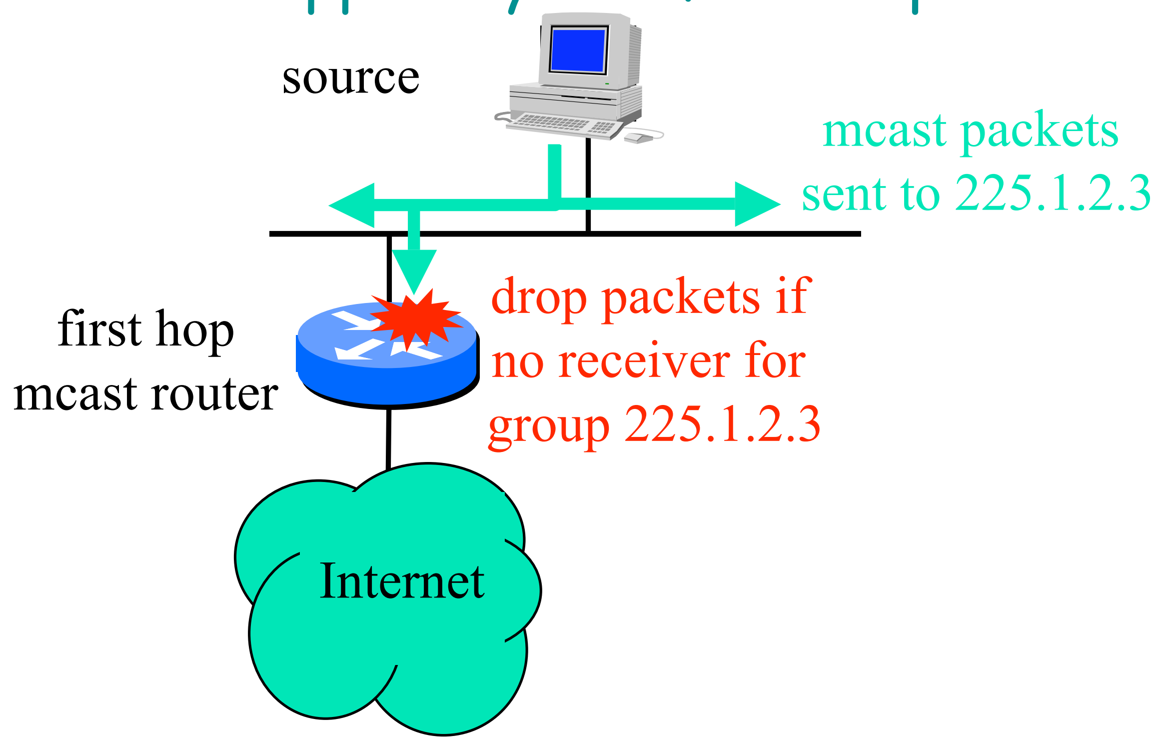
ALC… (cont')

- number of layers received is dynamic
  - depends on losses experienced
  - symbol scheduling must take it into account!
- example



ALC… (cont')

- How does it work… (cont')
  - sending to a multicast group with no receiver attached is not a problem…
  - packets are dropped by the first hop router !



The ALC PI... (cont')

- How does it work... (cont')
  - mix **randomly** all the data+FEC packets and send them on the various layers
  - required to counter the random losses and random layer addition/removal
  - other more intelligent organizations are possible (and can avoid duplications) but only work in an ideal world... (e.g. a LAN)
    - in practice losses, layer dynamic, layer de-synchronization lead to catastrophic performances...

The ALC PI... (cont')

- a transmission approach completely different from NORM/TRACK
  - file transmission with NORM/TRACK

source recvs:								NAK(2)						NAK(4)				
source sends:	0	1	2	3	4	5	6	FEC1	7	8	9	10	11	FEC2	12	13	14	END

  - file transmission with ALC (just an example!)

	Layer 3	F3	F12	F0	F1	F4	F11	F6	F5	F14	F7	F8	F2	F9	F10	F13	END
	Layer 2	2	4	10	8	5	9	11	14	7	3	0	12	1	6	13	END
	Layer 1	F12	F9	F2	F1	F10	F7	F6	F4	F13	F3	F5	F11	F14	F0	F8	END
source sends:	Layer 0	11	2	4	9	0	13	10	7	8	1	3	14	5	12	6	END

time →

What is ALC really good at?

- On-demand delivery mode
  - **yes**, this is the only RM solution supporting it!
- Streaming delivery mode
  - **yes**, partial reliability is possible too
- Push delivery mode
  - **no** for the general case, **yes** when there is no (or a very small) feedback channel (e.g. satellite)
- Scalability
  - **yes**, this is the only RM solution supporting it
- Heterogeneity
  - **yes**, this is the only RM solution supporting it
- Robustness
  - **yes**, reception can be stopped and restarted several times without any problem
  - a source is never impacted by the receiver behavior, neither are other receivers (in general)

ALC implementations

- Slides on ALC are from Vincent Roca (INRIA PLANETE)
- See Vincent Roca's web page on MCL
  - http://www.inrialpes.fr/planete/people/roca/mcl/mcl.html
  - MCL includes NORM and ALC

Conclusions on the « present »

- Standardization efforts
- Group management & routing
  - More security
  - Simpler communication models
- Reliability & congestion
  - Concerns for scalability and fairness