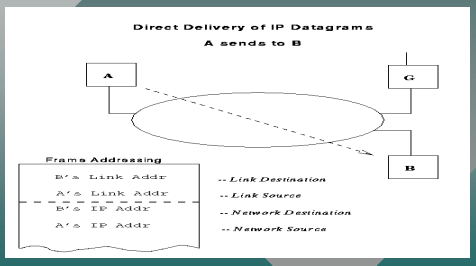
EECS 122, Lecture 11

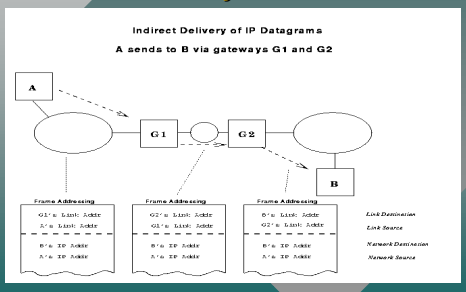
Kevin Fall

kfall@cs.berkeley.edu

Direct Delivery (no router)



Indirect Delivery



Direct Delivery (summary)

- Sender acquires receiver's IP address (e.g. through DNS or other mechanism)
- Sender determines receiver is on same network (by comparing network prefixes)
- Sender performs ARP query to obtain receiver's MAC address
- Sender encapsulates IP packet in local frame destined for receiver's MAC addr

Indirect Delivery (summary)

- Same as direct, except sender determines receiver is on different net
- Sender queries routing table to determine correct next hop router
- Encapsulates IP packet in local frame destined for router's MAC address
- Routers repeat this procedure

IP Options

- Option space limited to 40 bytes due to 4-bit IHL and 20 byte min IP header
- Zero or more options per datagram
- Different option encoding formats:
  - single byte (option type)
  - variable, starting with (type, length)

Option Types

- Contains 3 sub-fields
  - copied on fragmentation bit
  - option class number (2 bits)
  - option number (5 bits)
- Option Classes
  - control, reserved, debugging
- Simple options: EOL, nop (padding)

Source Routing

- header contains "pointer" and list of IP addresses indicating routers to be used for transit
- destination IP address is replaced by the IP address in the source routing list
- pointer is updated to next address
- IP header size remains constant

Record Route

- sender specifies size of IP header and sets "pointer" to indicate first (empty) 4-byte entry in option space
- each forwarder fills in its own [outgoing] IP address and increments pointer
- if full, just forwards
- *issue*: only 40 bytes for both option and its storage space, so 9 hops max!

Record Route Example

```
prompt (26)> ping -R www.cs.berkeley.edu
PING hyperion.CS.Berkeley.EDU (169.229.60.105): 56 data bytes
64 bytes from 169.229.60.105: icmp_seq=0 ttl=249 time=9.013 ms
RR:     ir40gw.lbl.gov (131.243.128.40)
 er1gw.lbl.gov (198.128.16.1)
 lbl-1c1-1.es.net (198.128.16.11)
 f8-0.inr-666-eva.Berkeley.EDU (128.32.2.2)
 f4-0.inr-107-eva.Berkeley.EDU (128.32.120.107)
 f3-0.inr-181-soda.Berkeley.EDU (128.32.40.1)
 169.229.60.65
 hyperion.CS.Berkeley.EDU (169.229.60.105)
 128.32.40.202
64 bytes from 169.229.60.105: icmp_seq=1 ttl=249 time=9.541 ms  (same route)
64 bytes from 169.229.60.105: icmp_seq=2 ttl=249 time=193.611 ms       (same route)

prompt (27)> traceroute www.cs.berkeley.edu
traceroute to hyperion.CS.Berkeley.EDU (169.229.60.105), 30 hops max, 40 byte packets
 1  ir40gw.lbl.gov (131.243.1.1)  0.758 ms  0.615 ms  0.695 ms
 2  er1gw.lbl.gov (131.243.128.11)  0.945 ms  1.104 ms  1.189 ms
 3  lbl-1c1-1.es.net (198.128.16.11)  30.318 ms  197.045 ms  1.555 ms
 4  f5-0.inr-666-eva.berkeley.edu (198.128.16.21)  4.349 ms  2.695 ms  2.170 ms
 5  f1-0-0.inr-107-eva.Berkeley.EDU (128.32.2.1)  1.400 ms  3.824 ms  2.172 ms
 6  f1-0.inr-181-soda.Berkeley.EDU (128.32.120.181)  3.244 ms  2.752 ms  1.679 ms
 7  128.32.40.202 (128.32.40.202)  1.731 ms  2.342 ms  2.587 ms
 8  hyperion.CS.Berkeley.EDU (169.229.60.105)  2.598 ms  2.663 ms  2.337 ms
```

Time Stamp

- Facility to record routers 'notions of time, and optionally their IP addresses
- Options contains "pointer", overflow counter [4 bits], and flag [4 bits]
  - overflow: # of IP modules that could not fit their addresses into the header
  - flag: times only, times + RR, or selected times (list of address/zero pairs)

The Time Value

- TS Options use the number of milliseconds since midnight UT
- This is a loose time requirement, so not very useful for precise measurement
- Also: setting high-order bit in time allows for non-standard time values

Source and Record Route Options

- Loose Source & Record Route (LSRR):
  - "loose" source routing: list of IP addresses need not be exact; multi-hop routes may be used between each entry
- Strict Source & Record Route (SSRR):
  - "strict" source routing: list of IP addresses need to be 1-hop away from each other

Internet Control Message Protocol (ICMP)

- IP provides no direct way of discovering the fate of an IP packet
- Want a mechanism for error reporting and information exchange
- ICMP Protocol (RFC792)
  - logically part of IP module, but is actually encapsulated within IP

ICMP Operation

- Provides IP module to IP module message delivery
- Error and information reporting only
  - *queries*: client/server info request/resp
  - *errors:* reports of error conditions
- Restrictions are placed on the generation of ICMP messages to avoid cascades

ICMP Restrictions

- ICMP messages are not allowed to be sent in response to (RFC1812):
  - an ICMP error message (ok for queries)
  - datagrams failing header validation tests
  - broadcast or multicast IP datagrams
  - link-layer broadcast or multicast frames
  - invalid src address or zero net prefix
  - any fragment other than the first

IP Header Validation Tests

- To be a valid IP header:
  - link-layer must indicate frame is long enough
  - IP checksum must be correct
  - IP version number must be 4
  - IP IHL field must be at least 5
  - IP total len must be at least (IHL*4)

ICMP Error Message Data

- Historically, ICMP errors returned the offending IP header and 1st 8 data bytes
- No longer adequate with more complicated headers like IP in IP
- New rules say should contain as much as original datagram as possible, without the length of ICMP datagram being > 576 bytes (standard Internet min size)

ICMP Header

- Encapsulated as IP payload
- Type field is 1 of 15 message types
- Code indicates special sub-types
- Checksum covers entire ICMP message

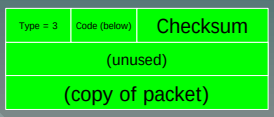
ICMP Error Message Types

- 3 = Destination Unreachable
- 4 = Source Quench
- 5 = Redirect
- 11 = Time Exceeded
- 12 = Parameter Problem

ICMP Query Message Types

- 0 = Echo Reply ("ping response")
- 8 = Echo Request ("ping query")
- 9 = Router Advertisement (RFC 1256)
- 10 = Router Solicitation (RFC 1256)
- 13 = Time Stamp Request
- 14 = Time Stamp Reply
- 17 = Address Mask Request
- 18 = Address Mask Reply

ICMP Destination Unreachable



- Unreachable things:
  - 0:network, 1:host, 2:protocol, 3:port
  - 4: frag needed, but DF set [may incl MTU]
  - 5: source route failed
  - (there are others defined in RFC 1122)

Unreachable Destinations

- Network Unreachable
  - generated by router lacking any route to destination
- Host Unreachable
  - last hop router cannot contact destination
- Protocol Unreachable
  - host lacks a layer-4 protocol implementation
- Port Unreachable
  - no process bound to port (usually with UDP--later)

Fragmentation Needed

- Code 4 indicates the datagram required fragmentation but the DF bit was set
- Newer implementations replace (unused) 2nd word of ICMP header with next MTU
- MTU info returned to host, where it can subsequently alter its packet size to avoid fragmentation (path MTU discovery)

ICMP Source Quench

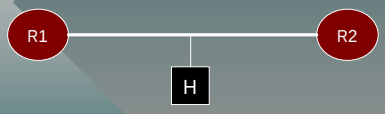
- Initial idea was that routers could generate "slow down" messages
- Problem is generating more traffic during periods of high traffic is not attractive
- Currently, routers should not generate source quench ICMP messages

ICMP Redirect

Type = 5	Code (below)	Checksum
IP Address of Router		
(copy of packet)		

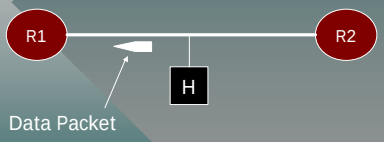
- Indicates wrong router on network is being used as first hop. Redirect indicates which router to use instead.
- Code field values:
  - 0:network, 1:host
  - 2:TOS & Network, 3: TOS & Host

ICMP Redirect

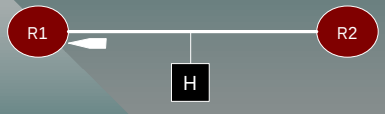


- H's routing table indicates R1 is proper first-hop router for its packet

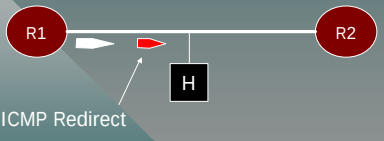
ICMP Redirect



ICMP Redirect

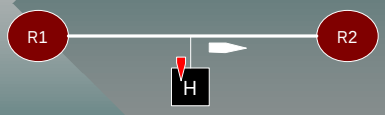


ICMP Redirect

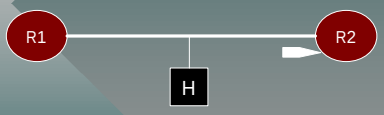


- R1's routing table indicates R2 (attached to same network prefix) is the correct router for the data packet

ICMP Redirect



ICMP Redirect



- H's routing table is now updated to indicate R2 is the proper next-hop router
- R2 will forward packet normal way

ICMP Time Exceeded

Type = 11	Code (below)	Checksum
(unused)		
(copy of packet)		

- Indicates IP packet's delivery time has been exceeded
- Code field values:
  - 0: TTL exceeded in transit
  - 1: fragment reassembly time exceeded

ICMP Parameter Problem

Type = 12	Code (below)	Checksum
Pointer	(unused)	
(copy of packet)		

- General catch-all for any delivery error not otherwise covered
- Pointer indicates the byte offset of the error (relative to beginning of IP header)

ICMP Echo Response/Reply

Type = 0/8	Code (below)	Checksum
Identifier		Sequence Number

data ...

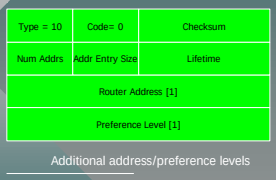
- Typically used to quickly indicate connectivity ("ping program"). Also can indicate loss, duplication, and re-ordering using the sequence number.
- Identifier allows for matching up requests with responses

ICMP Router Solicitation

Type = 10	Code= 0	Checksum
Reserved		

- Sent by hosts (during init) to find nearby routers. May be sent from address 0.0.0.0 or known IP address. Sent to multicast 224.0.0.2 [all routers] or local broadcast IP address.

ICMP Router Advertisement

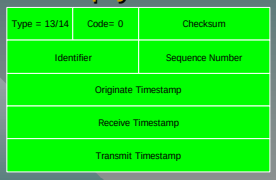


- Sent by routers quasi-periodically to indicate default routes to hosts. Sent to multicast 224.0.0.1 [all systems] or local broadcast.

ICMP Router Advertisement

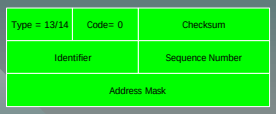
- "Num Addrs" field gives the number of address blocks in advertisement message
- "Addr Entry Size" field gives # of words in each address block
- "Lifetime" is # of seconds to believe the info
- One way to get a default route [but today DHCP is more popular]

ICMP Timestamp Request/Reply



- Originate: when sender last touched data
- Receive: when receiver first received data
- Transmit: when echoer last touched data

ICMP Address Mask Request/Reply (RFC 950)



- Used to obtain network prefix (subnet mask) using ICMP
- Hosts may send during init (to broadcast address using 0.0.0.0 as source)
- Typically provided by DHCP now

Special Uses for ICMP

- Path MTU discovery
  - determine the smallest MTU along a path
- Route tracing
  - use ICMP error messages to "trace the route" of packets

Path MTU Discovery

- RFC 1191, common but not universal
- Start with packet size *p* <= *local MTU*
  - send all packets with DF = 1
  - if frag required, router sends ICMP Dest Unreach, and may send the next MTU
  - set *p* to be this MTU, or search common sizes
  - periodically try to increase (up to orig. *p*)

Route Tracing using ICMP

- "traceroute" ("tracert") tool:
  - send UDP packet to destination host
  - start with TTL = 1, send 3, bump TTL and repeat
  - each router generates ICMP time exceeded, with its source address (provides route)
  - host generates ICMP port unreachable for bad UDP port in probe packet
- May be erroneous for changing and asymmetric routes