Inside Linux Router

Dr. Ying-Dar Lin
High Speed Network Lab.
Department of Computer Information Science
National Chiao Tung University
May 15, 1999
Content

- What’s inside the router box?
- Linux Architecture
- Modules and Daemons
- Protocols and Algorithms
- Packet Flows: User/Control-plane
- New feature: QoS
What’s inside the router box?

- Much the same as the PC with “LINUX” inside!!
Modules & Daemons

User Space
- routed (RIP) / gated (RIP, OSPF, IS-IS, BGP, ...etc)

Kernel Space
- Routing Table
- Protocol Driver
- Adapter Driver
- Adapter Driver

Control packet
Data packet

Copyright © 林盈達@交大資訊科學系
Protocols and Algorithms

- **Standard**
  - Protocol: ICMP, RIP, OSPF, IS-IS, BGP, etc.
  - Algorithm:
    - Shortest path computation
      - Distance Vector: Bellman-Ford
      - Link State: Dijkstra

- **Non-standard**
  - Protocol: IGRP (Cisco)
  - Algorithms:
    - checksum computation
    - Routing Table lookup
Distance Vector Routing (Bellman-Ford)

For destination F

\[(\equiv, -) (\equiv, -) (11, C) (9, D) (9, D)\]

\[(\equiv, -) (4, F) (4, F) (4, F) (4, F)\]

\[(\equiv, -) (6, C) (6, C) (6, C) (6, C)\]

\[(\equiv, -) (7, F) (7, F) (7, F) (7, F)\]

\[(\equiv, -) (0, F) (0, F) (0, F) (0, F)\]
Problems with Distance Vector

- No link-bandwidth consideration
  - only cares instantaneous queue length
  - instability & oscillation

- Only rapidly to Good News
  - travel at the rate of one hop per exchange

- But leisurely to Bad News
  - count to infinite
  - No router ever has a value a few more higher than the minimum of all its neighbors
Link State Routing (Dijkstra)

For destination A
Link State Routing

- Ex: IS-IS, OSPF
- Learn neighbors & their network addresses
  - (HELLO packet)
- Measure link state
  - (ECHO packet)
- Building link state packets
  - (router id, sequence, age, (neighbors, cost), ....)
- Distribute link state packets to all other routers
  - check and update the table
    - (source router, sequence, age, send flags, ACK flags)
- Compute new routes
  - run Dijkstra’s algorithm locally
Packet Flows - User/Control Plane

User

Kernel

IP

MAC driver

Data Plane

Control Plane

Notation

Copyright © 林盈達@交大資訊科學系
Control-Plane : Shortest Path(1/2)

- RIP in routed - Bellman-Ford

<table>
<thead>
<tr>
<th>RIP Routing Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RIP Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 7 8 15 16 31</td>
</tr>
<tr>
<td>Command</td>
</tr>
<tr>
<td>Address Family</td>
</tr>
<tr>
<td>IP Address</td>
</tr>
<tr>
<td>*Subnet Mask</td>
</tr>
<tr>
<td>*Next Hop IP Address</td>
</tr>
<tr>
<td>Metric</td>
</tr>
<tr>
<td>Repeat of previous 20 bytes</td>
</tr>
</tbody>
</table>

* Only in RIP-2
Control-Plane: Shortest Path (2/2)

- OSPF in gated - Dijkstra (Dynamic Programming)

### OSPF Routing Table

<table>
<thead>
<tr>
<th>Destination</th>
<th>Next Hop</th>
<th>Distance metric</th>
</tr>
</thead>
</table>

### OSPF Header

<table>
<thead>
<tr>
<th>0</th>
<th>7</th>
<th>8</th>
<th>15</th>
<th>16</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>Type</td>
<td>Packet Length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Router ID</td>
<td>Area ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checksum</td>
<td>Authentication Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 林盈達@交大資訊科學系
User-Plane Processing (1/2) - Linux Router

- Receive a packet
  - dest. IP addr. = this port IP addr.?  
    - Yes: Control-plane
    - No: User-plane

  IP Protocol Field
    - ICMP, TCP, UDP

- TTL=0?
  - Yes: Send ICMP to packet source
  - No: Routing Table Lookup

- Packet Modification
- Forward the packet
User-Plane Processing (2/2) - Layer 3 switch

- Layer 2 switching (using MAC DA)
  - DA MAC learned?
    - Yes
      - MAC DA = this router port MAC address?
        - Yes
          - Send the packet to CPU
        - No
          - Dest. IP = this router port IP address?
            - Yes
              - Routing table lookup to find next hop IP address
                - ARP table lookup to find next hop MAC address
                  - Send the packet to the next hop
            - No (flood)
              - No (ICMP)
                - No (TTL ≥ 1?)
                  - Yes
                    - MAC DA / SA replacement
                    - TTL decrement, recalculate IP checksum & CRC
                  - No
                    - Send the packet to CPU

Copyright © 林盈達@交大資訊科學系
User-Plane : Table Lookup (1/2)

- Routing table in Linux kernel
  - organized as a hash table with linked lists
User-Plane : Table Lookup (2/2)

- Routing table in phase-2 router code
  - organized as hash table with trees

- Methodology
  - Hash to each tree by IP mask
  - Binary search with IP address
  - Not Found : Search another tree via forward pointer
User-Plane : Packet Modification

Packet modification summary

<table>
<thead>
<tr>
<th></th>
<th>MAC DA</th>
<th>MAC SA</th>
<th>TTL</th>
<th>Checksum</th>
<th>CRC (Org. Vtag)</th>
<th>CRC(Vtag Changed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same subnet</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>Recalculate</td>
</tr>
<tr>
<td>L3 unicast</td>
<td>Next hop</td>
<td>Router</td>
<td>•</td>
<td>•</td>
<td>Recalculate</td>
<td>Recalculate</td>
</tr>
<tr>
<td>L3 subnet directed BC</td>
<td>•</td>
<td>Router</td>
<td>Decrement</td>
<td>Recalculate</td>
<td>Recalculate</td>
<td>Recalculate</td>
</tr>
<tr>
<td>L3 multicast</td>
<td>•</td>
<td>Router</td>
<td>Decrement</td>
<td>Recalculate</td>
<td>Recalculate</td>
<td>Recalculate</td>
</tr>
</tbody>
</table>

These two may occur at the same time if subnet directed broadcast is supported.

These two may also occur at the same time in a multi-layer switch.
New Feature: QoS

- InterServ: RSVP

Signaling Protocol

Traffic Control
QoS Modules & Daemons

User Space
- routed
- update

Kernel Space
- Routing Table
- Protocol Driver
- Traffic Control Module
- Adapter Driver

rsvpd
- update

Control packet
Data packet

Copyright © 林盈達@交大資訊科學系

20
Packet Flows - User/Control Plane

User
- OSPF
- RIP
- RSVP
- BGP

Kernel
- ICMP
- UDP
- TCP

IP
- Data Plane
- Control Plane
- Notation

MAC driver

Copyright © 林盈達@交大資訊科學系
New Control-Plane Module

- **rsvpd**
  - by Information Sciences Institute (ISI)
    - Use CBQ as traffic scheduler
    - link aggregation for CL service
    - no traffic control modules
  - patched by Alexey Kuznetsov
    - traffic control function:
      - TC_AddFlowspec()
      - TC_ModFlowspec()
      - TC_DelFlowspec()
      - TC_AddFilter()
      - TC_DelFilter()
      - TC_Advertise()
      - Needs admission control to admit
New User-Plane Modules

- **Scheduler:**
  - CBQ (Class-based Queuing)
  - CSZ (Clark-Shanker-Zhang)
  - PRIO (n-band priority queue)

- **Rate estimator**
  - a base for statistical multiplexing for CL service

- **Classifier:**
  - Routing table based
  - Firewall based
  - U32
Known Bugs

- **rsvpd compilation**
  - don’t use IPv6

- **kernel modules**
  - don’t support auto-load yet
    - sch_cbq.o
    - cls_u32.o