# MOOSE: Addressing the Scalability of Ethernet

Malcolm Scott
Jon Crowcroft

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

## Ethernet

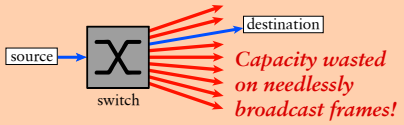
Continues to be used for new network deployments

- Simple, easy to implement, ubiquitous hardware
- Layer-2; support multiple layer-3 (IP) networks over single common infrastructure
- Convenient "lowest common denominator" protocol

But scalability issues are becoming increasingly significant

### *Heavy use of broadcast*

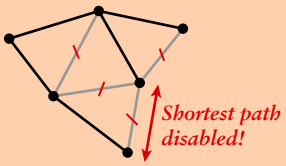
Broadcast ARP required for interaction with IP




*Capacity wasted on needlessly broadcast frames!*

On large networks, broadcast can overwhelm slower links e.g. wireless

### *Inefficient routing: Spanning Tree*



*Shortest path disabled!*

### *Switches' address tables*

| MAC address | Port |
|---|---|
| 01:23:45:67:89:ab | 12 |
| 00:a1:b2:c3:d4:e5 | 16 |
| … | … |

- Maintained by every switch
- Automatically learned
- Table capacity ~8000 addresses
- *Full table results in unreliability, or at best heavy flooding*

Underlying problem: the MAC address namespace is *unstructured*
(as far as switches are concerned: a MAC address gives no hint as to its location)

## The solution: MOOSE

*Multi-level Origin-Organised Scalable Ethernet*

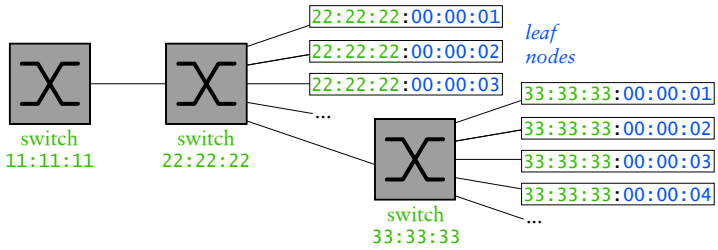
- Modified switch performs address rewriting on ingress
- Introduce two (or more) level hierarchy to MAC addresses:

`ss:ss:ss:nn:nn:nn`
*switch ID* *node ID*

Switch ID identifies the node's local switch
- *Configured by administrator, or negotiated automatically*
- *This may itself be hierarchical*

Node ID is allocated by the local switch

- New source address remains in frame when passed to destination

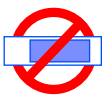



Now, switches need only store the locations of other switches
*Above, switch 11:11:11 only needs two address table entries!*

Completely compatible with unmodified standard Ethernet devices

## Previous solutions

*(and why they don't solve all the problems)*



*Most proposed solutions have so far relied on encapsulation of the Ethernet frame. This does not improve scalability.*

**Multiprotocol Label Switching (MPLS):** [1]
- Becoming popular throughout both the Internet and private LANs
- Label Edge Router must convert frames' destinations into a label ID
  Equivalent to Ethernet's address table problem

**Ethernet-in-Ethernet encapsulation:** [2]
- Edge switch must convert frames' destinations into another MAC address
  Again equivalent to Ethernet's address table problem

**Broadcastless Ethernet:** [3]
- Fundamental redesign of the protocol
- Perhaps a good idea, but it is incompatible with existing Ethernet devices
- (Also, the link-state protocol must again track every MAC address)

*The important difference:*

- MOOSE leaves the hierarchical address as the frame's source when delivering it to its destination
- Hosts only ever see other hosts' hierarchical addresses: therefore no destination address rewriting is ever needed
- *Subverts computers' ARP tables to look up hierarchical addresses!*

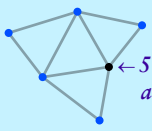
[1] Rosen, E., Viswanathan, A. and Callon, R., "Multiprotocol label switching architecture", RFC 3031, January 2001
[2] Hadžić, I, "Hierarchical MAC address space in public Ethernet networks", IEEE GLOBECOM '01, November 2001
[3] Myers, A., Ng, E. and Zhang, H., "Rethinking the service model: scaling Ethernet to a million nodes", ACM HotNets-III, November 2004

## Hierarchical MAC addresses

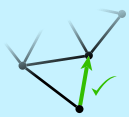
*and their many benefits*

### *Address tables*

- Not only do we reduce the table size to 1-5%
- But we crucially also make it *deterministic*



*← 5 table entries, always*

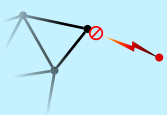


### *Best-path routing and resilience*

- Switches can participate in a routing protocol
- Many benefits
- This can now route prefixes, rather than individual addresses

### *Security and isolation*

- Source addresses rewritten ⇒ source spoofing ineffective
- Duplicate MAC addresses don't matter




### *Minimisation of broadcast traffic*

- No longer need any broadcast for node location
- Could use a distributed directory service on the switches for ARP
- Or could make the IP addressing hierarchy correspond to the MAC address hierarchy and dispense with ARP entirely



Email: Malcolm.Scott@cl.cam.ac.uk

# MOOSE: addressing the scalability of Ethernet

*Poster abstract*

Malcolm Scott
Malcolm.Scott@cl.cam.ac.uk
University of Cambridge
Computer Laboratory

Jon Crowcroft
Jon.Crowcroft@cl.cam.ac.uk
University of Cambridge
Computer Laboratory

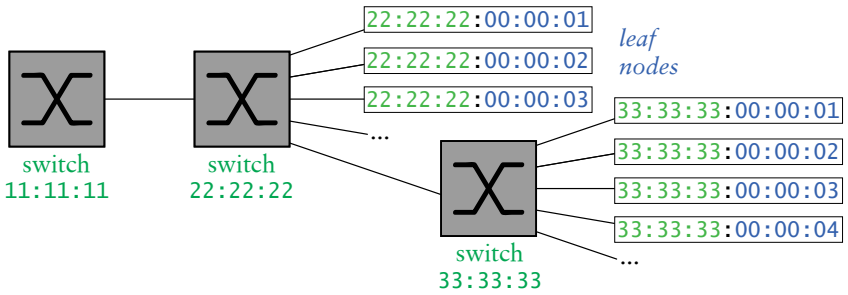
Ethernet, despite being an old protocol, continues to be used for new network deployments both in the core and at the edges of the Internet. It is desirable due to its simplicity, its ability to support multiple independent higher-layer (e.g. IP) networks over a single infrastructure, and its compatibility with the vast majority of currently-available products. Ethernet is a convenient "lowest common denominator" protocol.

However, it is well-known that Ethernet suffers from a few severe scalability problems. For example:

- It constrains packets to a spanning tree, by completely disabling links not on that tree. Shortest-path routing would make much better use of available resources.
- It makes heavy use of broadcast, for example when handling packets destined for unknown addresses and for higher-layer address resolution (ARP). Clearly the quantity of broadcast traffic increases with the number of systems connected to the network, and on large networks the broadcast traffic may even exceed the capacity of individual edge links.
- Each switch on a network must learn and store the location of every MAC address in use on the network, as a system's MAC addresses give no hint as to its location.

Previous attempts to solve the problem have generally relied on encapsulation of Ethernet frames within another protocol — MPLS, for example — but as I illustrate they do not completely address the issue or have unwanted side-effects.

In this poster I present our work to help to eradicate these problems from Ethernet. Our extension to Ethernet, MOOSE (Multi-level Origin-Organised Scalable Ethernet), is a modified switch which performs in-place rewriting of the source MAC address in Ethernet frames entering the network. The replacement MAC address is hierarchical, which means that switches are no longer required to maintain an address database for the entire network. Furthermore this also permits easier implementation of shortest-path routing.




MOOSE has several desirable properties when compared with MPLS-based or similar solutions. In particular, rewriting of destination addresses is not necessary: as the hierarchical source address remains in the packet when it reaches the destination end system, hierarchical addresses make their way into that system's ARP table and replies will be sent to the relevant hierarchical address directly. As a result, no lookup service to convert MAC addresses to hierarchical addresses is required; the requirement for such a lookup service is the downfall of several previous solutions. Hence, MOOSE is less prone to increasing broadcast traffic than alternative solutions.

This work forms part of the Intelligent Airport (TINA) project.[1]

[1] http://intelligentairport.org.uk/