

Networking for Embedded Switches and Routers: A Primer

From dedicated switches and routers to multi-layer switches, content switches and Web switches, manufacturers are advancing their technologies to keep up with the ever growing demands in traffic and sophistication in today's networks.

by Nauman Arshad, CWCEC and
S. Rajesh Kumar, Aricent

Ethernet is today's dominant network medium of choice, offering the best mix of simplicity, reliability and cost-effectiveness, across residential, business, enterprise and carrier networks. As more and more devices become networked and as the need to access, disseminate and share information continues to increase, networks and networking protocols have evolved to handle the complexity and the size of the information.

To begin, it is important to understand the difference between embedded switches and routers. This requires familiarity with the Open Systems Interconnection (OSI) model. The OSI is a specification published by the International Organization for Standardization (ISO) in 1984 that defines a seven-layered model, which simplifies complex network interactions by breaking them into simple modular elements (Figure 1). In this framework, each OSI layer can only communicate with the layer directly above it, below it, and with its peer layer on another device.

Following the typical flow of data between two communicating nodes in a network, a source node sends encapsulated data down the seven-layer OSI stack, across the

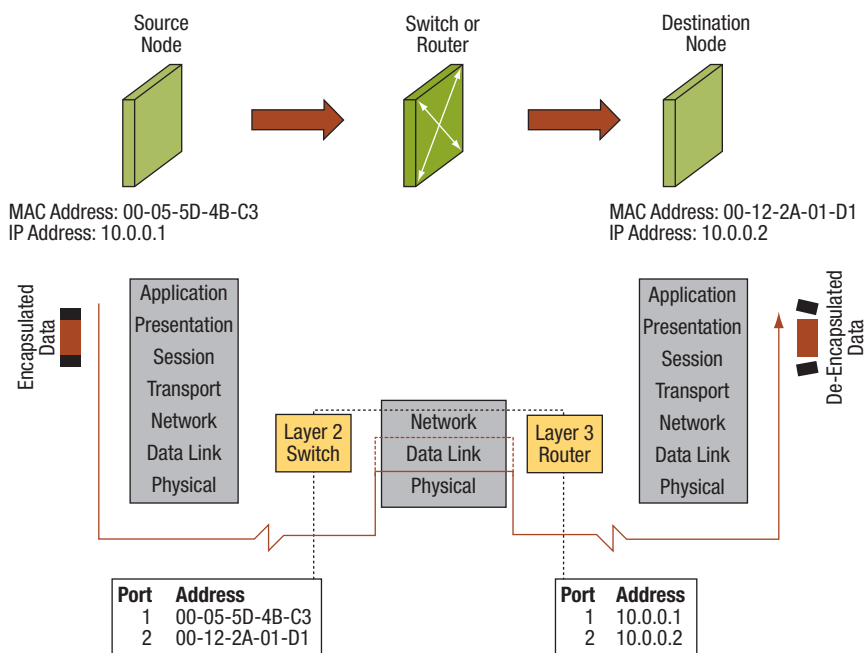


Figure 1 Illustration of seven OSI layers with function and example protocols discussed in this article, including sender, receiver, and where the physical cable, switches and routers fit. It also shows a MAC address and IP address. The layers discussed are:

- Data Link—Layer 2: VLAN, Link Aggregation, Port Mirroring, Spanning Tree Protocols
- Network—Layer 3: IPv4/v6, RIP, OSPF, IP Multicasting
- Transport—Layer 4: UDP/TCP
- Application—Layer 7: SNMP

network, and back up the stack to the destination node where it is de-encapsulated.

If the two nodes are connected in a network through a switch then the data first flows from the source node to an ingress port on the switch. The switch looks up the destination address embedded in the incoming data frame with an internal switching table. Upon a positive match, the switch forwards the data frame to the appropriate egress port on the switch, which is connected to the destination node. In Figure 1, the address the switch looks up is a unique 48-bit Media Access Control (MAC) address assigned to each node on the network. Switches typically operate at Layer 2 on the OSI stack.

Similarly, if the two nodes are connected in a network using a router, rather than the 48-bit MAC address, either a 32-bit IPv4 (Internet Protocol version 4) address or a 128-bit IPv6 (Internet Protocol version 6) address is used, with a forwarding policy to send the data to the right destination port. Routers typically operate at Layer 3 on the OSI stack.

Over the years switches have evolved to where the switching silicon can now switch at higher layers of the OSI stacks (e.g. Layers 3+) making the lines blurrier between switches and routers. These switches are also known as multi-layer switches, content switches or web switches.

LANs and VLANs

A Local Area Network (LAN) is a set of physically connected nodes in a small geographic area (e.g., a network in a home, office, group of buildings, vehicle etc.). A LAN can be divided into multiple “network segments” through the use of switches and routers for security purposes and to improve traffic flow by filtering out packets that are not destined for that segment.

A Virtual Local Area Network (VLAN) is the logical grouping of nodes (or switch ports) that behaves as though the nodes were connected on the same LAN segment regardless of their actual physical location. For example, in an office, the Finance Department data may be on one VLAN whereas the Engineering Department data resides on another. On a commercial aircraft, the in-flight entertainment data may be on one VLAN and data communications may be on another.

VLANs can be spread across switches.

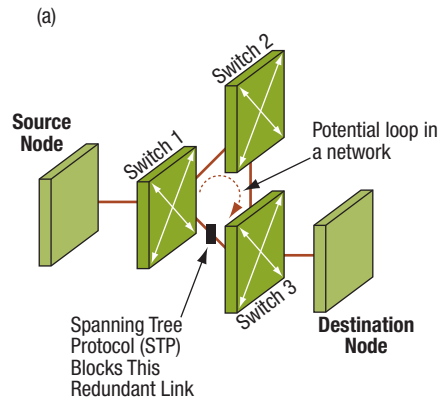


Figure 2 Spanning Tree Protocol (STP) – (a) shows how STP can be used to block the link between Switch 1 and Switch 3 to avoid a loop in the network; (b) shows how VLANs can be used with STP for efficient load balancing across unused links.

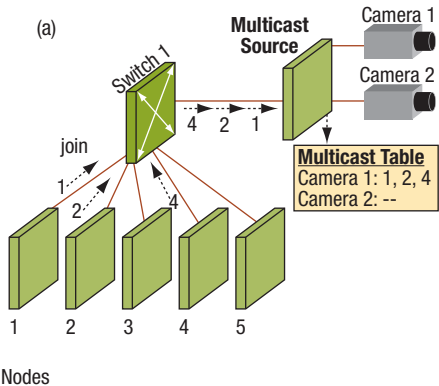
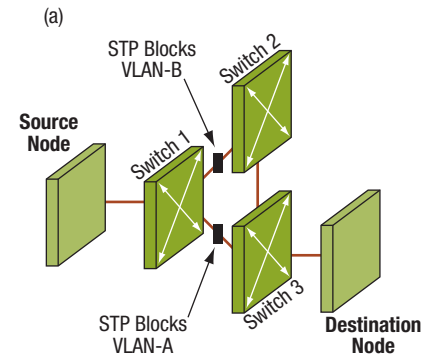
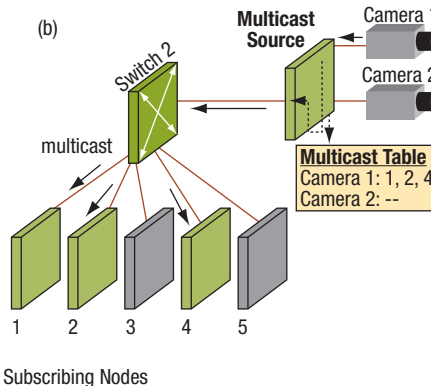


Figure 3 IP Multicasting (a) shows interested nodes joining a Multicast group for Camera 1; (b) shows Camera 1 multicasting data sent only to the subscribing nodes.



When the number of nodes in a network is large, or spread over an area that cannot be reached by a single switch, multiple switches are used. In such cases, it is quite possible that multiple paths exist between switches. Multiple paths can create loops in the network, which can be harmful because data packets may get duplicated and may potentially circulate on the network indefinitely, severely degrading network performance or at worst bringing down the network altogether.

Spanning Tree Protocols

The spanning tree protocol (STP) is a mechanism that enables switches to detect and prevent loops in a network. In Figure 2a, Switch 3 blocks the port connected to Switch 1 when the protocol detects that there are two paths connecting Switch 3 to Switch 1.

This blocking prevents the loop in the network. If the link between Switch 2 and Switch 3 fails, the STP automatically unblocks the link between Switch 1 and Switch 3. In this way, the STP also provides network resiliency. The original STP can require between 30 seconds and multiple minutes to restore alternate paths in a network, which makes it too slow for today's high-speed and mission-critical networks. The rapid spanning tree protocol (RSTP) is an enhanced variation of STP that offers 10X faster convergence time, enabling the network to recover much more rapidly from changes or failures. RSTP convergence times are typically less than one second for smaller networks.

With either RSTP or STP, there are unused links in the network. To better utilize the available capacity, the multiple STP provides a way to prevent loops, cre-

ate resiliency and increase network utilization by load balancing traffic across the alternate paths available. In Figure 2b, Switch 3 blocks the link to Switch 1 for VLAN A while Switch 2 blocks the link to Switch 1 for VLAN B. So, between the two nodes, traffic that is constrained in VLAN A is carried through the links connected to Switch 2 while the traffic constrained within VLAN B is carried over the link between Switch 1 and Switch 3, which provides load balancing and more effectively utilizes network capacity.

When multiple switches are used to connect nodes together in a network, the paths between the switches may require higher capacity than any individual link's speed. Also, in this case, the links between switches become more critical because they are concentrated points of failure. Link aggregation is a technique that helps mitigate the risk of failure by grouping multiple physical links into a logical channel. The resulting logical channel's capacity is thus the sum of the capacities of the individual physical links. Using link aggregation, if one link in the channel fails, the other links continue to function and traffic is moved to the working links, which provides redundancy and fault tolerance. Other common terms used to refer to an aggregated group are "port channel" or a "trunk."

One advantage of switches is that they enable network operators to monitor specific traffic via a technique called "mirroring." With mirroring, all traffic that requires monitoring is sent to a specific port on the switch. The network operator can monitor the network by connecting a monitoring node to that specific port to analyze the traffic going through the switch. Mirroring can be based on ports (port mirroring) or on specific flows (flow-based mirroring). Traffic across multiple ports can be monitored simultaneously using a single port. To ensure data is not lost, filtering is used on the source ports to ensure the data does not exceed the capacity of the port that is doing the monitoring.

IPv4/v6

IP (Internet Protocol) is a network layer protocol. Nodes are assigned IP addresses that are used to communicate with each other. The IP address has a subnet part and a host part. A subnet is a

group of nodes that is logically separated from other groups (much like a VLAN). Nodes within a subnet can directly communicate with each other. Routers are used to interconnect subnets. A popular switching scheme is to map IP subnets to VLANs. When nodes in different subnets need to communicate, they must go through a router. A Layer 3 switch is essentially a router. IPv4, the older version of IP, continues to be widely deployed and used today. IPv4 uses a 32-bit IP address for nodes. With the proliferation of IP-based devices, the 32-bit space was deemed to be insufficient. In response, a new version, IPv6, was developed, which uses a 128-bit addressing scheme. One of the key new features it offers over IPv4 is a much larger address space, enabling a greater proliferation of IP-based devices.

IP Multicasting

Multicast is a very common way to send information like TV or video streaming or public addresses, from one sender to many receivers. In its simplest form, a switch sends a multicast frame to all of the VLAN ports that the multicast came in on. The network operator can set up controls to restrict the multicast to specific ports if required.

A more popular approach for multicasting is for nodes to select what multicasts they want to receive. This is accomplished with the Internet Group Management Protocol (IGMP). Nodes that want to receive a multicast for a certain group send IGMP messages to join the group. By analyzing these IGMP join messages the multicast source is able to send the multicast only to the ports for which there are interested receiving nodes (Figure 3).

An implementation of a deployable switch that supports IP Multicasting is Curtiss Wright's SMS-682 SwitchBox II using Aricent's protocol stacks. With up to 24 ports of Gigabit Ethernet (GbE) and 2 ports of 10 GbE, the SwitchBox II has full support for IP Multicast. In addition, SwitchBox II has support for snooping on the IGMP messages ("IGMP snooping") sent from the receiver nodes to the sender node, enabling a switch to obtain information about which ports a multicast frame needs to be switched to.

RIP and OSPF

Layer 3 switches are used to interconnect IP subnets. To know which neighboring switch must be accessed to reach a particular subnet, the switches need to be configured with the reachability information. This information is called a "routing table." The greater the number of switches or subnets the more cumbersome and error prone the configuration becomes. As a remedy, Layer 3 switches use routing protocols to both advertise their own routing tables and learn the routing tables of other switches, which results in greater connectivity.

One such routing protocol is Routing Information Protocol (RIP). Using RIP, each switch simply sends updates to inform its neighboring switches which subnets it's directly attached to. Upon receiving an update, the neighboring switch adds the new data into its own routing table. It then sends its own information, along with the updates, to other neighbors, along with a "hop count." The hop count is the number of other switches that must be gone through to reach a subnet, and is limited to 16. Reliance on the hop count limits the scalability of RIP. It cannot handle large networks and is slow to react to changes in network topology.

An alternative approach, The Open Shortest Path First (OSPF) routing protocol overcomes the limitations of RIP. With OSPF, each switch sends its connectivity information to its neighbors. This information is sent in the form of a link state advertisement. Each switch adds the received link state advertisements to its own directly attached link states and sends all of these to its own neighbors. In this manner, every switch in a large network is able to learn the topology of the entire network and calculate the routing table. Because every switch knows the complete network topology, OSPF is extremely scalable and capable of handling large and complex networks.

TCP and UDP

Every node in the network, although having a single IP address, may support many applications that need to use network facilities. For example, computer

users on the network may have several Web browsing sessions, some gaming sessions and an Internet chat session, all using a single IP address. This is made possible through the use of transport layer protocols such as Transport Control Protocol (TCP) and User Datagram Protocol (UDP), which enable applications to share an IP interface. These transport layer protocols layer logical connections using logical port numbers on top of the same IP address. For example, the Web browser may use TCP port 1 and the gaming application may use TCP port 2, but with the same IP address. The difference between TCP and UDP is that TCP provides a more reliable communication channel by requiring acknowledgements for every transmitted packet. In comparison UDP is more unreliable but much faster.

SNMP

A popular method of managing network elements is through the Simple Network Management Protocol (SNMP). An SNMP manager program installed on a server or a monitoring node sends SNMP requests to the switches that the network operator wishes to manage. The SNMP requests may be messages to change the configuration of the switches or requests for the switches to provide information, such as traffic statistics, existing configuration, etc. The switches to be managed have an SNMP agent program that can process the request from the SNMP Manager. The switches, upon receiving SNMP requests, send back SNMP Responses with the requested information. The SNMP Responses also indicate if the manager's request was successful or unsuccessful. SNMP has 3 versions—version

1, version 2 and version 3. The latest version, SNMPv3, provides added security to the management scheme to protect the network from intentional attacks by hackers or inadvertent, but still costly damage from inexperienced users. ▲

Curtiss-Wright Controls Embedded Computing
Ottawa, Ontario.
(613) 599-9199.
[www.cwembedded.com].

Aricent
Palo Alto, CA.
(650) 391-1088.
[www.aricent.com].