z/OS TCP/IP Routing

Linda Harrison

lharriso@us.ibm.com

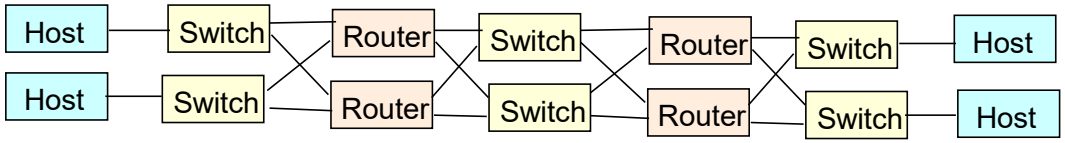
IBM Washington Systems Center (WSC)

Agenda

- Routing Basics
- Virtual LAN (VLAN)
- Sending and Receiving Data
- ARP & IP Routing
- Maximum Transmission Unit (MTU)
- OMPRoute
- Multipath and Cisco® ACI®
- More Information

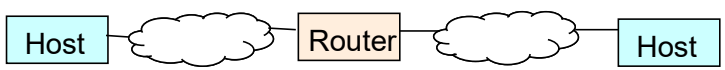
Routing Basics

Physical and Logical Network



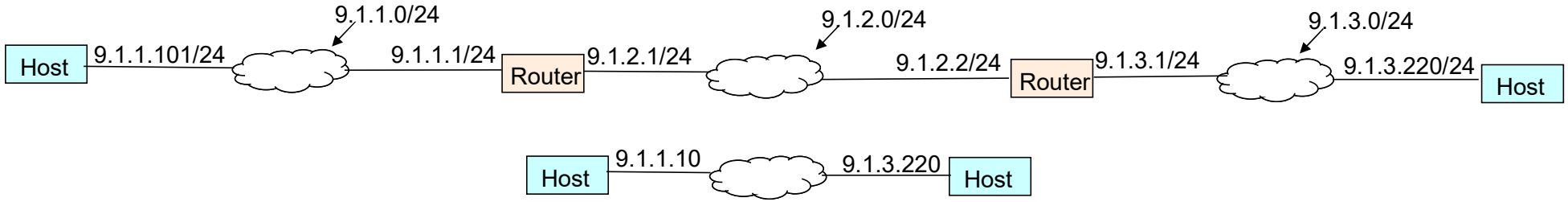
Layer 5	Application
Layer 4	Transport
Layer 3	Network
Layer 2	Data Link
Layer 1	Physical

- There is some physical network.
- It can be shared between TCP/IP and other protocols concurrently.
- Devices are connected with cables. End devices may be connected using WiFi.
- Each device network connection has a Media Access Control (MAC) address that is used for sending and receiving messages.



- There are logical network views that use the underlying physical network.
  - The level of detail depends upon the discussion.
- Each device network connection has a Logical address that is used for sending and receiving messages.

TCP/IP Network

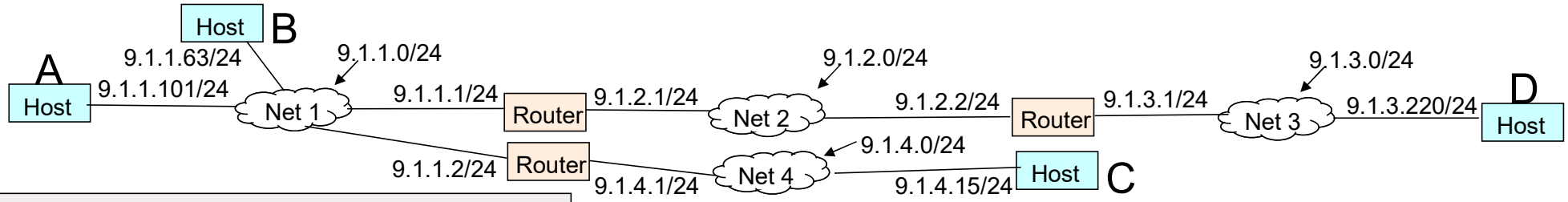


- IPv4 Address has 32 bits 1111 1111 1111 1111 1111 1111 1111 1111
- Dotted Decimal IPv4 Address 9.1.1.101 = 0000 1001 0000 0001 0000 0001 0110 0101
- Subnet Mask 24 indicates that the first 24 bits identify the subnetwork and only the last 8 bits identify the host address.
- 24 = 1111 1111 1111 1111 1111 1111 0000 0000
  - Also referred to as 255.255.255.0

Network, Subnet, and Supernet

- IPv4 Networks are defined in different Classes
  - Class A addresses are 0.0.0.0 through 127.255.255.255 have a mask of 255.0.0.0 (aka mask /8)
  - Class B addresses are 128.0.0.0 through 191.255.255.255 have a mask of 255.255.0.0 (aka mask /16)
  - Class C addresses are 192.0.0.0 through 223.255.255.255 have a mask of 255.255.255.0 (aka mask /24)
  - Class D addresses are 224.0.0.0 through 239.255.255.255 have a mask of 255.255.255.255 (aka mask /32)
  - Class E addresses are 240.0.0.0 through 247.255.255.255 have a mask of 255.255.255.255 (aka mask /32)
- Those networks can be divided into smaller subnets
  - Class A Network 9.0.0.0/8 (mask 255.0.0.0) is assigned to IBM
  - It has been divided into different subnets, ie. 9.82.1.0/24 (aka mask 255.255.255.0)
- Supernet is the opposite of subnet
  - Routers can reduce the number of routes being sent by combining them.
  - The following four routes:
    - ROUTE 9.82.0.0/18 9.15.6.1 NIC521 MTU 1492
    - ROUTE 9.82.64.0/18 9.15.6.1 NIC521 MTU 1492
    - ROUTE 9.82.128.0/18 9.15.6.1 NIC521 MTU 1492
    - ROUTE 9.82.192.0/18 9.15.6.1 NIC521 MTU 1492
  - Can be combined into single route:
    - ROUTE 9.82.0.0/16 9.15.6.1 NIC521 MTU 1492

TCP/IP Routing



Host A:
Direct Route to 9.1.1.0/24
Indirect Route to 9.1.4.0/24 via 9.1.1.2
Host Route to 9.1.4.15 via 9.1.1.2
Default Route to 9.1.1.1

```
ROUTE 9.1.1.0/24  =  OSA472 MTU 1492
ROUTE 9.1.4.0/24 9.1.1.2 OSA472 MTU 1492
ROUTE 9.1.4.15 HOST 9.1.1.2 OSA472 MTU 1492
ROUTE DEFAULT  9.1.1.1 OSA472 MTU 1492
```

- Direct Routes
  - Subnetworks that this host connects to.
- Indirect Routes
  - Subnetworks that this host can reach by routing through a router.
  - Host route is a special type of Indirect Route (see next page).
  - When the destination is not on a directly attached subnet but a host on a directly attached subnet does have a route to the destination.
- Default Routes
  - Where to send packets when there is no explicit (direct or indirect) route.

z/OS Routing Table

Subnet mask of 32 indicates Host Routes, the most specific type of route possible. More specific routes always take precedence over less specific routes.
Subnet or Network routes, in this case subnet route with mask of 24, are less specific than host routes but more specific than default routes.
Default routes are the least specific routes possible and are therefore only used when no other route matches the destination.

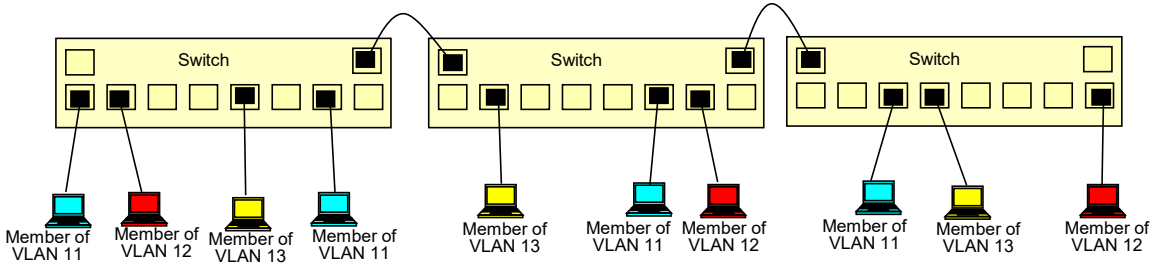
Host A:
Direct Route to 9.1.1.0/24
Indirect Route to 9.1.4.0/24 via 9.1.1.2
Host Route to 9.1.4.15 via 9.1.1.2
Default Route to 9.1.1.1

```
      IP Addr        Gateway Interface      MTU
ROUTE 9.1.1.0/24     =       OSA472     MTU 1492
ROUTE 9.1.4.0/24     9.1.1.2 OSA472     MTU 1492
ROUTE 9.1.4.15 HOST  9.1.1.2 OSA472     MTU 1492
ROUTE DEFAULT   9.1.1.1 OSA472          MTU 1492
```

- The destination IP Address, IP subnet, or IP network is in the second column above.
  - The destination IP Address and mask may be written with a decimal subnet mask, ie. 9.1.1.0/24, or a dotted decimal subnet mask, ie. 9.1.1.0 255.255.255.0, or in the case of a route to a single destination, may be depicted using the keyword HOST, ie. 9.1.4.15 HOST.
  - DEFAULT is a special destination keyword indicating where to send packets when the destination does not match any other route in the table.
- Gateway IP Address is in the third column above. An equal sign, =, indicates there is no next hop gateway needed because the destination is on the same subnet as the interface.
  - A route without a next hop gateway is known as a direct route.
  - A route with a next hop gateway is known as an indirect route.
- The interface name or link name is in the fourth column above to indicate which interface the packet is to be sent over.
- Each route in the table may be read as “If this destination, then send packet to this gateway over this interface”.

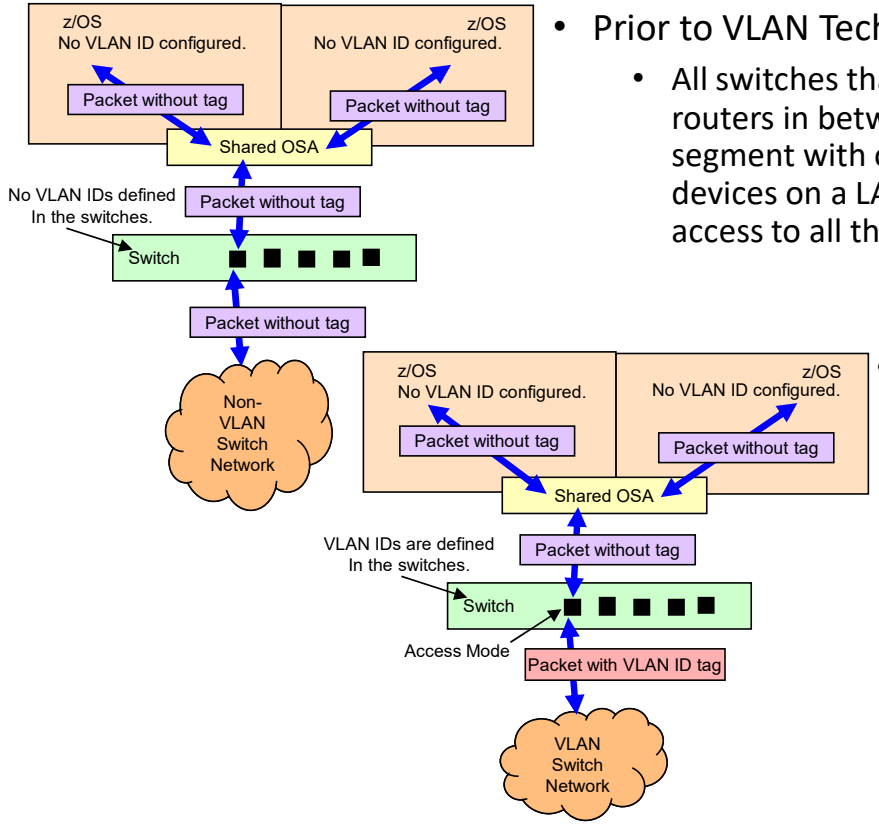
Virtual LAN (VLAN)

What is a VLAN?



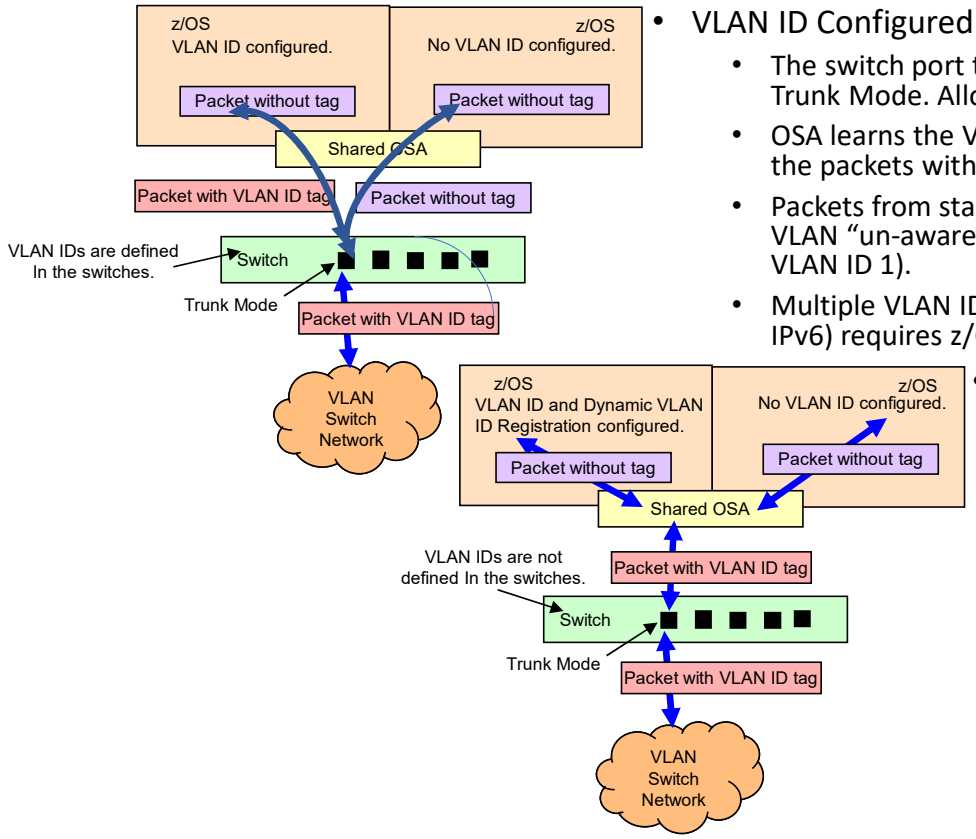
- A VLAN is a switched network that is logically segmented on an organizational basis, by functions, project teams, or applications rather than on a physical or geographical basis.
- Reconfiguration of the network can be done through software rather than by physically unplugging and moving devices or wires.
- A VLAN can be thought of as a broadcast domain that exists within a defined set of switches.
- A VLAN consists of a number of end systems, either hosts or network equipment (such as bridges and routers), connected by a single bridging domain.
- VLANs are created to provide the segmentation services traditionally provided by routers in LAN configurations.
  - None of the switches within the defined group will bridge any frames, not even broadcast frames, between two VLANs.
  - IP Router is needed to communicate between VLANs.

When z/OS is VLAN “un-aware”



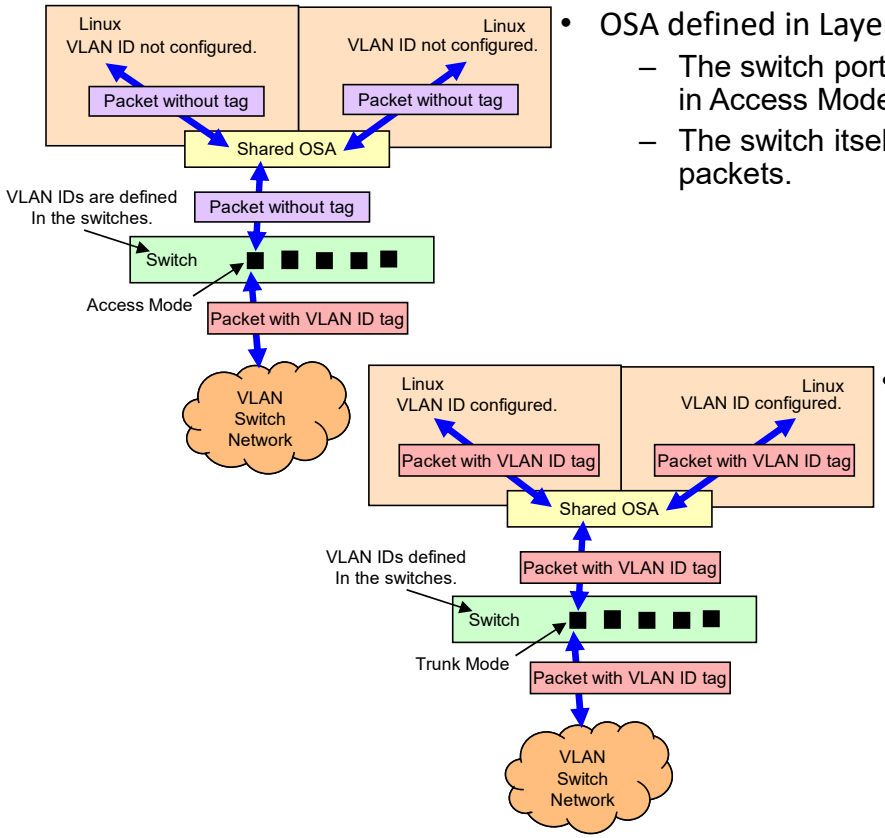
- Prior to VLAN Technology
  - All switches that were attached together (without routers in between them) formed one physical LAN segment with one IP subnet assigned to them. All devices on a LAN segment could potentially have access to all the packets flowing on the segment.
- VLANs Defined on Switches
  - z/OS may still be VLAN “un-aware”.
  - The switch port that OSA attaches to should be configured in Access Mode with a certain VLAN ID assigned.
  - The switch itself manages the VLAN ID tagging of the packets.

When z/OS is VLAN “aware”



- VLAN ID Configured on LINK/INTERFACE
  - The switch port that OSA attaches to must be configured in Trunk Mode. Allowed VLAN IDs are defined.
  - OSA learns the VLAN ID from the stack and manages tagging the packets with the appropriate VLAN ID.
  - Packets from stacks that do not configure a VLAN ID (still VLAN “un-aware”) are part of the default VLAN ID (usually VLAN ID 1).
  - Multiple VLAN IDs per stack/OSA port per IP version (IPv4 or IPv6) requires z/OS V1.10+ and VMACs.
- VLAN ID and Dynamic VLAN ID Registration Defined on LINK/INTERFACE
  - Rather than define the “allowed” VLAN IDs on the switch, the switch learns the VLAN IDs for the port from the OSA.
  - This does require the switch to be configured to support dynamic VLAN registration.

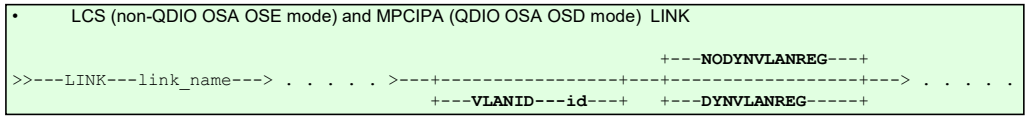
Linux on Z



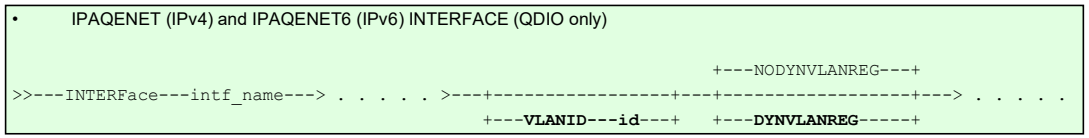
- OSA defined in Layer 3 mode is like “VLAN unaware” z/OS
  - The switch port that OSA attaches to should be configured in Access Mode with a certain VLAN ID assigned.
  - The switch itself manages the VLAN ID tagging of the packets.

- OSA defined in Layer 2 mode is a little like “VLAN aware” z/OS
  - The switch port that OSA attaches to must be configured in Trunk Mode. Allowed VLAN IDs are defined.
  - The difference is that the VLAN ID is included in the packet between the OSA and the Linux operating system

QDIO VLAN Support



DYNVLANREG
Requires Software:
z/OS V1.8 or later



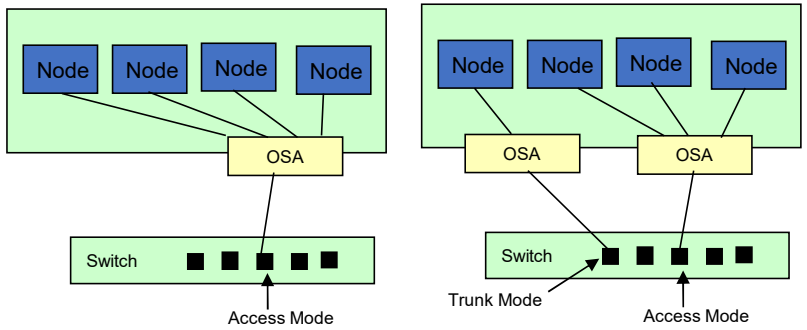
VLANID Requires Software: z/OS V1.5 or later.
Prior to z/OS V1.10:
Limited to one VLAN ID per IPv4 or IPv6 per stack/OSA port.
Other stacks may define different VLANIDs for same port.
z/OS V1.10 and later:
Multiple VLAN IDs per IP version per stack/OSA port (Interface Only – not supported on Link)
-Maximum of 8 VLAN IDs per IP version (IPv4 or IPv6) per OSA port per stack.
-Different VMACs are required.

- VLANID id_number
  - Specifies the VLAN ID tag for this link.
- DYNVLANREG/NODYNVLANREG
  - Dynamic registration of VLAN ID (GVRP).
    - Dynamic registration of VLAN IDs is handled by OSA and switch. Both must be at a level with the hardware support for dynamic VLAN ID registration.
  - DYNVLANREG specifies that if a VLAN ID is configured for this link, it is dynamically registered with physical switches on corresponding LAN.
    - This parameter is only applicable if a VLAN ID is specified.
  - NODYNVLANREG specifies that if VLAN ID is configured, it must be manually registered with switches on corresponding LAN. This is the default.
  - DYNVLANREG must match between LINK and INTERFACE for the same OSA.
- **VMAC is required to define multiple VLAN IDs for IPv4 or IPv6, from a single stack for a given OSA port.**

z/OS Support of VLAN IDs

- z/OS TCP/IP supports configuring the VLAN ID to be used on OSA connections.
  - z/OS may configure the VLAN ID but it is OSA that adds/removes the VLAN ID tag to the packets.
  - Conforms to the IEEE 802.1Q standard
- A Switch may configure a port in Trunk mode or Access mode.
  - Trunk mode
    - VLAN ID is defined by the end device, either configured on z/OS or defaulted by the OSA.
    - Requires VLAN ID tagged packets.
  - Access mode
    - VLAN ID is controlled by the switch rather than the end device. Any VLAN ID configured by z/OS is ignored.
- z/OS VLAN Rules:
  1. An OSA should either be:
     - Attached to a switch port in trunk mode if any of the stacks that share the OSA have a VLAN ID configured, or
     - Attached to a switch port in access mode and each stack that shares the OSA should not have a VLAN ID configured.
  2. As with any IP network, separate VLANs should be treated like separate physical networks and have separate subnets assigned.
  3. Some switch vendors use VLAN ID 1 as the default value when a VLAN ID value is not explicitly configured. It is recommended that you avoid the value of 1 when configuring a VLAN ID value.
  4. When a TCP/IP stack has access to multiple OSA ports that are on the same physical LAN, and a VLAN ID is configured on any of the OSA ports, it is recommended that this stack configure a VLAN ID for all OSA ports on the same physical LAN. Do not mix VLAN and no-VLAN on the same physical network accessed by a single stack through multiple OSA ports.
  5. When multiple INTERFACE statements are defined on a single stack for a single OSA port and a single IP version (IPv4 or IPv6), the VLAN IDs must be unique, and the INTERFACE definition will be rejected if the VLAN ID is omitted.
     - The VLAN ID, VMAC, and IP subnet values must be unique per IP version (IPv4 or IPv6) for multiple INTERFACE statements for a single OSA port defined on a single TCP/IP stack.
     - For parallel interfaces into the same IP subnet/VLAN ID from a single TCP/IP stack, multiple OSA ports are required.
  6. The requirement for a unique VLAN ID per INTERFACE statement rule only applies within a single stack. Each stack on a shared OSA port is completely independent of other stacks sharing the OSA port. Multiple stacks may define the same VLAN ID or different VLAN IDs for the same shared OSA port.

VLAN Migration

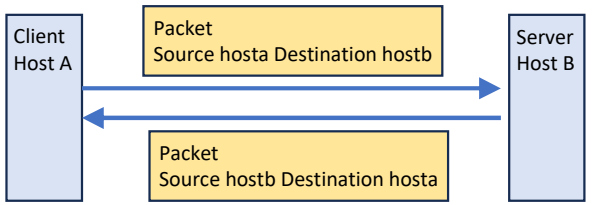


- Migration of z/OS VLAN “unaware” to z/OS VLAN “aware”
- Switch port defined in Access Mode
  - Operating Systems should define OSA without VLAN (VLAN “unaware”)
- Switch port defined in Trunk Mode
  - Operating Systems should define OSA with VLAN (VLAN “aware”)
- An OSA port attaches to a Switch in either Access Mode or Trunk Mode. If multiple LPARs share an OSA port attached to a switch in Access Mode and one of those LPARs wants to start to use VLAN configuration either:
  - Use a second OSA port in Trunk Mode.
  - Or change the Switch port from Access Mode to Trunk Mode. All traffic to/from the LPARs that do not define VLAN will be sent using the Default VLAN ID.
    - If the Default VLAN ID uses a different subnet, then all the LPARs IP addresses will have to change.

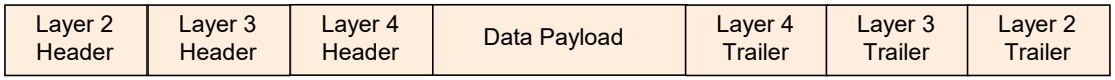
Sending and Receiving Data

Source and Destination IP Addresses

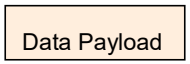
- TCP/IP Application Session
  - One partner initiates the connection and sends a connection request packet.
  - For TCP connections the host that sends the connection request is the client.
    - What destination IP address to connect to?
      - *If hostname is given then IP address is resolved by DNS or Local Host Name file (IPNodes).*
    - What source IP address to use in the initiate packet?
      - *Hierarchy of source IP address is detailed in z/OS Communication Server (CS) IP Configuration Guide, SC31-8775.*
        - *Sendmsg( ) using the IPV6_PKTINFO ancillary option specifying a nonzero source address (RAW and UDP sockets only)*
        - *Setsockopt( ) IPV6_PKTINFO option specifying a nonzero source address (RAW and UDP sockets only)*
        - *Explicit bind to a specific local IP address*
        - *bind2addrsel socket function (AF_INET6 sockets only)*
        - *PORT profile statement with the BIND parameter*
        - *SRCIP profile statement (TCP connections only)*
        - *TCPSTACKSOURCEVIPA parameter on the IPCONFIG or IPCONFIG6 profile statement (TCP connections only)*
        - *SOURCEVIPA: Static VIPA address from the HOME list or from the SOURCEVIPAINTERFACE parameter*
        - *HOME IP address of the link over which the packet is sent*
  - The other partner receives the connection request packet and responds.
  - For TCP connections the host that receives the connection request is the server.
    - The host that receives the connection request takes the source IP address from the connection packet and uses that for the destination IP address in the packet that it sends back.
    - The host that receives the connection request takes the destination IP address from the connection packet and uses that for the source IP address in the packet that it sends back.
    - These source and destination assignments are usually used for the life of the connection.



Sending a Packet



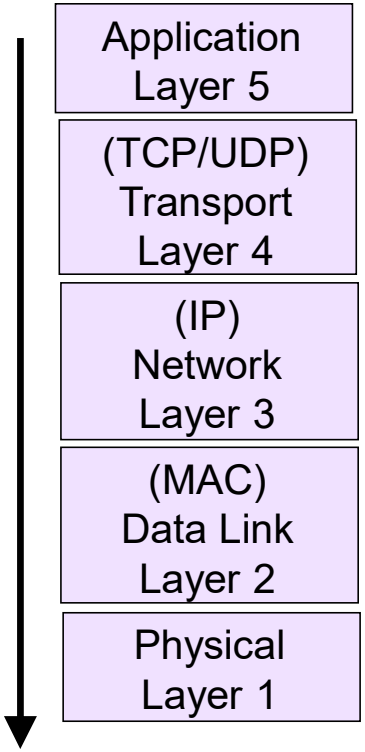
- Application Layer 5 sends some data



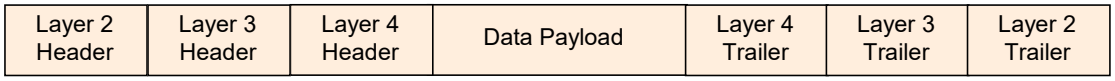
- TCP/IP stack Transport Layer 4
  - It puts some header information in front of the data.
  - It might put some trailer information after the data.

- TCP/IP stack Network Layer 3
  - It puts some header information in front of the data.
  - It might put some trailer information after the data.

- OSA Transport Layer 2
  - It puts some header information in front of the data.
  - It might put some trailer information after the data.



Receiving a Packet



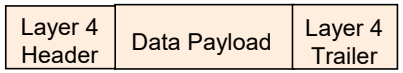
- OSA Transport Layer 2

  - It strips off header and option trailer and passes packet to TCP/IP Stack Layer 3.

- TCP/IP stack Network Layer 3

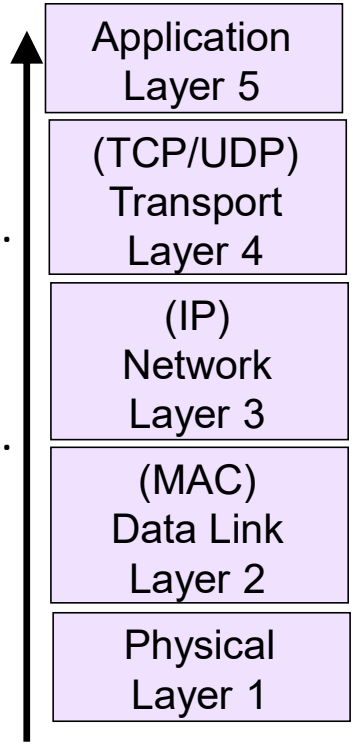
  - It strips off header and option trailer and passes packet to TCP/IP Stack Layer 4.

- TCP/IP stack Transport Layer 4

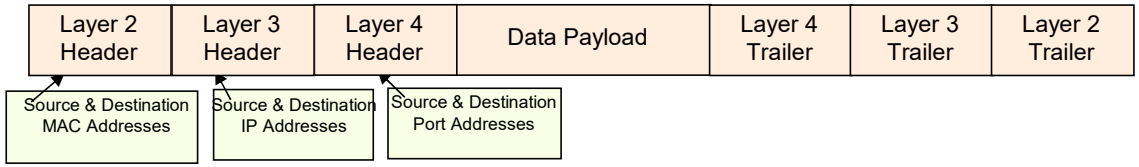


  - It strips off header and option trailer and passes data to Application Layer 5.

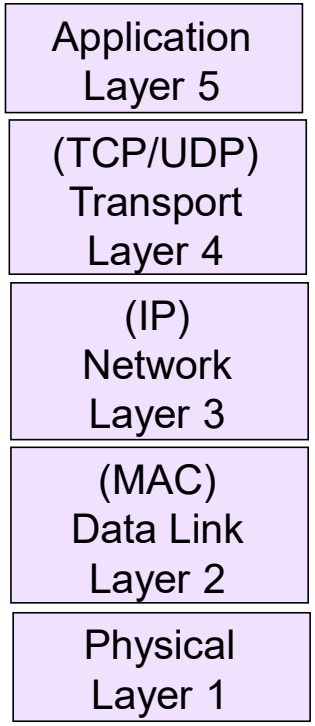
- Application Layer 5 receives some data



What is Layer 2



- z/OS Operating System only supports Layer 3 protocol transmission from QDIO OSA (CHPID type OSD). The Layer 2 Header and Trailer have been stripped off the packet before it is passed to z/OS.
- Linux on System z supports Layer 2 protocol transmission from QDIO OSA. Data received by OSA is passed to Linux with the Layer 2 Header and Trailer in place.
  - Layer 2 is also referred to as protocol agnostic since the Layer 3 protocol type does not matter (it could be IP, SNA, Apple Talk, anything).
- Can z/OS communicate to Linux even though z/OS is using Layer 3 LAN attachment and Linux is using Layer 2 attachment? YES
- A Layer 2 LAN can refer to a LAN between devices that does not have an IP router (also referred to as a Layer 3 router) in the communication path, all devices are in the same IP subnet.
  - z/OS can attach to a Layer 2 LAN.



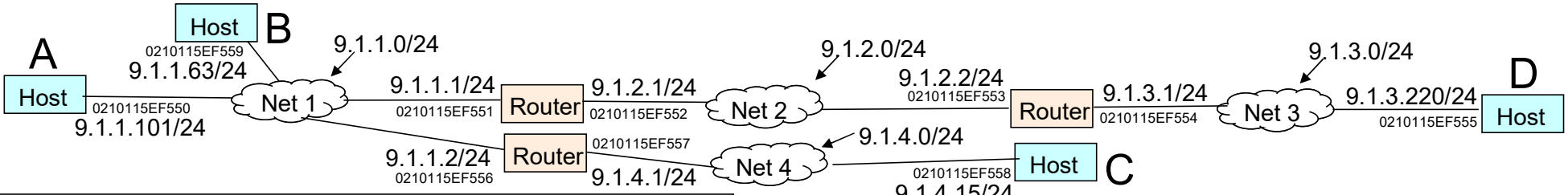
VLAN Priority and TOS

- Packet priority may be defined both in the Layer 2 header and the Layer 3 header. Packets with higher priority are sent ahead of packets with lower priority, if the devices the packets are routed through support that layer packet priority.
- At Layer 2, VLAN Priority may be defined.
  - VLAN Priorities from lowest to highest are 0 to 7.
  - For packets in the same subnet, the VLAN priority is used to prioritize the traffic through the devices that support it.
- At Layer 3, a IP Type of Service (TOS) may be defined.
  - The TCP/IP RCF defines the TOS field in the header but it does not define the number of hardware queues (OSA has four) or how those hardware queues are mapped to the TOS.
- Intrusion Detection Services policies may be used to define the VLAN Priority and TOS.

```
SetSubNetPrioTosMask
        {
         SubnetTosMask       11100000
         PriorityTosMapping  1 1110000  7
         PriorityTosMapping  1 1100000  6
         PriorityTosMapping  2 1010000  5
         PriorityTosMapping  2 1000000  4
         PriorityTosMapping  3 0110000  3
         PriorityTosMapping  3 0100000  2
         PriorityTosMapping  4 0010000  1
         PriorityTosMapping  4 0000000  0
        }                    ^       ^      ^
                             |       |      + This is the VLAN Priority
                             |       + This is the TOS value.
                             + This is the OSA Queue. Each OSA has 4 queues.
```

ARP & IP Routing

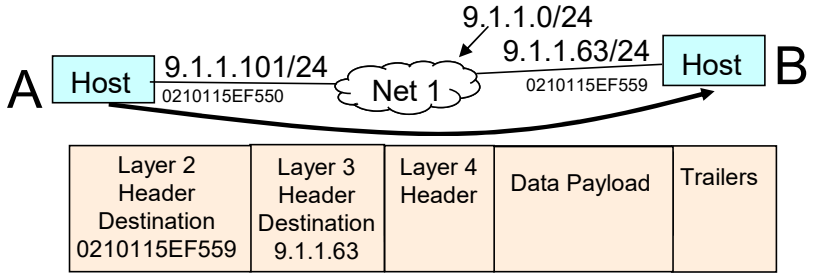
Address Resolution Protocol (ARP)



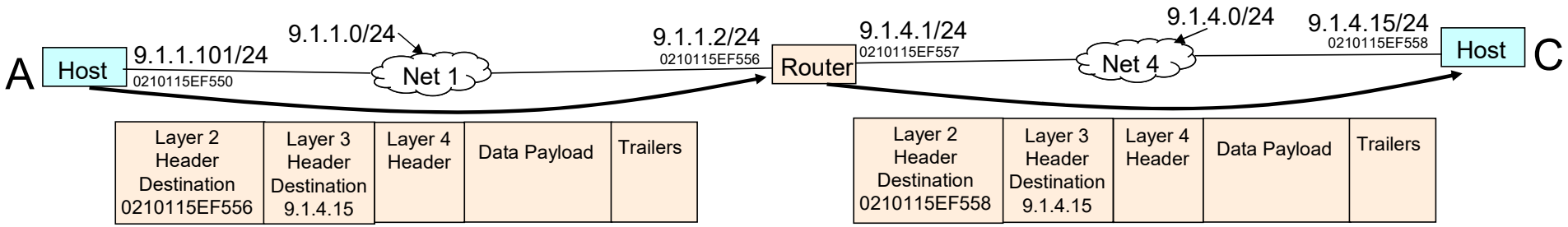

```
ROUTE 9.1.1.0/24  =  OSA472 MTU 1492
ROUTE 9.1.4.0/24 9.1.1.2 OSA472 MTU 1492
ROUTE 9.1.4.15 HOST 9.1.1.2 OSA472 MTU 1492
ROUTE DEFAULT  9.1.1.1 OSA472 MTU 1492
```

- Address Resolution Protocol (ARP) is used to discover the MAC address associated with an IP address.
  - Source and Destination MAC address are in the Layer 2 Header of the packet.
  - Source and Destination IP address are in the Layer 3 Header of the packet.
- For Host A to send data to Host B, 9.1.1.63, on the subnet that it is attached to (Direct Route) it uses ARP to learn the MAC address, 0210115EF559. The Layer 2 header is created with the destination 0210115EF559.
- For Host A to send data to Host C, 9.1.4.15, on remote subnet Net 4, the IP routing table indicates to send the packet to 9.1.4.15 (Indirect Route). ARP is used to learn the MAC address, 0210115EF557. It is put in the Layer 2 header as the destination.
- For Host A to send data to Host D, 9.1.3.220, on remote subnet Net 3, the IP routing table indicates to send the packet to 9.1.1.1 (Default Route). ARP is used to learn the MAC address, 0210115EF551. It is put in the Layer 2 header as the destination.
- Whenever an adapter comes up on an Ethernet LAN the device sends out a Gratuitous ARP (unsolicited ARP) message. An ARP Table stores MAC and IP address pairing but if there is no traffic to/from a host in a certain amount of time (ie. 20 minutes) then the pair is removed from the table. At a later time, if a packet is to be sent to an IP address that is not listed in the ARP Table, then an ARP Request must be sent to re-discover the MAC address.

Direct and Indirect Routes

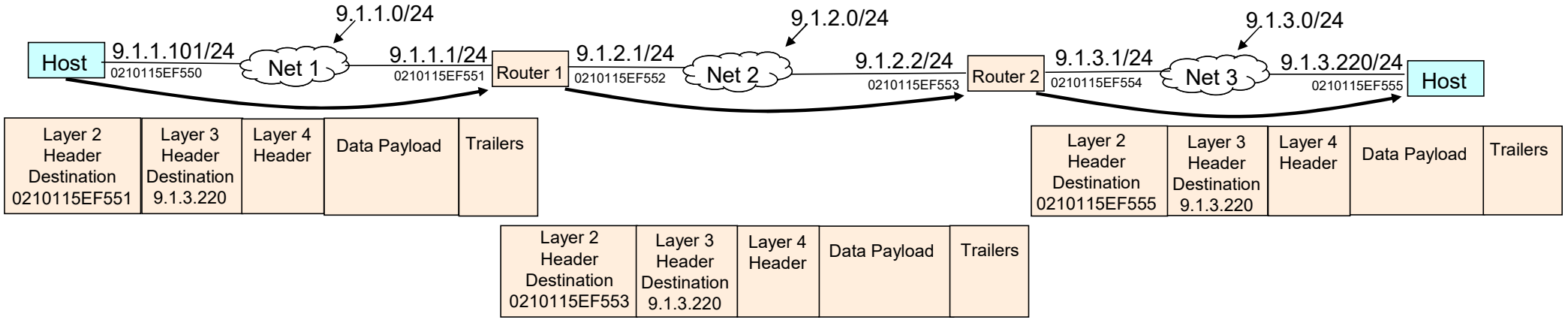


- For Host A to send data to Host B, 9.1.1.63, on the subnet that it is attached to (Direct Route), it uses ARP to learn the MAC address, 0210115EF559. The Layer 2 header is created with the destination 0210115EF559.



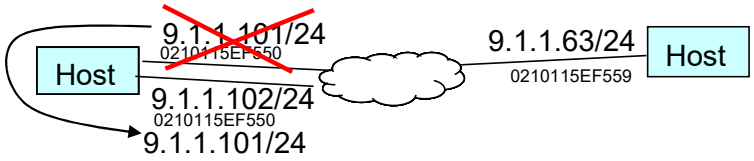
- For Host A to send data to Host C, 9.1.4.15, on remote subnet Net 4, the IP routing table indicates to send the packet to 9.1.4.15 (Indirect Route). ARP is used to learn the MAC address, 0210115EF557. It is put in the Layer 2 header as the destination.
- When the Router receives the packet it strips off the Layer 2 header, at Layer 3 the Router IP routing table determines the packet destination 9.1.4.15 is on the direct attached network. ARP is used to learn the MAC address, 0210115EF558. It is put in the new Layer 2 header as the destination.

Default Route

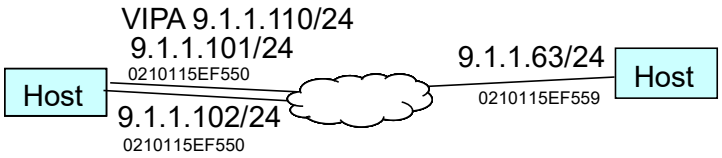


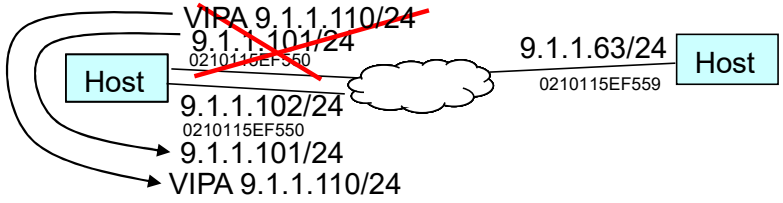
- For Host A to send data to Host D, 9.1.3.220, on remote subnet Net 3, the IP routing table indicates to send the packet to 9.1.1.1 (Default Route). ARP is used to learn the MAC address, 0210115EF551. It is put in the Layer 2 header as the destination.
- When the Router1 receives the packet, it strips off the Layer 2 header, at Layer 3 the Router IP routing table determines the packet destination 9.1.3.220 should be sent to 9.1.2.2. ARP is used to learn the MAC address, 0210115EF553. It is put in the new Layer 2 header as the destination.
- When the Router2 receives the packet, it strips off the Layer 2 header, at Layer 3 the Router IP routing table determines the packet destination 9.1.3.220 is on the direct attached network. ARP is used to learn the MAC address, 0210115EF555. It is put in the new Layer 2 header as the destination.

Gratuitous ARP Failover



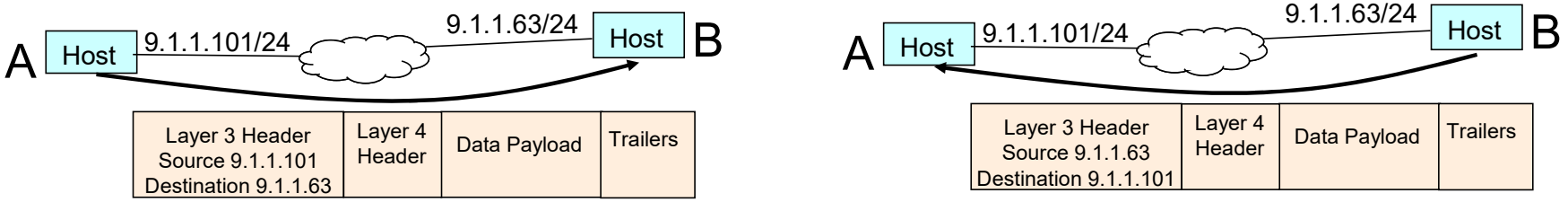
- The z/OS TCP/IP stack detects when multiple OSA ports activate to the same subnet. If one OSA fails the working OSA sends out a Gratuitous ARP (unsolicited ARP) message so that the failed OSA IP Address is associated with the working OSA MAC Address. In that case a single OSA owns two IP addresses associated with the single MAC Address. If the failed OSA recovers the IP Address moves back.
  - Dynamic Backup between OSA without Dynamic Routing (OSPF)
  - The support requires the Gratuitous ARP from the second OSA originally joining the LAN must be received over the first OSA so that the TCP/IP stack marks them as being in the same subnet. The OSAs show up as being in the same LAN Group on the Netstat Devlinks display.
  - If static routing is used, then automation should be configured to check that the OSAs are in the same LAN Group.





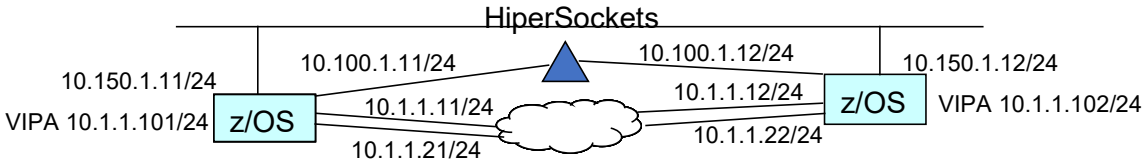
- If a VIPA is defined in the same subnet as multiple OSA ports, the first OSA that activates sends out a Gratuitous ARP so that the VIPA is associated with it. The OSA owns two IP addresses associated with the single MAC Address. If the OSA fails, both the OSA IP address and VIPA move to another OSA. Now all three IP addresses are associated with a single MAC Address. If the original OSA recovers the OSA port's IP address moves back but the VIPA remains on the second OSA port.

Packet Source and Destination IP Addr



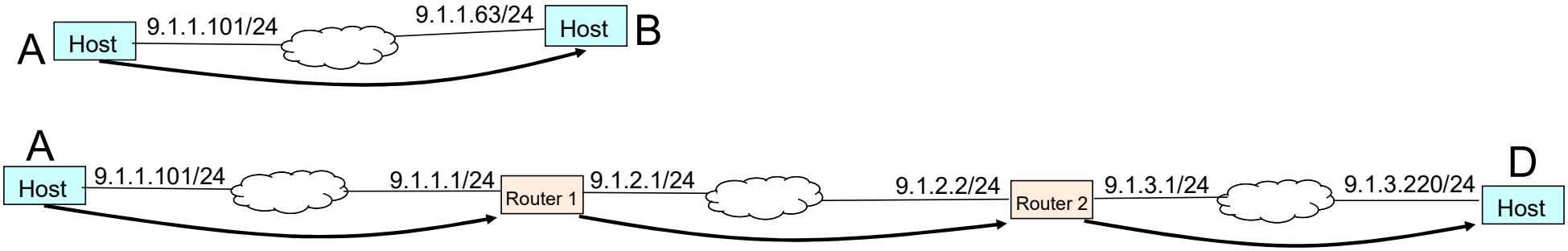
- The Source and Destination IP Addresses in the initial packet of a TCP session are determined as detailed on a previous page.
- Typically, the server reverses those addresses for the response. The Source in the initial packet is the Destination in the response packet. The Destination in the initial packet is the Source in the response packet.
- Typically, those IP Addresses remain the same for the duration of the connection.

Multiple Network Connections



- A Workstation, Laptop, and Linux on System z typically only have a single network connection defined to them. z/OS usually has multiple network connections:
  - Multiple OSA ports (10.1.1.0/24)
  - VIPA (10.1.1.0/24)
  - DynamicXCF (10.100.1.0/24)
  - HiperSockets (10.150.1.0/24)
- OSPF Dynamic Routing
  - OSAs should be in different subnets.
    - OSAs on other LPARs may be in the same subnet as an OSA on the local host.
  - VIPAs may be in a single subnet but that subnet should be different than the OSAs.
- Static Routing
  - OSAs and VIPAs should all be in the same subnet.
- When OSPF is being used and OSAs are in the same subnet, both OSPF Failover, and Gratuitous ARP Failover both occur, which can cause problems.

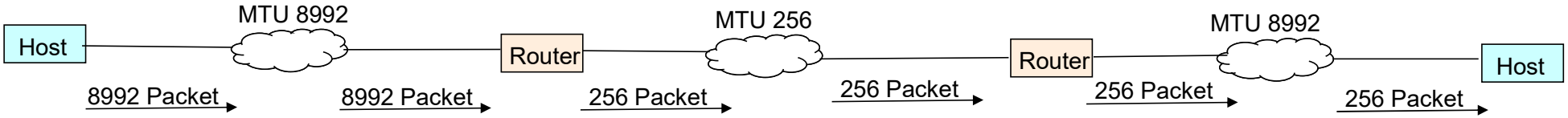
Routing Overhead



- There is an overhead associated with routing. All things being equal, performance is typically better between two hosts on the same subnet (sometimes referred to as a Layer 2 connection (or Flat network)), rather than two hosts on different subnets that connect using Layer 3 IP Routing.

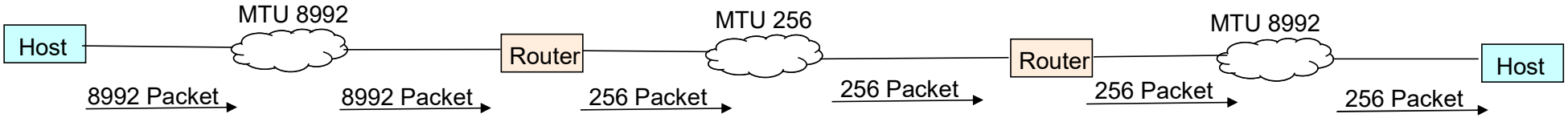
Maximum Transmission Unit (MTU)

Maximum Transmission Unit (MTU)



- Applications send data to TCP/IP to be sent out onto the network.
  - After the outbound interface is chosen using the Routing Table, the data is divided into pieces such that after all layer headers and trailers are added, the final packet is no larger than the defined MTU.
- All devices on the same subnet should all define the same MTU.
- The smallest MTU that may be defined is 256 bytes.
- If a packet is sent, but there is a smaller MTU subnet in the path to the destination, the router will divide the packet into smaller pieces such that they "fit" into the smaller MTU size. This is called Fragmentation.
  - If a 8992 byte packet is sent, but there is a 256 byte subnet in the path, the router will fragment the packet.
  - The message is reassembled at the destination host.

Jumbo Frames



- Reassembly of fragmented packets has processor overhead associated with it.
- When routers fragment packets, they send a TCP/IP Internet Control Message Protocol (ICMP) message back to the sender, indicating the smaller size so that their packet size can be lowered to the destination.
- Path MTU Discovery (IPCONFIG PATHMTUDISCOVERY) uses ICMP messages to discover the smallest MTU size in the path prior to sending any data to the destination.
- MTU 8992 bytes is also referred to as Jumbo Frames.
  - Path MTU Discovery is recommended when defining Jumbo Frames.
- Unfortunately, some networks have their routers block ICMP message flows.

OMPRoute

Static and Dynamic Routing

- Static routing is defined in the PROFILE.TCPIP file in the BEGINROUTE/ENDROUTE statement.
- Dynamic routing is defined in OMPROUTE configuration.
  - Routing Information Protocol (RIP)
    - RIP uses a distance vector algorithm to calculate the best path to a destination based on the number of hops in the path. RIP has several limitations. Several limitations that existed in RIP version 1 are resolved by RIP version 2.
    - RIP version 2 expands RIP version 1. Among the improvements are support for multicasting and variable subnetting. Variable subnetting allows the division of networks into variable size subnets.
  - Open Shortest Path First (OSPF)
    - OSPF uses a link state or shortest path first algorithm. OSPF's most significant advantage compared to RIP is the reduced time needed to converge after a network change. In general, OSPF is more complicated to configure than RIP and might not be suitable for small networks.

RIP Dynamic Routing

- RIP V2 is based on a distance vector algorithm.
- RIP V2 protocol extensions provide features such as the following items:
  - Route tags are used to separate internal RIP routes (routes for networks within the RIP routing domain) from external RIP routes, which might have been imported from an EGP (external gateway protocol) or another IGP (internal gateway protocol). OMPROUTE does not generate route tags, but preserves them in received routes and readvertises them when necessary.
  - Variable length subnet masks are included in routing information so that dynamically added routes to destinations outside subnetworks or networks can be reached.
  - Next hop IP addresses, when applicable, are included in the routing information. Their purpose is to eliminate packets being routed through extra hops in the network. OMPROUTE will not generate immediate next hops, but will preserve them if they are included in RIP packets.
  - An IP multicast address 224.0.0.9, reserved for RIP version 2 packets, is used to reduce unnecessary load on hosts that are not listening to RIP version 2 messages. RIP version 2 multicasting is dependent on interfaces that are multicast-capable.
  - Authentication keys can be included in outgoing RIP version 2 packets for authentication by adjacent routers as a routing update security protection. Likewise, incoming RIP version 2 packets are checked against local authentication keys. The authentication keys are configurable on a router-wide or per-interface basis.
  - Configuration switches are provided to selectively control which versions of RIP packets are to be sent and received over network interfaces. You can configure them router-wide or per-interface.
  - The supernetting feature is part of the Classless InterDomain Routing (CIDR) function. Supernetting provides a way to combine multiple network routes into fewer supernet routes. Therefore, the number of network routes in the routing tables becomes smaller for advertisements. Supernet routes are received and sent in RIP V2 messages.

OSPF Dynamic Routing

- The OSPF protocol is based on link-state or shortest path first technology.
- OSPF routing tables contain details of the connections between routers, their status (active or inactive), their cost (desirability for routing), and so on.
- Updates are broadcast when a link changes status, and consist merely of a description of the changed status. OSPF can divide its network into topology subsections, known as areas, within which broadcasts are confined.
- Features of OSPF are as follows:
  - OSPF supports variable length subnetting.
  - OSPF can be configured so that all its protocol exchanges are authenticated.
    - Only trusted routers can participate in an AS that has been configured with authentication.
  - Least-cost routing allows you to configure path costs based on any combination of network parameters. Bandwidth, delay, and metric cost are several examples.
  - There are no limitations to the routing metric. Although RIP restricts the routing metric to 16 hops, OSPF has virtually no restrictions.
  - Multipath routing is allowed. OSPF supports multiple paths of equal cost that connect the same points. These paths are then used for network load distribution, resulting in more use of the network bandwidth.
  - OSPF's area routing capability provides an additional level of routing protection and a reduction in routing protocol traffic.
- When defining OSPF, static routes may be defined as "replaceable", such that they will only be used if OSPF routes to the same destination do not exist.

Dynamic VIPA and Sysplex Distributor

- OSPF is preferred over RIP and Static Routing
  - OSPF recovery is faster
- OSPF Stub Area is recommended to avoid the overhead associated with routing.
  - If the first hop routers filter the routes sent to the mainframe so that only the DEFAULT route is sent, then Stub Area is not needed.
- Static Routing will work, and especially for simple networks, may be preferrable because it avoids the complication of implementing and supporting OSPF.
  - Gratuitous ARP Failover requires that the OSAs are detected in the same subnet.
  - Automation should be implemented.
    - D TCPIP,,NETSTAT,DEVLINKS
    - If the OSAs do not show the same LAN Group then they will not failover.
  - The first hop routers must implement VIPA backup.

Multipath and Cisco ACI

Multipath

- z/OS does not support network interface bonding. Where multiple interfaces are used for sending and receiving data, like a single device, with a single IP address assigned.
- Instead, z/OS supports load balancing outbound traffic using Multipath. This is independent of the source IP Address in the packet. In order to prevent out of order packets, all packets for any particular partner will be sent over a single OSA, but traffic to different partners will be “load balanced” (distributed in a round robin fashion) across all the available interfaces.

Cisco® Application Centric Infrastructure® (ACI®)

- Cisco® ACI® uses data-plane endpoint learning. It monitors ARP pairing (IP Address to MAC address). This can be helpful to notice problems in network traffic. However, it can be problematic when used with OSA static routing and Multipath. Because the source IP address in the packet can be associated with multiple MAC addresses.
  - The IP address appears to move from one MAC address to another and back again, referred to as “flapping”, which to Cisco® ACI® is interpreted as a problem.
  - Endpoint learning works well for relatively simple data hosts. However, VIPA, sysplex, and other z/OS networking functions introduce issues.

Cisco® ACI® Solution

- With Static Routing the following should be done:
  - On the z/OS side, Multipath may be disabled.
    - IPCONFIG NOMULTIPATH
      - Be aware that turning off Multipath limits OSA bandwidth, because only 1 OSA is used at any given time.
      - There may still be a “flapping” issue with Sysplex Distributor because traffic from the IP address outbound may be associated with multiple MAC addresses.
  - A better solution, that works with or without Sysplex Distributor, is to disable ACI® Endpoint Learning on the Cisco® Switch.
    - On the z/OS side, all traffic should use a VIPA.
      - Source VIPA for outbound traffic and inbound traffic directed to a VIPA.
    - On the Cisco® Switch the IP Data-plane Endpoint Learning option, under the VRF, should be disabled for z/OS VIPA address ranges.
- OSPF Dynamic Routing
  - ACI should work with OSPF.
  - If OSPF for dynamic routing, as provided by our OMPROUTE application, is defined instead of static routing, then the "flapping" will not be an issue because Cisco® ACI® sees the OSPF routes as outside the ACI® fabric.

More Information

Manuals

- z/OS Communications Server IP Configuration Guide, SC27-3650
- z/OS Communications Server IP Configuration Reference, SC27-3651
- Redbook IBM z/OS V2R1 Communications Server TCP/IP Implementation Volume 1: Base Functions, Connectivity, and Routing, SG24-8096
- Redbook TCP/IP Tutorial and Technical Overview, GG24-3376-05

End