

# 考点

2019年6月8日 23:39

## 1 典型网络安全事件剖析

- [1.1 常见的电信诈骗分几类](#)
  - [1.2 典型的网络安全事件有哪些?](#)
  - [1.3 网络安全趋势有哪些](#)
- ## 2 网络与信息安全体系结构介绍
- [2.1 国内信息安全包括哪些?](#)
  - [2.2 国外信息安全包括哪些?](#)
  - [2.3 PPDRR模型包括哪些?](#)
  - [2.4 常见的网络安全技术包括那些?](#)

## 3 互联网发展热点剖析

- [3.1 互联网发展的热点有哪些?](#)

## 4 大型网络应用剖析

- [4.1 大型网络应用发展的历程](#)
- [4.2 通信协议设计](#)

## 5 负载均衡策略与算法剖析

- [5.1 负载均衡按功能的分类](#)
- [5.2 负载均衡部署的方式有哪些?](#)
- [5.3 为什么需要负载均衡](#)
- [5.4 常见的负载均衡策略有哪些? 原理是什么?](#)
- [5.5 常见的负载均衡算法有哪些? 原理是什么?](#)

## 6 防火墙基础

- [6.1 防火墙分类有哪些?](#)
- [6.2 防火墙的弱点有哪些?](#)
- [6.3 常见的防火墙基本结构有哪些?](#)
- [6.4 攻击防火墙的主要方法包括?](#)

## 7 防火墙技术

- [7.1 包过滤防火墙的优缺点?](#)
- [7.2 网络地址翻译技术解决的问题是什么?](#)

## 8 防火墙安全策略及其配置

- [8.1 网络审计的内容包括哪些?](#)
- [8.2 翻转掩码计算](#)
- [8.3 创建访问控制列表的步骤是什么?](#)
- [8.4 子网掩码计算](#)

## 9 入侵检测

- [9.1 Anderson在报告中定义了三种恶意用户?](#)
- [9.2 常见的入侵检测数据源包括哪些?](#)
- [9.3 入侵检测各种分类方法](#)
- [9.4 通用的入侵检测模型](#)
- [9.5 基于网络数据源的优势是什么?](#)
- [9.6 基于主机/网络入侵检测的优缺点?](#)

## 10 基于主机的入侵检测技术

- [10.1 四种用于入侵检测统计模型](#)
- [10.2 审计数据预处理工作包括哪些?](#)
- [10.3 审计记录格式要求要点是什么?](#)
- [10.4 文件完整性检查的必要性有哪些?](#)

## 11 基于网络的入侵检测技术

- [11.1 OSI参考模型](#)
- [11.2 TCP/IP模型](#)
- [11.3 网络数据捕获的方法有哪些?](#)

## 12 先进的入侵检测技术

- [12.1 数据清洗的目的是什么?](#)
- [12.2 污染数据形成的原因是什么?](#)

## 1 典型网络安全事件剖析

- [1.1 常见的电信诈骗分几类](#)
- [1.2 典型的网络安全事件有哪些？](#)
- [1.3 网络安全趋势有哪些](#)

### 1.1 常见的电信诈骗分几类

常见的电信诈骗主要有以下四大类：

- 电话诈骗：电话交易诈骗、电话退税诈骗、电话中奖诈骗、电话冒充政府部门诈骗、电话推销诈骗、电话绑架诈骗、电话勒索诈骗、电话冒充熟人诈骗
- 网络诈骗：网络交易诈骗、网络冒充熟人诈骗、网络投资诈骗、网络中奖诈骗、网络推销诈骗、网络招工诈骗、网络办卡诈骗、网络交友诈骗、网络贷款诈骗
- 短信诈骗：短信贷款诈骗、短信中奖诈骗、短信汇款诈骗、短信交易诈骗、短信银行卡消费诈骗、短信冒充熟人诈骗、短信推销诈骗、短信办卡诈骗、短信退税诈骗
- 传统媒介诈骗：报刊交友诈骗、冒充诈骗、中奖诈骗、招工诈骗

### 1.2 典型的网络安全事件有哪些？

- 惊曝淘宝9900万账户信息遭窃
- OpenSSL水牢漏洞
- 国内部分网站存在Ramnit恶意代码攻击
- 跨境冒充公检法1.17亿电信诈骗案
- 2.7亿Gmail、雅虎和Hotmail账号遭泄露
- 全美互联网瘫痪
- 5家俄罗斯银行遭遇DDoS攻击
- 希拉里邮件门影响美国大选
- 电信诈骗导致高中生徐玉玉身亡
- 黑客利用恶意软件Mirai导致德国90万台路由器瘫痪

### 1.3 网络安全趋势有哪些

- 网络安全法律体系将加速形成
- 关键信息基础设施面临的网络安全风险不断攀升
- 联网智能终端引发的安全事件进一步升级
- 精准化的网络诈骗现象将更加突出
- 移动支付面临的安全形势更加严峻
- 网络可信身份的互联互通加速实现
- 安全可控信息产业将得到爆发式增长
- 优秀人才脱颖而出的环境将逐步具备
- 网络战威胁风险显著增加
- 双边和多边网络安全合作将持续深化

## 2 网络与信息安全体系结构介绍

- [2.1 国内信息安全包括哪些？](#)
- [2.2 国外信息安全包括哪些？](#)
- [2.3 PPDRR模型包括哪些？](#)
- [2.4 常见的网络安全技术包括那些？](#)

### 2.1 国内信息安全包括哪些？

- 可以把信息安全保密内容分为：实体安全、运行安全、数据安全和管理安全四个方面。（沈昌祥）
- 计算机安全包括：实体安全、软件安全、运行安全、数据安全。（教科书）
- 计算机信息人机系统安全的目标是着力于实体安全、运行安全、信息安全和人员安全维护。安全保护的直接对象是计算机信息系统，实现安全保护的关键因素是人。（等级保护条例）

### 2.2 国外信息安全包括哪些？

- 信息安全是使信息避免一系列威胁，保障商务的连续性，最大限度地减少商务的损失，最大限度地获取投资和商务的回报，涉及的是机密性、完整性、可用性。（BS7799）
- 信息安全就是对信息的机密性、完整性、可用性的保护。（教科书）
- 信息安全涉及到信息的保密性、完整性、可用性、可控性。综合起来说，就是要保障电子信息的有效性。（信息安全重点实验室）

## 2.3 PPDR模型包括哪些？

## 2.4 常见的网络安全技术包括那些？

### 安全策略 (Policy) 的前沿技术

- 风险分析、安全评估
  - 如何评估系统处于用户自主、系统审计、安全标记、结构化、访问验证等五个保护级的哪一级？
- 漏洞扫描技术
  - 基于关联的弱点分析技术
  - 基于用户权限提升的风险等级量化技术
- 网络拓扑结构的发现，尤其是Peer-to-Peer网络拓扑结构的发现
  - 拓扑结构综合探测技术（发现黑洞的存在）
  - 基于P2P的拓扑结构发现技术（解决局域性问题）
- 态势预测与分析
  - 如何评估当前系统或网络环境所处的安全状态、未来的发展趋势
    - 热点舆论事件论坛、博客日本地震、海啸、核辐射中国抢盐
    - 热点网络事件蠕虫、僵尸网络、DDoS、LDDoS、数字大炮等

### 系统防护 (Protection) 的前沿技术

- 病毒防护，侧重于网络制导、移动终端防护。
  - 病毒将始终伴随着信息系统而存在。随着移动终端的能力增强，病毒必将伴随而生
- 隔离技术
  - 基于协议的安全岛技术：协议的变换与解析
  - 单向路径技术：确保没有直通路径
- 拒绝服务攻击的防护
  - DoS是个致命的问题，需要有解决办法
- 访问控制技术
  - 家庭网络终端(电器)、移动终端的绝对安全
  - 多态访问控制技术

### 入侵检测 (Detection) 的前沿技术

- 基于IPv6的入侵检测系统
  - 侧重于行为检测
- 向操作系统、应用系统中进行封装
- 分布式入侵检测
  - 入侵检测信息交换协议
  - IDS的自适应信息交换与防攻击技术
- 特洛伊木马检测技术
  - 守护进程存在状态的审计
  - 守护进程激活条件的审计
- 预警技术
  - 基于数据流的大规模异常入侵检测

### 应急响应 (Response) 的前沿技术

- 快速判定、事件隔离、证据保全
  - 紧急传感器的布放，传感器高存活，网络定位
- 企业网内部的应急处理
  - 企业网比外部网更脆弱，强化内部审计
- 蜜罐技术
  - 漏洞再现及状态模拟应答技术
  - 沙盒技术，诱捕攻击行为
- 僵尸技术
  - 动态身份替换，攻击的截击技术
  - 被攻系统躲避技术，异常负载的转配

### 灾难恢复 (Restore) 的前沿技术

- 基于structure-free的备份技术
  - 构建综合备份中心IBC (Internet Backup Center)
  - 远程存储技术

- 数据库体外循环备份技术
- 容侵 (intrusion-tolerant) 技术
  - 受到入侵时甩掉被攻击部分
  - 防故障污染
- 生存 (容忍) 技术
  - 可降级运行, 可维持最小运行体系

### 3 互联网发展热点剖析

- [3.1 互联网发展的热点有哪些?](#)

#### 3.1 互联网发展的热点有哪些?

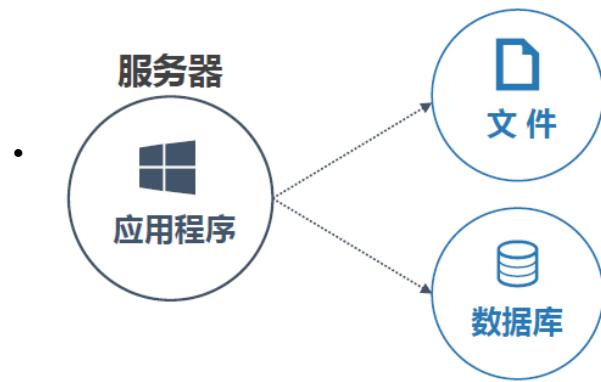
- 连接一切
- “互联网+” 创新涌现
- 开放的协作
- 消费者参与决策
- 数据成为资源
- 顺应潮流的勇气
- 连接一切的风险

### 4 大型网络应用剖析

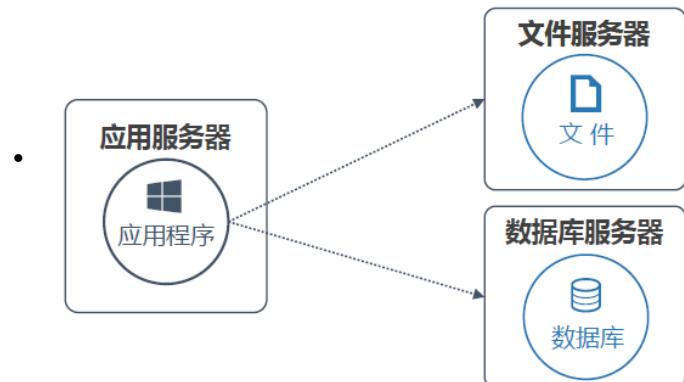
- [4.1 大型网络应用发展的历程](#)
- [4.2 通信协议设计](#)

#### 4.1 大型网络应用发展的历程

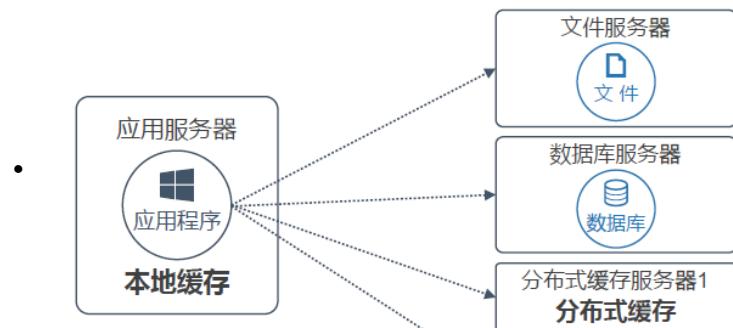
##### 第一代：最开始的网站架构



##### 第二代：应用、数据、文件分离

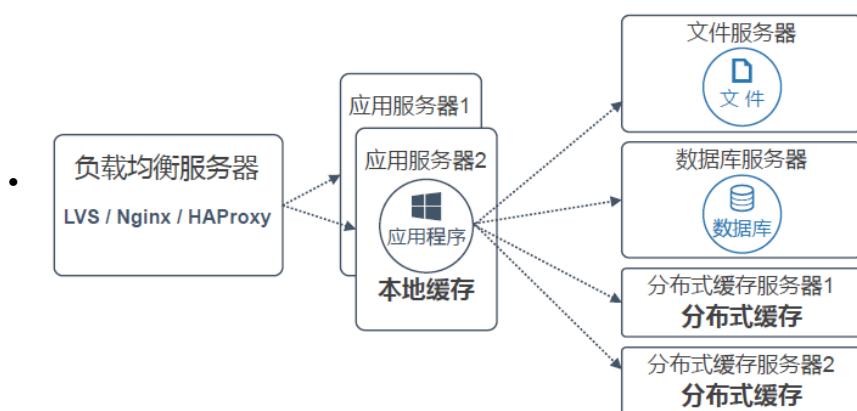


##### 第三代：利用缓存改善网站性能

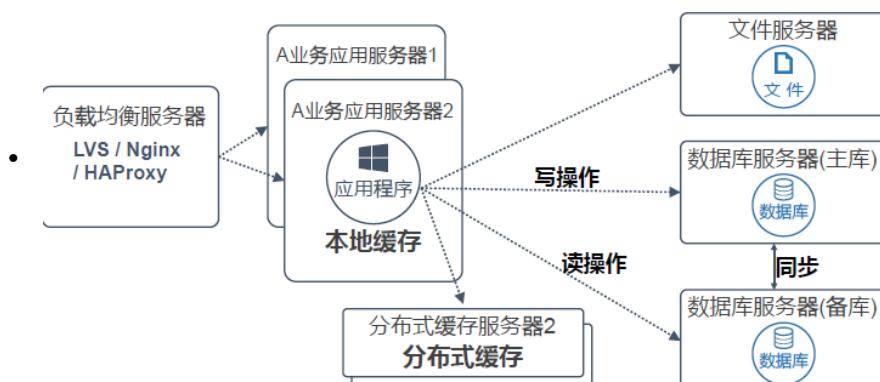


分布式缓存服务器2  
分布式缓存

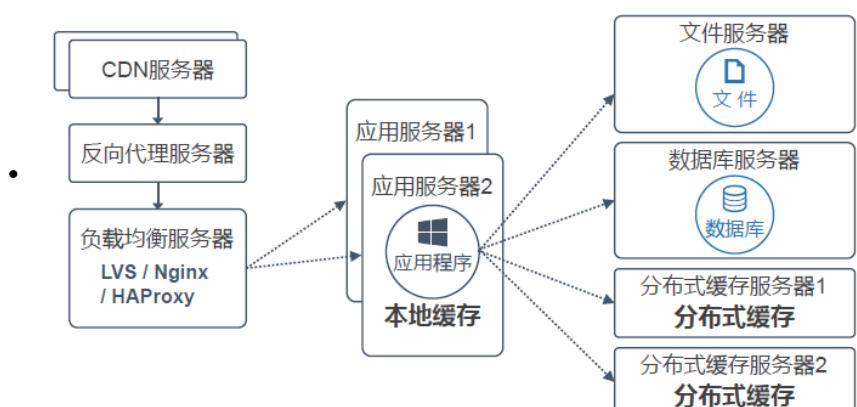
## 第四代：使用集群改善应用服务器性能



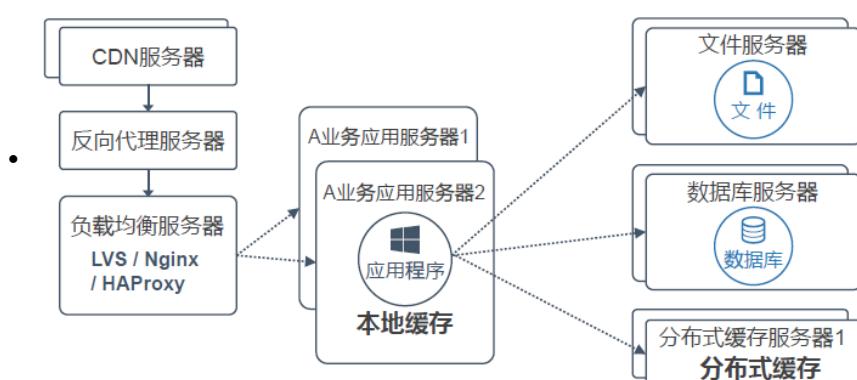
## 第五代：数据库读写分离和分库分表



## 第六代：使用CDN和反向代理提高网站性



## 第七代：使用分布式文件系统

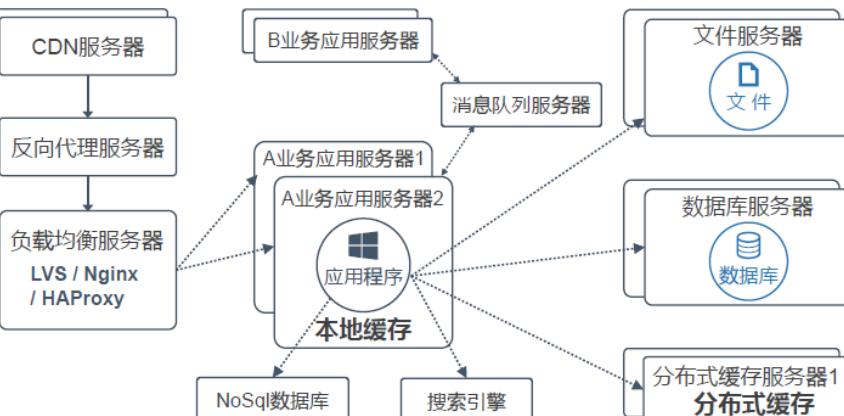


## 第八代：使用NoSql和搜索引擎

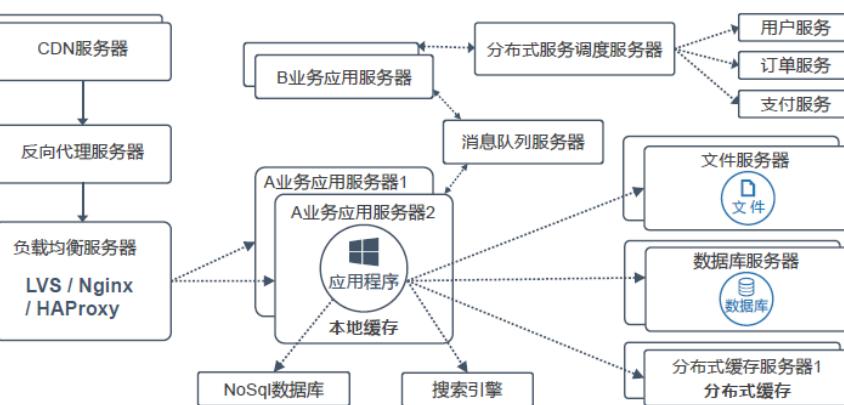




## 第九代：将应用服务器进行业务拆分

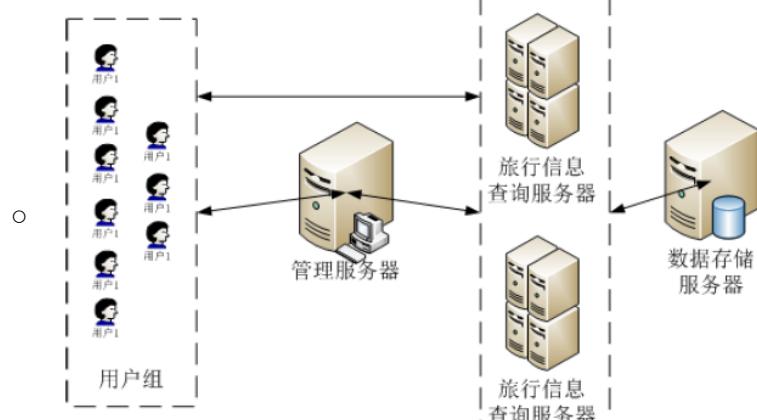


## 第十代：搭建分布式服务



### 4.2 通信协议设计

- 分布式系统通信协议
  - 双方实体完成通信或服务所必须遵循的规则和约定。
  - 通过通信信道和设备互连起来的多个不同地理位置的数据通信系统，要使其能协同工作实现信息交换和资源共享，它们之间必须具有共同的语言
  - 交流什么、怎样交流及何时交流，都必须遵循某种互相都能接受的规则。这个规则就是通信协议。
  - 组成的三要素：
    - 语法：“如何讲”，数据的格式、编码和信号等级。
    - 语义：“讲什么”，数据内容、含义以及控制信息。
    - 定时规则（时序）：明确通信的顺序、速率匹配和排序。
- 通信协议设计举例--火车/航班/轮船信息查询系统通信协议设计



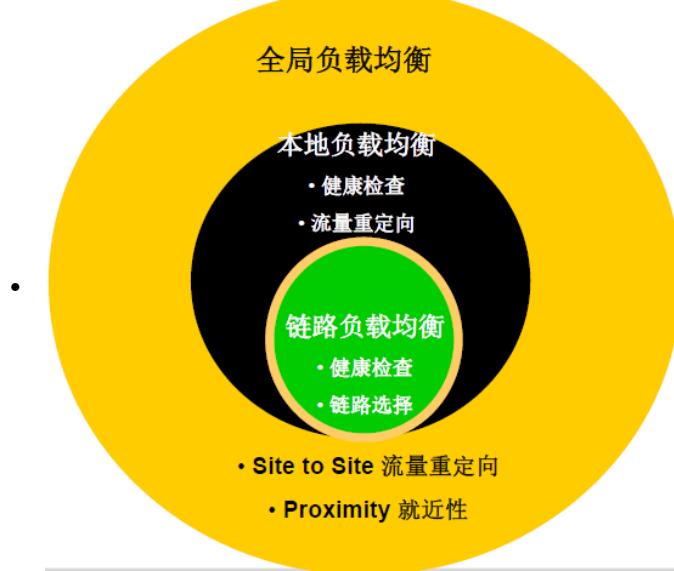
- 注册（用户名、密码、手机号、Email）
  - 是否有重名？如何返回
- 登录（用户名、密码、时间）

- 密码错误？防止攻击？
- 查询（旅行方式、起点、终点、时间）
  - 有结果、无结果（网络？确实无数据？）、如何返回？返回什么？

## 5 负载均衡策略与算法剖析

- [5.1 负载均衡按功能的分类](#)
- [5.2 负载均衡部署的方式有哪些？](#)
- [5.3 为什么需要负载均衡](#)
- [5.4 常见的负载均衡策略有哪些？原理是什么？](#)
- [5.5 常见的负载均衡算法有哪些？原理是什么？](#)

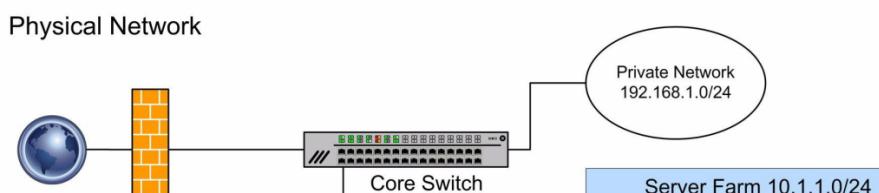
### 5.1 负载均衡按功能的分类

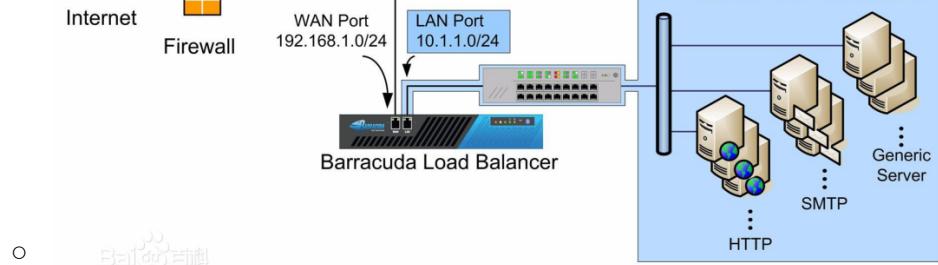


- 负载均衡从其应用的地理结构上分为本地负载均衡（LocalLoadBalance）和全局负载均衡（GlobalLoadBalance，也叫地域负载均衡），本地负载均衡是指对本地的服务器群做负载均衡，全局负载均衡是指对分别放置在不同的地理位置、有不同网络结构的服务器群间作负载均衡。
- 本地负载均衡能有效地解决数据流量过大、网络负荷过重的问题，并且不需购置性能卓越的服务器，充分利用现有设备，避免服务器单点故障造成数据流量的损失。
  - 其有灵活多样的均衡策略把数据流量合理地分配给服务器群内的服务器共同负担。扩充升级方便，增加一个新的服务器到服务群中，而不需改变现有网络结构、停止现有的服务。
- 全局负载均衡主要用于在一个多区域拥有自己服务器的站点，为了使全球用户只以一个IP地址或域名就能访问到离自己最近的服务器，从而获得最快的访问速度，也可用于子公司分散站点分布广的大公司通过Intranet（企业内部互联网）来达到资源统一合理分配的目的。
  - 实现地理位置无关性，能够远距离为用户提供完全的透明服务
  - 可避免服务器、数据中心等单点失效，可避免由于ISP专线故障引起的单点失效。
  - 解决网络拥塞问题，提高服务器响应速度，服务就近提供，达到更好的访问质量

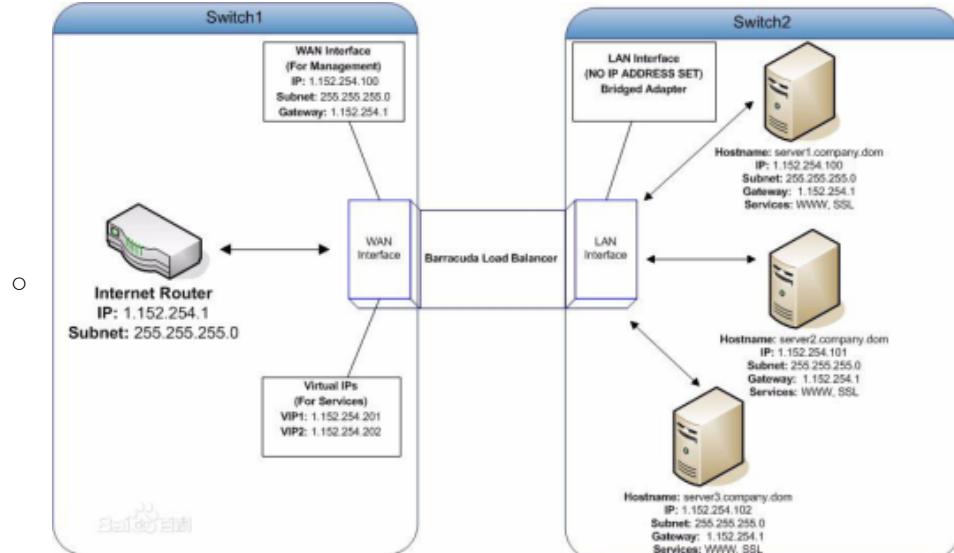
### 5.2 负载均衡部署的方式有哪些？

- 负载均衡有三种部署方式：路由模式、桥接模式、服务直接返回模式。
  - 路由模式部署灵活，约60%的用户采用这种方式部署；
  - 桥接模式不改变现有的网络架构；
  - 服务直接返回（DSR）比较适合吞吐量大特别是内容分发的网络应用。约30%的用户采用这种模式
- 路由模式：路由模式的部署方式如右图。服务器的网关必须设置成负载均衡机的LAN口地址，且与WAN口分署不同的逻辑网络。因此所有返回的流量也都经过负载均衡。这种方式对网络的改动小，能均衡任何下行流量

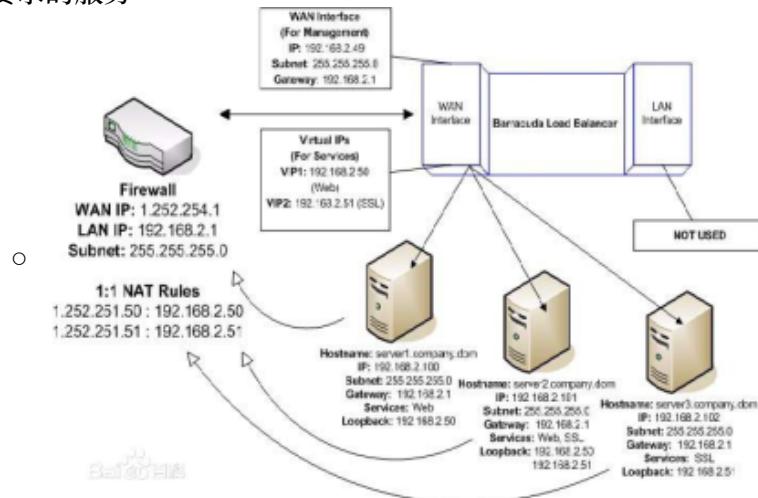




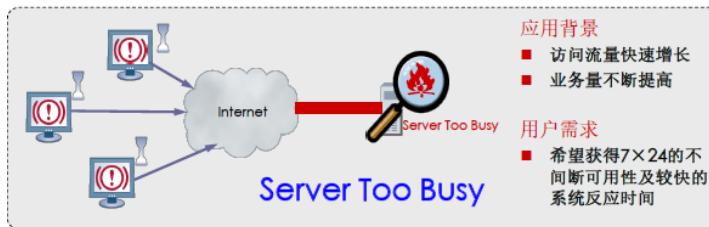
- 桥接模式：配置简单，不改变现有网络。负载均衡的WAN口和LAN口分别连接上行设备和下行服务器。LAN口不需要配置IP（WAN口与LAN口是桥连接），所有的服务器与负载均衡均在同一逻辑网络中。参见右图：由于这种安装方式容错性差，网络架构缺乏弹性，对广播风暴及其他生成树协议循环相关联的错误敏感，因此一般不推荐这种安装架构



- 服务直接返回模式：这种安装方式负载均衡的LAN口不使用，WAN口与服务器在同一个网络中，互联网的客户端访问负载均衡的虚IP（VIP），虚IP对应负载均衡机的WAN口，负载均衡根据策略将流量分发到服务器上，服务器直接响应客户端的请求。因此对于客户端而言，响应他的IP不是负载均衡机的虚IP（VIP），而是服务器自身的IP地址。返回的流量不经过负载均衡。因此这种方式适用大流量高带宽要求的服务



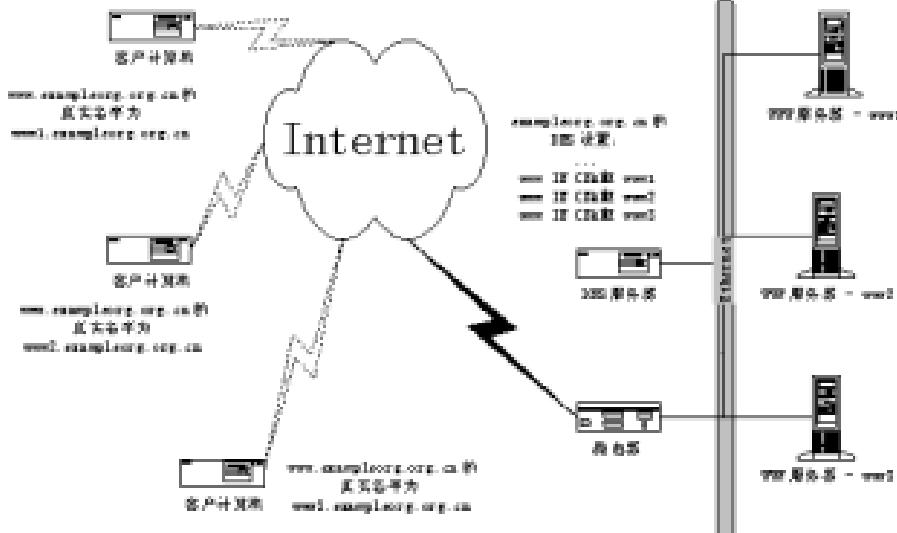
### 5.3 为什么需要负载均衡



## 5.4 常见的负载均衡策略有哪些？原理是什么？

### 基于DNS的负载均衡

- 实现原理：一个域名绑定多个IP，通过DNS服务中的随机域名解析来实现



- 优点：
  - 实现简单、实施容易、成本低、适用于大多数TCP/IP应用。
- 问题：
  - 一旦某个服务器出现故障，即使及时修改了DNS设置，还是要等待足够的时间（刷新时间）才能发挥作用，在此期间保存了故障服务器地址的客户计算机将不能正常访问服务器。
- 缺陷：
  - DNS负载均衡无法得知服务器之间的差异，它不能做到为性能较好的服务器多分配请求，也不能了解到服务器的当前状态，甚至会出现客户请求集中在某一台服务器上的偶然情况。

### 基于反向代理的负载均衡

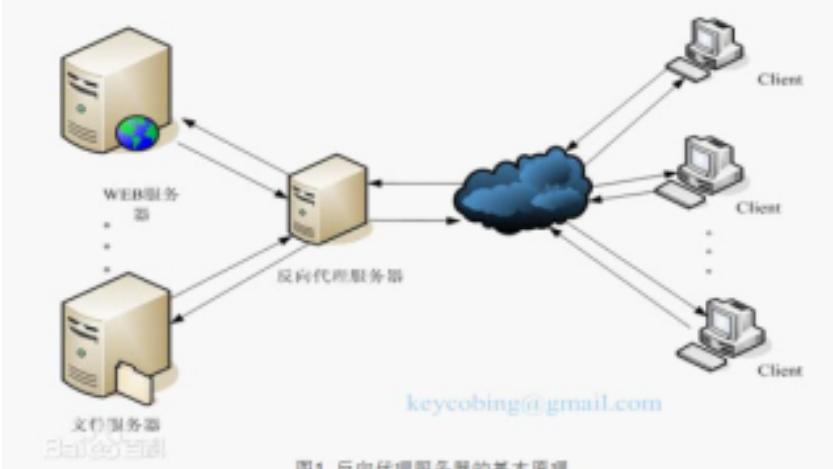
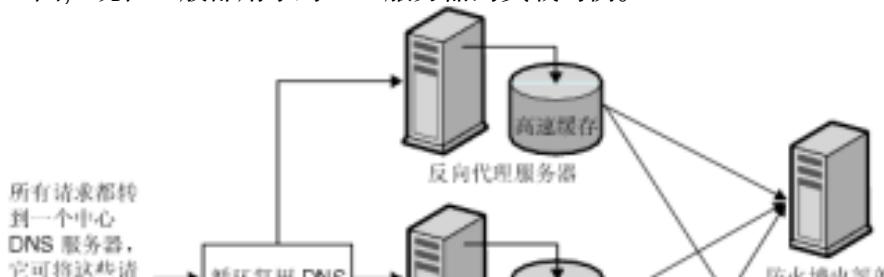
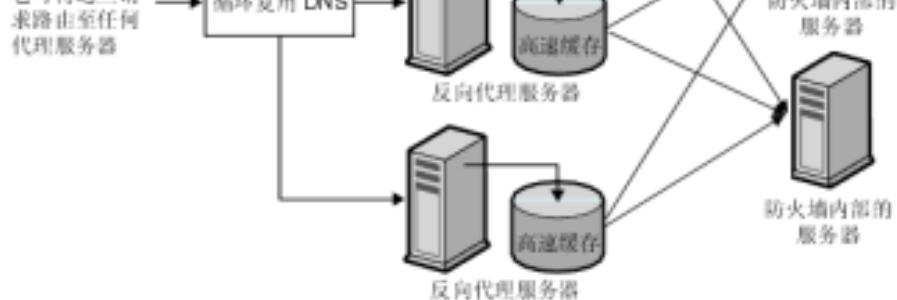


图1 反向代理服务器的基本原理

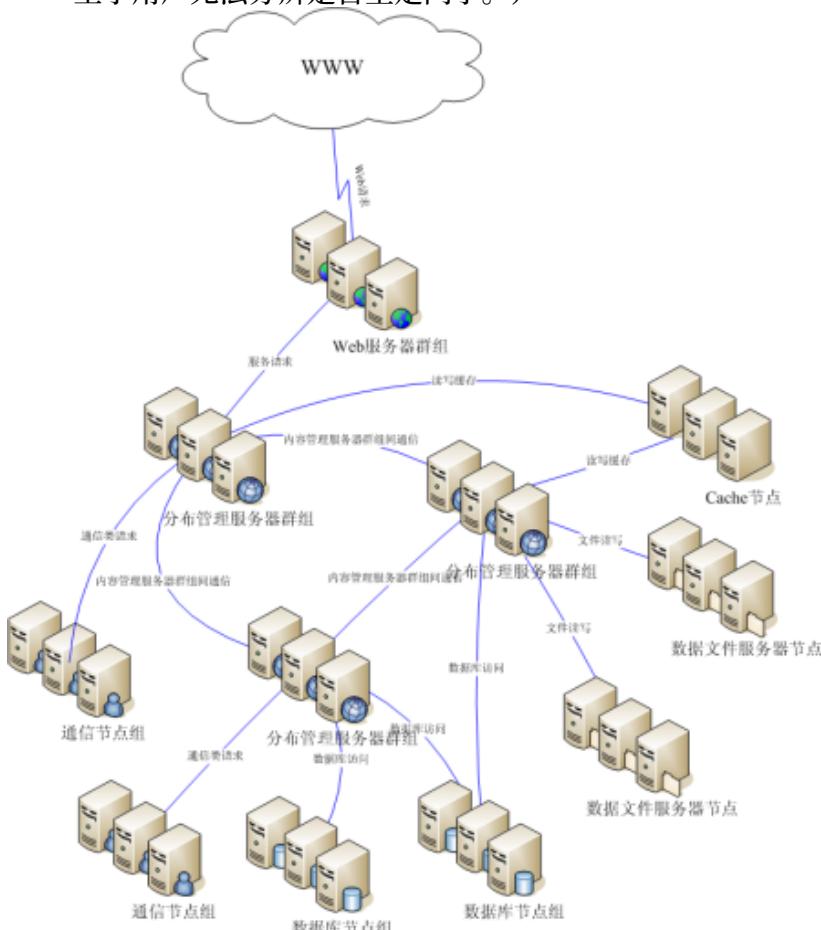
- 优点：
  - 自带高速缓存，可减轻内容服务器压力，提速网络访问效率。
- 问题：
  - 针对每一次代理，代理服务器就必须打开两个连接，一个对外，一个对内，因此在并发连接请求数量非常大的时候，代理服务器的负载也就非常大，最后代理服务器本身可能会成为服务的瓶颈。
- 缺陷：
  - 反向代理是处于OSI参考模型第七层应用的，所以就必须为每一种应用服务专门开发一个反向代理服务器，这样就限制了反向代理负载均衡技术的应用范围，现在一般都用于对web服务器的负载均衡。





#### 基于特定服务器软件的负载均衡

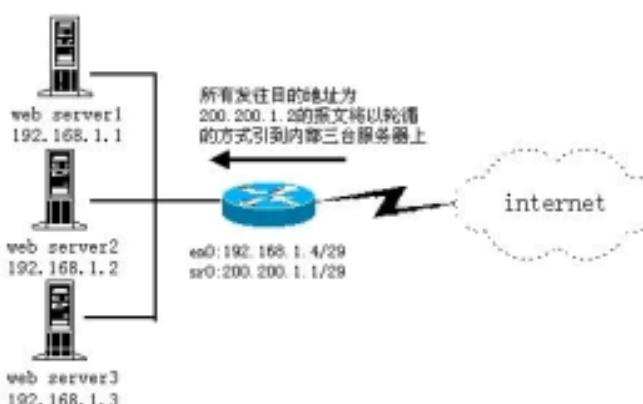
- 实现原理：
    - 利用网络协议的重定向功能来实现。
    - HTTP重定向：服务器无法处理浏览器发送过来的请求（request），服务器告诉浏览器跳转到可以处理请求的url上。（浏览器会自动访问该URL地址，以至于用户无法分辨是否重定向了。）



- 优点：
    - 服务可定制，可依据底层服务器的性能及实况进行负载调控。
  - 问题：
    - 需要改动软件，成本较高。

基于NAT的负载均衡

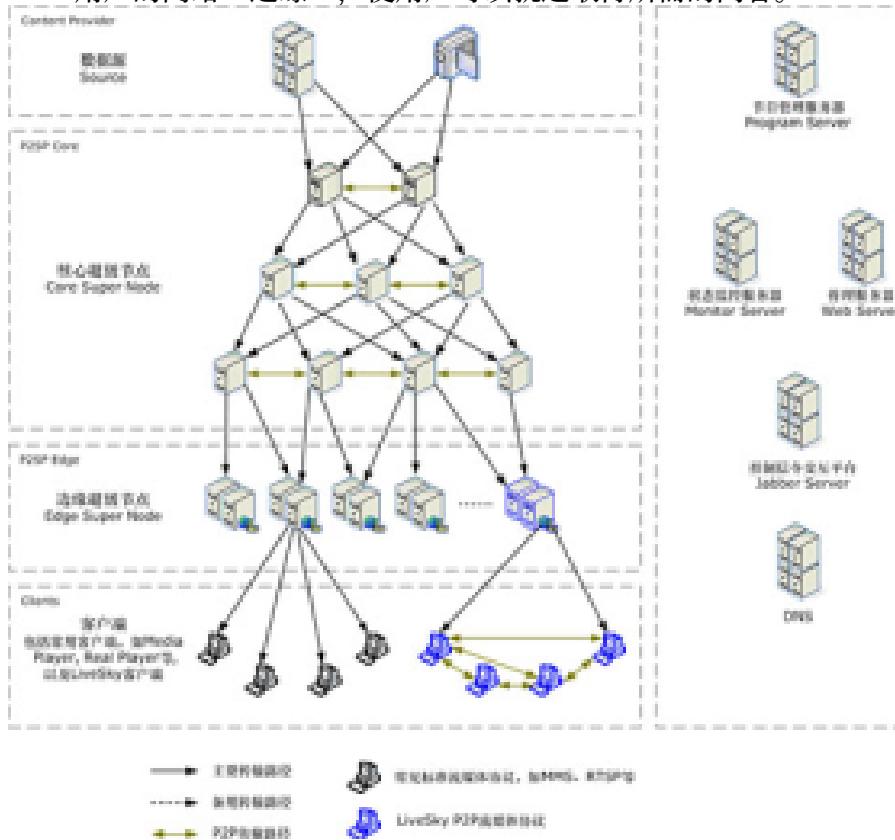
- 实现原理：
    - 将一个外部IP地址映射为多个内部IP地址。



- 优点：
  - 比较完善的负载均衡技术，均衡算法也较灵活，如随机选择、最少连接数及响应时间等来分配负载
- 问题：
  - 伸缩能力有限，当服务器结点数目过多时，调度器本身有可能成为系统的新瓶颈。

### 基于CDN的负载均衡

- 实现原理：
  - 通过在现有的Internet中增加一层新的网络架构，将网站的内容发布到最接近用户的网络“边缘”，使用户可以就近取得所需的内容。



- 优点：
  - 用户访问就近服务器，提高访问速度

## 5.5 常见的负载均衡算法有哪些？原理是什么？

### 静态负载均衡算法

- 轮询 (RoundRobin)：顺序循环将请求一次顺序循环地连接每个服务器。当其中某个服务器发生第2到第7层的故障，就将其从顺序循环队列中拿出，不参加下一次的轮询，直到其恢复正常。
- 比率 (Ratio)：给每个服务器分配一个加权值为比例，根据这个比例，把用户的请求分配到每个服务器。当其中某个服务器发生第2到第7层的故障，就将其从服务器队列中拿出，不参加下一次的用户请求的分配，直到其恢复正常。
- 优先权 (Priority)：给所有服务器分组，给每个组定义优先权，用户的请求分配给优先级最高的服务器组（在同一组内，采用轮询或比率算法，分配用户的请求）；当最高优先级中所有服务器出现故障，才将请求送给次优先级的服务器组。这种方式，实际为用户提供一种热备份的方式。

### 动态负载均衡算法

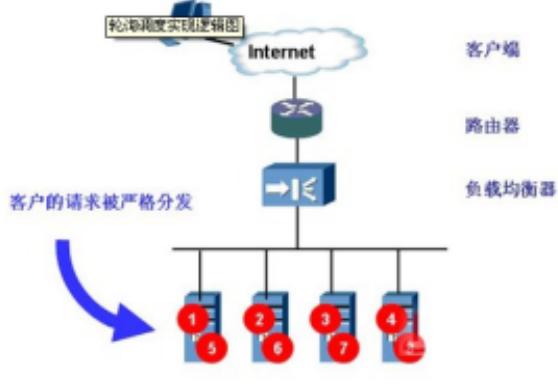
- 最少的连接方式 (LeastConnection)：传递新的连接给那些进行最少连接处理的服务器。当其中某个服务器发生第2到第7层的故障，就将其从服务器队列中拿出，不参加下一次的用户请求的分配，直到其恢复正常。
- 最快模式 (Fastest)：传递连接给那些响应最快的服务器。当其中某个服务器发生第2到第7层的故障，就将其从服务器队列中拿出，不参加下一次的用户请求的分配，直到其恢复正常。
- 观察模式 (Observed)：连接数目和响应时间以这两项的最佳平衡为依据为新的请求选择服务器。当其中某个服务器发生第2到第7层的故障，就将其从服务器队列中拿出，不参加下一次的用户请求的分配，直到其恢复正常。
- 预测模式 (Predictive)：利用收集到的服务器当前的性能指标，进行预测分析，选择一台服务器在下一个时间片内，其性能将达到最佳的服务器相应用户的请求。

- 动态性能分配(DynamicRatio-APM): 收集到的应用程序和应用服务器的各项性能参数, 动态调整流量分配。
- 动态服务器补充(DynamicServerAct): 当主服务器群中因故障导致数量减少时, 动态地将备份服务器补充至主服务器群。
- 服务质量(QoS) : 按不同的优先级对数据流进行分配。
- 服务类型(ToS): 按不同的服务类型 (在TypeofField中标识) 负载均衡对数据流进行分配。
- 规则模式: 针对不同的数据流设置导向规则, 用户可自行设置

## 轮询算法

- 实现原理:
  - 每一次把来自用户的请求轮流分配给内部中的服务器, 从1开始, 直到N(内部服务器个数), 然后重新开始循环

负载均衡算法—轮询



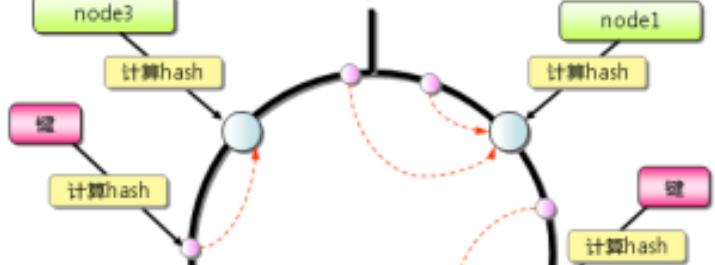
- 优点:
  - 简洁, 无状态调度。
- 缺点:
  - 轮询调度算法假设所有服务器的处理性能都相同, 不关心每台服务器的当前连接数和响应速度。当请求服务间隔时间变化比较大时, 轮询调度算法容易导致服务器间的负载不平衡。
- 适用:
  - 服务器组中的所有服务器都有相同的软硬件配置并且平均服务请求相对均衡的情况

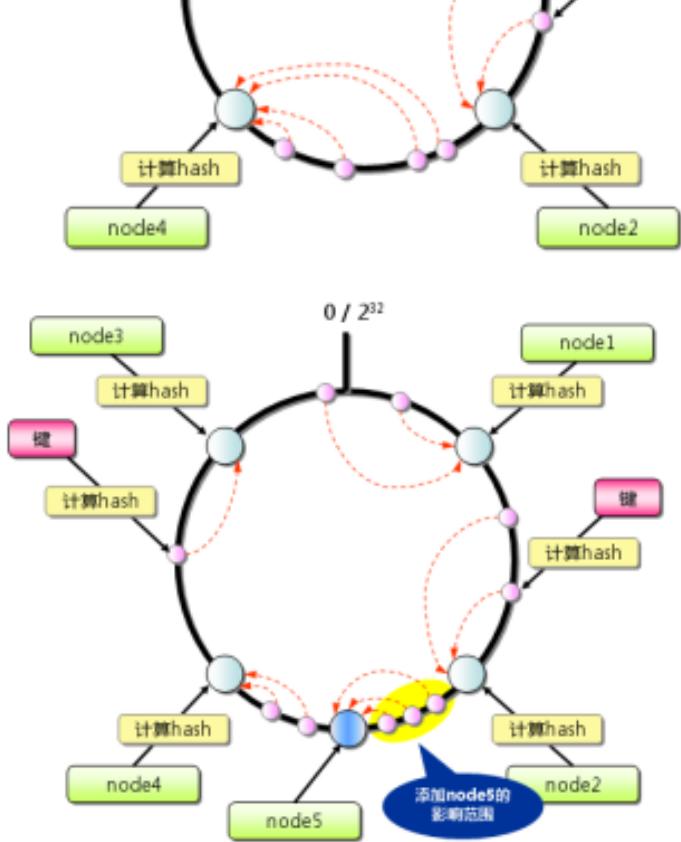
负载均衡算法—权重轮询调度



## Hash散列算法

- MD5
- 一致性Hash算法
- 各种经典Hash算法
- 自定义Hash算法

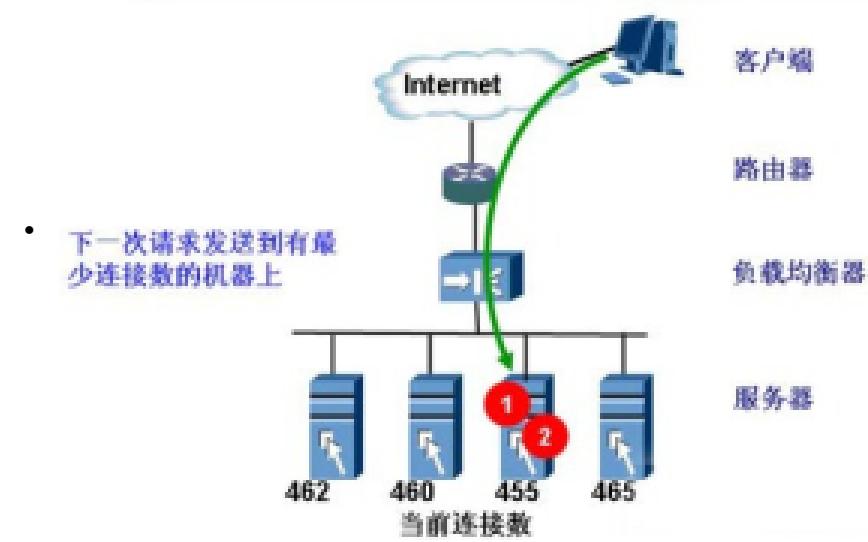




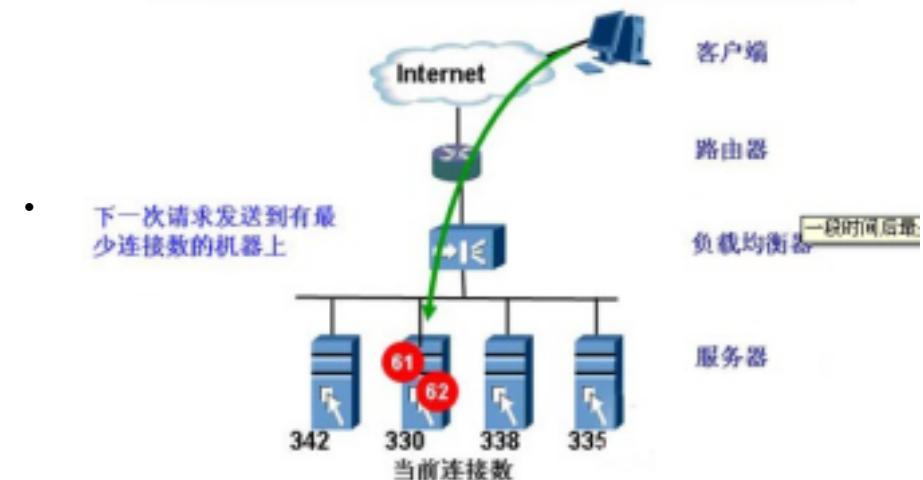
### 最少链接算法

- 实现原理:
  - 将请求分配至当前链接数最少的服务器

### 负载均衡算法—最少连接数



### 负载均衡算法—最少连接数—一定时间后



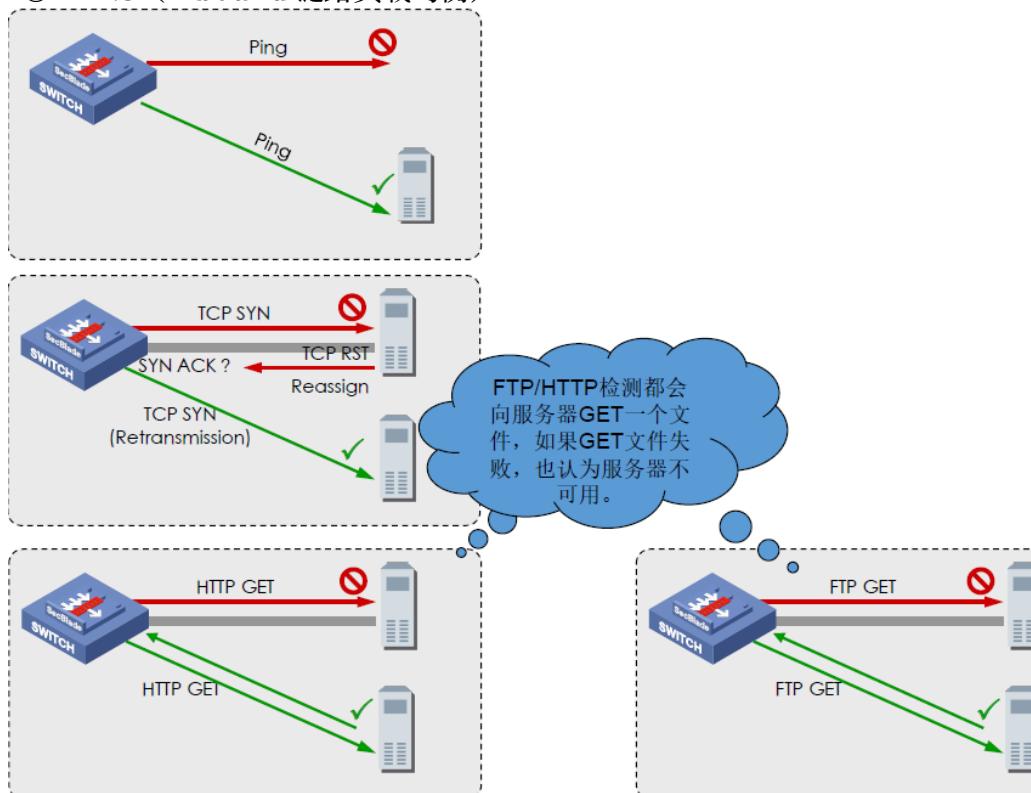
- 优点:
  - 实现起来比较简洁，在大多数情况下非常有效。
- 缺点:
  - 当各个服务器的处理能力不同时，该算法并不理想。
- 适用:
  - 需要长时处理的请求服务，如FTP等应用

## 最快链接算法

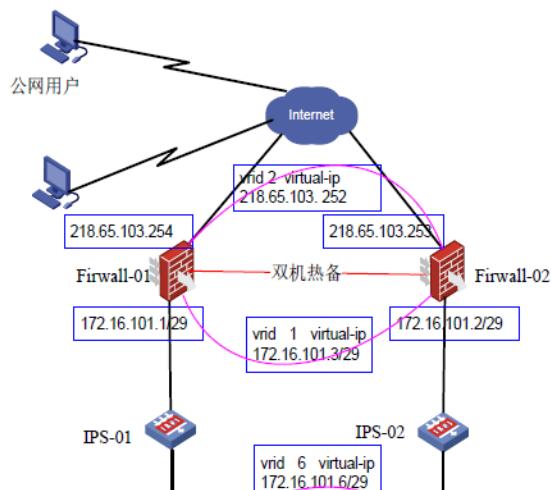
- 实现原理
  - 均衡器记录自身到每一个集群节点的网络响应时间，并将下一个到达的连接请求分配给响应时间最短的节点。
- 适用:
  - 基于拓扑结构重定向的高级均衡策略。

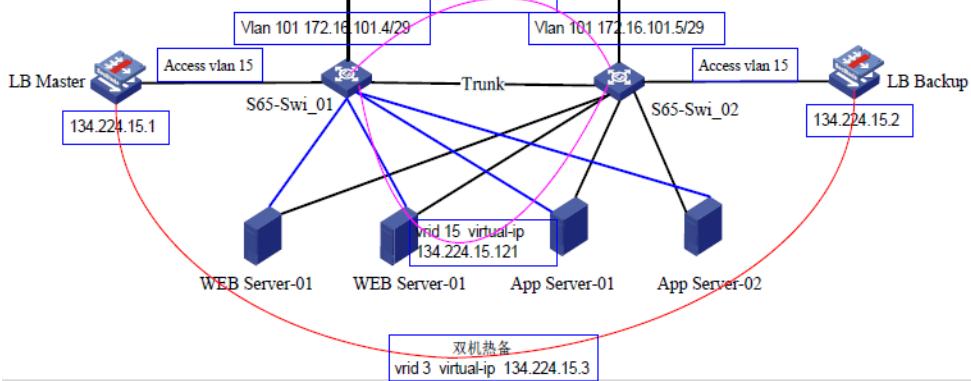
## 负载均衡产品中的关键指标：健康性检查算法

- 健康性检查算法的目的:
  - 通过某种探针机制，检查服务器群中真实服务器的健康情况，避免把客户端的请求分发给出现故障的服务器，以提高业务的HA能力。
- 目前常用的健康性检查算法:
  - Ping (ICMP)
  - TCP
  - HTTP
  - FTP
  - DNS (inbound 链路负载均衡)



## 负载均衡技术典型组网应用（双机热备）





## 6 防火墙基础

- [6.1 防火墙分类有哪些?](#)
- [6.2 防火墙的弱点有哪些?](#)
- [6.3 常见的防火墙基本结构有哪些?](#)
- [6.4 攻击防火墙的主要方法包括?](#)

### 6.1 防火墙分类有哪些?

- 包过滤防火墙
  - 第一代也是最基本形式的防火墙。根据所建立的一套规则，检查每一个通过的网络包，或者丢弃，或者放行，这称为包过滤防火墙。目的是放行正常的数据包，截住有危害的数据包
  - 例：
    - 规则1:阻断FTP、TELNET、SMTP、POP3等连接；
    - 规则2:允许HTTP连接；
    - 查看网络包的目的地址端口号，目的端口为21、23、25、110的丢弃，目的端口为80的放行。
- 状态/动态检测防火墙
  - 包过滤防火墙见到的每一个网络包都是孤立存在的，包中没有包含任何描述它在信息流中的位置的信息，该包被认为是无状态的，包中没有防火墙所关心的历史信息或未来状态。
  - 一个有状态包检查防火墙跟踪的不仅是包中包含的信息。为了跟踪包的状态，防火墙还记录有用的信息以帮助识别包，例如已有的网络连接、数据的传出请求等。
  - 状态/动态检测防火墙是在使用基本包过滤防火墙的通信上，跟踪通过防火墙的网络连接和包，以使用一组附加的标准，确定允许或拒绝通信。
  - 例1：对于传入的TCP包，只有它是在响应一个已建立的连接时，才会被允许通过。
  - 例2：对于传入的UDP包，若它所使用的地址和UDP包携带的协议与传出的连接请求相匹配，则该包就被允许通过。
- 应用程序代理防火墙
  - 应用程序代理防火墙实际上并不允许它连接的网络之间直接通信。相反，它接受来自内部网络特定用户应用程序的通信，然后建立与公共网络服务器单独的连接。网络内部的用户不直接与外部服务器通信，所以外部服务器不能直接访问内部网的任何一部分。同时，如果不为特定的应用程序安装代理程序，则这种服务是不会被支持的，不能建立任何连接。这种建立方式拒绝任何没有明确配置的连接，从而提供了额外的安全性和控制性。
  - 例：一个用户的Web浏览器可能在80端口或8080端口连接到了内部网络的HTTP代理防火墙。防火墙会接受这个连接请求并把它转到所请求的Web服务器。这种连接和转移对该用户是透明的，因为它完全是由代理防火墙自动处理的。
  - 代理防火墙通常支持的一些常见的应用程序有：HTTP，HTTPS/SSI，SMTP，POP3，IMAP，NNTP，TELNET，FTP，IRC。
- 个人防火墙
  - 个人防火墙是一种能够保护个人计算机系统安全的软件，一般是应用程序级的。它可以直接在用户的计算机上运行，使用与状态/动态检测防火墙相同的方式保护一台计算机免受攻击。
  - 通常，个人防火墙安装在计算机网络接口的较低级别上，使其可以监视传入传出网卡的所有网络通信。

- 例: 用户安装了一台个人Web服务器, 个人防火墙可能将第一个传入的Web连接加上标志, 并询问用户是否允许它通过, 用户可能允许所有的Web连接, 也可能只允许来自某些特定IP地址范围的连接等, 个人防火墙会把这条规则应用到所有传入的Web连接。

## 6.2 防火墙的弱点有哪些?

- 防火墙不能防范来自内部网络的攻击

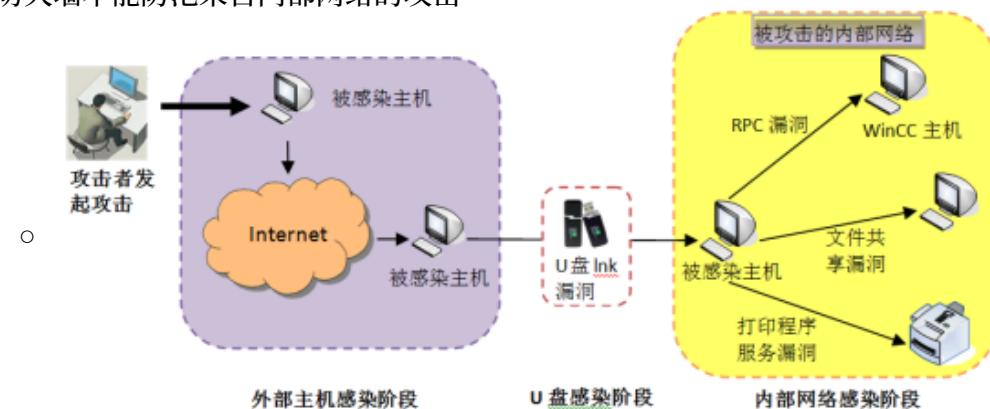
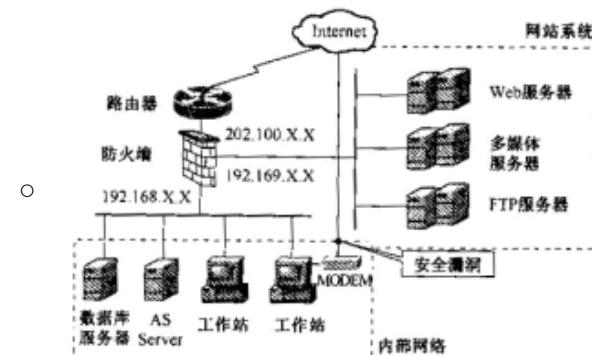


图 2: “震网”病毒的传播感染过程

- 防火墙不能防范不经由防火墙的攻击



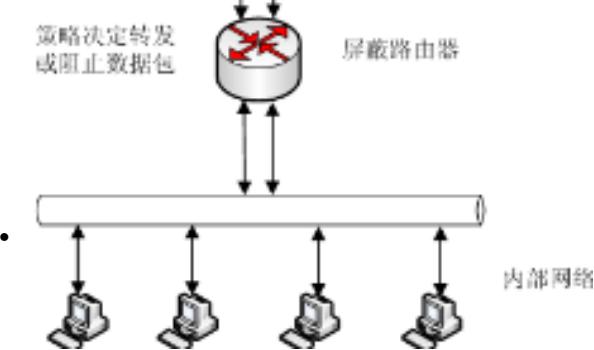
- 防火墙不能防范感染病毒的软件或文件的传输
- 防火墙不能防范数据驱动式攻击
  - 当有些表面看来无害的数据邮寄或复制到内部网的主机上并被执行时, 可能会发生数据驱动式的攻击。
- 防火墙不能防范利用标准网络协议中的缺陷进行的攻击
  - 一旦防火墙准许某些标准网络协议, 就不能防止利用该协议中的缺陷进行的攻击。
- 防火墙不能防范利用服务器漏洞进行的攻击
  - 防火墙不能防止黑客通过防火墙准许的访问端口对该服务器的漏洞进行攻击。
- 防火墙不能防范新的网络安全问题
  - 防火墙是一种被动式的防护手段, 它只能对现在已知的网络威胁起作用。随着网络攻击手段的不断更新和一些新的网络应用的出现, 不可能依靠一次性的防火墙设置来解决永远的网络安全问题。
- 防火墙可能限制有用的网络服务
  - 为提高被保护网络的安全性, 防火墙可能限制或关闭了一些有用但存在安全缺陷的网络服务。

## 6.3 常见的防火墙基本结构有哪些?

### 2.1 屏蔽路由器

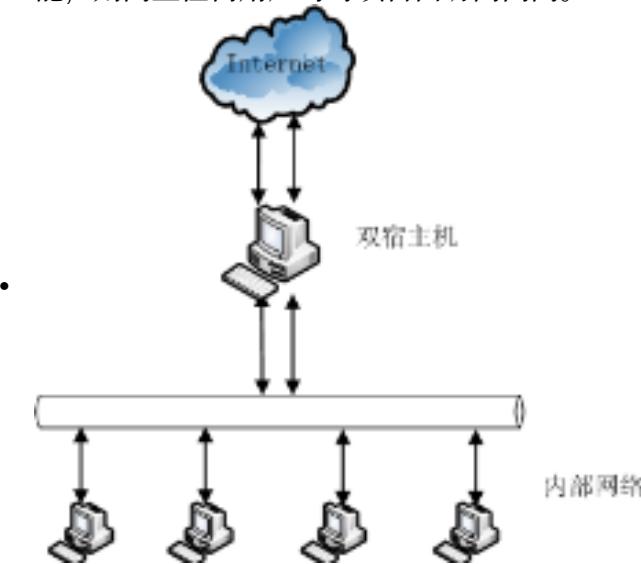
- 屏蔽路由器是防火墙最基本的构件。
  - 屏蔽路由器作为内外连接的惟一通道, 要求所有的报文都必须在此通过检查。
  - 路由器上可以安装基于IP层的报文过滤软件, 实现报文过滤功能。
  - 许多路由器本身带有报文过滤配置选项, 但一般比较简单。
  - 单纯由屏蔽路由器构成的防火墙的危险区域包括路由器本身及路由器允许访问的主机。
  - 屏蔽路由器的缺点是路由器一旦被控制后很难发现, 而且不能识别不同的用户。





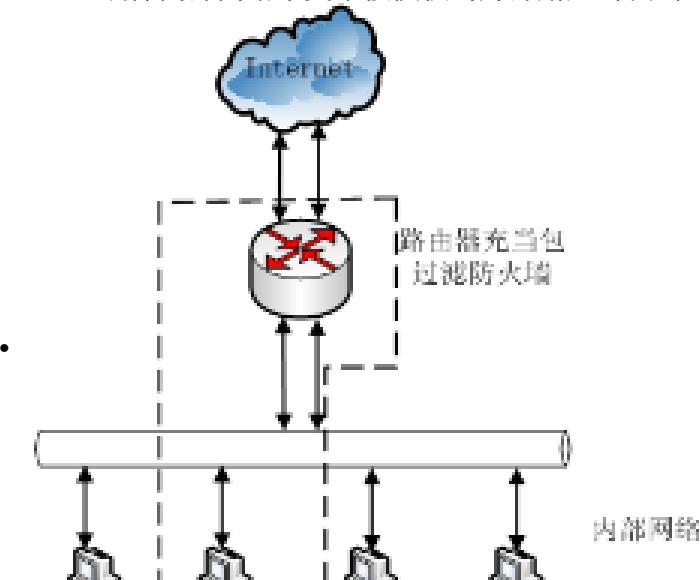
## 2.2 双宿主主机防火墙

- 用一台装有两块网卡的堡垒主机做防火墙。
  - 两块网卡分别与内部网和外部网相连。
  - 堡垒主机上运行着防火墙软件，可以转发应用程序，提供服务等。
  - 内部网和外部网之间的直接通信被完全阻止。
- 双宿主机防火墙优于屏蔽路由器的方面：
  - 堡垒主机的系统软件可用于维护系统日志、硬件复制日志或远程日志。
  - 日志对于日后的检查很有用，但不能帮助网络管理者确认内网中哪些主机可能已被黑客人侵。
- 双宿主机防火墙的一个致命弱点：一旦入侵者侵入堡垒主机并使其只具有路由功能，则网上任何用户均可以自由访问内网。



## 2.3 屏蔽主机防火墙

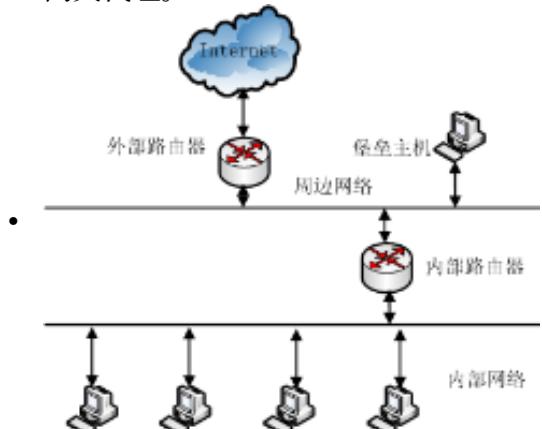
- 屏蔽主机防火墙易于实现也很安全，因此应用广泛。
  - 一个分组过滤路由器连接外部网络
  - 一个堡垒主机安装在内部网络上
  - 通常在路由器上设立过滤规则
  - 堡垒主机成为从外部网络惟一可直接到达的主机
  - 确保内部网络不受未被授权的外部用户的攻击





## 2.4 屏蔽子网防火墙

- 在内部网络和外部网络之间建立一个被隔离的子网（周边网络），用两台分组过滤路由器将这一子网分别与内部网络和外部网络分开。
- 在很多实现中，两个分组过滤路由器放在子网的两端，在子网内构成一个非军事区（DMZ）。
- 有的屏蔽子网中还设有一个堡垒主机作为唯一可访问点，支持终端交互或作为应用网关代理。

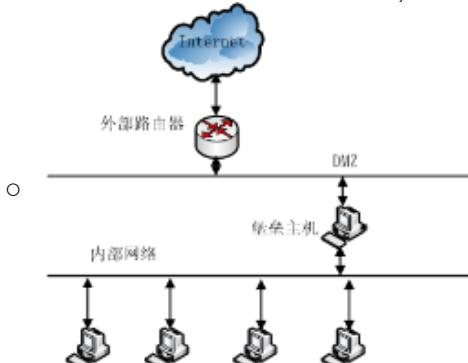


- 屏蔽子网体系结构防火墙的危险区域包括堡垒主机、子网主机及所有连接内网、外网和屏蔽子网的路由器。
- 如果攻击者试图完全破坏防火墙，他必须重新配置连接3个网的路由器，既不切断连接又不要把自己锁在外面，同时又不使自己被发现，这样的攻击还是可以完成的。
- 若禁止网络访问路由器或只允许内网中的某些主机访问，则攻击会变得很困难。在这种情况下，攻击者得先侵入堡垒主机，然后进入内网主机，再返回来破坏屏蔽路由器，而且整个过程中不能引发警报。

## 2.5 其他防火墙结构

- 一个堡垒主机和一个非军事区

- 堡垒主机一个网络接口接到非军事区（DMZ）另一个网络接口接到内部网络，过滤路由器一端接到因特网，另一端接到非军事区。



- 配置过滤路由器，只有过滤路由器规则允许的网络流量才能转发给堡垒主机。
  - 入侵者必须首先穿过过滤路由器，然后还必须穿过或者控制堡垒主机。
  - 在非军事区内没有主机。因为只有两个网络接口，所以它可以用专用的点对点连接代替，这就使得通过协议分析来获取这个连接变得更加困难。
  - 在这种结构中，堡垒主机使用双宿主机，提高了系统的安全性，可以防止入侵者绕过堡垒主机，入侵到内部网络中。

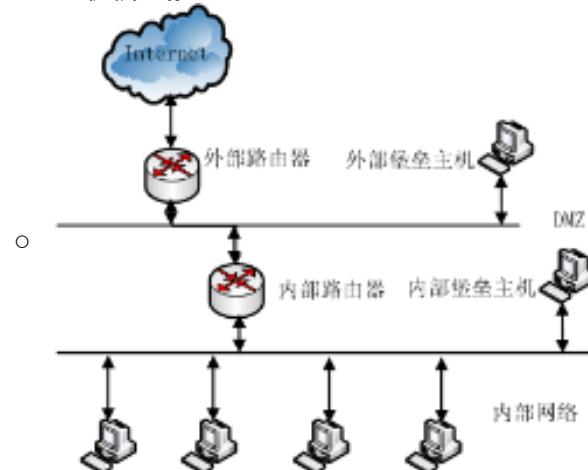
- 两个堡垒主机和两个非军事区

- 这种结构使用两台双宿堡垒主机，有两个非军事区，并在网络中分成了4个部分：内部网络、外部网络、内部非军事区和外部非军事区。
  - 过滤路由器和外部堡垒主机是外部非军事区上仅有的两个网络接口。
  - 内部非军事区受到过滤路由器和外部堡垒主机的保护，具有一定的安全性，可以把一些相对而言不是很机密的服务器放在这个网络上，并把敏感的主机隐藏在内部网络中。

- 两个堡垒主机和一个非军事区

- 使用两个具有单一网络接口的堡垒主机，加上一个内部过滤路由器作为阻塞器。

- 内部过滤路由器位于DMZ和内部网络之间。
- 该结构中，必须保证堡垒主机不被越过，还应保证两个过滤路由器使用静态路由方式。
- 这种结构要比标准的屏蔽主机的结构更为安全。因为内部网络受到双重保护，入侵者即使控制了第一个堡垒主机也不能为所欲为，还需设法攻破第二道堡垒主机防线。



## 6.4 攻击防火墙的主要方法包括？

踩点->扫描->攻击

- 踩点信息收集

- SNMP协议

- 简单网络管理协议（SNMP）允许你从网络主机上查询相关的信息。例如，你可以收集TCP/IP方面的信息，还有在路由器、工作站和其它网络组件上运行的服务情况。SNMP由网络管理系统（NMS）和代理Agent组成。NMS通常安装在一台工作站上，再将代理安装在任何需要接受管理和配置的主机上。

- Traceroute程序

- 用于路由追踪，如判断从你的主机到目标主机经过哪些路由器、跳计数、响应时间如何、是否有路由器宕掉等。大多数操作系统，包括UNIX, Novell和Windows NT，若配置了TCP/IP协议的话都会有自己版本的traceroute程序

- Whois协议

- [类似于finger] 是一种internet的目录服务，whois提供了在Internet上一台主机或某个域的所有者的信息，如管理员的姓名、通信地址、电话号码和Email地址等信息，这些信息是在官方网站whoisserver上注册的，如保存在InterNIC的数据库内。Whois命令通常是安全审计人员了解网络情况的开始。一旦你得到了Whois记录，从查询的结果还可得知primary和secondary域名服务器的信息。

- DNS服务器

- Finger协议

- 服务使你可以获取远程服务器上的用户信息。使用Finger，你可以得到：用户名，服务器名，E-mail账号，用户当前是否在线，用户登录时间等。

- Ping实用程序

- 一个公司的Web服务器可帮助你获得该公司所使用的IP地址范围。一旦你得知了HTTP服务器的IP地址，你可以使用Ping扫描工具Ping该子网的所有IP地址，这可以帮助你得到该网络的地址图。

- 扫描漏洞侦测

- 使用自制工具

- 使用专用工具SATAN等

- SATAN是为UNIX设计的，它主要是用C和Perl语言编写的。
    - SATAN用于扫描远程主机的许多已知的漏洞，可写的FTP目录，Sendmail

- 攻击

- 建立帐户

- 安装远程控制器

- 发现信任关系全面攻击

- 获取特权

## 7 防火墙技术

- [7.1 包过滤防火墙的优缺点？](#)
- [7.2 网络地址翻译技术解决的问题是什么？](#)

### 7.1 包过滤防火墙的优缺点？

- 优点：
  - 包过滤防火墙是两个网络之间访问的惟一途径，防火墙可对每个传入和传出网络的数据包实行低水平控制。
  - 每个IP包的字段都被检查，如源地址、目的地址、协议、端口等。防火墙将基于这些信息应用过滤规则。
  - 防火墙可以识别、丢弃带欺骗性源IP地址的包。
  - 包过滤通常被包含在路由器中，不需要额外的系统来处理。
- 缺点：
  - 访问控制列表的配置和维护困难
  - 包过滤防火墙难以详细了解主机之间的会话关系，容易受到欺骗
  - 基于网络层和传输层实现的包过滤防火墙难以实现对应用层服务的过滤

### 7.2 网络地址翻译技术解决的问题是什么？

- 最初设计目的是用来增加私有组织的可用地址空间和解决将现有的私有TCP/IP网络连接到互联网上的IP地址编号问题。
- 网络地址翻译技术并非为防火墙而设计，它在解决IP地址短缺的同时提供了内部主机地址隐藏功能，使其成为防火墙实现中经常采用的核心技术之一
- NAT技术还可以实现负载均衡

## 8 防火墙安全策略及其配置

- [8.1 网络审计的内容包括哪些？](#)
- [8.2 翻转掩码计算](#)
- [8.3 创建访问控制列表的步骤是什么？](#)
- [8.4 子网掩码计算](#)

### 8.1 网络审计的内容包括哪些？

- 网络审计的主要内容：
  - 网络连接审计
  - 协议审计
  - 端口审计
  - 拨号连接审计
  - 个人帐户审计
  - 文件访问审计
  - 数据审计
  - 流量统计审计
  - 数据库审计
  - WEB服务器审计
  - 安全事件再现审计
  - 键盘审计
  - 屏幕审计
  - 系统统计分析

### 8.2 翻转掩码计算

- 在CiscoISO中，网络地址的辨别和匹配不是通过子网掩码，而是通过翻转掩码。与子网掩码类似，翻转掩码是由0和1组成的32位二进制数字，分成4段。32位中的每一位正好可以和二进制IP地址的相应位对应。

十进制网络地址	192.168.9.1/255.255.255.0			
二进制IP地址	11000000	10100100	00001001	00000001
二进制子网掩码	11111111	11111111	11111111	00000000
二进制翻转掩码	00000000	00000000	00000000	11111111

### 8.3 创建访问控制列表的步骤是什么？

- 创建访问控制列表就是向某个访问控制列表中添加命令的过程。命令的一般格式是：命令+访问控制列表编号+操作+条件

- 创建某个访问控制列表并添加一条命令语句
  - Router(config)#access-list access-list-number[permit|deny]测试条件
- 删除某个访问控制列表
  - Router(config)#noaccess-list access-list-number
- 说明:
  - 可以向同一个访问控制列表写入多条语句；
  - 访问控制列表的配置命令比较繁琐，可以使用文本文件事先将命令编辑好，再复制粘贴至IOS中；
  - 使用“noaccess-list access-list-number”命令将删除整个访问控制列表。在标准和扩展访问控制列表中，不能删除访问控制列表中的某一条命令语句，只能一次删除整个访问控制列表。

## 8.4 子网掩码计算

## 9 入侵检测

- [9.1 Anderson在报告中定义了三种恶意用户？](#)
- [9.2 常见的入侵检测数据源包括哪些？](#)
- [9.3 入侵检测各种分类方法](#)
- [9.4 通用的入侵检测模型](#)
- [9.5 基于网络数据源的优势是什么？](#)
- [9.6 基于主机/网络入侵检测的优缺点？](#)

### 9.1 Anderson在报告中定义了三种恶意用户？

- 外部入侵者：系统的非授权用户
- 内部入侵者：超越合法权限的系统授权用户
- 违法者：在计算机系统上执行非法程序的合法用户

### 9.2 常见的入侵检测数据源包括哪些？

- 操作系统的审计记录
- 操作系统日志
- 应用程序日志
- 基于网络数据的信息源
- 其他安全产品提供的数据
- 网络设备提供的数据
- “带外”信息源

### 9.3 入侵检测各种分类方法

按照信息源的分类

- 基于主机的入侵检测
  - 基于主机的入侵检测通常从主机的审计记录和日志文件中获得所需的主要数据源，并辅之以主机上的其他信息，在此基础上完成检测攻击行为的任务
- 基于网络的入侵检测
  - 基于网络的入侵检测通过监听网络中的数据包来获得必要的数据来源，并通过协议分析、特征匹配、统计分析等手段发现当前发生的攻击行为

按照检测方法的分类

- 滥用入侵检测
  - 滥用入侵检测的技术基础是分析各种类型的攻击手段，并找出可能的“攻击特征”集合
- 异常入侵检测
  - 通常都会建立一个关于系统正常活动的状态模型并不断进行更新，然后将用户的当前活动情况与这个正常模型进行对比，如果发现了超过设定阈值的差异程度，则指示发现了非法攻击行为
- 滥用入侵检测与异常入侵检测的比较
  - 滥用入侵检测比异常入侵检测具备更好的确定解释能力
  - 滥用入侵检测具备较高的检测率和较低的虚警率
  - 开发规则库和特征集合方便、容易
  - 滥用检测只能检测到已知的攻击模式，模式库只有不断更新才能检测到新的攻击方式

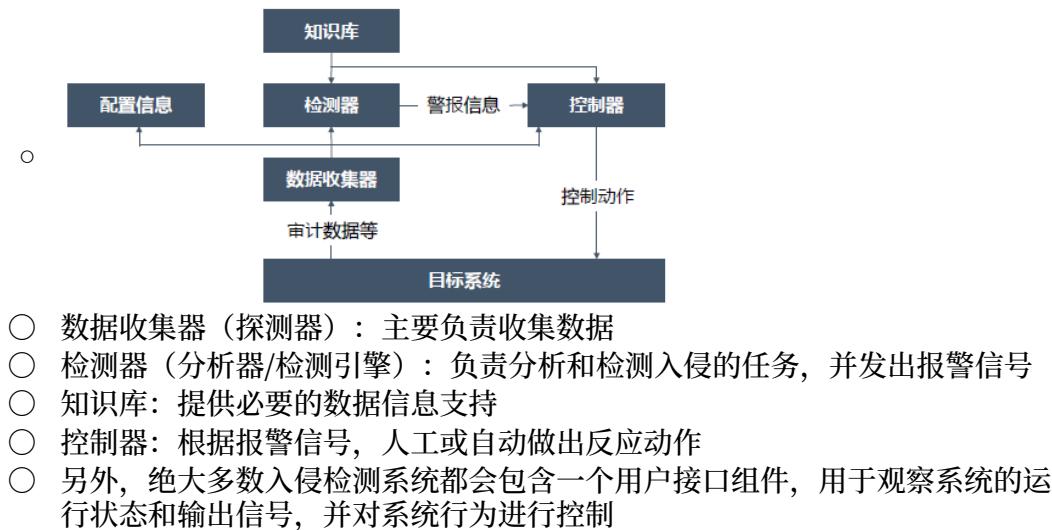
- 异常检测可以检测到未知的入侵行为

其他的分类标准

- 非实时处理系统
  - 通常在事后收集的审计日志文件基础上，进行离线分析处理，并找出可能的攻击行为踪迹，目的是进行系统配置的修补工作，防范以后的攻击
- 实时处理系统
  - 根据用户需求而定的变量，系统分析和处理的速度处于用户需求范围内

## 9.4 通用的入侵检测模型

- 入侵检测与通用入侵检测系统模型



## 9.5 基于网络数据源的优势是什么？

- 采用网络数据源有以下优势：
  - 通过网络被动监听方式获取网络数据包，作为入侵检测系统输入信息源的工作过程，对目标监控系统的运行性能几乎没有任何影响，并且通常无须改变原有网络的结构和工作方式
  - 嗅探器模块在工作时，可以采用对网络用户透明的模式，因而降低了其本身遭到入侵者攻击的概率
  - 基于网络数据的输入信息源，可以发现许多基于主机数据源所无法发现的攻击手段，例如基于网络协议的漏洞发掘过程
  - 网络数据包的标准化程度，相对主机数据源而言要高许多，例如目前几乎大部分网络协议都采用了TCP/IP协议族

## 9.6 基于主机/网络入侵检测的优缺点？

- 基于主机的入侵检测
  - 优点：
    - 能够较为准确地监测到发生在主机系统高层的复杂攻击行为
  - 缺点：
    - 无法移植
    - 影响性能
    - 无法对网络环境下发生的大量攻击行为做出及时的反映
- 基于网络的入侵检测
  - 优点：
    - 能够实时监控网络中的数据流量，并发现潜在的攻击行为和做出迅速的响应
    - 可移植性好
    - 不影响宿主机性能
  - 缺点：
    - 准确率
    - 发生在应用进程级别的攻击行为无法依靠基于网络的入侵检测来完成

## 10 基于主机的入侵检测技术

- [10.1 四种用于入侵检测统计模型](#)
- [10.2 审计数据预处理工作包括哪些？](#)
- [10.3 审计记录格式要求要点是什么？](#)

## [10.4 文件完整性检查的必要性有哪些？](#)

### 10.1 四种用于入侵检测统计模型

- 操作模型
  - 该模型主要关心对系统中所发生事件的计数度量情况，例如观察在特定时间间隔内发生的失败登录事件的次数等
  - 通常的模型操作包括将所关心的特定事件计数值与某个阈值进行比较，如果超过，则指示发生了异常情况
  - 该模型可以同时应用到异常入侵检测和滥用入侵检测技术中
- 均值和标准偏差模型
  - 该模型成立的一个假设基础：
    - 系统当前状态特征可以采用数据的均值和标准偏差两个度量参数来刻画
  - 在检测过程中，如果当前用户行为超出了可信任的区间范围，则表示为异常行为
  - 信任区间的定义通常采用参数度量与其平均值的标准偏差值
- 多元模型
  - 该模型是对均值和标准偏差模型的扩展，其主要思想是：
    - 在多个参数度量之间进行相关分析，从而摆脱单纯依赖单个度量值来判断系统当前状态的限制因素
- 马尔可夫过程模型
  - 四个统计模型中最复杂，其基础思想是：
    - 将每个审计记录中不同类型事件的出现视为随机变量的不同取值
    - 采用随机过程模型来刻画入侵检测系统的输入事件流
    - 模型中采用状态转移矩阵来表示不同事件序列出现的概率值，如果当前审计事件按照正常的状态转移矩阵所计算出的发生概率小于某个阈值，则指示为异常行为
- 四个模型比较：
  - 第一个模型“操作模型”的典型应用包括入侵检测中的阈值检测或者启发式阈值检测技术等
  - 第二和第三个模型在经典的IDES系统中得到很好的实现
  - 第四个模型在TIM系统中得到较好的体现

### 10.2 审计数据预处理工作包括哪些？

- 主机入侵检测所要进行的主要工作就是审计数据的预处理工作，包括映射、过滤和格式转化等操作。预处理工作体现在以下几个方面：
  - 不同目标系统环境的审计记录格式各不相同，对其进行格式转化的预处理操作形成标准记录格式后，将有利于系统在不同目标平台系统之间的移植；同时，有利于形成单一格式的标准审计记录流，便于后继的处理模块进行检测工作
  - 对于审计系统而言，系统中发生的所有可审计活动都会生成对应的审计记录，因此对某个时间段而言，审计记录生成速度非常快，而其中往往大量充斥着对于入侵检测而言无用的事件记录，所以需要对审计记录进行必要的映射和过滤等操作

### 10.3 审计记录格式要求要点是什么？

- 以IDES系统为例，IDES审计记录格式是基于若干考虑设计的：
  - 它必须通用程度足够高，以便能够表示目标监控系统的所有可能事件类型
  - 它应该是机器中最有效的数据表示格式，以将处理开销降低到最小程度
  - 记录格式应该按标准化设计，使IDES能够从多个不同类型的机器处接收输入记录，而无须进行任何的数据转换
  - IDES的功能取决于其所接收到的信息，因此每一个审计记录中应该尽可能地提供更多的信息。在IDES审计记录中所有能够被填入相关信息的字段都应被填写
  - 并不是所有的可检测事件都需要报告给IDES，例如SunOS系统中就不将STAT(2)系统调用报告给IDES，其原因为：
    - 这些系统调用很频繁，并通常是冗余的
    - STAT(2)调用的量很大，忽略它将能够显著提高系统运行性能

### 10.4 文件完整性检查的必要性有哪些？

- 检查文件系统完整性的必要性包括以下几个方面：
  - 攻击者在入侵成功后，经常在文件系统中安装后门或木马程序，以方便后继的攻击活动

- 攻击者还可能安装非授权的特定程序，并且替换掉特定的系统程序，以掩盖非授权程序的存在
- 为了防止攻击活动痕迹的暴露，攻击者还可能删除若干重要系统日志文件中的审计记录
- 入侵者还可能为了达成拒绝服务攻击目的或破坏目的，恶意修改若干重要服务程序的配置文件或者数据库数据，包括系统安全策略的配置信息等

## 11 基于网络的入侵检测技术

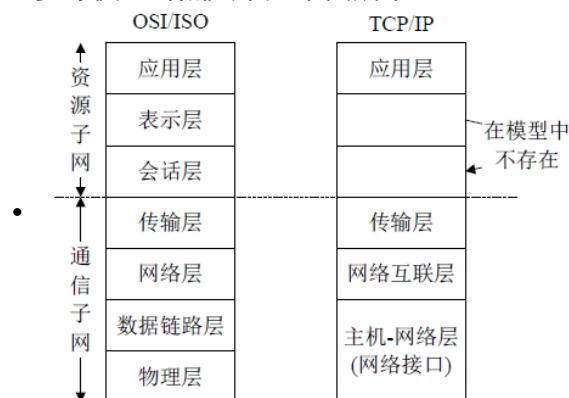
- [11.1 OSI参考模型](#)
- [11.2 TCP/IP模型](#)
- [11.3 网络数据捕获的方法有哪些？](#)

### 11.1 OSI参考模型

- OSI参考模型分为7层：
  - 应用层：提供用户网络分布信息服务的接口，如文件传送、电子邮件服务等
  - 表示层：提供两个应用层协议之间数据表示的语法，如加、解密算法等
  - 会话层：提供应用层实体会话通道的建立和清除以及会话过程的维护等
  - 传输层：提供上面面向应用的高三层和下面面向网络的低三层之间的接口，为会话层提供与具体网络无关的可靠的端对端通信机制。主要有面向连接的服务（字节流）和无连接的服务（数据报）两种类型服务
  - 网络层：建立传输层实体之间的网络（WAN或LAN）连接，包括路由选择等服务
  - 数据链路层：建立于特定网络（LAN）的物理连接上，为网络层提供可靠的传送通道，提供传输错误检测与数据重发
  - 物理层：提供网络端设备接口的物理和电气接口，与物理传输介质直接相连

### 11.2 TCP/IP模型

TCP/IP协议模型从更实用的角度出发，形成了具有高效率的四层体系结构，TCP/IP和OSI参考模型对照关系如下图所示：



TCP/IP协议模型包括四层：

- 主机—网络（网络接口）层
  - TCP/IP模型中的主机—网络（网络接口）层与OSI/ISO的物理层、数据链路层以及网络层的一部分相对应。该层中所使用的协议大多数是各通信子网固有的协议
- 网络互连层（IP层）
  - 是TCP/IP模型的关键部分。功能是使主机把分组发往任何网络，并使各分组独立地传向目的地，这种信息传送的类型称为数据报方式。这些分组到达的顺序可能和发送的顺序不同，因此当需要按顺序发送和接收时，高层必须对分组进行排序。
  - 使用IP协议，它把传输层送来的消息组装成IP数据报文，并把IP数据报文传递给主机—网络层
  - 网络互连层的主要任务是：
    - 为IP数据报分配一个全网唯一的IP地址，实现IP地址的识别与管理
    - IP数据报的路由机制
    - 发送或接收时使IP数据报的长度与通信子网所允许的数据报长度相匹配
- 传输层
  - 传输层作为应用程序提供端到端通信功能，和OSI/ISO中的传输层相似。该层协议处理网络互连层没有处理的通信问题，保证通信连接的可靠性，能够自动

适应网络的各种变化。传输层主要有两个协议——传输控制协议（TCP）和用户数据报协议（UDP）

- 应用层

- 位于传输层之上的应用层包含所有的高层协议，为用户提供所需要的各种服务，主要包括：
  - 远程登录（Telnet）：用户可以使用异地主机
  - 文件传输（FTP）：用户可以在不同主机之间传输文件
  - 电子邮件（SMTP）：用户可以通过主机和终端相互发送信件
  - Web服务（HTTP）：发布和访问具有超文本格式的HTML的各种信息
  - 域名系统（DNS）：把主机名映射成网络地址
- TCP/IP模型中的应用层与OSI/ISO中的应用层有较大的差别，它不仅包括了会话层及上面的三层的所有功能，而且还包括了应用进程本身在内

### 11.3 网络数据捕获的方法有哪些？

- 网络数据包截获方法
  - 利用以太网络的广播特性
  - 通过设置路由器的监听口或者镜像口来实现
- 以太网环境下的数据包截获方法：
  - 将网卡的工作模式置于混杂（Promisc）模式
  - 直接访问数据链路层，截获相关数据
  - 由应用程序而非上层协议如IP和TCP协议对数据进行过滤处理
  - 截获到流经网卡的所有数据
  - 不同的操作系统提供的接口功能并不相同，因此直接采用套接字设备的编程代码在不同系统平台上不能通用。这个问题的解决办法之一是使用由美国洛伦兹伯克利国家实验室（LawrenceBerkeleyNationalLaboratory）编写的专用于数据报截获功能的API函数库“Libpcap”
- 交换网络环境下的数据包截获方法：
  - 常用的方法是利用交换机或者路由器上设置监听口或者镜像口，此时所有的网络信息数据包除按照正常情况转发外，将同时转发到镜像端口，从而达到截获所有网络流量的目的。在实际工作中存在两个问题：
  - 随着交换带宽的不断增长，并非所有的网络流量都会反映在镜像口上
  - 并非所有的交换设备都提供类似的镜像口
  - 很多IDS系统会选择挂接在流量通常最大的上下行端口上，用来截获进出内外网的数据流量

## 12 先进的入侵检测技术

- [12.1 数据清洗的目的是什么？](#)
- [12.2 污染数据形成的原因是什么？](#)

### 12.1 数据清洗的目的是什么？

- 数据清理通过填写空缺值，平滑噪声数据，识别删除孤立点，并解决不一致来清理数据
- 污染数据的普遍存在，使得在大型数据库中维护数据的正确性和一致性成为一个及其困难的任务

### 12.2 污染数据形成的原因是什么？

- 滥用缩写词
- 数据输入错误
- 数据中的内嵌控制信息
- 不同的惯用语
- 重复记录
- 丢失值
- 拼写变化
- 不同的计量单位
- 过时的编码
- 含有各种噪声