Floodless in SEATTLE:
A Scalable Ethernet Architecture for Large Enterprises

Full paper available at
http://www.cs.princeton.edu/~chkim

Changhoon Kim, Matthew Caesar,
and Jennifer Rexford

Outline of Today's Lecture

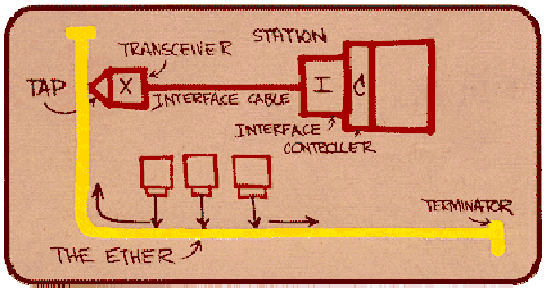
- Review Ethernet bridging
- New challenges to Ethernet
  - Control-plane scalability
  - Data-plane efficiency
- SEATTLE as a solution

2

Quick Review of Ethernet

Ethernet

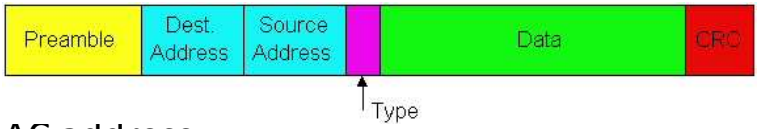
- Dominant wired LAN technology
  - Covers the first IP-hop in most enterprises/campuses
- First widely used LAN technology
- Simpler, cheaper than token LANs, ATM, and IP
- Kept up with speed race: 10 Mbps – 10+ Gbps



Metcalfe's Ethernet sketch

4

Ethernet Frame Structure

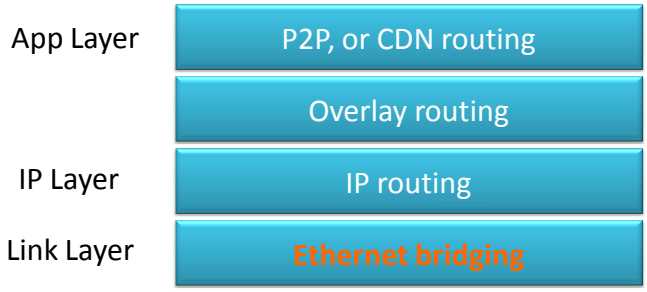


- MAC address
  - Flat, globally unique, and permanent 48-bit value
  - Adaptor passes frame to network-level protocol
    - If destination address matches the adaptor
    - Or the destination address is the broadcast address
  - Otherwise, adapter discards frame
- Type: indicates the higher layer protocol
  - Usually IP

5

Ethernet Bridging: Routing at L2

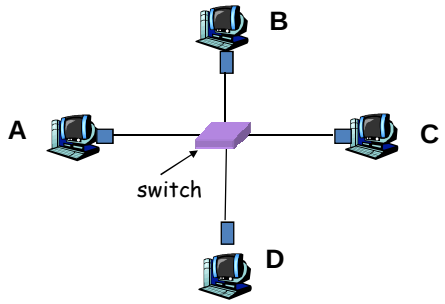
- Routing determines paths to destinations through which traffic is forwarded
- Routing takes place at any layer (including L2) where devices are reachable across multiple hops



6

Ethernet Bridges Self-learn Host Info

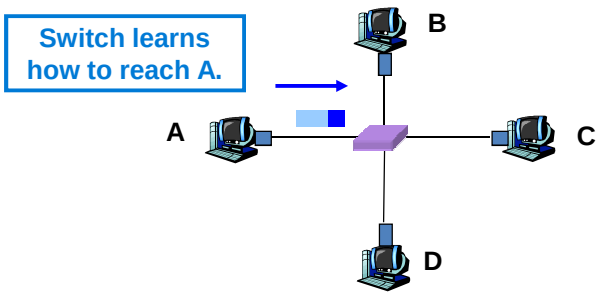
- Bridges (switches) forward frames selectively
  - Forward frames only on segments that need them
- Switch table
  - Maps destination MAC address to outgoing interface
  - Goal: construct the switch table automatically



7

Self Learning: Building the Table

- When a frame arrives
  - Inspect the *source* MAC address
  - Associate the address with the *incoming* interface
  - Store the mapping in the switch table
  - Use a timeout to eventually forget the mapping

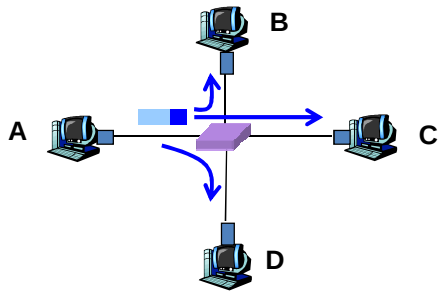


8

Self Learning: Handling Misses

- Floods when frame arrives with unfamiliar destination or broadcast address
  - Forward the frame out all of the interfaces
  - … except for the one where the frame arrived
  - Hopefully, this case won't happen very often

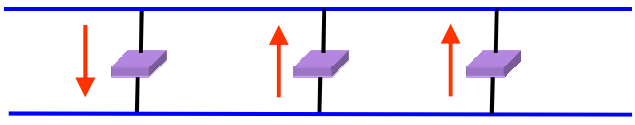
When in doubt, shout!



9

Flooding Can Lead to Loops

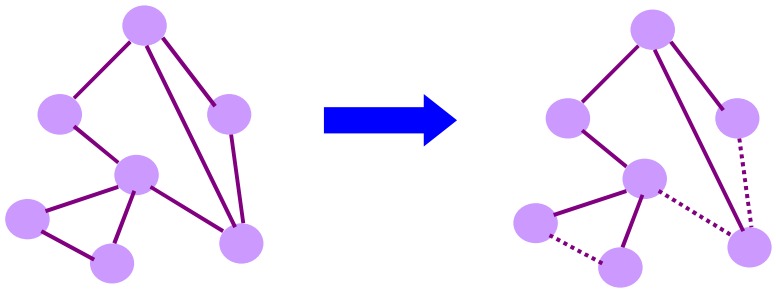
- Flooding can lead to forwarding loops, confuse bridges, and even collapse the entire network
  - E.g., if the network contains a cycle of switches
  - Either accidentally, or by design for higher reliability



10

Solution: Spanning Trees

- Ensure the topology has no loops
  - Avoid using some of the links when flooding
- Spanning tree
  - Sub-graph that covers all vertices but contains no cycles
  - Links not in the spanning tree do not forward frames



11

Interaction with the Upper Layer (IP)

- Bootstrapping end hosts by automating host configuration
  - DHCP (Dynamic Host Configuration Protocol)
  - Broadcast DHCP discovery and request messages
- Bootstrapping each conversation by enabling resolution from IP to MAC addr
  - ARP (Address Resolution Protocol)
  - Broadcast ARP requests
- Both work via Ethernet-layer broadcasting

12

Broadcast Domain and IP Subnet

- Ethernet broadcast domain
  - A group of hosts and switches to which the same broadcast or flooded frame is delivered
  - Broadcast domain != Collision domain
- Broadcast domain == IP subnet
  - Uses ARP to reach other hosts in the same subnet
  - Uses default GW to reach hosts in different subnets
- Too large a broadcast domain leads to
  - Excessive flooding and broadcasting overhead
  - Insufficient security/performance isolation

13

New Challenges, and SEATTLE as a solution

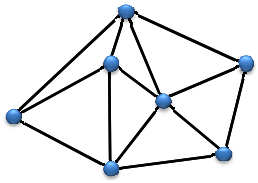
Ethernet in Enterprise Nets?

- Ethernet has substantial benefits
  - Simplifies network management, greatly reducing operational expense
  - Naturally supports host mobility
  - Enhances network flexibility
- Why do we still use IP routing inside a single network?

15

Ethernet Doesn't Scale!

- Reasons for poor scalability
  - Network-wide flooding
  - Frequent broadcasting
  - Unbalanced link utilization, low availability and throughput due to tree-based forwarding

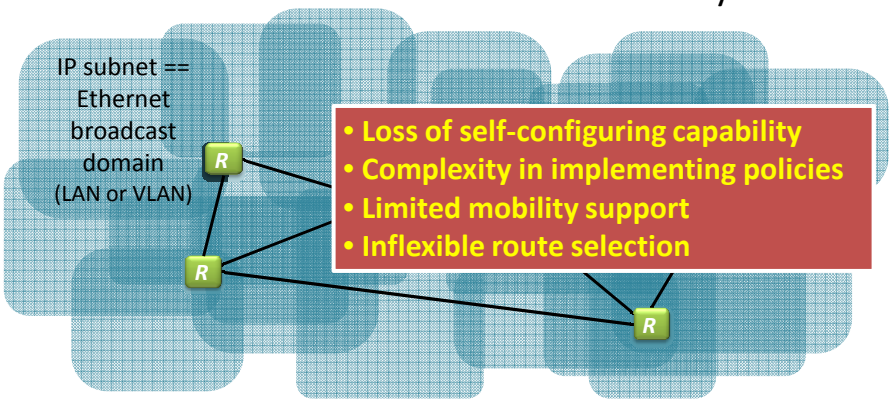


- Limitations quickly growing with network size
- Scalability requirement is growing very fast
  - 50K ~ 1M hosts

16

Current Practice

A hybrid architecture comprised of several small Ethernet-based IP subnets interconnected by routers



Sacrifices Ethernet's simplicity and IP's efficiency only for scalability

17

Key Question and Contribution

- Can we maintain the same properties as Ethernet, yet **scales** to large networks?
- SEATTLE: The best of IP and Ethernet
  - Two orders of magnitude more scalable than Ethernet
  - Broadcast domains in **any size**
  - Vastly simpler network management, with host mobility and network flexibility
  - Shortest path forwarding

18

Objectives and Solutions

Objective	Approach	Solution
1. Avoiding flooding	Never broadcast unicast traffic	Network-layer one-hop DHT
2. Restraining broadcasting	Bootstrap hosts via unicast	
3. Reducing routing state	Populate host info only when and where it is needed	Traffic-driven resolution with caching
4. Shortest-path forwarding	Allow switches to learn topology	L2 link-state routing maintaining only switch-level topology

*** Meanwhile, avoid modifying end hosts**

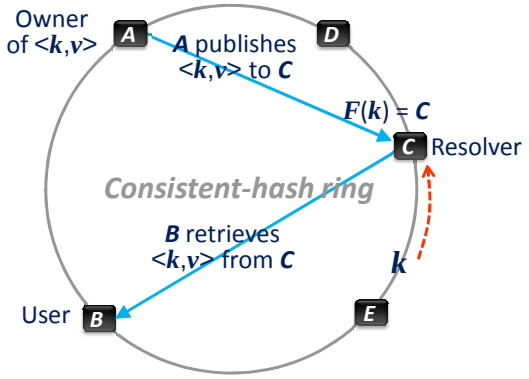
19

Network-layer One-hop DHT

- Switches maintain <***key, value***> pairs by commonly using a hash function F
  - F: Consistent hash mapping a key to a switch
  - F is defined over the live set of switches
  - LS routing ensures each switch knows about all the other live switches, enabling one-hop DHT operations

20

One-hop DHT Details and Benefits

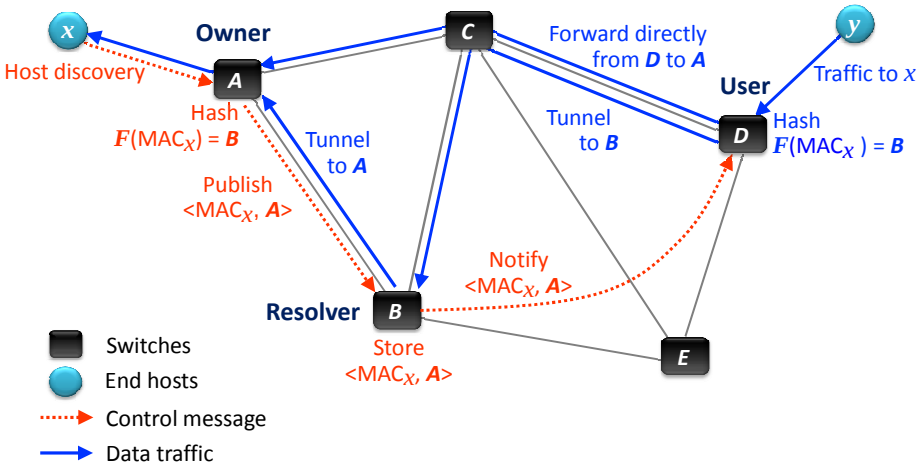


- Benefits
  - Fast and efficient reaction to changes
  - Reliability and capacity naturally growing with network size

21

Location Resolution

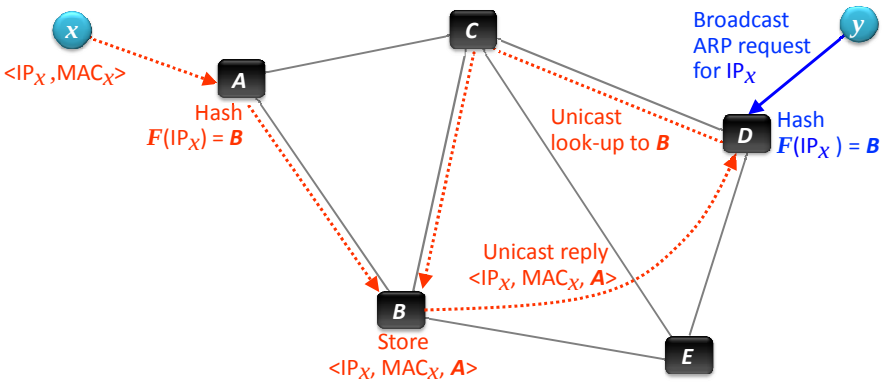
<key, val> = <MAC addr, location>



22

Address Resolution

<key, val> = <IP addr, MAC addr>



Traffic following ARP takes a shortest path without separate location resolution

23

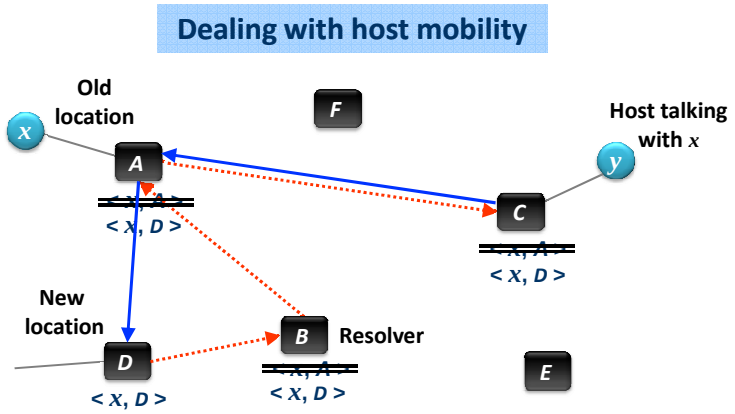
Handling Network Dynamics

- Events not modifying the set of live switches
  - E.g., most link failure/recovery
  - LS routing simply finds new shortest paths
- Events modifying the live set of switches
  - E.g., switch failure/recovery
  - F works differently after a change
  - Two simple operations ensure correctness
    - If $F_{new}(k)$!= $F_{old}(k)$, owner re-publishes to $F_{new}(k)$
    - Remove any <k,v> published by non-existing owners

24

Handling Host Dynamics

- Host location, MAC-addr, or IP-addr can change



MAC- or IP-address change can be handled similarly

Ensuring Ethernet Compatibility

- Scalable host bootstrapping using the DHT
  - Hashes a pre-determined string (e.g., "*DHCP_SERVER*") to resolve DHCP server
- Group: Scalable and flexible alternative of VLAN
  - A "group" is a highly scalable location-independent broadcast domain
  - Resolver controls inter-group access
  - Broadcast frames in each group are forwarded along a multicast tree

Further Enhancements

- **Goal**: Dealing with switch-level heterogeneity
- **Solution**: Virtual switches
- **Goal**: Attaining very high availability of resolution
- **Solution**: Replication via multiple hash functions
- **Goal**: Dividing administrative control to sub-units
- **Solution**: Multi-level one-hop DHT
  - Similar to OSPF areas
  - Contains local resolution within a region

27

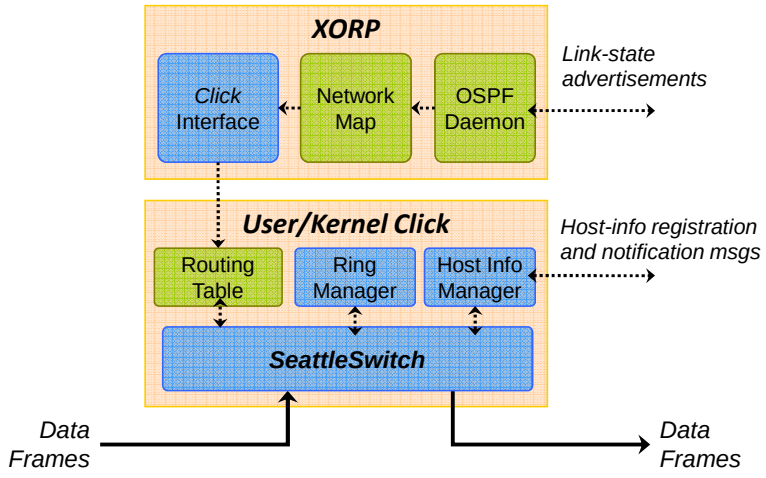
Performance Evaluation

- Large-scale packet-level simulation
  - Event-driven simulator optimized for control-plane evaluation
  - Synthetic traffic based on real traces from LBNL
    - Inflated the trace while preserving original properties
  - Real topologies from campus, data centers, and ISPs
- Emulation with prototype switches
  - Click/XORP implementation

28

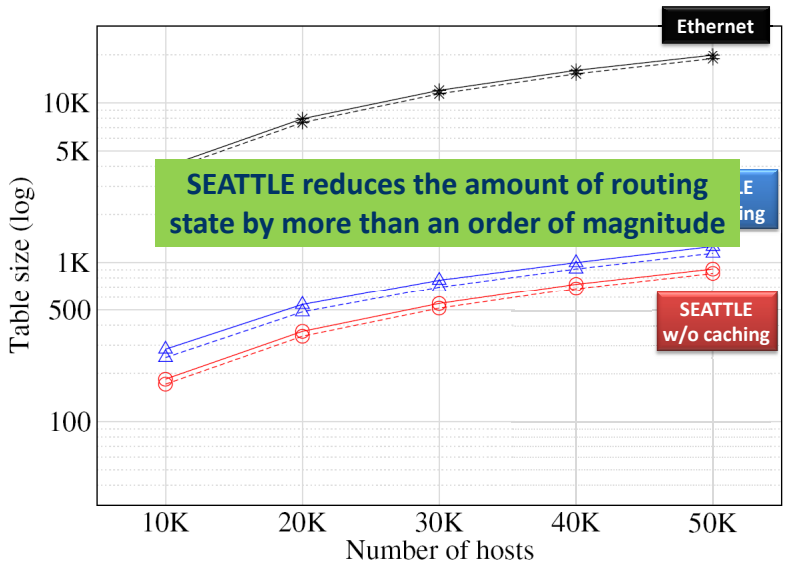
Prototype Implementation

- Link-state routing: ***XORP OSPFD***
- Host-info management and traffic forwarding: ***Click***



29

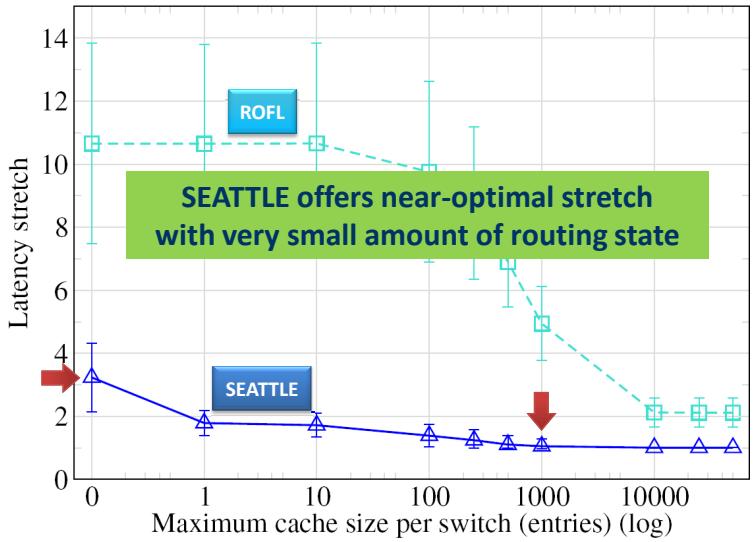
Amount of Routing State



30

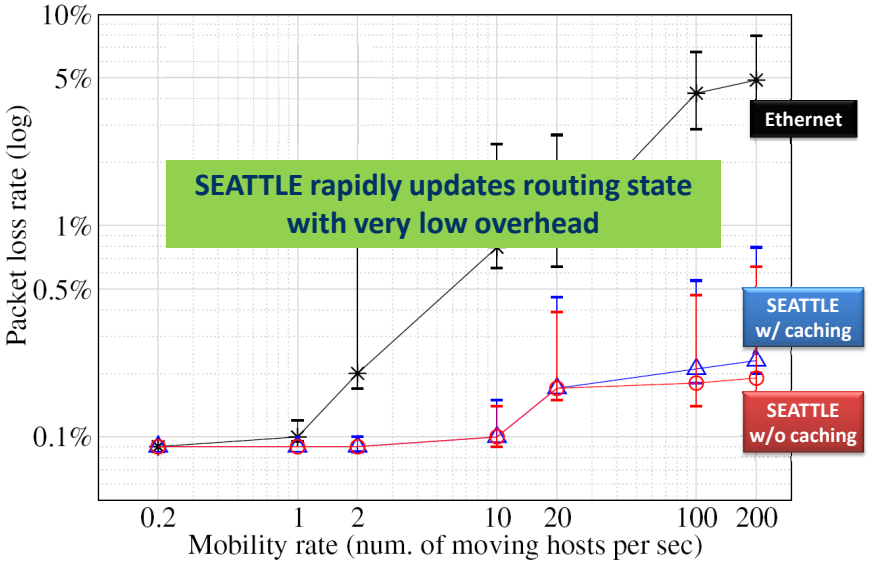
Cache Size vs. Stretch

Stretch = actual path length / shortest path length (in latency)



31

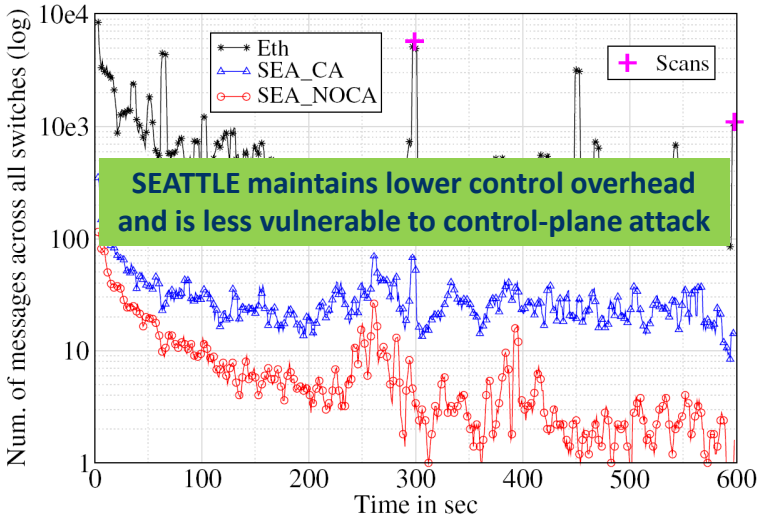
Sensitivity to Mobility



32

Control Overhead

Instrumentation results from Emulab test



33

Conclusion and Future Work

- SEATTLE is a plug-and-playable network architecture ensuring both scalability and efficiency
- Enabling design decisions
  - One-hop DHT tightly coupled with LS routing
  - Reactive location resolution and caching
  - Shortest-path forwarding
- Future work
  - Using SEATTLE to improve network security
  - Utilizing indirect delivery for load balancing
  - Optimizations when end hosts can be changed

34