Rethinking the Service Model: Scaling Ethernet to a Million Nodes

Andy Myers acm@cs.cmu.edu

T. S. Eugene Ng eugeneng@cs.rice.edu

Hui Zhang hzhang@cs.cmu.edu

Vision: More Ethernet Switches Fewer IP Routers

- Today's world: IP routers + Ethernet PHY
  - Ethernet is the dominant PHY layer
  - Large number of IP routers connecting small Ethernet networks
    - E.g. CMU campus networks
- More Ethernet switches/fewer IP routers ➔ large Ethernet networks
  - Enterprise/campus networks
  - Broadband access networks
  - Data center networks

Why Large Ethernet Networks?

- Ethernet switches
  - simple, cheap, fast
  - Last fully automatic network
    - No host configuration
    - No switch configuration
  - Seamless mobility
  - Should be used to connect in the same network
- IP routers
  - Complex, expensive
  - Should be left to connect different networks

Why Not? Reasons Listed In Textbooks

- Flat addressing doesn't scale
- Need to link different L2's
- Spanning tree
  - No multi-path
  - Slow fail-over
- Broadcast overhead

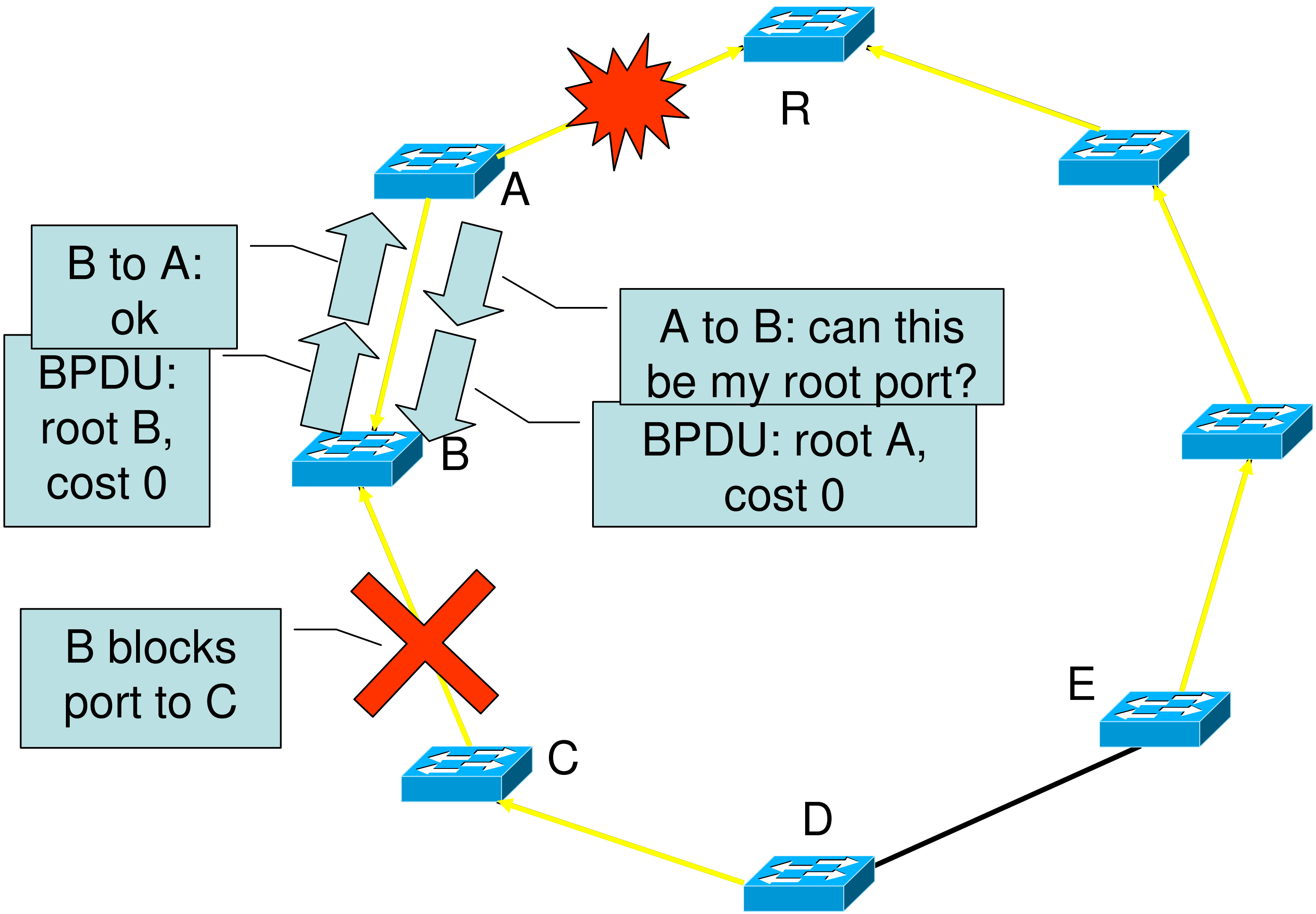
Current Reality

- Flat addressing doesn't scale
  - Bridges with 500K-1M MAC capacity ship today
- Need to link different L2's
  - Ethernet is the only L2 left
- Spanning tree
  - ??
- Broadcast overhead
  - ??

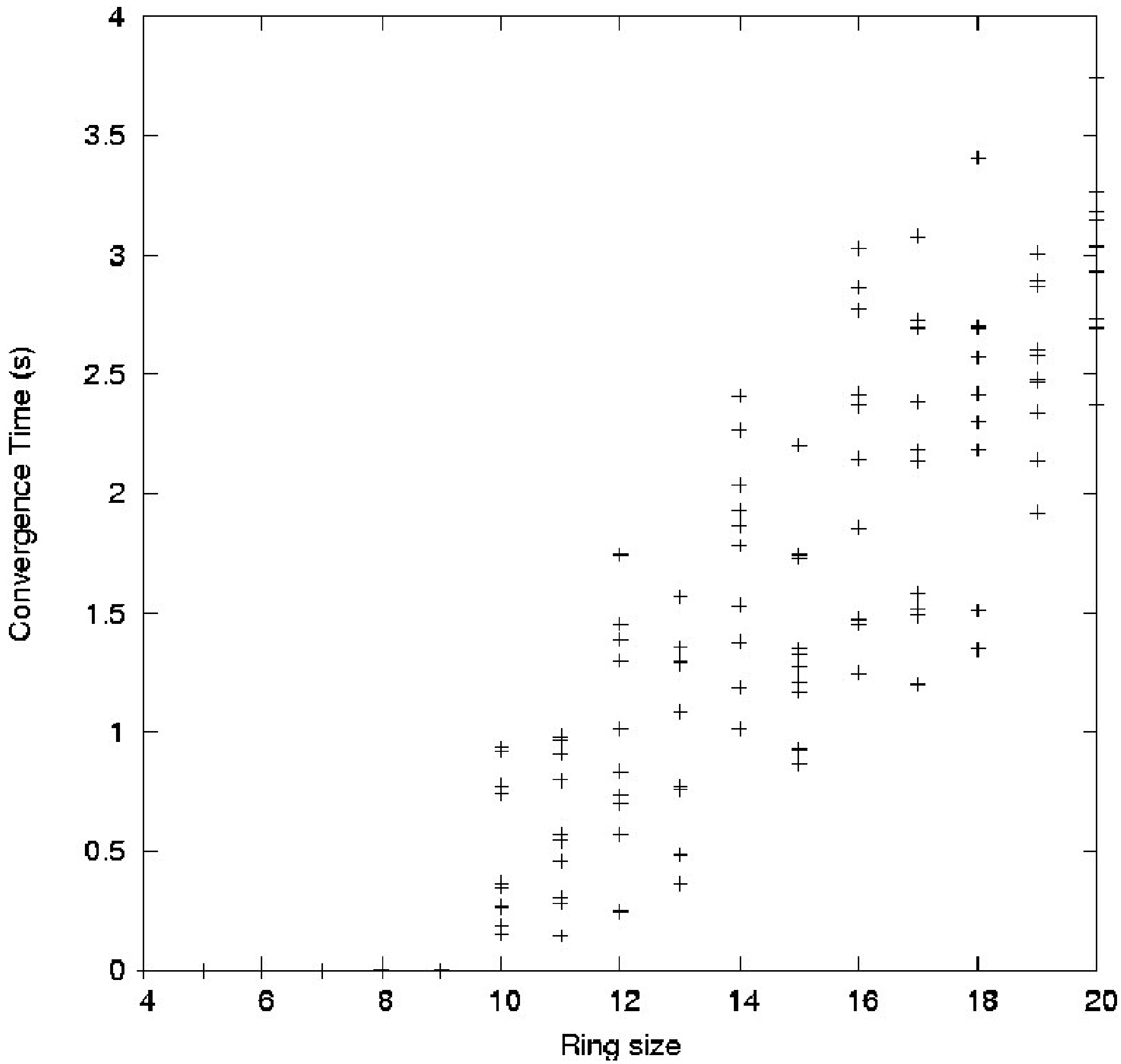
Outline

- Study Ethernet's flaws
  - Spanning Tree
  - Broadcast
- Identify the root cause
  - Broadcast service model
- Propose a solution
  - Turn off broadcast
  - Replace Ethernet's control plane

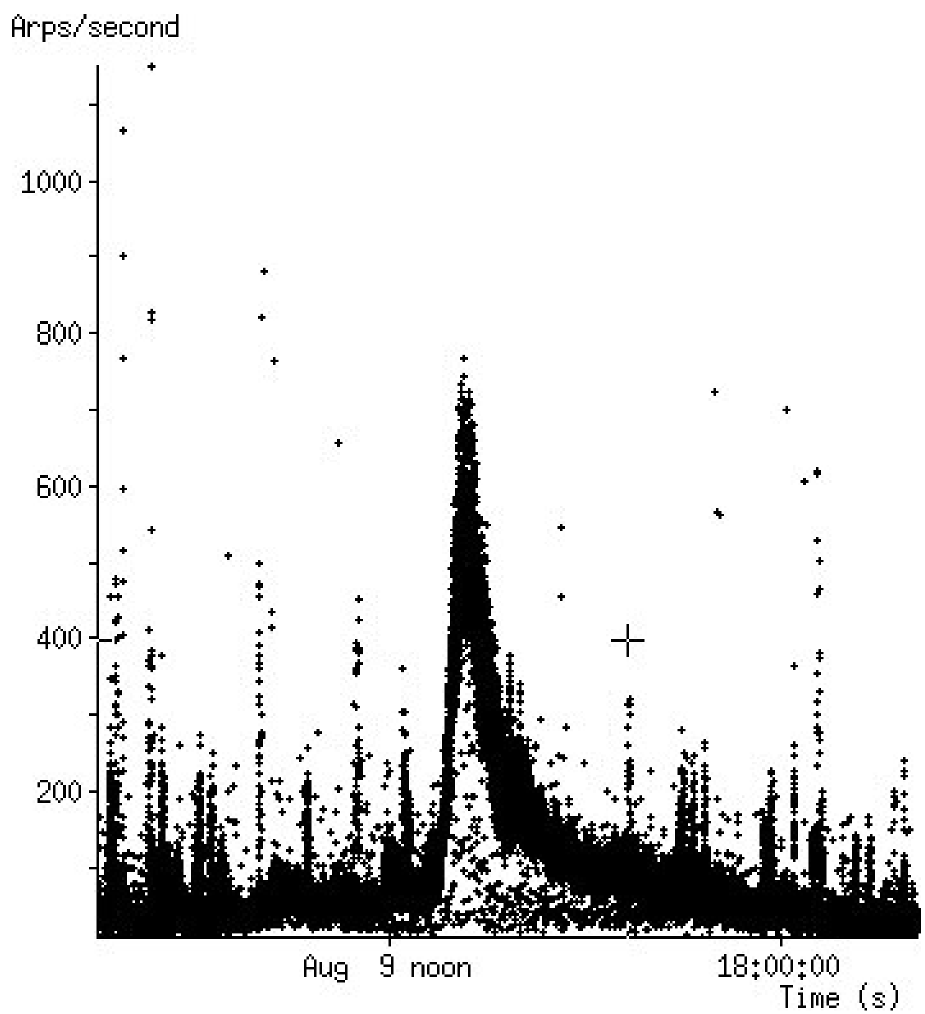
RSTP


R
A
B to A: ok
BPDU: root B, cost 0
A to B: can this be my root port?
BPDU: root A, cost 0
B
B blocks port to C
C
D
E

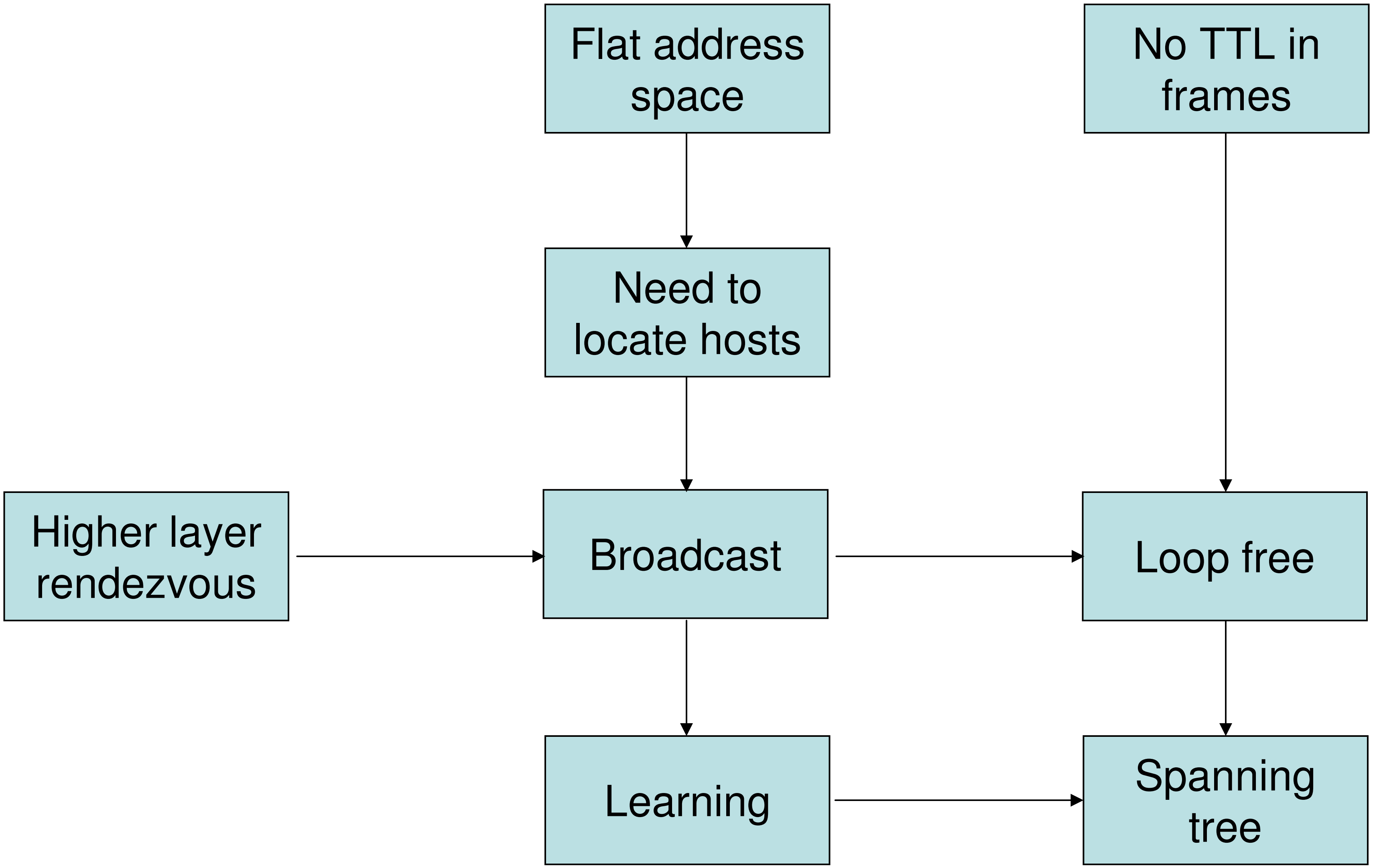
RSTP Convergence (ring)



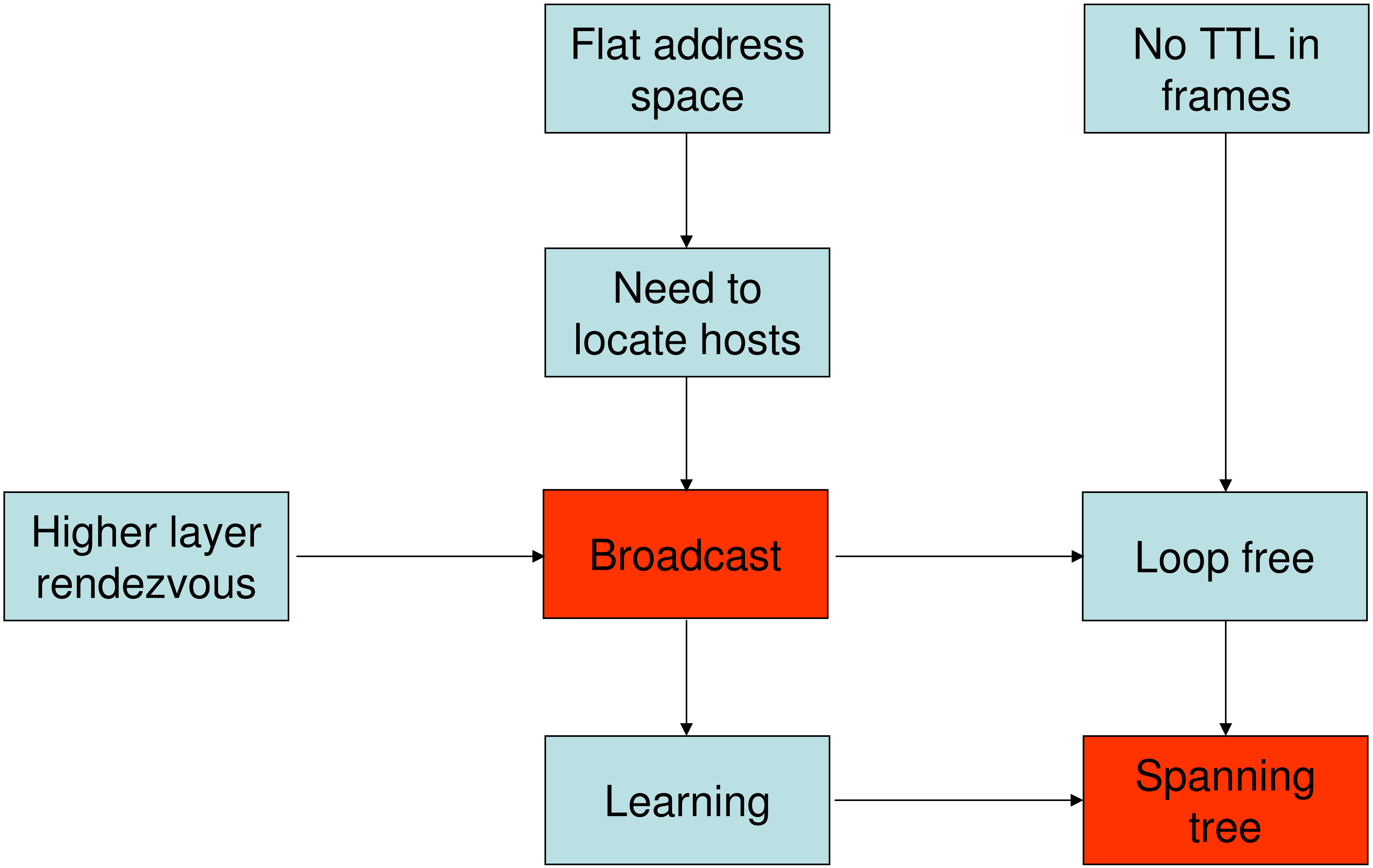
Broadcast (ARP)



Ethernet's Features



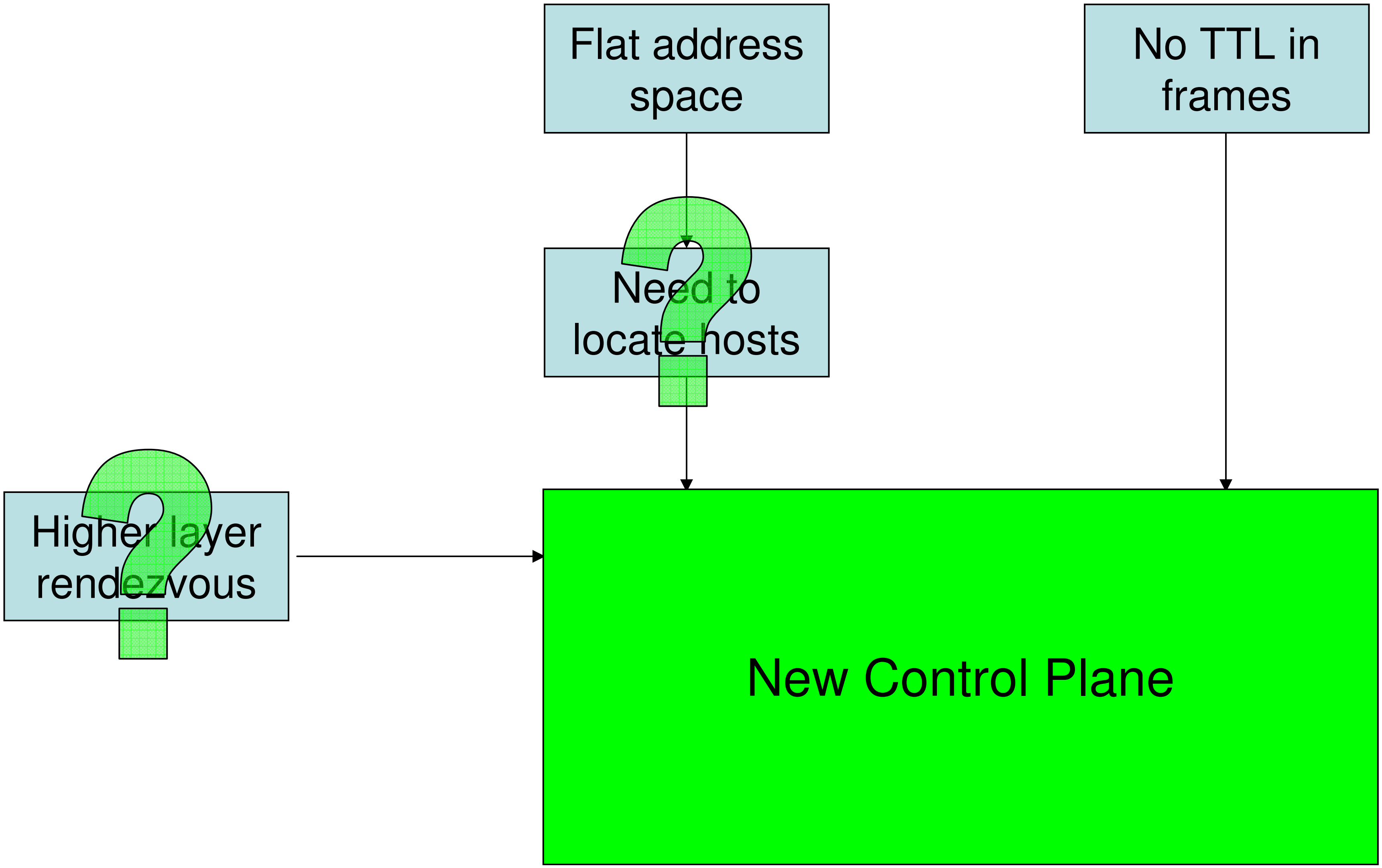
Ethernet's Features



Breaking the Broadcast/RSTP Dependency

- Change the service model: Turn off broadcast
  – Eliminates security risk
  – Improves scalability
  – Removes exponential packet copying
- Can eliminate RSTP
  – Unicast packets may loop, but no blowup
  – Network doesn't overload during transient loops

Fixing Ethernet



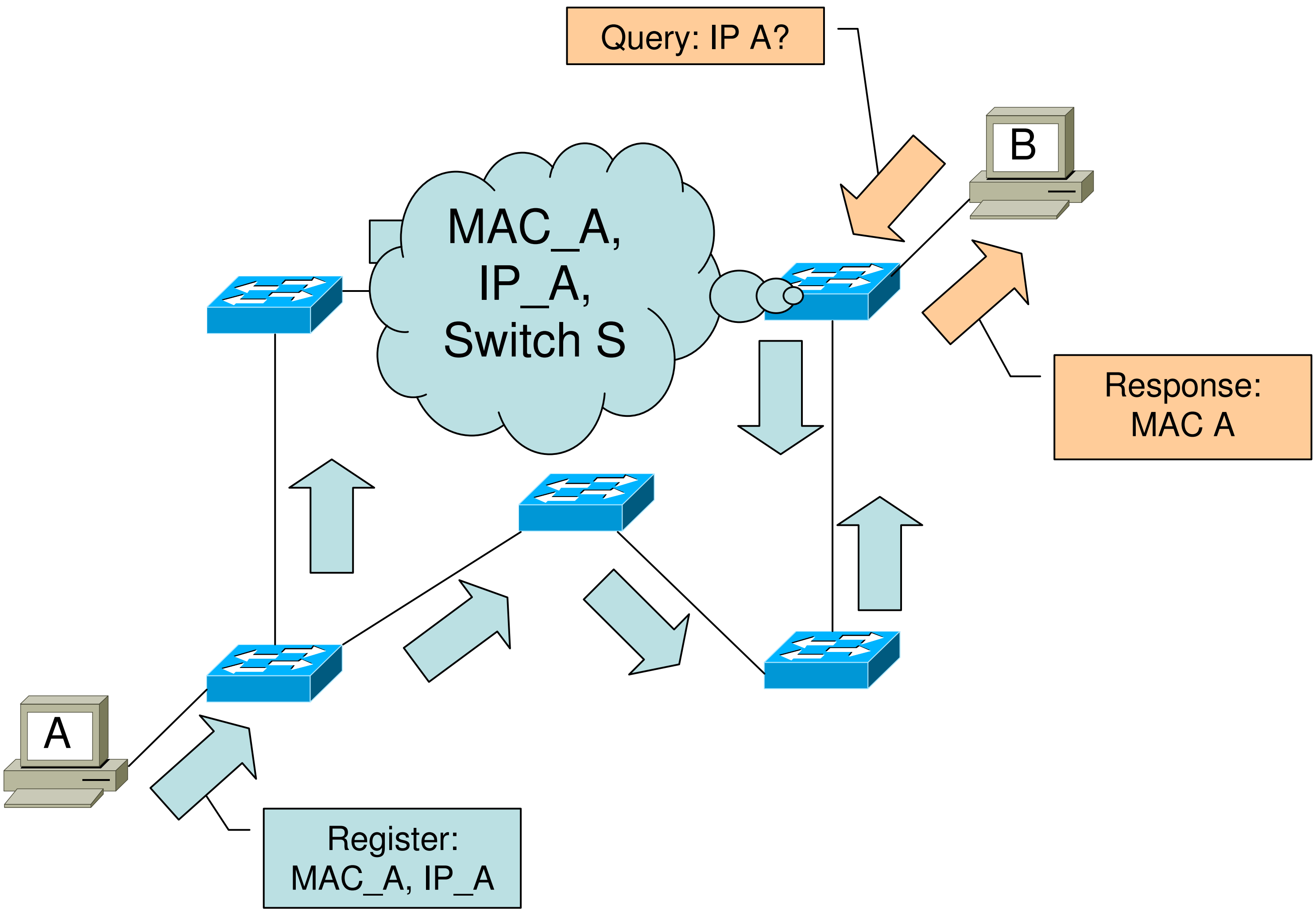
Why Replace the Control Plane?

- Fix what's broken
- Enable extensibility
  - Faster convergence (MAN)
  - Traffic engineering (SAN)
  - Isolation (Access net)
- Two control planes to consider
  - Fully distributed
  - Thin control plane

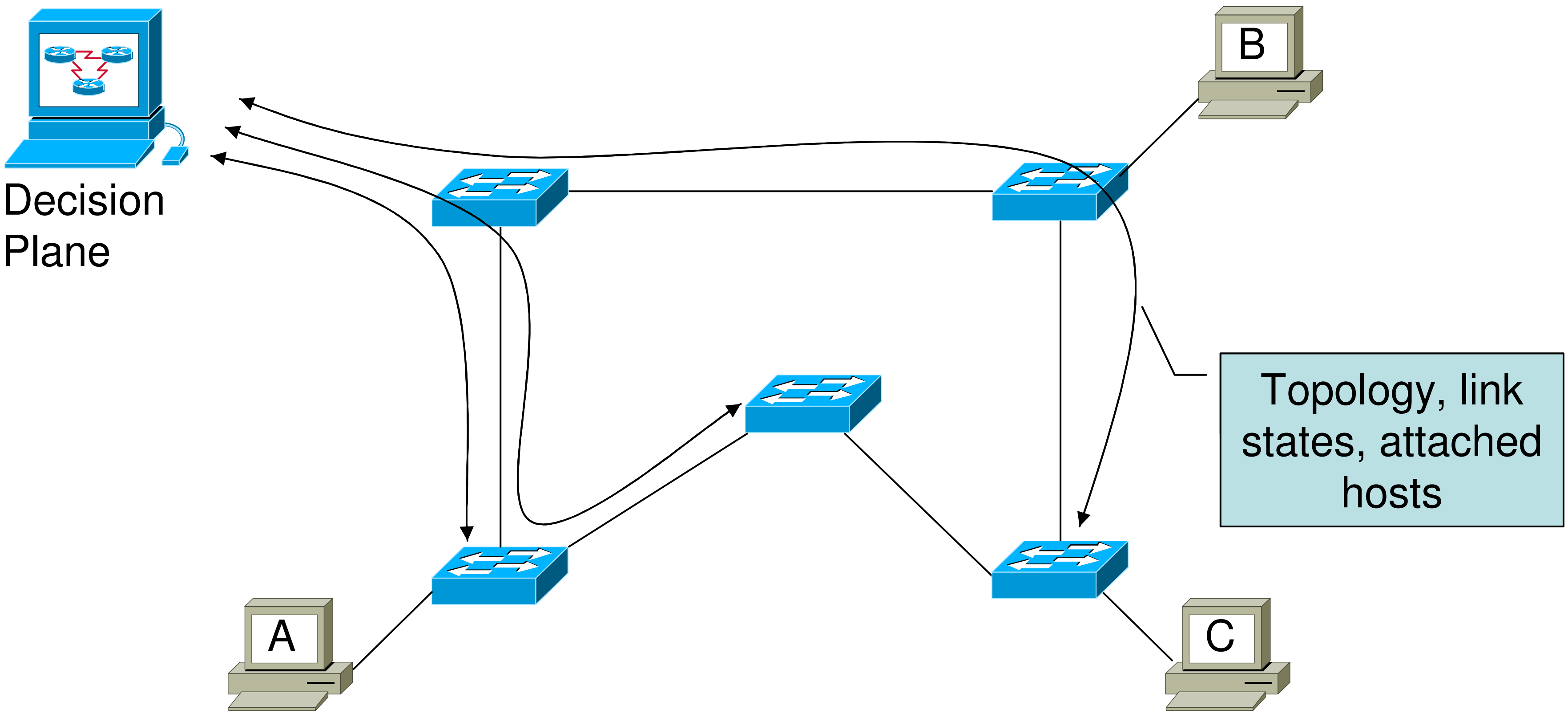
Fully Distributed Control Plane

- Link state computation of forwarding paths
  - Fast convergence
  - Multiple paths, not just a spanning tree
- Distributed directory replicated at all bridges
  - Provides IP to MAC mapping
  - Also used for service location
- Hosts register with local switch

Distributed Directory Example



Thin Control Plane



Thin Control Plane Advantages

- Switches remain simple
- Decisions made with global view of network
  - Multi-path forwarding
  - Directory service
- Can introduce new services
  - Traffic engineering
  - Pre-planned failure response

Related Work

- Control plane
  - OSI's CLNP/ESIS
  - Rexford04's Thin Control Plane
- Multi-path forwarding with [R]STP
  - SmartBridge00, STAR02, Pellegrini04, Viking04
- Replacing spanning tree with link state
  - Garcia03 ("LSOM")
  - Perlman04 ("RBridges")
    - Adds header with TTL for links between bridges
    - No host registration needed

Summary

- Vision: More switches, fewer routers
  - Ethernet switches are cheaper, less complex than IP routers
  - Leads to larger Ethernet networks
  - Many potential application scenarios
- To realize
  - Eliminate broadcast
  - New control plane to enable practical L2