

Gigabit Campus Network Design— Principles and Architecture

Introduction

The availability of multigigabit campus switches from Cisco presents customers the opportunity to build extremely high-performance networks with high reliability. Gigabit Ethernet and Gigabit EtherChannel® provide the high-capacity trunks needed to connect these gigabit switches. If the right network design approach is followed, performance and reliability are easy to achieve. Unfortunately, some alternative network design approaches can result in a network with lower performance, reliability, and manageability. With so many features available, and with so many permutations and combinations possible, it is easy to go astray. This paper is the result of Cisco's experience with many different customers and it represents a common sense approach to network design that will result in simple, reliable, manageable networks.

The conceptual approach followed in this paper has been used successfully in routed and switched networks around the world for many years. This hierarchical approach is called the "multilayer design." The multilayer design is modular and capacity scales as building blocks are added. A multilayer campus intranet is highly deterministic, which makes it easy to troubleshoot as it scales. Intelligent Layer 3 services reduce the scope of many typical problems caused by misconfigured or malfunctioning equipment. Intelligent Layer 3 routing protocols such as Open Shortest Path First (OSPF) and Enhanced Interior Gateway Routing Protocol (EIGRP) handle load balancing and fast convergence.

The multilayer model makes migration easier because it preserves the existing addressing plan of campus networks based on routers and hubs. Redundancy and fast convergence to the wiring closet are provided by Hot Standby Router Protocol (HSRP). Bandwidth scales from Fast Ethernet to Fast EtherChannel and from Gigabit Ethernet to Gigabit EtherChannel. The model supports all common campus protocols.

The multilayer model will be described, along with two main scalability options appropriate for building-sized networks up to large campus networks. Five different backbone designs with different performance and scalability are also presented. In this paper the term backbone is used to represent the switches and links in the core of the network through which all traffic passes on its way from client to server.

Structured Design with Multilayer Switching

The development of Layer 2 switching in hardware several years ago led to network designs that emphasized Layer 2 switching. These designs are characterized as "flat" because they avoid the logical, hierarchical structure and summarization provided by routers. Campus-wide virtual LANs (VLANs) are also based on the flat design model.

Layer 3 switching provides the same advantages as routing in campus network design, with the added performance boost from packet forwarding handled by specialized hardware. Putting Layer 3 switching in the distribution layer and backbone of the campus segments the campus into smaller, more manageable pieces. Important multilayer services such as broadcast suppression and protocol filtering are used in the Layer 2 switches at the access layer. The multilayer approach combines Layer 2 switching with Layer 3 switching to achieve robust, highly available campus networks.

It is helpful to analyze campus network designs in the following ways:

Public

Copyright © 1999 Cisco Systems, Inc. All Rights Reserved.

Page 1 of 21

Failure Domain

A group of Layer 2 switches connected together is called a Layer 2 switched domain. The Layer 2 switched domain can be considered as a failure domain because a misconfigured or malfunctioning workstation can introduce errors that will impact or disable the entire domain. A jabbering network interface card (NIC) may flood the entire domain with broadcasts. A workstation with the wrong IP address can become a black hole for packets. Problems of this nature are difficult to localize.

The scope of the failure domain should be reduced by restricting it to a single Layer 2 switch in one wiring closet if possible. In order to do this, the deployment of VLANs and VLAN trunking is restricted. Ideally one VLAN (IP subnet) is restricted to one wiring-closet switch. The gigabit uplinks from each wiring-closet switch connect directly to routed interfaces on Layer 3 switches. One way to achieve load balancing is to configure two such VLANs in the wiring-closet switch, which is shown later.

Broadcast Domain

Media Access Control (MAC)-layer broadcasts flood throughout the Layer 2 switched domain. Use Layer 3 switching in a structured design to reduce the scope of broadcast domains. In addition, intelligent, protocol-aware features of Layer 3 switches will further contain broadcasts such as Dynamic Host Configuration Protocol (DHCP) by converting them into directed unicasts. These protocol-aware features are a function of the Cisco IOS® software, which is common to Cisco Layer 3 switches and routers.

Spanning-Tree Domain

Layer 2 switches run spanning-tree protocol to break loops in the Layer 2 topology. If loops are included in the Layer 2 design, then redundant links are put in blocking mode and do not forward traffic. It is preferred to avoid Layer 2 loops by design and have the Layer 3 protocols handle load balancing and redundancy, so that all links are used for traffic.

The spanning-tree domain should be kept as simple as possible and loops should be avoided. With loops in the Layer 2 topology, spanning-tree protocol takes between 30 and 50 seconds to converge. So, avoiding loops is especially important in the mission-critical parts of the network, such as the campus backbone. To prevent spanning-tree protocol convergence events in the campus backbone, ensure that all links connecting backbone switches are routed links, not VLAN trunks. This will also constrain the broadcast and failure domains as explained previously.

Use Layer 3 switching in a structured design to reduce the scope of spanning-tree domains. Let a Layer 3 routing protocol, such as Enhanced IGRP or OSPF, handle load balancing, redundancy, and recovery in the backbone.

Virtual LAN

A VLAN is also an extended Layer 2 switched domain. If several VLANs coexist across a set of Layer 2 switches, each individual VLAN has the same characteristics of a failure domain, broadcast domain, and spanning-tree domain, as described above. So, although VLANs can be used to segment the campus network logically, deploying pervasive VLANs throughout the campus adds to the complexity. Avoiding loops and restricting one VLAN to a single Layer 2 switch in one wiring closet will minimize the complexity.

One of the motivations in the development of VLAN technology was to take advantage of high-speed Layer 2 switching. With the advent of high-performance Layer 3 switching in hardware, the use of VLANs is no longer related to performance. A VLAN can be used to logically associate a workgroup with a common access policy as defined by access control lists (ACLs). Similarly, VLANs can be used within a server farm to associate a group of servers with a common access policy as defined by ACLs.

IP Subnet

An IP subnet also maps to the Layer 2 switched domain; therefore, the IP subnet is the logical Layer 3 equivalent of the VLAN at Layer 2. The IP subnet address is defined at the Layer 3 switch where the Layer 2 switch domain terminates. The advantage of subnetting is that Layer 3 switches exchange summarized reachability information, rather than learning the path to every host in the whole network. Summarization is the key to the scalability benefits of routing protocols, such as Enhanced IGRP and OSPF.

In an ideal, highly structured design, one IP subnet maps to a single VLAN, which maps to a single switch in a wiring closet. This design model is somewhat restrictive, but pays huge dividends in simplicity and ease of troubleshooting.

Policy Domain

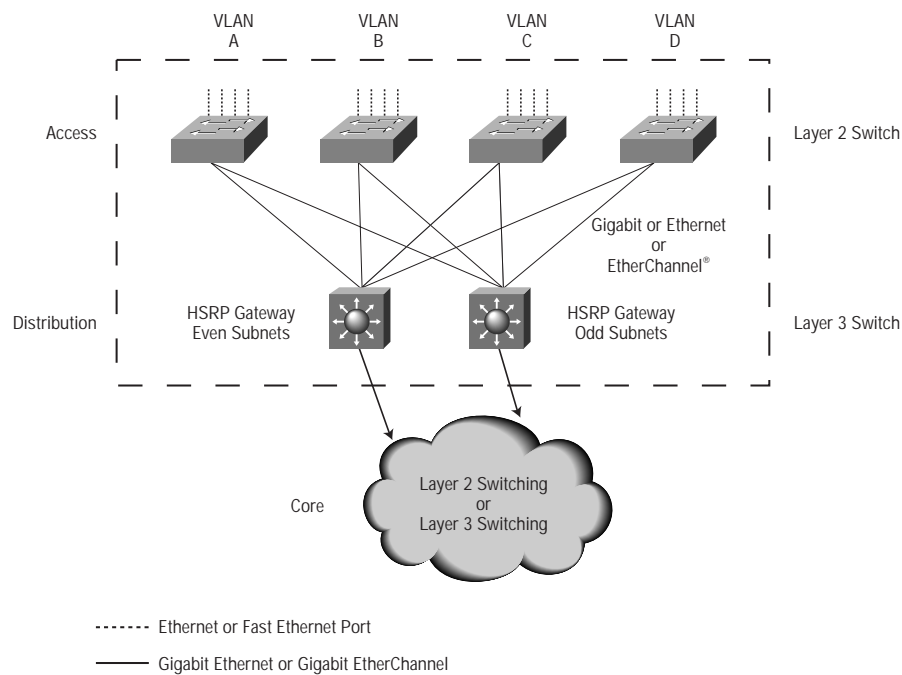
Access policy is usually defined on the routers or Layer 3 switches in the campus intranet. A convenient way to define policy is with ACLs that apply to an IP subnet. Thus, a group of servers with similar access policies can be conveniently grouped together in the same IP subnet and the same VLAN. Other services, such as DHCP are defined on an IP subnet basis.

A useful new feature of the Catalyst® 6000 family of products is the VLAN access control list (VACL). A Catalyst 6000 or Catalyst 6500 can use conventional ACLs as well as VACLs. A VACL provides granular policy control applied between stations within a VLAN.

The Multilayer Model

Two versions of the multilayer design model are discussed. Both versions are based on the building blocks shown in Figure 1. The building design is appropriate for a building-sized network with up to several thousand networked devices. The campus design is appropriate for a large campus consisting of many buildings. Both are based on a simple building block or module, which is the foundation of the modular design. The building design is described first because it is also used for each building within the campus design. To scale from the building model to the campus model, a campus backbone is added. Each building block or module connects to the campus backbone. Five different campus backbone designs are presented in the following section.

Figure 1 Building Block



Modular Design

The multilayer design is based on a redundant building block, also called a module. Gigabit Ethernet trunks connect Layer 2 switches in each wiring closet to a redundant pair of Layer 3 switches in the distribution layer. A single module is the basis of the building network as described in the next section. The modular concept can also be applied to server farms and WAN connectivity.

Redundancy and fast failure recovery is achieved with HSRP configured on the two Layer 3 switches in the distribution layer. HSRP recovery is 10 seconds by default, but can be tuned down as required. The cost of adding redundancy is on the order of 15 to 25 percent of the overall hardware budget. The cost is limited because only the switches in the distribution layer and the backbone are fully redundant. This extra cost is a reasonable investment when the particular building block contains mission-critical servers or a large number of networked devices.

In the model shown in Figure 1, each IP subnet is restricted to one wiring-closet switch. This design features no spanning-tree loops and no VLAN trunking to the wiring closet. Each gigabit uplink is a native routed interface on the Layer 3 switches in the distribution layer. Although this model is rugged, it is not the most general solution. If it is required that one VLAN span more than one wiring-closet switch, refer to the model described in the “Alternative Building-Block Design” section. The solution described in that section is more general and also supports distributed workgroup servers attached to the distribution-layer switches.

An optimal design features load balancing from the wiring-closet switch across both uplinks. Load balancing within the module can be achieved in several ways. For example, two IP subnets (two VLANs) can be configured on each wiring-closet switch. The distribution-layer switch on the left is designated the HSRP primary gateway for one subnet and the distribution-layer switch on the right is designated the HSRP primary gateway for the other subnet. A simple convention to follow is that the distribution-layer switch on the left is always HSRP primary for even-numbered subnets (VLANs) and that the distribution-layer switch on the right is always HSRP primary for odd-numbered subnets (VLANs).

An alternative way to achieve load balancing is to use Multigroup HSRP (MHSRP). With MHSRP, a single IP subnet is configured on a wiring-closet switch, but two different gateway router addresses are used. The Layer 3 switch on the left acts as the gateway router for half of the hosts in the subnet and the Layer 3 switch on the right acts as the gateway router for the other half.

With this design as described, packets from a particular host will always leave the building block via the active HSRP gateway. Either Layer 3 switch will forward returning packets. If symmetric routing is desired, configure a lower routing metric on the wiring closet VLAN interface of the HSRP gateway router. This metric will be forwarded out to Layer 3 switches in the backbone as part of a routing update, making this the lowest-cost path for returning packets. For a description of symmetric routing, see the section by the same name.

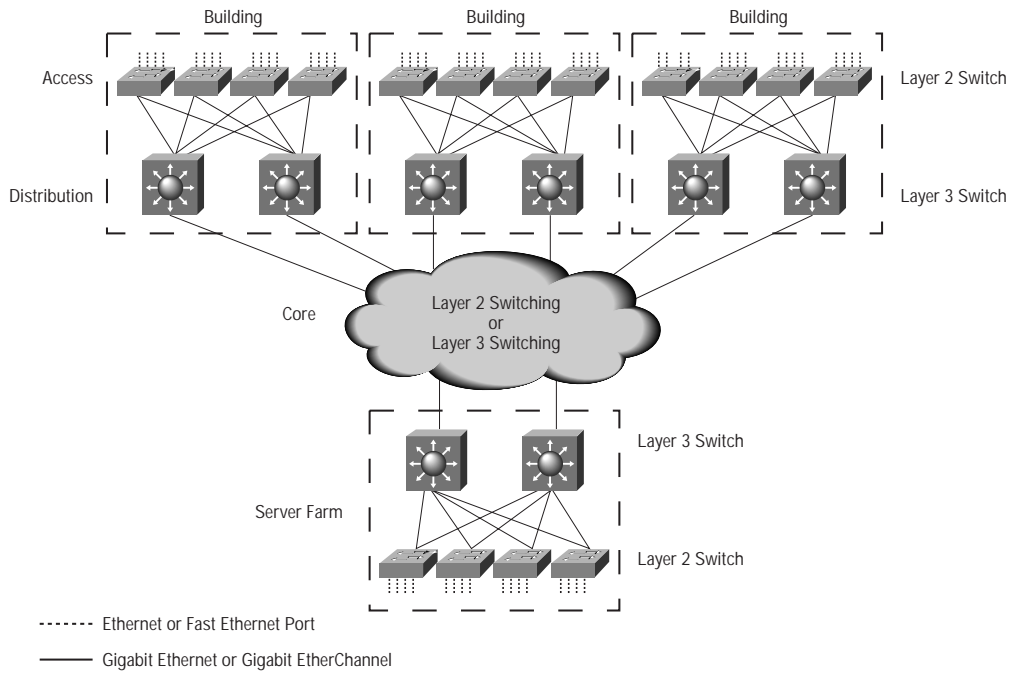
The Building Design

The building design shown in Figure 4 comprises a single redundant building block as defined in the previous section. The two Layer 3 switches form a collapsed building backbone. Layer 2 switches are deployed in the wiring closets for desktop connectivity. Each Layer 2 switch has redundant gigabit uplinks to the backbone switches. Alternatively, if one VLAN must span more than one wiring closet, or if distributed servers are to be attached to the distribution layer switches, please refer to the section “Alternative Building-Block Design,” which describes the more general building-block solution.

An optimization for the building design is to turn off routing protocol exchanges through the wiring closet subnets. To do this, use the passive interface command on the distribution-layer switches. In this configuration the distribution switches only exchange routes with the core switches and not with each other across the wiring closet VLANs. Turning off routing protocol exchanges reduces CPU overhead on the distribution-layer switches. Other protocol exchanges, such as Cisco Discovery Protocol (CDP) and HSRP, are not affected.

In the building design, servers can be attached to Layer 2 switches or directly to the Layer 3 backbone switches, depending on performance and density requirements.

Figure 2 Generic Campus Design



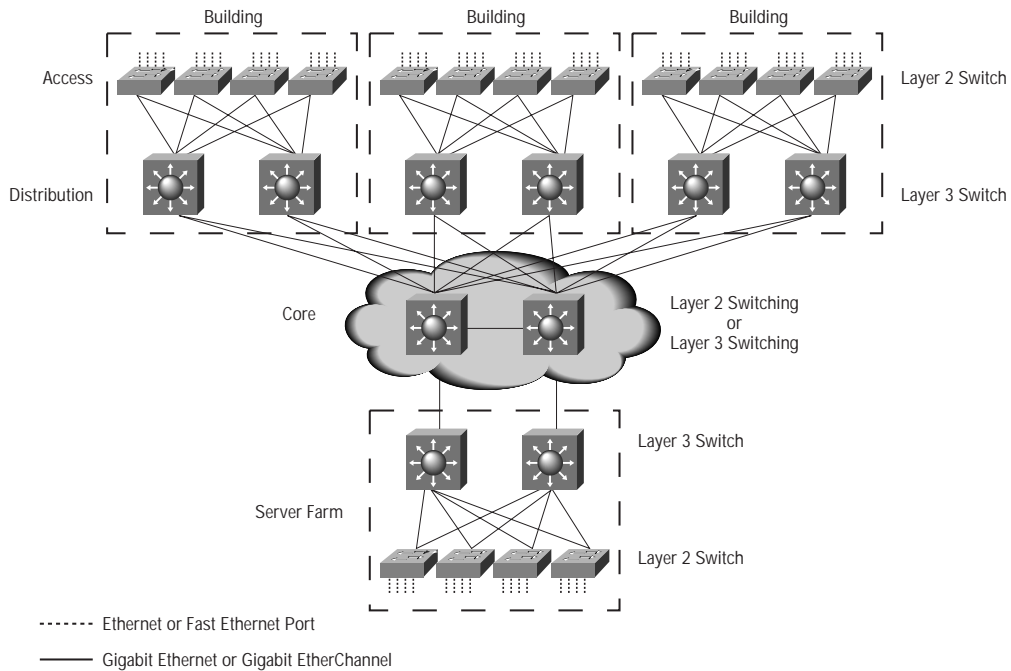
The Multilayer Campus Design

The multilayer campus design consists of a number of building blocks connected across a campus backbone. Several alternative backbone designs are discussed in the next section. See Figure 2 to view the generic campus design. Note the three characteristic layers: access, distribution, and core. In the most general model, Layer 2 switching is used in the access layer, Layer 3 switching in the distribution layer, and Layer 3 switching in the core.

One advantage of the multilayer campus design is scalability. New buildings and server farms can be easily added without changing the design. The redundancy of the building block is extended with redundancy in the backbone. If a separate backbone layer is configured, it should always consist of at least two separate switches. Ideally, these switches should be located in different buildings to maximize the redundancy benefits.

The multilayer campus design takes maximum advantage of many Layer 3 services including segmentation, load balancing, and failure recovery. IP multicast traffic is handled by Protocol Independent Multicast (PIM) routing in all the Layer 3 switches. Access lists are applied at the distribution layer for granular policy control. Broadcasts are kept off the campus backbone. Protocol-aware features such as DHCP forwarding convert broadcasts to unicasts before packets leave the building block.

Figure 3 Dual Path for Fast Recovery



In the generic campus model in Figure 2, each module has two equal cost paths to every other module. See Figure 3 for a more highly redundant connectivity model. In this model, each distribution-layer switch has two, equal cost paths into the backbone. This model provides fast failure recovery, because each distribution switch maintains two equal cost paths in the routing table to every destination network. When one connection to the backbone fails, all routes immediately switch over to the remaining path in about one second after the link failure is detected.

An alternative design that also achieves high-availability is to use the design model in Figure 2, but use EtherChannel links everywhere. EtherChannel achieves a high-degree of availability as well as load balancing across the bundle. A benefit of the Catalyst 6x00 is that availability can be improved further by attaching the links to different line cards in the switch. In addition, the Catalyst 6X00 and Catalyst 8500 products support IP-based load balancing across EtherChannel. The advantage of this approach, versus the design in Figure 3, is that the number of routing neighbors is smaller. The advantage of the design in Figure 3 is the greater physical diversity of two links to different switches.

Campus Backbone Design--Small Campus Design

Five alternative campus backbone designs will be described. These five differ as to scalability, while maintaining the advantage of Layer 3 services.

Collapsed Backbone—Small Campus Design

The collapsed backbone consists of two or more Layer 3 switches as in the building network. This design lends itself well to the small- to medium-sized campus network or a large building network, but is not recommended for a larger campus network. Scalability is limited primarily by manageability concerns. It is also a consideration that the Layer 3 switches in the backbone must maintain Address Resolution Protocol (ARP) entries for every active networked device in the campus. Excessive ARP activity is CPU-intensive and can affect overall backbone performance. From a risk and performance point-of-view it is desirable to break larger campus networks into several smaller collapsed modules and connect them with a core layer.

Figure 4 Building Design or Collapsed Backbone Model

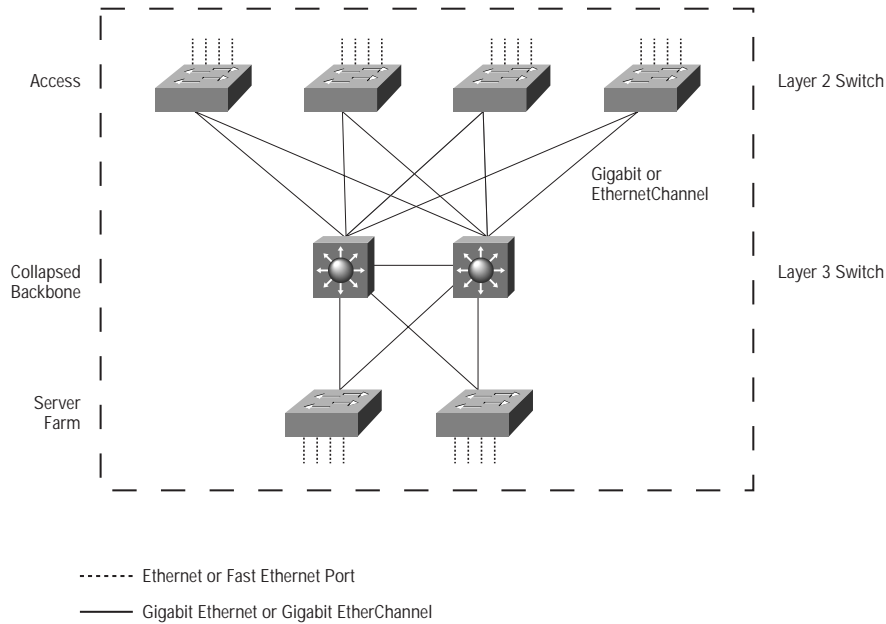
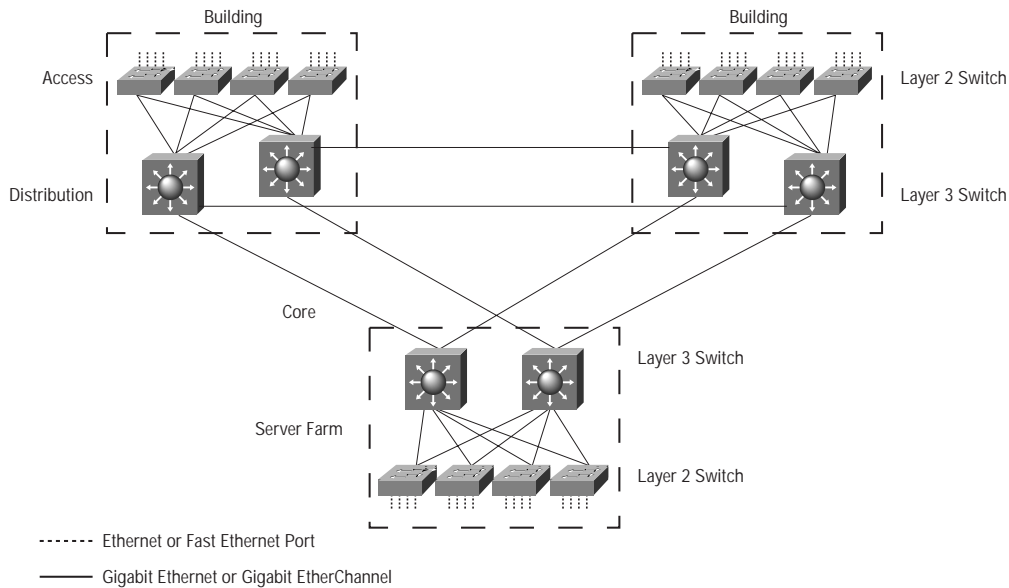


Figure 4 illustrates the collapsed backbone model. The server farm is incorporated directly into the collapsed backbone. Use the passive interface command on wiring closet subnet interfaces of the backbone switches to reduce routing protocol overhead.

Full-Mesh Backbone—Small Campus Design

Figure 5 Small Campus Network with Full-Mesh Backbone

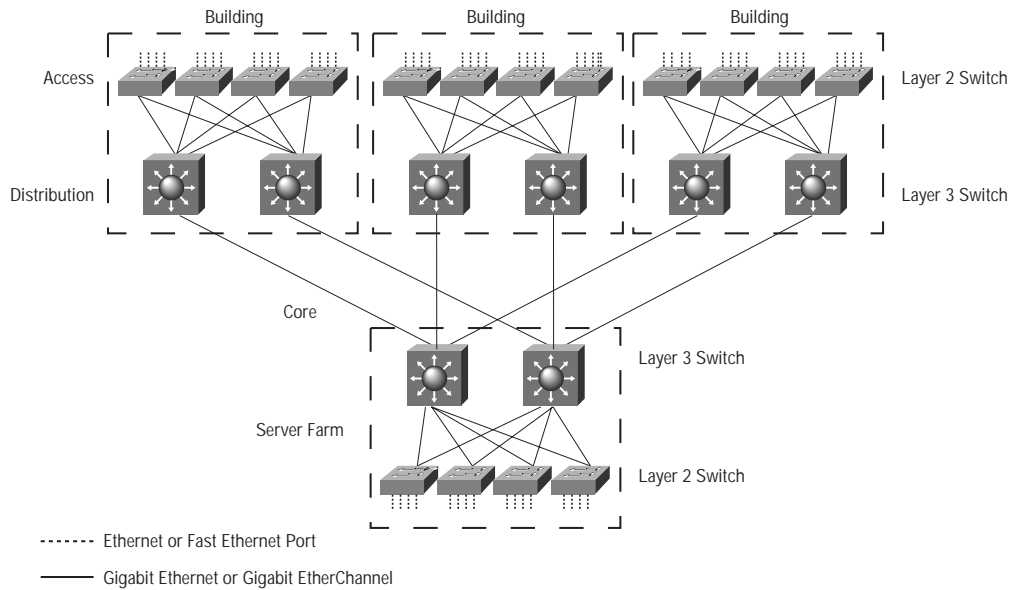


A full-mesh backbone consists of up to three modules with Layer 3 switches linked directly together forming a full-connectivity mesh. Figure 5 shows a small campus network with the full-mesh backbone. The full-mesh design is ideal for connecting two or three modules together. However, as more modules are added, the number of links required to maintain a full-mesh rises as the square of the number of modules. As the number of links increases, the number of subnets and routing peers also grows and the complexity rises.

The full-mesh design also makes upgrading bandwidth more difficult. To upgrade one particular module from Fast Ethernet links to Gigabit Ethernet links, all the other modules must be upgraded at the same time where they mesh together. Therefore, upgrades and changes are required everywhere. This approach is in contrast to using a dedicated Layer 2 or Layer 3 core to interconnect the distribution modules.

Partial Mesh—Small Campus Design

Figure 6 Partial-Mesh Campus Backbone

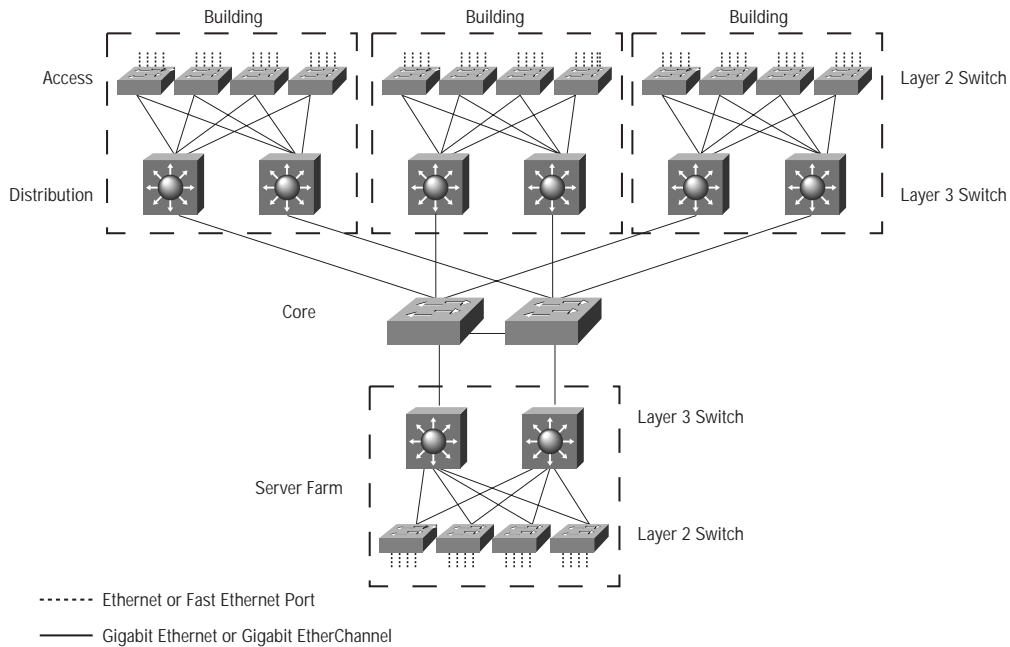


The partial-mesh backbone is similar to the full-mesh backbone with some of the trunks removed. Figure 6 depicts the partial-mesh campus backbone. The partial-mesh backbone is appropriate for a small campus where the traffic predominately goes into one centralized server farm module. Place high-capacity trunks from the Layer 3 switches in each building directly into the Layer 3 switches in the server farm.

One minor consideration with the partial-mesh design is that traffic between client modules requires three logical hops through the backbone. In effect, the Layer 3 switches at the server-farm side become a collapsed backbone for any client-to-client traffic.

Layer 2 Switched Backbone

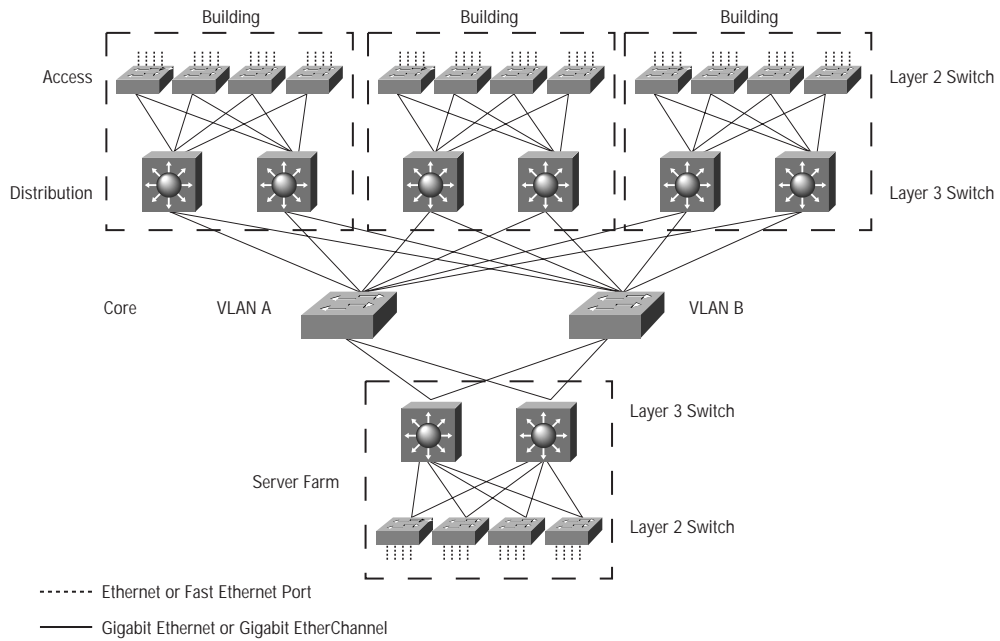
Figure 7 Layer 2 Switched Backbone—Single VLAN



The Layer 2 switched backbone is appropriate for a larger campus with three or more buildings to be connected. Adding switches in the backbone reduces the number of connections and makes it easier to add additional modules. Refer to Figure 7. The backbone is actually a single Layer 2 switched domain VLAN with a star topology. A single IP subnet is used in the backbone and each distribution switch routes traffic across the backbone subnet. Because there are no loops, spanning-tree protocol does not put any links in blocking mode, and spanning-tree protocol convergence will not affect the backbone. To prevent spanning-tree protocol loops, the links into the backbone should be defined as routed interfaces, not as VLAN trunks.

It is easy to avoid spanning-tree protocol loops with just two Layer 2 switches in the backbone as shown in Figure 7. However, this restriction limits the ultimate scalability of the Layer 2 backbone design. Another limitation is that all broadcasts and multicasts flood the backbone. Gigabit EtherChannel can be used to scale bandwidth between backbone switches without introducing a loop.

Figure 8 Split Layer 2 Campus Backbone

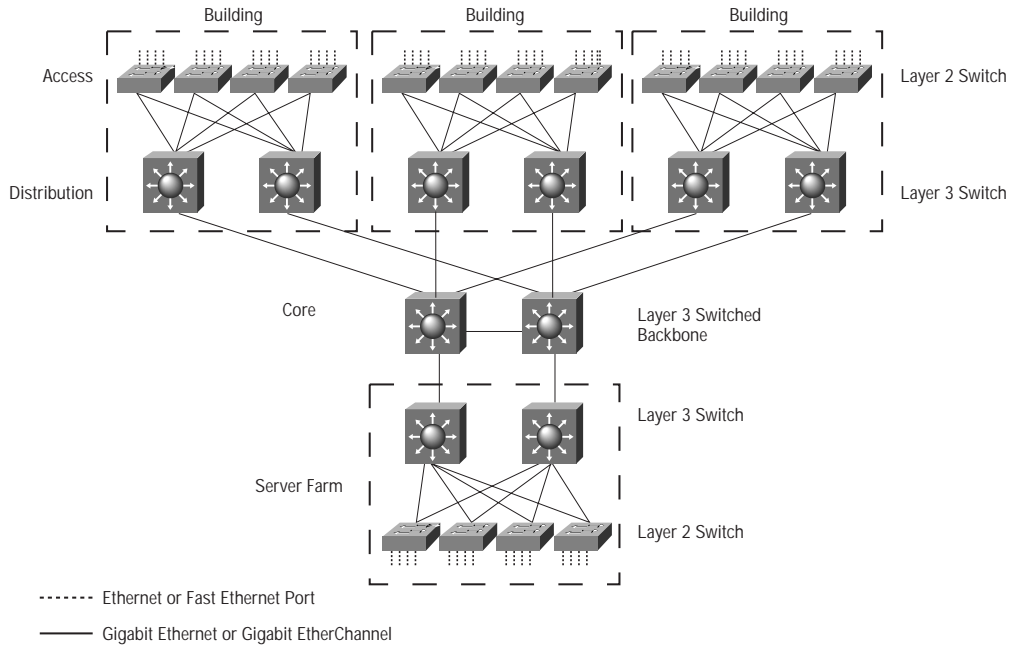


An alternative design with higher availability and higher capacity is shown in Figure 8. Here, two Layer 2 switched VLANs form two totally separate redundant backbones. Note that there is no trunk linking the switch marked VLAN A to the switch marked VLAN B. Each Layer 3 switch in the distribution layer now has two distinct equal-cost paths to every other distribution-layer switch, based on entries in the Layer 3 routing table. If the VLAN A path is disconnected, the Layer 3 switch will immediately route all traffic over VLAN B.

The advantage of the Split Layer 2 backbone design is that two equal-cost paths provide fast convergence. This high-availability advantage is possible because it is not limited by any protocol mechanisms, such as periodic hello packets. The extra cost of the dual-backbone design is associated with the extra links from each distribution switch to each backbone switch.

Layer 3 Switched Backbone

Figure 9 Layer 3 Switched Campus Backbone



The most flexible and scalable campus backbone consists of Layer 3 switches, as shown in figure 9. The backbone switches are connected by routed Gigabit Ethernet or Gigabit EtherChannel links. Layer 3 switched backbones have several advantages:

- Reduced router peering
- Flexible topology with no spanning-tree loops
- Multicast and broadcast control in the backbone
- Scalability to arbitrarily large size

Figure 10 Layer 3 Backbone—Dual Paths for Fast Recovery

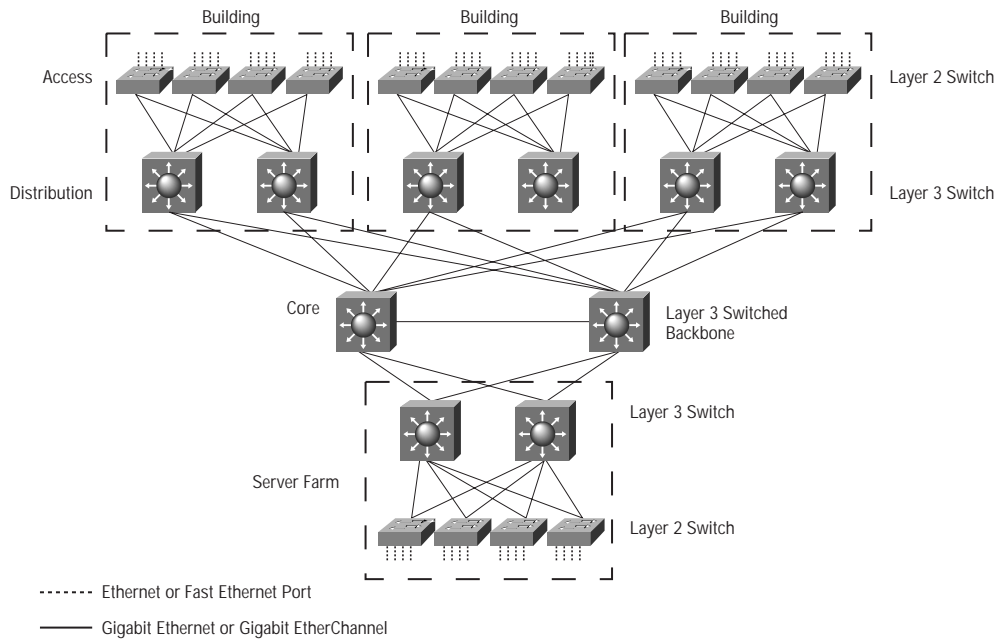
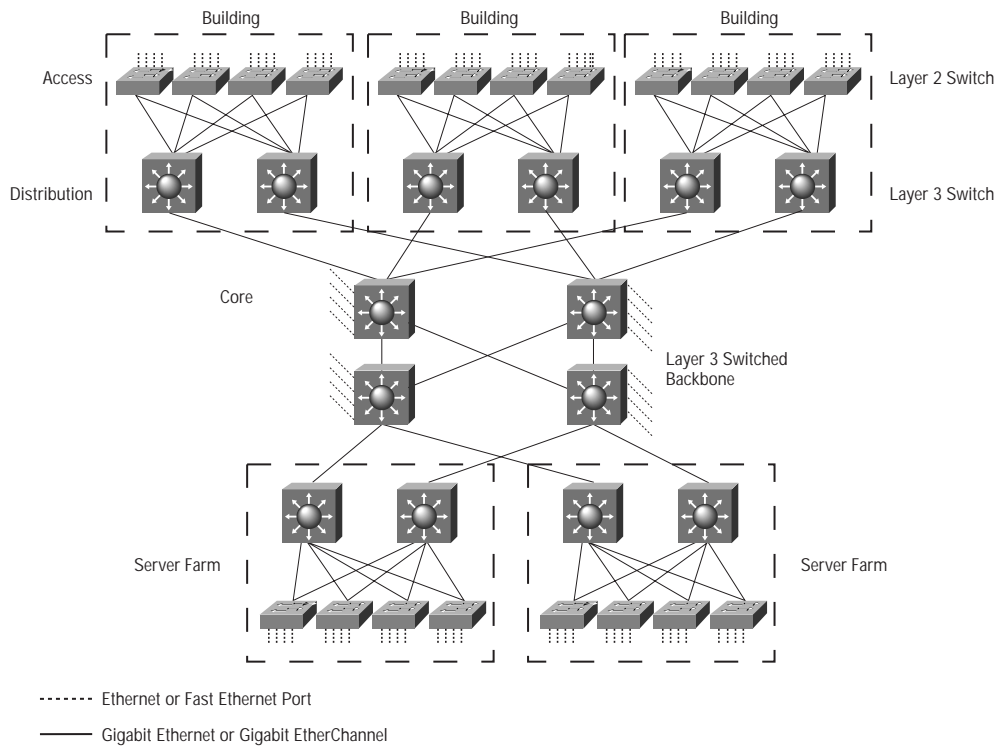


Figure 10 shows the Layer 3 switched campus backbone with dual links to the backbone from each distribution-layer switch. The main advantage of this design is that each distribution-layer switch maintains two equal-cost paths to every destination network, so recovery from any link failure is fast. This design also provides double the trunking capacity into the backbone.

Figure 11 Large-Scale Layer 3 Switched Campus Backbone






Figure 11 shows the Layer 3 switched campus backbone on a large scale. The Layer 3 switched backbone has the advantage that arbitrary topologies are supported because a sophisticated routing protocol such as Enhanced IGRP or OSPF is used pervasively. In Figure 11, the backbone consists of four Layer 3 switches with Gigabit Ethernet or Gigabit EtherChannel links. All links in the backbone are routed links, so there are no spanning-tree loops. The diagram suggests the actual scale by showing several gigabit links connected to the backbone switches. Note that a full mesh of connectivity between backbone switches is possible but not required. Consider traffic patterns when allocating link bandwidth in the backbone.

Scalable Bandwidth

Upgrading to EtherChannel can provide increased bandwidth and redundancy. An EtherChannel bundle is a group of up to eight Fast or Gigabit Ethernet links. The bundle functions as a single, logical connection between switches and routers. EtherChannel is convenient because it scales the bandwidth without adding to the complexity of the design. Spanning-Tree Protocol treats the EtherChannel bundle as a single link, so no spanning-tree loops are introduced. Routing protocols also treat the EtherChannel bundle as a single, routed interface with a common IP address, so no additional IP subnets are required, and no additional router peering relationships are created. The load balancing is handled by the interface hardware.

Use EtherChannel to link backbone switches, to connect the backbone to the distribution layer, and to join the distribution layer to the wiring closet.

High-Availability Considerations

High availability is a function of the application as well as the whole network between a client workstation and a service located in the network. While the mean time between failure of individual components is a factor, network availability is determined mostly by the network design. Adding redundancy to the distribution layer and the campus backbone is highly recommended and adds on the order 15 to 25 percent to the overall hardware budget.

Determinism is an important design goal. For the network to be deterministic, the design must be as simple and highly structured as possible. Recovery mechanisms must be considered as part of the design process. Recovery timing is determined in part by protocol messages such as hellos and keepalives, and these may need to be tuned to achieve recovery goals.

The following is an example of a multilayer design configured with all relevant mechanisms enabled for a high availability solution. Other white papers from Cisco provide a more detailed explanation of how these features work together to provide a high availability solution. The scenario below will give you an idea of the convergence numbers that can be achieved and the features that play a primary roll in each failure scenario.

Figure 12 High Availability and Recovery

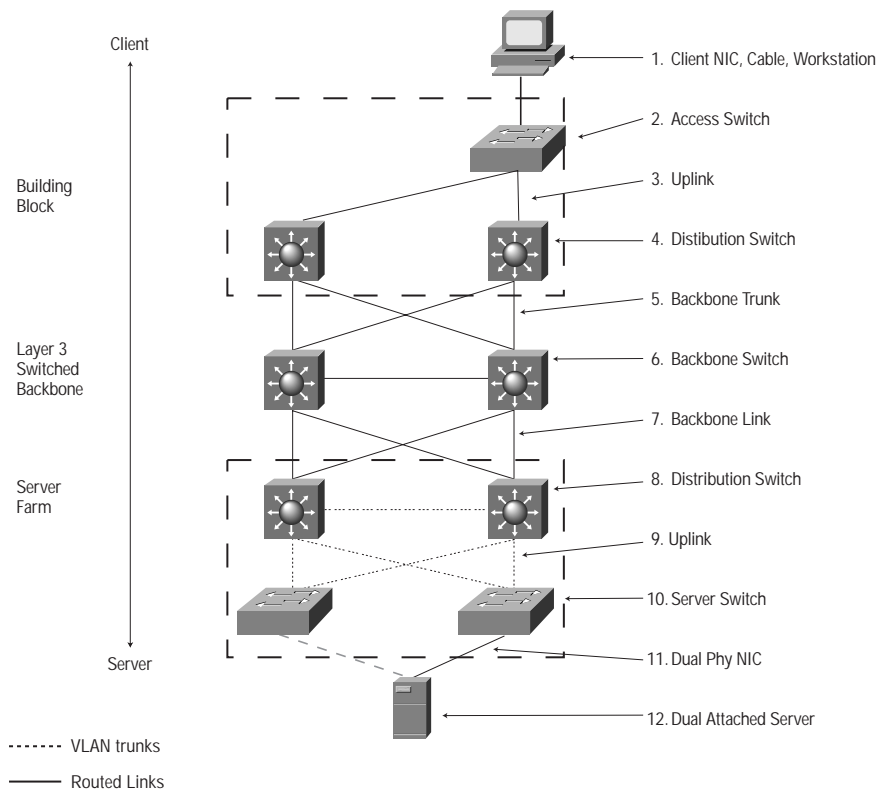



Figure 12 shows a two-way data flow between a client and a server. Different recovery mechanisms apply depending on the type of failure. The nonredundant parts are the client workstation, the NIC card on the workstation, the Ethernet cable from the workstation, as well as the dedicated port on the access switch. It is not generally considered cost-effective and practical to provide redundancy for these points in the design.

If the uplink number 3 fails, recovery is provided by HSRP, which can be tuned to recover in less than one second if required. If the distribution switch number 4 fails, the recovery is also HSRP. If one backbone link number 5 fails, equal-cost routing automatically uses the remaining link for all traffic to the core. If both uplinks to the core are lost, a feature called HSRP-track will move the default gateway services over to the neighboring distribution switch. If the backbone switch number 6 fails, the recovery is by the routing protocol Enhanced IGRP or OSPF. The hello timers for Enhanced IGRP can be tuned down to one second, giving recovery of backbone failures in about five seconds.

For most failures in the backbone, traffic will not be affected for more than one second as long as two equal-cost paths exist to any destination subnet. Equal-cost links are shown connecting each Layer 3 switch in Figure 12.

If one backbone link number 7 to the server farm is lost, equal-cost routing recovers immediately. If both backbone links are lost, HSRP-track recovers the gateway router function to the backup distribution-layer switch. Again HSRP can be tuned to recover in less than one second. If the distribution switch number 8 fails, several things occur simultaneously. The distribution switch is the root of odd-numbered spanning trees so spanning-tree protocol convergence is initiated. HSRP can be tuned to recover in less than one second. The routing protocol will converge in the backbone for any subnets in the server farm; however, recovery should be within one second as long as an alternative equal-cost path exists. If the uplink number 9 fails, the Layer 3 recovery is HSRP and the Layer 2 recovery is about three seconds with UplinkFast configured at the wiring-closet switch. If the server switch number 10 fails, the dual-phy NIC card fails over in about one second. If the cable to the server number 11 fails, the dual-phy NIC card also fails over in about one second.

PortFast enables the server switch to give immediate access to the port without waiting for spanning-tree protocol listening and learning to take place. PortFast should be enabled on all client and server ports.



Quality of Service for Voice and Video

Quality of service (QoS) for voice over IP (VoIP) consists of providing low-enough packet loss and low-enough delay so that voice quality is not affected by conditions in the network. The brute force solution is to simply provide sufficient bandwidth at all points in the network so that packet loss and queuing delay are small. A better alternative is to apply congestion management and congestion avoidance at oversubscribed points in the network.

A reasonable design goal for end-to-end network delay for VoIP is 150 milliseconds. At this level, delay is not noticeable to the speakers. To achieve guaranteed low delay for voice at campus speeds, it is sufficient to provide a separate outbound queue for real-time traffic. The bursty data traffic such as file transfers is placed in a different queue from the real-time traffic. Because of the relative high speed of switched Ethernet trunks in the campus, it does not matter much whether the queue allocation scheme is based on weighted round robin, weighted fair, or strict priority.

If low delay is guaranteed by providing a separate queue for voice, then packet loss will never be an issue. Weighted random early detection (WRED) is used to achieve low packet loss and high throughput in any queue that experiences bursty data traffic flows.

QoS maps very well to the multilayer campus design. Packet classification is a multilayer service that applies at the wiring-closet switch, which is the ingress point to the network. VoIP traffic flows are recognized by a characteristic port number. The VoIP packets are classified with an IP type of service (ToS) value indicating “low delay voice.” Wherever the VoIP packets encounter congestion in the network, the local switch or router will apply the appropriate congestion management and congestion avoidance based on the ToS value.

Multicast Routing and Control

The multilayer campus design is ideal for control and distribution of IP multicast traffic. The Layer 3 multicast control is provided by PIM routing protocol. Multicast control at the wiring closet is provided by Internet Group Membership Protocol (IGMP) snooping or Cisco Group Multicast Protocol (CGMP). Multicast control is extremely important because of the large amount of traffic involved when several high-bandwidth multicast streams are provided.

The Layer 3 campus backbone is ideal for multicast because Protocol Independent Multicast (PIM) runs on the Layer 3 switches in the backbone. PIM routes multicasts efficiently to their destinations along a shortest path tree. In a Layer 2 switched backbone on the other hand, all multicast traffic is flooded.

At the wiring closet, IGMP snooping and CGMP are multilayer services that prune the multicast traffic back to the specific client ports that join a multicast group. Otherwise all multicast traffic floods all ports, interrupting every client workstation.

One decision to be made in designing IP multicast routing is whether to use dense mode or sparse mode. Sparse mode is more efficient because dense mode periodically floods multicasts throughout the network and then prunes back based on client joins. The characteristic feature of sparse mode is a router acting as rendezvous point to connect multicast servers to multicast clients, which in turn establish a shortest-path tree. Routers use the rendezvous point to find sources of multicast traffic, and the rendezvous point instructs the last-hop designated router how to build multicast trees toward those sources. A rendezvous point and a backup rendezvous point should be chosen. With the Cisco auto rendezvous point feature, all other routers will automatically discover the rendezvous point. PIM should be configured pervasively on all IP routers in the campus.

A design goal for multicast networks is to place the rendezvous point and backup rendezvous point in the shortest path. If this is accomplished, there is no potential for sub-optimal routing of multicast traffic. Put the rendezvous point and the backup rendezvous point on the Layer 3 distribution switches in the server farm close to the multicast sources. This allows all state information to be prepopulated in the backup rendezvous point. If you have a redundant rendezvous point configured on the interior of your network, recovery is much slower and more CPU-intensive. Put loopback interfaces on each server distribution switch and configure them for multicast rendezvous point functionality.

A logical rendezvous point is a pair of Layer 3 switches or routers configured to act as a single redundant rendezvous point. Define a loopback interface on both switches with the same IP address. Both switches see all multicast packets for all sources and create a state for them. If the primary rendezvous point fails for any reason, the backup rendezvous point just needs to throw the right interfaces into forwarding to resume operation of the multicast network. For this to work, a well-defined Layer 3 topology is required. Each router in the network must

create just one entry in its routing table for this redundant loopback address. With the right topology and addressing, recovery takes under 10 seconds for most failures, and fallback is even less. The logical rendezvous point address for the group is advertised and the unicast routing protocol takes care of the rest.

Alternative Building-Block Design

Figure 13 Workgroup Server Design Detail

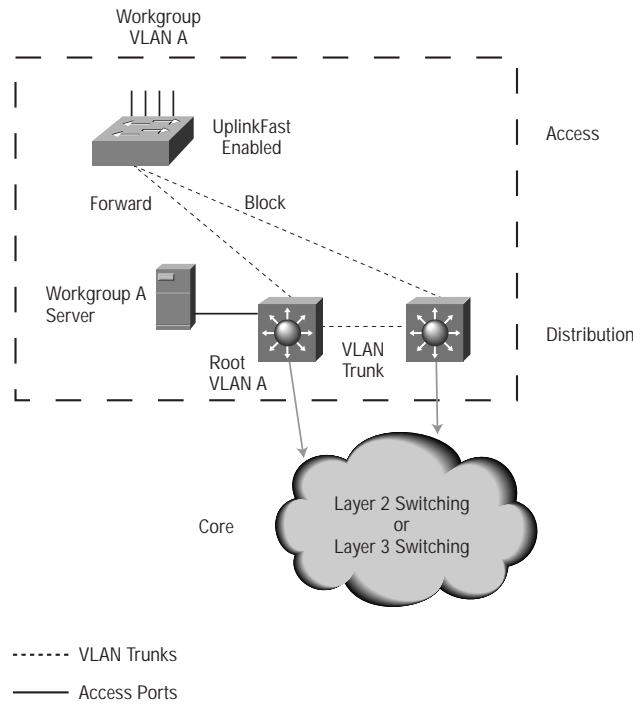


Figure 13 shows a more general building-block design with a workgroup server attached to the distribution-layer switch. This design assumes that the customer wishes to have the server for workgroup A in the same subnet and VLAN as the client workstations for policy reasons. To accomplish this safely, a VLAN trunk is placed between the distribution-layer switches. The VLAN for workgroup A now forms a triangle; hence spanning-tree protocol will put one link in blocking mode as shown. The triangle topology is required to maintain the integrity of the VLAN should one of the uplinks fail; otherwise a discontinuous subnet would result. A discontinuous subnet is a persistent traffic black hole that cannot be resolved and causes unreachability. In this case, the VLAN trunk becomes the backup recovery path at Layer 2. The distribution-layer switch on the left is made the spanning-tree root switch for all the even-numbered VLANs, and the distribution-layer switch on the right is made the spanning-tree root for all the odd-numbered VLANs.

It is important that the distribution-layer switch on the left is also the HSRP primary gateway for even-numbered VLANs so that symmetry between Layer 2 and Layer 3 is maintained. Fast Layer 2 spanning-tree recovery is achieved by enabling the UplinkFast feature on each wiring-closet switch as shown. If the forwarding uplink is broken, UplinkFast will put the blocking uplink into forwarding mode in about two seconds.

Figure 14 VLANs Span Wiring Closets

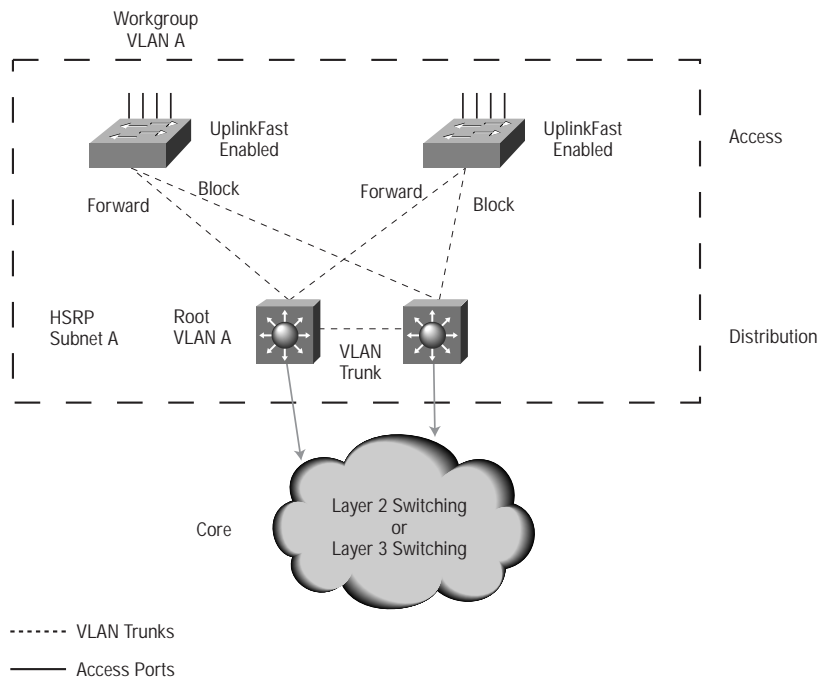


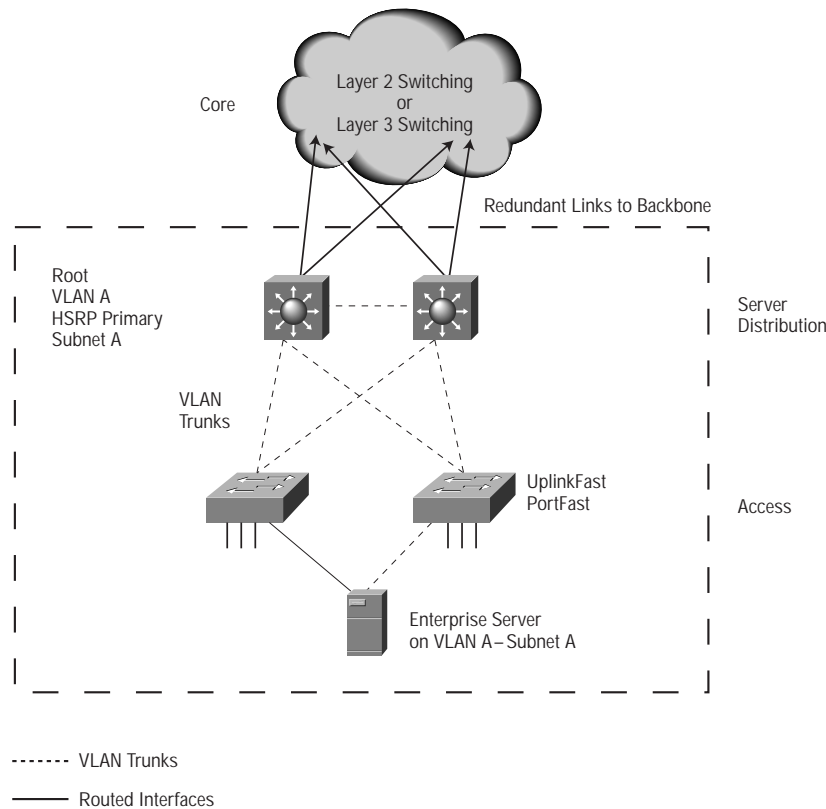
Figure 14 shows an even more general building-block design with workgroup VLAN A spanning several wiring-closet switches. To enforce traffic-flow determinism, the distribution-layer switch on the left is designated spanning-tree root for all even-numbered VLANs and the distribution-layer switch on the right is designated spanning-tree root for all the odd-numbered VLANs. The secondary STP root is also configured at the other distribution-layer switch so that a wiring-closet switch never becomes the root. Again HSRP must be configured so that Layer 2 traffic flows align with the Layer 3 configuration. As in the previous example, UplinkFast is configured on the wiring-closet switches to achieve fast recovery if a link is disconnected.

Server Farm Design

A server farm is implemented as a high-capacity building block attached to the campus backbone. The modular design approach described in the section “Multilayer Model” applies. One difference is that the server farm is an aggregation point for much of the traffic from the whole campus. As such, it makes sense to design the server farm with less oversubscription of switch and trunk resources than the normal building block. In fact, a nonblocking or non-oversubscribed design may be appropriate. For example, if the other campus building blocks connect to the backbone with Fast Ethernet, then the server farm should probably attach to the backbone with Gigabit Ethernet.

It is true that the server farm is the consolidation point for the entire enterprise as it hosts enterprise applications and therefore should have a lower contention ratio. However, one must consider not only the server-link bandwidth but also the actual ability of the server to transmit traffic. Although server manufacturers support a variety of NIC connection rates such as Gigabit Ethernet, the underlying network operating system may not be able to transmit at maximum capacity. As such, oversubscription ratios may be raised thereby reducing the overall cost of the server farm.

Figure 15 Dual-Attached Server



Server connectivity may be achieved in several different ways. For example, a server may attach by a single Fast Ethernet or by two. If the server is dual attached, one interface may be active while the other is in hot standby. Figure 15 illustrates this type of attachment. Using multiple single port or multiport NICs in the servers allows dual homing using various modes. Connecting servers in a dual-homed fashion allows the extension of the highly available server farm to the server itself. The server farm design presented in this document provides Layer 2 dual homing. Dual homing at Layer 3 introduces many complexities related to the network operating system (NOS), and will not be covered here.

Ports follow the design model described in the section “Alternative Building-Block Design” because the server attaches to the same VLAN A on two separate access switches. Note that one distribution-layer switch is designated as the primary HSRP gateway router for subnet A and also the spanning-tree root for VLAN A. Configure the access-layer switches with the PortFast feature.

Within the server farm, multiple VLANs are used to create multiple policy domains as required. If one particular server has a unique access policy, you may wish to create a unique VLAN and subnet for that server. If a group of servers has a common access policy, you may wish to place the whole group in a common VLAN and subnet. ACLs are applied on the interfaces of the Layer 3 switches.

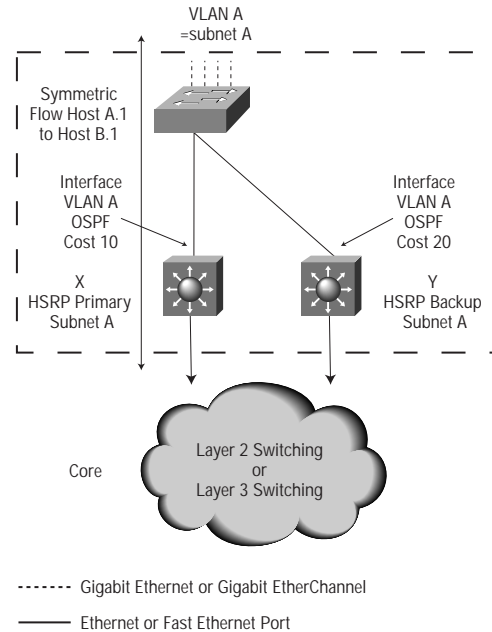
There are two main thresholds that must be considered when designing a server farm. Based on the traffic patterns and number of users of enterprise applications, packets will be generated at an average frequency and an average size. Interactive applications such as conferencing tend to generate high packet rates, but with small packet sizes. For these applications, the packet-per-second (pps) limit of the Layer 3 switches may be more critical than the throughput in terms of bandwidth. Applications involving large movements of data such as file repositories transmit a high percentage of full-length packets. For these applications, uplink bandwidth and oversubscription ratios become key factors in the overall design. Actual switching capacities and bandwidths will vary based on the mix of applications.

Symmetric Routing in the Campus Network

With redundant design traffic flows may follow two or more paths. The packets travelling from A to B may follow a different path than packets travelling back from B to A. In most cases this is of no particular concern. In some cases a high degree of determinism may be desired. The symmetric routing configuration shown in Figure 16 can be used to achieve this.

Routing protocol metrics are tuned to ensure that packets leaving a building block follow the same path as packets returning to the building block. Packets flow from station A.1 on VLAN A through its default gateway which is Layer 3 switch X. On Layer 3 switch X the routing metric on interface VLAN A is adjusted to make this path more favorable than the alternate return path through switch Y. If OSPF is the routing protocol, the interface cost metric is adjusted. If EIGRP is the routing protocol, the interface delay metric is adjusted.

Figure 16 Symmetric Routing Example



Product Capabilities

Refer to the matrix in Table 1 to match the capabilities of some of the Cisco campus switching products to the different parts of the multilayer design. As product features are constantly changing, please work with your Cisco account team to select the right products for your requirements.

Table 1 Cisco Products for Large-Scale Designs

Product Family	Wiring Closet	Distribution Layer	Server Distribution	Server Farm Fan Out	Backbone (Layer 2 or Layer 3)
Catalyst 2948G/ Catalyst 2980G	Medium density fixed configuration Layer 2				
Catalyst 3500	Stackable Layer 2 with multilayer services				
Catalyst 4000	Modular Layer 2 with Layer 3/QoS downlinks			Modular Layer 2	
Catalyst 2948G-L3/ Catalyst 4908G-L3		Wirespeed fixed configuration multilayer (L2/3) switching	Multigigabit Layer 3		
Catalyst 5500	High density modular multilayer switch	High density modular multilayer switch with RSM or RSFC			
Catalyst 6000	Modular Layer 2 with a multilayer services in wiring closet	Modular Layer 3 with MSFC		Modular Layer 2	Modular Layer 2 or Layer 3 with MSM or MSFC
Catalyst 6500		Modular multigigabit Layer 3 with MSFC	Multigigabit Layer 3/4 with MSFC		Multigigabit Layer 3 with MSFC
Catalyst 8500					Multigigabit Layer 3 switching and ATM switching



Corporate Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
<http://www.cisco.com>
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters

Cisco Systems Europe s.a.r.l.
Parc Evolic, Batiment L1/L2
16 Avenue du Quebec
Villebon, BP 706
91961 Courtaboeuf Cedex
France
<http://www-europe.cisco.com>
Tel: 33 1 69 18 61 00
Fax: 33 1 69 28 83 26

Americas Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
<http://www.cisco.com>
Tel: 408 526-7660
Fax: 408 527-0883

Asia Headquarters

Nihon Cisco Systems K.K.
Fuji Building, 9th Floor
3-2-3 Marunouchi
Chiyoda-ku, Tokyo 100
Japan
<http://www.cisco.com>
Tel: 81 3 5219 6250
Fax: 81 3 5219 6001

Cisco Systems has more than 200 offices in the following countries. Addresses, phone numbers, and fax numbers are listed on the Cisco Connection Online Web site at <http://www.cisco.com/offices>.

Argentina • Australia • Austria • Belgium • Brazil • Canada • Chile • China • Colombia • Costa Rica • Croatia • Czech Republic • Denmark • Dubai, UAE Finland • France • Germany • Greece • Hong Kong • Hungary • India • Indonesia • Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Singapore Slovakia • Slovenia • South Africa • Spain • Sweden • Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela



Refer to the matrix in Table 1 to match the capabilities of some of the Cisco campus switching products to the different parts of the multilayer design. As product features are constantly changing, please work with your Cisco account team to select the right products for your requirements.