Network Scalability for Green Computing

Malcolm Scott
Jon Crowcroft

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

EPSRC
Pioneering research and skills

The scenario

Migrate a computational load **to follow power availability**

- *Save energy by moving data rather than power*
- *Multiple data centres*
  near renewable and non-renewable power sources
- *Geographically-diverse network*

For this to work, you need...

- Virtualisation (with live migration):
  move a running system image to another physical server
  - Applications keep running, connections stay open
    - Host must keep the same address wherever it goes
      - Therefore, layer-2 network spanning all data centres supporting **millions of hosts** on a non-tree topology

Scope for protocol redesign is limited

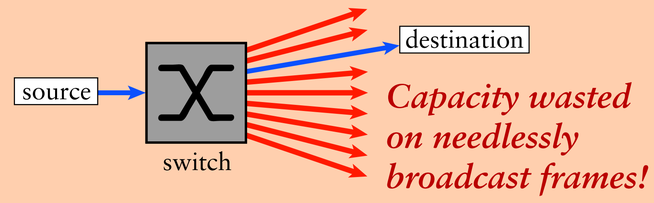
- Must support off-the-shelf servers and software
  - So we're stuck with ***Ethernet and IP***
    (Ubiquitous; too late to change)
- Currently, falls over with a few 10s of thousands of hosts
- But **intelligent network infrastructure** can boost scalability

Ethernet's scalability problems

Heavy use of broadcast

Broadcast ARP required for interaction with IP

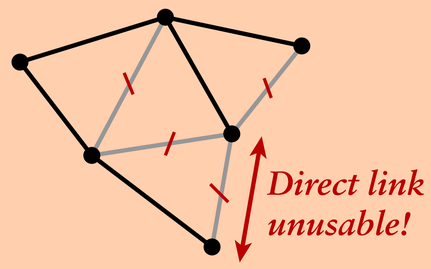
Higher-layer protocols use broadcast for discovery



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

Inefficient forwarding:

RSTP **disables links** to form a spanning tree



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 ~16000 addresses
- *Full table results in unreliability, or at best heavy flooding*

Underlying problem:

the MAC address space is *flat*

(looking at a MAC address gives you no hint as to the location of its owner)

The solution: MOOSE

Multi-level Origin-Organised Scalable Ethernet

- Introduce hierarchy to MAC addresses:

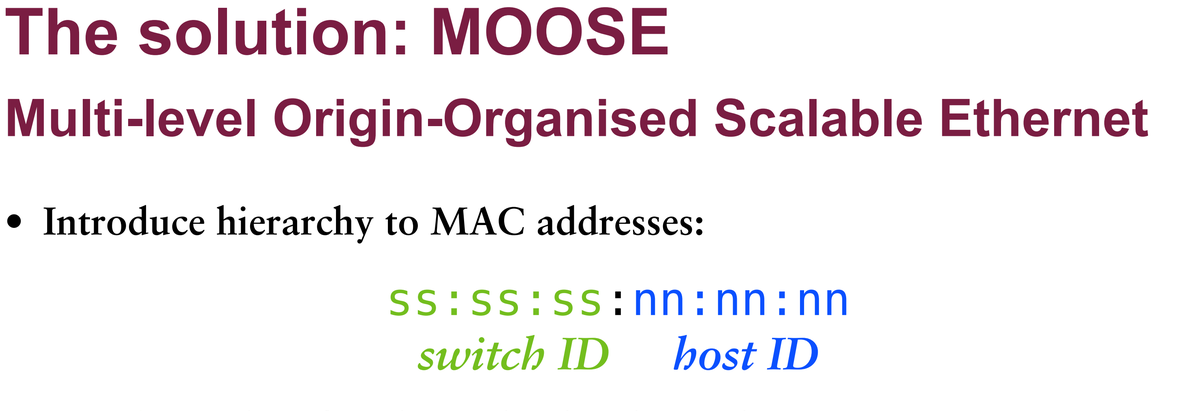
  `ss:ss:ss:nn:nn:nn`
  switch ID *host ID*

  Switch ID identifies the node's local switch

  Host ID is allocated by the local switch

- Modified switch performs source address rewriting on ingress
  Hierarchy automatically enforced

- New source address remains in frame when passed to destination



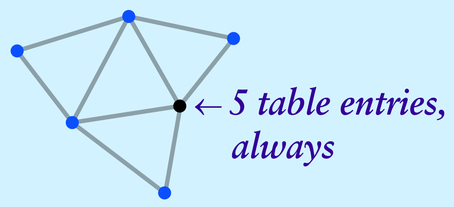
Now, switches need only store the locations of other switches
Above, switch 11:11:11 only needs two address table entries

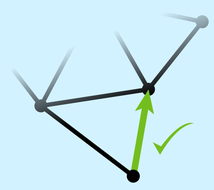
Compatible with unmodified standard Ethernet devices

Benefits of Hierarchical Addresses

Address tables

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



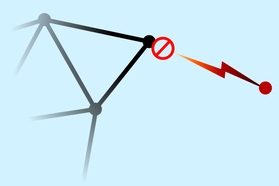


Best-path routing and resilience

- Switches can participate in a routing protocol
- Remove bottlenecks
- Better resilience
- Route prefixes: blocks of adjacent addresses

Security and isolation

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



Spam

Minimisation of broadcast traffic

- No longer need any broadcast for node location
- ELK: distributed directory service converting ARP into unicast
- Optimise broadcast traffic by inferring multicast groups?
  (early work)

http://www.cl.cam.ac.uk/~mas90/MOOSE/

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