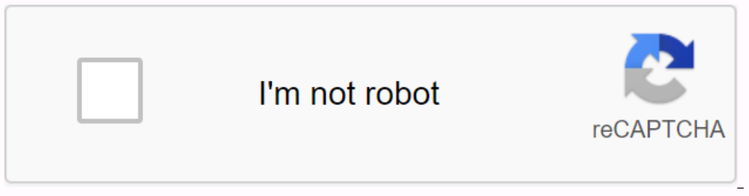

Continue

Data center design overview pdf

This section discusses the following topics: Types of server farms and data centers Data center topology Completely unnecessary Layer 2 and Layer 3 designs No superfluous Level 2 and Level 3 designs with services This section focuses on three main features of the data center architecture: scaling, flexibility, and high availability. Data centres are developing rapidly to meet higher growth, consolidation and security expectations. While traditional Layer 2 and Layer 3 designs have not changed drastically over the past few years, strict requirements for insanity and service availability, along with new technologies and protocols make design efforts more difficult and demanding. Scaling, flexibility, and high availability needs can be summarized as follows: Scaling – Data center must support fast and smooth growth without major disruptions. Flexibility – The data centre must support new services without a major change in its infrastructure. High availability – the data centre must not have a single failure point and must offer predictable uptime (for serious failures). NOTE A serious failure is a failure when the component needs to be changed to return to working constant state. Scaling turns into an opportunity to maintain rapid productivity growth, the number of devices under the auspices of the data center, and the amount and quality of services offered. Higher productivity means a tolerance to very short-term changes in flow patterns without loss of packages and long-term plans linking growth trends to data center capacity. Scaling the number of hosted devices means that you can seamlessly add more ports to servers, routers, switches, and other service devices, such as server load balancers, firewalls, IDS, and SSL dump devices. Higher density also includes the density of slots, as the number of slots ultimately leads to potential system growth. Flexibility turns into designs that have been adapted to new service offers without requiring a complete overhaul of architecture or drastic changes, except for normal maintenance periods. The flexibility approach is a modular design that contains known characteristics of modules, and the steps to add more modules are simple. High availability means a completely unnecessary architecture in which all possible serious disturbances are predictable and deterministic. This means that the failure of each potential component has a predetermined failover and fallback time, and that the worst-case scenario failure condition still meets acceptable failover limits and meets the requirements measured by the app's availability. This means that while the time of failure and recovery of the network component should be predictable and known, the more important the time involves the user's perception of the time taken to restore the application service. NOTE After failure, recovery the perspective of a Level 2 environment (covering wood) or layer 3 perspective (directed network) can be measured, but the availability of the app is ultimately important for the user. If the failure is such that the user connection time-out bing, then, regardless of the convergence time, the network convergence does not meet the requirements of the program. In the data center design, it is important to measure recovery times from both a network and an application point of view to ensure predictable user network recovery times (application service). Figure 4-1 provides an overview of the Data Center, which, as a device, contains several components and components of the larger enterprise's network architecture. These books relate in particular to the engineering of the application environment and their integration into the rest of the enterprise network. Different types of server farms support application environments, but this book focuses on understanding, designing, deploying, and maintaining server farms that support intranet application environments. The actual engineering of different types of server farms – internet, extranet and intranet server farms – is no different from type to type; but their integration with the rest of the architecture is different. Design choices that differ for each type of server farm are the result of their main functional purpose. This determines the specific location of their location, for safety reasons, overeffectability, scalability and efficiency. In addition to server farm concepts, a brief discussion about server farm types explains these points in more detail. NOTE The numbers in this section include a wide range of Cisco icons. Link to the icons used in this book (before the introduction) icon list and their descriptions. Figure 4-1 Overview of server farms and data center data center data center topology types As shown in Figure 4-1, there are three different types of server farms: internet extranet intranet All three types live in a data center and are often in the same data center tool, commonly referred to as a business data center or enterprise data center. If the sole purpose of a data center is to support online applications and server farms, the Data Center is called an Internet Data Center. Server farms are located in the center of the data center. In fact, data centers are built to support at least one type of server farm. Although different types of server farms have many architectural requirements, their objectives vary. Thus, a specific set of data center requirements depends on which type of server farm is to be maintained. Each type of server farm has separate infrastructure, security, and management requirements that must be addressed when building a server farm. Although each server farm design and its specific topology may vary, the design guidelines all of them. The following sections provide server farms. Web server farms As their name suggests, web server farms are exposed to the Internet. This means that users who use server farms are primarily somewhere on the Internet and use the Internet to access the server farm. Web server farms are available to the internet community in general and support enterprise user services. Typically, internal users also have access to web server farms. Server farm services and their users rely on the use of web interfaces and web browsers, so they spread in the Web environment. There are two different types of web server farms. Figure 4-2 As a rule, the main function of the business is based on the presence of the Internet or internet marketing. In general, dedicated web server farms maintain the company's e-business goals. Architecturally, these server farms adhere to the data center architecture entered in Section 1 data center overview, but the data for each layer and required layers is determined according to the requirements of the application environment. Security and scalability are the main concern of this type of server farm. On the one hand, most users who connect to the server farm are online, thus presenting a higher security risk; on the other hand, the number of potential users is very high, which can easily cause scaling problems. A data center that supports this type of server farm is often referred to as an Internet Data Center (IDC). IdCs is developing both companies to support their e-business infrastructure and service providers selling hosting services, thus enabling businesses to merge e-business infrastructure on the provider's network. Another type of Web server farm, shown in Figure 4-3, is designed to support online applications, as in addition to Internet access from the company. This means that the infrastructure that supports server farms is also used to maintain internet access for the company's users. These server farms are usually located in a demilitarized zone (DMZ) because they are part of the enterprise's network, but they are accessible from the Internet. These server farms are called DMZ server farms to separate them from dedicated web server farms. Figure 4-2 Dedicated web server farms These server farms support services such as e-commerce and are the door of access to portals for more generic applications used by both Internet and intranet users. Scalability aspects depend on the extent to which the intended user base is projected. Security requirements are also very strict, as security policies aim to protect server farms from external users while maintaining the security of the company's network. Please note that according to this model, the company's network private farm for WAN and intranet servers. NOTE that Figure 4-3 shows a small number of servers in the segment from the firewall. Depending on the requirements, a small number of servers can become hundreds or thousands, which would change the topology to include a set of 3-layer switches and as many layers of 2 switches for server connection as needed. Figure 4-3 DMZ Server Farms Intranet Server Farms Client/Server Model Evolution and Extensive Web Application Adoption on the Internet have been the basis for intranet creation. Intranet server farms are similar to web server farms for easy access, but they are only available to the company's internal users. As described earlier in this section, intranet server farms include most of the computer resources that are important for the enterprise that support business processes and internal applications. This list of critical resources includes mid-range and large computer systems that support a wide range of applications. Figure 4-4 illustrates the holding of intranet servers. Note that the intranet server farm module is connected to the main switches that form part of the company's main base and provide a connection between private WAN and Internet edge modules. Users who use an intranet server farm are on the university campus and on a private WAN. Internet users are generally denied access to the intranet; however, internal users using the Internet as transportation have access to the intranet using virtual private network (VPN) technology. Figure 4-4 Intranet Server Farms Internet Edge Module supports several features that include: Enterprise Network Protection Internet Access Management from intranet Internet access server farm management Data Center provides additional security to further protect intranet server farm data. This is achieved by applying security policies to the edge of the data center, as well as for the application tiers that apply when you try to harden the connection between different tiered servers. The security design applied to each step depends on the architecture of the applications and the desired level of security. Enterprise user access requirements dictate the size and architecture of server farms. The increasing number of users, as well as the increased load imposed by rich programs, increase the demand on the server farm. This requires scale to become a critical design criterion, as well as high accessibility, security and management. Extranet server farms Functionally extranet server farms sit between internet and intranet server farms. Extranet server farms continue to use web applications, but unlike Internet or intranet server farms, they are accessed only by a selected user group that is neither online nor intranet-based. Extranet server farms are mostly available partners who are considered to be external but reliable users. The main objective of extranets is to improve communication between businesses and businesses by enabling faster exchange of information in a convenient and secure environment. This reduces the cost of running the market and business. Communication between the company and its business partners, traditionally supported by special links, is quickly transferred to the VPN infrastructure due to the simplicity of setup, lower costs, and simultaneous support for voice, video and data traffic over the IP network. As explained above, the concept of extranet is analogous to IDC, since the server farm is located on the edge of the enterprise network. Since the purpose of the extranet is to provide server farm services to reliable external end users, there are specific safety aspects. These security considerations mean that business partners have access to a subset of business applications, but there is limited access to the network of other companies. Figure 4-5 shows the farm of extranet servers. Note that the extranet server farm is available to internal users, but access from the extranet on the intranet is protected or highly protected. Typically, access from an extranet to an intranet is restricted by using firewalls. Many factors need to be taken into account when developing extranet topology, including scalability, availability and safety. Dedicated firewalls and routers extranet is the result of a very secure and scalable network infrastructure partner connection, but if there are only a small number of partners to deal with, you can take advantage of the existing internet edge infrastructure. Some partners require direct communication or special private communications, while others expect secure connections through VPN links. The server farm architecture does not change whether you are creating Internet or intranet server farms. Design guidelines apply equally to all types of server farms, but the specifics of the design are determined by the requirements of the application environment. Figure 4-5 Extranet server farms This section discusses the types of data centres briefly mentioned in this section. Internet Data Center Internet Data Centers (IDCs) are traditionally designed and managed by service providers, but companies with a business model based on Internet commerce also create and manage IDCs. The company's IDC architecture is very similar to the SERVICE PROVIDER's IDCs, but scaling requirements are generally lower because the user base is generally smaller and there are fewer services compared to SPC, which have multiple customers. In fact, the IDC architecture is the same as shown in Figure 4-2. It is interesting to take into account the company's IDCs that if the business model requires, the premises used by the data center can be concentrated in the service provider's data center, it remains under the control of the undertaking. This is usually done to reduce the costs associated with building servers on the farm and reduce product time on the market, avoiding creating a data center inside from the ground up. Enterprise Data Center Corporate or Enterprise Data Centers supports many different features that enable different business models based on Internet services, intranet services, or both. Therefore, the support of internet, intranet, and extranet server farms is not uncommon. This concept was illustrated in Figure 4-1, where the data center device supports each type of server farm and is also connected to the rest of the company's network – private WAN, university, Internet Edge, etc. Support for intranet server farms is still the primary purpose of the company's data centers. The company's data centers are evolving, and this development is partly the result of new trends in the application environment, such as n-tier, web services, and network computing, but this is largely due to the criticality of data stored in data centers. This section discusses the typical topology used in the data center architecture. Page 2 This section discusses the topology of the data center and, in particular, the server farm topology. Initially, the discussion focused on traffic flow through network infrastructure (on general topology) from a logical point of view and then physically. Generic Layer 3/Layer 2 Designs Generic Layer 3/Layer 2 designs based on the common ways of deploying server farms. Figure 4–6 shows a general server farm topology that supports a large number of servers. NOTE that the distribution layer is now called the aggregate layer, which is due to becoming a concentration point for most, if not all, services beyond the traditional layer 2 and layer 3. Figure 4-6 General server farm design Topology highlights are aggregate-level switches that perform basic layer 3 and layer 2 functions, access level switches that provide connections to server servers, and connection between aggregation and access layer switches. The main Level 3 functions performed on the aggregation switches are as follows: Redirecting packages based on Level 3 information between the server farm and the rest of the network to keep the view of the route network, which should change dynamically as network changes take place support for default server farm gateways The main level 2 functions performed by aggregation switches are as follows: Covering Tree Protocol (STP) 802.1d between aggregation and access switches to create a non-looped forwarding topology. STP improvements after 802.1d that improves the default spanning tree behavior, such as 802.1s, 802.1w, Uplinkfast, Backbonefast, and Loopguard. For more information, see Chapter 12 basics of the Level 2 Protocol. VLANs logical separation of server farms. Other services, such as security of multicast and AAD for services such as QoS, tariff limitation, inhibition of broadcasting, etc. Access layer switches provide direct connection to the server farm. Server farm server types are shared servers, such as DNS, DHCP, FTP, and Telnet; hosts using SNA via IP or IP; database servers. Note that some servers have both internal drives (storage) and bar units, while others have external storage (usually SCSI). The connection between the two aggregation switches and the between the aggregation and the access switches is as follows: EtherChannel between aggregation switches. The channel is in bus mode, which allows physical links to support as many VLANs as needed (only 4096 VLAN on 12-bit VLAN ID). One or more links (EtherChannel, depending on how many over-expected links) from each access switch to each aggregation switch (uplinks). These links also contain suitcases, resulting in multiple VLANs within one physical path. Servers are twofold to different access switches for redundancy. The NIC used by the server is considered to have two configuration ports for active standby. When the primary port

fails, the wait takes over by using the same MAC and IP addresses that the active port was used for. For more information about two homed servers, see section 2, Server Architecture Overview. The normal configuration of the server farm environment has just been described in Figure 4-7. Figure 4-7 shows the location of the critical services required by the server farm. These services are clearly configured as follows: agg1 is clearly configured as an STP root directory. agg2 is clearly configured as a secondary root. agg1 is clearly configured as the primary default gateway. agg2 is clearly configured as a standby or secondary default gateway. Figure 4-7 Note that there is no single point failure architecture, and the roads are now deterministic. Other STP services or protocols, such as UplinkFast, are also clearly defined between aggregation and access layers. These services/protocols are used to reduce the convergence time under failover conditions from 802.d for approximately 50 seconds to 1-3 seconds. In this topology, servers are configured to use the agg1 switch as the primary default gateway, which means that sending traffic from servers follows a direct path to the agg1 switch. Inbound traffic can arrive on any aggregate switch, but traffic can reach servers holding only through agg1 because links from agg2 access switches are not forwarding (blocking). The incoming paths are specified by the dot arrows, and the sending path is indicated by a solid arrow. The next step is to predict and fallback behavior, which is much simpler when you have a deterministic primary and alternate path. This is achieved if each component has a primary path and the save and adjustment failover time of the backup component until the requirements are met. The same process must be carried out in order to return to the original parent unit. This is because failover and backup processes are not the same. In certain cases, the backup can be performed manually and not automatically to prevent certain unwanted conditions. NOTE For 802.1d. if the main STP root fails and the secondary takes over when it returns to the top, it automatically takes over because it has a lower priority. In an active server farm environment, you may not want to automatically change the STP topology, especially when the convergence time is 50 seconds. However, this is not the case when using 802.1w, when the fallback process takes only a few seconds. Whether using 802.1d or 802.1w, the process is automatic, unlike using HSRP, where the user can control the behavior of the primary HSRP equivalent nodes when it resumes using preemption. If the preemption is not used, the user will manually control when to return the mastership to the original master of the HSRP equivalent nodes. The use of STP is the result of a topology of 2 layer, which can have loops that need to detect the automatic mechanism and avoid. An important question is whether there is a need for layer 2 servers in a farm environment. This topic is discussed in this section. For more information about Level 2 design information, see Data Center Infrastructure Design 20. The need for Access Layer Access switches 2 layer has traditionally been 2 layer switches. This is also true of the campus network wiring cabinet. This discussion is focused strictly on the data center, because it has different and specific requirements, some similar to and some differ from the wire cabinets. The reason access switches in the data center have traditionally been level 2 is the result of the following requirements: When they share specific properties, servers are typically grouped on the same VLAN. These properties can be as simple as ownership of the same section or perform the same function (file and printing services, FTP, and so on). Some servers that perform the same function may need to communicate with each other, whether because of the grouping protocol or simply as part of the program functionality. This connection exchange should be on the same subnet and is sometimes only available on the same subnet if the clustering protocol heart or server-to-server application packages are not routable. Servers are usually two-sided so that each foot connects to another access switch due to overwork. If the adapter you are using has a standby interface that uses the same MAC and IP after failure, the active and standby interfaces shall be in the same VLAN (the same default gateway). The growth of the server farm takes place horizontally, which means that new servers are added to the same VLAN or IP subnets when there are other servers that perform the same functions. If layer 2 switches are hosted by servers run from ports, the same VLAN or subnets must be supported for a new set of layer 2 switches. This allows for flexibility in growth and prevents the connection of two access switches. Using the latest devices that provide services to server farms, such as load balancing and firewalls, these relationship devices expect the incoming and outgoing traffic to use the same path. They also need to constantly exchange the connection and session state information, which requires 2 layer adjacentness. For more information about these requirements, see access level, which is available in the Multi-step designs section. Using only layer 3 access level to avoid double-homing, Layer 2 adjacent servers for different access switches, and layer 2 adjacent service devices. However, if these requirements are not common in your server farm, you can consider a Level 3 environment in the access layer. Before deciding what's best, it's important to read the section Completely unnecessary layer 2 and level 3 design with services in the section. New service trends determine a new set of architectural requirements that need to be taken into account before deciding which strategy is best for your data center. The reasons for the design of the layer 2 access switch are motivated by the need to move away from the covered tree due to the slow convergence time and performance challenges of operation of a controlled beklote topology and troubleshooting loops when they occur. While this is true when using 802.1d, the environment that take advantage of 802.1w along with Loopguard has the following features: They do not suffer from the same problems, they are as stable as layer 3 environments, and they support a small convergence time. NOTE STP Standard 802.1d has limitations to address certain conditions in addition to its convergence at the time, but a number of covering tree-related problems are the result of improper configuration or rogue STP devices that appear on the network and bridge between level 2 domains. For more information on this topic, see section 12. The next section discusses an alternative topology solution with an encompassing tree that does not provide STP problems or restrictions. Alternate Layer 3/Layer 2 Designs Figure 4-8 presents an alternative Layer 3/Layer 2 design due to the need to address STP restrictions. Figure 4-8 Loopless Topology Fig. 4-8. Although STP works, its limitations do not pose a problem. This bed-free topology is achieved by removing or VLAN(s) used in access layer switches through the trunk between the two aggregate switches. This basically prevents the topology loop, while it supports the requirements for layer 2 required. In this topology, servers are configured to use the agg1 switch as the primary default gateway. This means that sending traffic from servers connected to acc2 crosses the connection between the two access switches. The incoming stream can either use a pool switch because both have active (non-blocking) paths to access switches. The incoming paths are specified by the dot arrows, and the sending path is indicated by strong arrows. This topology is not without its challenges. These tasks are discussed in the section when other information related to the deployment of the services has been received. Multi-step designs Most applications are in line with the client/server model or the n-stage model, which means that most networks and server farms support these application environments. The supported tiers of the data center infrastructure are based on specific applications and can be any range of applications from client/server to client/web server/application server/database server. When you set connection requirements between tiers, you can set up the specific network services that you need. Communication requirements between tiers are usually higher in scalability, efficiency and security. This can translate into load balancing between tier scaling and performance, or SSL between tiers of encrypted operations, or simply firewall and intrusion detection between web and application tiers for more security. Figure 4–9 presents a topology that helps illustrate the previous discussion. Please note that figures 4-9 are not clear. This means that the actual physical topology may be different. Layer separation simply indicates that different server functions can be physically separated. Physical separation could be the result of a project or specific requirements relating to the relationship between different tiers. For example, when we encounter web servers, the most common problem is changing the scale of the web tier to serve many concurrent users. This means deploying more web servers that have similar characteristics and the same content so that user requests can be executed in the same way by any of them. This, in turn, requires the use of a load equaliser against a server farm that hides a number of servers and virtualizes its services. For users, a specific service still supports one server, but the load balancer dynamically selects the server to execute the request. Figure 4-9 Multi-tier application environment Assume that you have multiple types of web servers that support different applications, and some of these n-step model. The server farm can be split along application or function lines. All Web servers, regardless of their supported application(s), may be part of the same server farm on the same subnet, and application servers can be part of a separate server farm on another subnet and another VLAN. After the same logic used for scaling the web tier, the load equalization tool logically can be a tier between the web tier and the application tier-scale application tier from a web tier perspective. One web server now has multiple application servers that can be accessed. The same set of arguments applies to security requirements in the web tier and a separate set of security aspects in the application tier. This means that the firewall and intrusion detection capabilities are different in each layer and are therefore tailored to the requirements of the application and database tiers. SSL unloading is another example feature that server farm infrastructure can support and can be installed in a web tier, application tier, and database tier. However, its use depends on the application environment by using SSL to encrypt client-server and server-to-server traffic. The extended discussion of multi-level design in the past leads to the introduction of several network services in the architect's concept. These services are provided in Figures 4-10 using icons representing the function or service of the network device. NOTE Figures 4 to 10 contain the icons used in this section to illustrate the services provided by the data center network devices. Different icons are placed in front of the servers for which they perform functions. At the aggregate level, you will find load balance tools, firewalls, SSL cargo, intrusion detection systems, and caches. These services are provided through service modules (line cards that can be inserted into the aggregate switch) or devices. The important thing to consider when it comes to service devices is that they provide scalability and high availability for server farm capacity, and that maintaining the basic premise is not a single failure point, to be used by at least two. If you have more than one (and given that you are associated with a perseveral application environment), failover and backup processes require special mechanisms to restore the connection context, in addition to level 2 and level 3 paths. This simple concept in the application of the exemption has a significant impact on network design. Figure 4-10 Network services icons The number of devices in these network services is replicated before the application layer to provide services to application servers. In Figures 4 to 10, please note that the server steps are physically separated. This separation is one of the alternatives to server farm design. Physical separation is used to services. The expanded design is more expensive because it uses more devices, but it allows for more control and better scalability because road devices handle only part of the traffic. For example, placing a firewall between tiers is considered a safer approach due to physical separation between layer 2 switches. This argument is correct, but it is likely to be much more closely linked to existing security policy than to a real threat. Having a logical rather than physical separation simply requires a consistent application of security policies to ensure that the expanded security zone is as safe logically as it is physically. This brings discussion to another alternative to designing multitier servers on the farm, an alternative that does not include physical separation, but rather the logical separation of tiers, as presented in the next section. The collapsed Multitier Design A collapsed multi-layer design is one in which all server farms are directly connected to the access layer to the aggregation switches, and there is no physical separation between 2 layer switches that support different tiers. Figure 4-11 shows the collapsed design. Figure 4-11 A multi-step design message is collapsed that in this project services are again concentrated in the aggregation layer, and service facilities are now used in the front stage and between tiers. When using a collapsed model, you don't need to have a load leveling or SSL offloaders for a certain tier set. This reduces costs, but device management is more complex and performance needs are higher. Service devices, such as firewalls, protect all server tiers from each other instead of in the data center. The Load Balancer can also be used to simultaneously load balance traffic from the client to web servers and traffic from web servers to application servers. Please note that figures 4-11 are not clear. Other collapsed designs can combine the same physical 2-layer switches into home web applications and database servers at the same time. This means that only this means that the servers are logically located in different IP subnets and VLANs, but service devices are still used simultaneously on the front and between tiers. Please note that service devices are always paired. Linking avoids a single point of failure throughout the architecture. However, both maintenance devices communicate with each other in the pair, which is part of a discussion on whether the access layer requires layer 2 or layer 3. Access level 2 requirement Each service device pair must have status information about the relationships handled by the pair. This requires a mechanism to set up the active device (primary) in another mechanism in order to regularly exchange communication status information. Purpose of the dual service facility is to ensure that in the event of a failure, the unnecessary device is not only able to continue service without interruption, but can also seamlessly failover without disturbing the current fixed connections. In addition to the requirements raised earlier about level 2 needs, this chapter deals in detail with requirements relating to service facilities: service facilities and server farms they serve are usually layer 2 side-by-side. This means that the service device has a foot sitting on the same subnet and the VLAN is used by servers that are used to communicate directly with them. Often, in fact, service devices provide the default gateway support for the server farm themselves. Service facilities must exchange heartbeats as part of their overflow protocol. Heartbeat packets may or may not be routable; If they are routable, you may not want the exchange to go through unnecessary Layer 3 hops. Service devices running during a state failover must exchange connection and session status information. For the most part, this exchange takes place through a VLAN joint between the two devices. Much like heartbeat packets, they may or may not be routable. If service devices provide default gateway support to the server farm, they must be adjacent to the servers. After considering all the requirements of layer 2 of the access layer, it is important to note that although a topology may be available, such as the topology in Figure 4-8, which supports layer 2 in the access layer, the topology shown in Figure 4-7 is preferred. Topology with loops is also supported if they have advantages in protocols such as 802.1w and features such as Loopguard. NOTE To date, the most common deployments use layer 2 at the access level and are based on the protocols of the covered trees and cisco improvements to reduce convergence time and achieve stability, as shown in Figure 4-7. Few use connected topology. The main reasons relate to whether it is possible to have an ununsolic topology, subject to the limitations imposed, and, where possible, whether the setup is simple enough for the reasons for support, maintenance and management. Double homing requires 2 layer adjacent access switches to perform the same VLAN, and unnecessary sublime service devices need 2 layer adjacent to proper operation. It is therefore important to take due account of the requirements for the development of the network infrastructure of the server farm. This section discusses topics related to server farm topology. On page 3 To this point, all topology that has been submitted is completely unnecessary. This section explains the various aspects of the design of an unnecessary and scale-up data centre, providing several possible design alternatives, highlighting sound practices and specifying what actions should be avoided. Need for redundancies Figures 4-12 Logical design steps shown in Figures 4 to 12 server farm infrastructure. The process starts with a Level 3 switch that provides direct server connectivity and kernel ports. A Level 2 switch is available, but the Level 3 switch limits broadcasts and floods to and from server farms. This is Figure 4-12. The main problem with the design is that there are several one-point failure problems: There is one NIC and one switch, and if the NIC or switch fails, the server and applications become unavailable. The solution is twofold: Make single switch components unnecessary, such as double power supplies and dual maintenance. Add a second switch. Excess components make one switch more tolerant, but if the switch fails, the server farm is unavailable. Option B indicates the next step that adds an unnecessary 3-layer switch. Figure 4-12 Multilayer unnecessary design with two layer 3 switches and spreading servers on both of them, you can achieve a higher level of release when the failure of one 3-layer switch does not completely harm the program's environment. The environment is not completely damaged when the servers are dual homed, so if one of the layer 3 switches fails, the servers can still recover using the connection to the second switch. In variants A and (b), the density of the port is limited to the capacity of the two switches. As more ports increase the needs of the server and other service devices, and when maximum capacity is reached, adding new ports becomes more complex, especially when you try to maintain level 2 adjacentness between servers. The mechanism used to grow the server farm is given in option (c). You add layer 2 access switches to the topology to provide a direct server connection. Figure 4-12 shows the 2-layer switches connected to both 3-layer aggregation switches. Two uplinks, one to each aggregation switch, provide overflow from access to aggregation switches, giving servers holding an alternative path to access layer 3 switches. The design described in option C still has problems. If the layer 2 switch fails, the servers lose their only connection feature. The solution is two home servers with two different Layer 2 switches as shown in figure 4-12 in variant d. NOTE In this book, the terms access level and access switches refer to the switches used to ensure port density. The maturity layer and the aggregate switches refer to the switches used to connect the flow to and from access switches and to connect service units (load balancing devices, SSL unloading devices, firewalls, tanks, etc.). The aggregate switches are 3-layer switches, which means that they have a built-in router that can transmit traffic at cord speed. Access switches are mainly 2-layer switches, but they can be 3 layer switches operating only in layer 2 mode Farms. Layer 2 and Layer 3 in option d of access level d 4-12 is described in detail in Figure 4-13. Figure 4-13 Data center layer 3 and layer 2 Fig. 4-13. Redequite in a Level 2 domain is mainly achieved using an spanning tree, and level 3 overflows are achieved through routing protocols. Historically, routing protocols have proven to be more stable than covering the tree, making one issue using Layer 2 instead of layer 3 wisdom access layer. This topic was discussed earlier in the Required level 2 access section. As shown in Figures 4 to 13. Figures 4-13 Flexibility is due to the fact that the design makes it easy to add service devices in the aggregation layer with minimal changes to the rest of the design. Simpler design, such as figure 4-13. Layer 2, Loops and covering tree Layer 2 domains should make you think immediately loops. Each network designer experienced Layer 2 loops in the network. When a layer of 2 loops appears, the packages are repeated endless times, reducing the net. Under normal conditions, the covering tree protocol retains logical topology without loops. Unfortunately, physical failures such as unidirectional links, incorrect wires, rogue bridge devices or bugs can lead to the appearance of loops. Fortunately, the introduction of 802.1w addresses many restrictions and features of the original covering tree algorithm, such as Loopguard, solves the problem of faulty transceiver or error. Still, deploying a legacy covering tree experience drives network designers to try to create a Layer 2 topology without loops. In the data center, this is sometimes possible. An example of this type of design is given in Figures 4 to 14. As you can see, layer 2 domain (VLAN) that contains subnet 10.0.0.x does not have a trunk between two aggregation switches, and neither is 10.0.1.x. Please note that GigE3/1 and GigE3/2 are not combined. Fig. 4-14 If subnets must include multiple access switches, you must have a looped topology. This is the case when you have dual servers because NIC cards configured for grouping typically use a floating IP and MAC address, which means that both interfaces belong to the same subnet. Keep in mind that a loop without a topology is not necessarily better. Specific requirements, e.g. requirements imposed by content in fact, an additional path provided by a loopthic may be needed. Also note that looped topology simply means that any Layer 2 device can access any other Layer 2 device from at least two different physical paths. This does not mean that you have a redirect cycle in which packages are repeated at infinite time: Embracing the tree prevents this from happening. In the loop in topology, malfunction switches can cause 2-layer loops. In a loop without topology, there is no chance for layer 2 loop, because there is no superfluous layer 2 paths. If the number of ports has to increase for any reason (dual servers, more servers, etc.), you can follow the step-by-step circuit 2-layer switch method as shown in Figure 4-15. Figure 4-15 Alternate loop-free Layer 2 design To help visualize Layer 2 loop without topology, figure 4-15 shows each aggregate switch broken down as a router and Layer 2 switch. The problem with topology is that terminating the connection between the two access switches would create a non-continuous potox – this problem can be solved by using EtherChannel between access switches. Another issue occurs when there are not enough port servers. If the number of servers is to be added to the same subnet 10.0.0.x, you cannot add a switch between the two existing servers as shown in option 4-15(b). This is because there is no workaround for the secondary switch failure that would create a split subnet. This design is not fundamentally wrong, but it is not optimal. Both topologys shown in Figures 4-14 and 4-15 should go to loop topology as soon as you have any of the following requirements: an increase in the number of servers on a particular subnet Dual-attached NIC cards existing servers spreading a certain subnet due to various access switches Inserting from state network service devices (e.g. load balance devices) operating in active/standby mode Functions a and b 4-16 figure shows how to enter additional access switches existing subnet creates loop topology creates looping topology. Both a and b, GigE3/1 and GigE3/2 are combined. Figure 4-16 Unnecessary topology with physical layer 2 loops If the requirement is to implement the topology, which brings layer 3 access layer, topology, which meets the requirements of dual connected servers, figure 4-17. Figures 4-17 This strain also carries the Layer 3 VLAN, which is basically used only to make two switches to neighbors from a route point of view. Figures 4-17 Figures 4-17 Please note that when installing each pair has a set of subnets, disconnected from any other pair of subnets. For example, one pair of access switches hosts subnets 10.0.1.x and 10.0.2.x; the other pair can not arrange the same subnets simply because it connects to the aggregation layer with 3 layer links. NOTE If you compare figure 4-17. These are the right dots, and the answer actually depends on the size of the data center. Note that the access layer is added for port density reasons, and the aggregate layer is mainly used to attach devices such as load balancing devices, firewalls, tanks, etc. So far, the debate has focused on unnecessary level 2 and level 3 designs. Layer 3 switch provides the default gateway server farms for all topology introduced so far. However, default gateway support can also be provided by other service devices, such as load balancing tools and firewalls. The next section examines alternatives. Page 4 After discussing the development of a completely unnecessary level 2 and level 3 topology and taking into account the basis of the data center, the focus becomes design problems with other data center services. These services aim to improve security and increase the provision of applications by bleeding processing from the server farm to the network. These services include security, load balancing, SSL unloading and cached storage; they must be supported by several network installations to be integrated into the infrastructure in accordance with design requirements. In addition, this chapter discusses trends in the application environment caused by technological advances or in applications, application infrastructure or network infrastructure. Additional services In addition to Level 2 and Level 3, the data center may need to support the following devices: Firewall Intrusion Detection System (IDS) Load Balancing SSL Migration Caches Important to discuss design issues while maintaining some of these devices. Service devices provide their own requirements that may change certain aspects of the design, such as exchange status or status information, the NAT function they perform at source or destination IP addresses, which forces them to be on the inbound and outgoing path, etc. Maintenance equipment can be used using service modules integrated into aggregate switches or as devices connected to aggregation switches. Both installations require a network connection and have predicted the actual traffic path. Firewalls and load balancer can support the default gateway function on behalf of server farms. Default gateway support has traditionally been provided by the router, so with two additional alternatives, you need which is the default gateway and in which the order flow is processed through multiple devices. Firewalls and load-balancing devices can provide excellent failover, supported by specific overflow protocols. Protocols specific to firewalls or load balancing methods shall be supported by design. SSL dump is typically used with load equalization tools and requires the same circumstances, with one exception: They do not support default gateway services. IDS is a transparent design, which means that they are well integrated with any existing design. The main aspect of idss is their destination, which depends on the choice of the location where the traffic is analyzed and the traffic types monitored. On the other hand, the cache is installed in reverse proxy cache mode. The cache display and mechanism to redirect traffic to them affect the design of the data center. Traffic redirection options include Web Cache Connection Protocol (WCCP) Level 2 or Level 3 switches and load-balancing to distribute the load between the cache cluster. In any case, the cache or cache cluster changes the primary traffic path to the server farm when used. This section provides several deployment options. Service deployment options Two options are available for installing data center services: using service modules integrated into the aggregation switch, and using devices connected to the aggregate switch. Figure 4 to 18 shows two options. Figures 4-18 The aggregate switch is represented by the router (Layer 3) and the switch (Layer 2) as the main components of the foundation (shown on the left) and firewall, load balancer, SSL module and IDS module (shown on the right as accessory services). Service modules communicate with chassis route and switching components through the backplane. Option B shows the design of the device. The aggregate switch provides routing and switching functions. Other services are provided by devices directly connected to aggregation switches. NOTE Also available designs that use modules and devices. Most often occur when using containers that are devices, both design options. Current trends in data center services towards integrated services. Evidence of this integration trend is the distribution of Catalyst 6500 family service modules and the use of blade servers and blade chassis to collapse multiple services on one device. A deliberate approach to design problems is required when choosing traffic flow on different devices, whether you are going to choose option a, (b) or any of the following: This means that you should clearly select the default gateway and the order in which packages from the client to the server are processed. Designs that use devices require more maintenance because interoperability and consistency of protocols. Design considerations with service facilities To date, several issues have been mentioned concerning the integration of service facilities into the design of the data centre. They relate to whether you run layer 2 or layer 3 in the access layer, whether you're using a device or modules, whether they're state-state or stateless, and whether they require you to change the default lock location from the router. When you change the default gateway location, you can determine the order in which the package is to be processed using the aggregation switch and service facilities. Figure 4-19 provides possible default alternatives to maintaining gateway using service modules. The implications of each alternative project are discussed below. Figure 4-19 shows the aggregate switch, catalyst 6500 using the firewall service module and the content switching module, in addition to routing and switching functions provided by the multi-layer switch function card (MSFC) and the maintenance module. One constant design factor is the location of the server-link switch; it is adjacent to the server farm. Figures 4-19. If the content switch acts as a router (path mode), it becomes the default gateway for the server farm. However, if it acts as a bridge (bridge mode), the default gateway would be a firewall. This configuration facilitates the creation of multiple instance firewalls and content switch combination segregation and load balancing for each server farm self-creation. Option B has a firewall pointing to the server farm, and the content switches between the router and the firewall. Whether it is in router mode or bridge mode, the firewall configuration must enable server health management (health probes) traffic from the content switching module to the server farm; this adds management and configuration tasks to the design. Note that in this design, the firewall provides default server farm gateway support. Option C displays a firewall pointing to the main IP network, a content switch pointing to the server farm, and a firewall module between the router and the content switching engine. To put a firewall on the edge of the farms of intranet servers, you need the firewall to have router-like routing capabilities to make it easier to integrate with the network you are pointing and at the same time to separate all server farms. This makes ensuring each server farm is self-made more difficult because the content switch and router can route packages between servers on the farm without going through the firewall. Depending on whether the content switching module is in router or bridge mode, the default gateway can be a content switch or router, as appropriate. Option D displays the firewall module pointing to the main IP network, the router in the server farm, and the content switching module between them. This option poses some of the same challenges as option c in terms of a firewall that supports the IPP and the inability to separate each server farm individually. However, the design has one main advantage: the router is the default gateway for the server farm. When using the router as the default gateway, server farms can use some basic protocols, such as HSRP, and features such as HSRP tracking, QoS, DHCP relay function, and so on, which are only available on routers. All previous design options are available— some are more flexible, some are more secure, and some are more complex. The choice should be based on knowing the requirements as well as the advantages and limitations of everyone. The various design issues related to viable options are discussed in the different chapters of Chapter 21 of Part V Integration of Security into infrastructure for network design in the context of the firewall. The main trends in the application environment are undoubtedly the most important trends in how programmes are developed and expected to work on the network. These trends can be ranked arbitrarily into two main areas: Trends in the architecture of the application architecture network infrastructure program include the evolution of the classic client/server model into a more specialized n-tier model, web services, specific application architectures, server and client software (operating systems), application clients, server and client hardware, and intermediate software used to integrate distributed applications into non-sub-environments. More visible trends in application architecture are the extensive application of web technologies combined with the use of the n-tier model to functionally segment different types of servers. Currently, web, application, and database servers are the main types, but they are combined in different ways (depending on the application vendor and how the buyer wants to implement it). This functional partition requires that the network be smarter due to the assurance and scaling of tiers independently. For example, the n-tier model web tier layer created the need for smaller and faster servers used to extend the front-end system functionality. This resulted in 1RU (rack unit) servers that offer proper performance of web servers at low cost and minimal infrastructure requirements (power and rack space). Web services have a service-oriented approach to the use of different and different distributed applications available through standard messages through Internet protocols. Web services are originally dependent on network transport and finally uses the network as an extension to provide computer capacity in a distributed application environment by unloading tasks into network hardware. NOTE The Web Consortium (W3C) defines a Web service as a URI-identifiable software application whose interfaces and mapping can be defined, described, and discovered by XML artifacts and support direct interaction with other software applications that use XML-based messages through online protocols. For more information about web services and its architecture, see the W3C . Grid computing is another trend that actually brings applications and the network closer together by treating servers as the CPU network in which applications use the most accessible CPU network. Other trends in network calculation include blade servers as an alternative to 1RU servers to ensure higher CPU density per rail company, lower power consumption per server and additional benefits of lower cable requirements. Blade servers are drive servers (or modules) inserted into the chassis, similar to network modules or line cards. Using blade servers can centralize server control functions on the blade chassis (one chassis instead of many servers are on the chassis), require fewer cables (one set per chassis instead of one server) and provide a higher computing and memory capacity per stand unit. However, disk server technology is still young, which explains the variety of flavors, architectures, connectivity options and features. An instance of intermediate software is software used to manage and manage distributed CPU on a computer grid that can be 1RU or blade servers. This specific proxy virtualizes cpu usage so that applications are provided with a CPU cycle from the CPU network instead of through the traditional way. Network infrastructure trends Network infrastructure grows smarter and more programmable, so it supports application environments, both by unloading some computer intensive tasks into the network (usually hardware) and by replacing some server functions that can be better managed by network devices. Load balancing is a good example of a feature performed by a network that replaces grouping protocols used by servers with high availability. Grouping protocols are usually software-based, difficult to manage, rather than very scalable by providing a function that the network performs well when using hardware. Trends, such as blade servers, provide new aspects of design. Most of the blade server chassis (blade chassis, briefly) market support both the ability to provide unnecessary Ethernet switches inside the chassis and as an opportunity to connect disk servers to the network using pass-through connections, with chassis simply providing at least twice as many uplinks as server chassis that possible dual homing. Figures 4 to 20 also provide alternatives to communication blade chassis. Figure 4-20 Blade Server Chassis Server Connectivity Option a Figure 4-20 shows the blade server chassis in which each blade server is connected to each blade chassis unnecessary layer 2 Ethernet switches. The Ethernet switch for each blade chassis provides several uplinks that can be routed to the IP network. Up links are usually less than the total number of links per server that require scheduling oversubscription, especially if the servers are Gigabit Ethernet attached. The middle plane is a fabric used to perform control tasks, that is, to control the traffic of the plane, such as the state of the switch. Figures 4-20 This option does not use Ethernet switches inside the chassis. Pass-through fabric is as simple as a patch panel that stores server NIC features, but it can also become smarter fabric, adding new features and allowing blade server vendors to distinguish their products. Any approach you take to connect blade servers to the network requires careful consideration of the consequences of short-term and long-term design. For example, if you choose to use unnecessary Ethernet switches on the drive chassis, you have the following design alternatives to consider: How to use unnecessary Ethernet switches to connect a connectivity connection Or connect the drive chassis to the access or aggregation switches What is the level of overreescent is tolerated in Figure 4-21, shows two connection choices that use the directions for unnecessary ethernet switches. In the case of overflow, two switches shall be used by uplinking connections from the blade chassis. Switches A and B, small clouds in the IP network cloud, provide unnecessary network tissue for the drive drive to avoid a single fault problem point. Figure 4-21 Up-directional connectivity of the blade chassis, Fig. 4-21. This allows you to point upside. On the contrary, Figures 4 to 21 This is an advantage to have a direct connection to switch A or B, thus avoiding unnecessary hops. In addition, if each blade landing gear Ethernet switch supports more than two uplinks, they can also be routed to switches A and B to ensure greater overstrum and higher bandwidth. The next step is to determine whether to connect the blade chassis to the access layer switches, as traditionally done with servers, or aggregation layer switches. Figure 4 to 22 shows the options for connecting other hop switches from the blade chassis. Figure 4-22 Blade chassis other hop switch) (figure 4-22). access layer switches. This design selection corresponds to connecting Layer 2 access switches to layer 2 access switches. The draft shall take into account the tree recommendations covering the trees, which, in accordance with Figure 4-22. If the blade-chassis ethernet switches support 802.1w, the convergence time remains within two to three seconds; However, if the support is strictly 802.1d, the convergence time returns to a typical interval of 30-50 seconds. Other aspects of the design relate to whether the average aeroplane is used more than control and switching control traffic communication functions. If for any reason the midplane is also used for bridge VLAN (forward bridge protocol data units, or BPDUs) STP topology shall be carefully considered. Design goals continue to make the topology predictable and deterministic. This means that you need to clearly prioritize roots and bridges and analyze possible failure scenarios to make sure they support application requirements. Figures 4-22 This is the most appropriate alternative as it can be more deterministic and support lower convergence times. Similarly to the previous version, if the blade-chassis Ethernet switches do not support 802.1w or some STP improvements, such as Uplinkfast and Loopguard, the convergence time would be between 30 and 50 seconds. Topology still needs to be deterministic and predictable by clearly setting root and bridge priorities and checking failure scenarios. As a farm for scale blade servers is another reward. Scaling servers in the environment are done simply by adding pairs of access switches to the overflow and connecting them to the aggregation switches as shown in Figure 4-23. Fig. 4-23 The total number of servers can be X*Y. Depending on the density of the access switch port and the density of the aggregation switch socket, it can grow to thousands of servers. The number of scalable disk servers may require a slightly different policy. Because the blade chassis with Ethernet switches is an access layer, the number of blade servers is limited to the number of slots and ports per slot when the aggregate switches. Figures 4-23 Note that the resizable module is now an aggregate switch along with a certain number of blade chassis. This is because the totalization switch must limit the which can be used for the blade chassis. In addition, line cards used to support blade server uplinks now receive shared server traffic, so it requires less oversubscription. This results in fewer ports used on the line card. Thus, the total number of blade servers is slightly limited due to the density of the socket and port. While this design alternative can support hundreds of blade servers and meet the requirements of a fast-growing server farm environment, you must have a plan for what to do if you need to increase your server farm beyond what the current design supports. Figure 4 to 24 shows this alternative. Figure 4-24 Core Layer Within the Data Center Fig. 4-24. The main layer is used to assemble as many server blade modules as needed, but the number is limited to port and socket capacity with aggregation switches. The access option may not require so much planning because there are no unnecessary Ethernet switches in the blade chassis. The up links are connected to an access layer that is equivalent to current designs in which servers are dual homed into an unnecessary set of access switches. To eliminate connectivity, port density, slot density, and scaling, other areas, such as overdirection, uplink capacity and service deployment options, design and testing may be required before you set up the data center architecture. Additional trends include server dual management, migration from Fast Ethernet to Gigabit Ethernet, application firewalls, and the use of slide network service devices. Application firewalls are firewalls that are more compatible with the operation of the program than normal firewalls, so the firewall process is more detailed for the app information, and only network or transport layer information. For example, an application firewall may be able to determine not only that the package contains TCP and that tcp traffic information is HTTP, but also that the request comes from a specific high-priority user and is a SQL query for sensitive payroll information that requires a higher security service level. Transparent network services include firewall, load balancing, SSL unloading, etc. These services are provided by network devices with minimal interoperability problems, which keep existing designs unchanged. These transparent services could be used for traditional network services, such as load balancing and firewalls, but they are implemented to reduce environmental disruptions and changes in the application environment. This method may include the use of physical devices as if they were different logical objects providing services to different server farms at the same time. This means that the administration of these services, such as configuration changes or troubleshooting efforts, is separate from a specific logical service. Think of it as one physical firewall that is installed to support server farms at the same time that access to CLI and configuration commands is only available to users who have access to the firewall service for a specific server farm. This appears to be a user as a completely separate firewall. Some of these trends are continuing, and some are barely beginning. Some will need special designs and architectural considerations, and some will be accepted smoothly. Others won't take long enough to worry. Page 5 Data Centers are very dynamic environments that have multiple types of server farms that support basic business applications. The design of the data centre covers a wide range of aspects related to how applications are architectureed, how they are deployed and their network infrastructure. A sound approach to design includes a combination of architectural principles such as scalability, flexibility and high accessibility, as well as the application of these principles to environmental requirements. The result should be a structure that meets current needs, but it is flexible enough to meet the needs of short- and long-term trends. The robust design base of the data center is based on an unnecessary, scalable and flexible layer 2 and 3 infrastructure, where operation is predictable and deterministic. The infrastructure should also be adapted to service facilities with

key functions to expand or ensure the environment of the programmes. You need to plan carefully to install service equipment, such as firewalls, load balancing tools, SSL unloading devices, and tanks. Planning efforts must ensure that the desired behavior is achieved in the following areas: overflow protocols between service facilities, exchange of communication and session information between state-of-the-range devices, default gateway services location, and traffic path through data center infrastructure from device to device. Additional considerations require an architectural approach to address environmental trends and requirements for network infrastructure. Subsequent chapters of the book dig deeper into the specifics of the design of the data center and server farms. Design.