探寻架构设计中的第六感

研发中心—架构部　　李晓栋

个人介绍

- 工作经验：在新浪有5年以上的系统、网络、安全相关经验
- 擅长技术领域：Linux系统内核、服务性能评估和优化
- 获奖情况：
  新浪公司2008年度技术创新奖
  所带团队与网络架构团队联合获得新浪公司2008年度优秀团队奖
- 所负责的团队：基础技术组
  - 新浪网络设备研发

    软件负载均衡系统、DDOS防火墙、IPSec-VPN

    网页挂马检测系统、流量快速分析系统
  - 新浪操作系统研发、优化
  - 技术培训

    09年Q2至今开展培训7次，有近400人次参加

目录

童年的回忆

假如我有。。。。。。

假如我是。。。。。。

童年的回忆

长大以后我就成了你。。。。。

童年的回忆

PK

第六感小宇宙燃烧吧！

省钱大比拼！

君子博学而日三省乎己！

省钱大比拼！

新浪软件负载均衡系统-现状

所用开源软件

- LVS
- Haproxy
- Keepalived

部署规模

- 从07年9月开始筹划
- 07年12月~08年7月上线百台以上

服务质量

- 历经07股票牛市、512汶川地震、2008北京奥运、改革开放30年、2009日全食等重大历史事件的突发流量考验

成本收益

- 为公司节约成本1千万元（含人力、机架、端口等）
- 日常运维：1名网络工程师兼顾LVS运维

了新浪网络设备研发、Linux内核研究的序幕

省钱大比拼！

新浪软件负载均衡系统-初期遇到困难

推广问题

- 谁愿意成为“白老鼠”？
- 如何与业务部门达成共识？

运维问题

- 研发人员和网络工程师如何互相取长补短？
- 7×24小时的技术支持由谁来做？
- 比F5等设备配置繁琐如何解决？

成本控制

- 人力成本随部署规模的增长而大幅度增加
- 机架、交换机端口等方面的成本问题

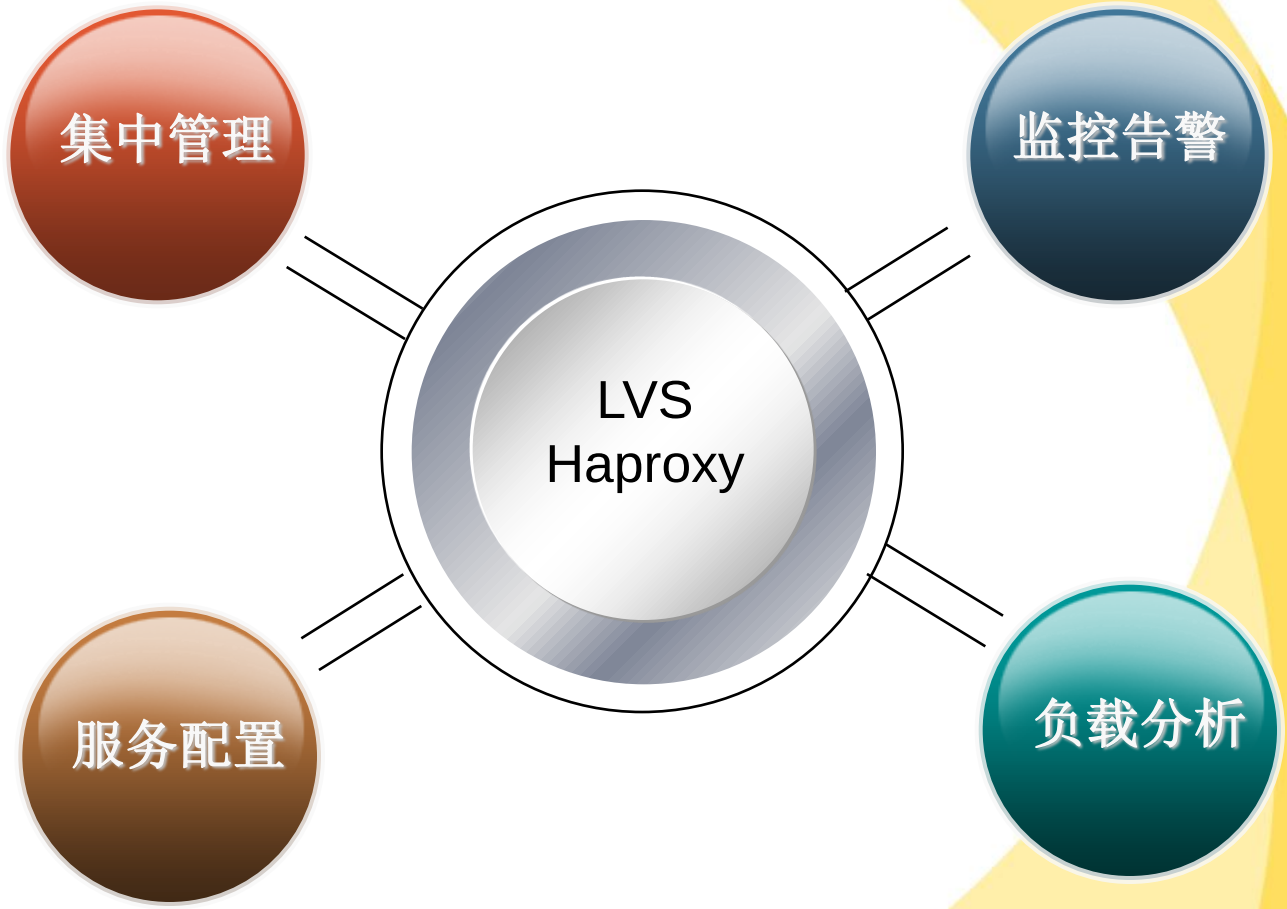
风险管理

- 粗心导致的上线及配置变更错误带来的风险
- 未知BUG导致的宕机风险

处理好 风险和 工程师积极性方面 的关系

省钱大比拼！

新浪软件负载均衡系统-成功的秘密1



省钱大比拼！

- 集中管理

文件同步规则表 | 添加规则

	名称
⊞	所有服务器文件同步规则
⊞	Haproxy配置文件
⊞	配置管理系统相关同步
⊞	VPN专用同步规则
⊞	防火墙专用规则

已关联组
- 所有LVS/Haproxy

报警规则表 | 添加规则

	名称	描述
⊞	default	default notification rule will be used if no rules specified.
⊞	confsync	this rule is used for confsync program only.
⊟	misc	email and sms. [redacted]
	报警方法	• email • sms
	自动关联	Yes
	级别	0
	联系人	• [redacted]@staff.sina.com.cn • [redacted]@staff.sina.com.cn • [redacted]@staff.sina.com.cn • [redacted]@staff.sina.com.cn • [redacted]@staff.sina.com.cn • 138[redacted] • 13[redacted] • 13[redacted] • 138[redacted]
	标题	
	次数限制	2/5000000

负载分析策略表 | 添加策略

	名称	
⊞	default	filter the data between 0 and 8 for a
⊟	finance_SH	filter the data between 16 and 23 fo
	应用对象	hour
	是否开启	1
	开始时间	16
	结束时间	23

省钱大比拼！

•监控告警：

在长期软件负载均衡运维中，根据经验精选了10余种监控告警脚本
低误报率、避免视觉疲劳

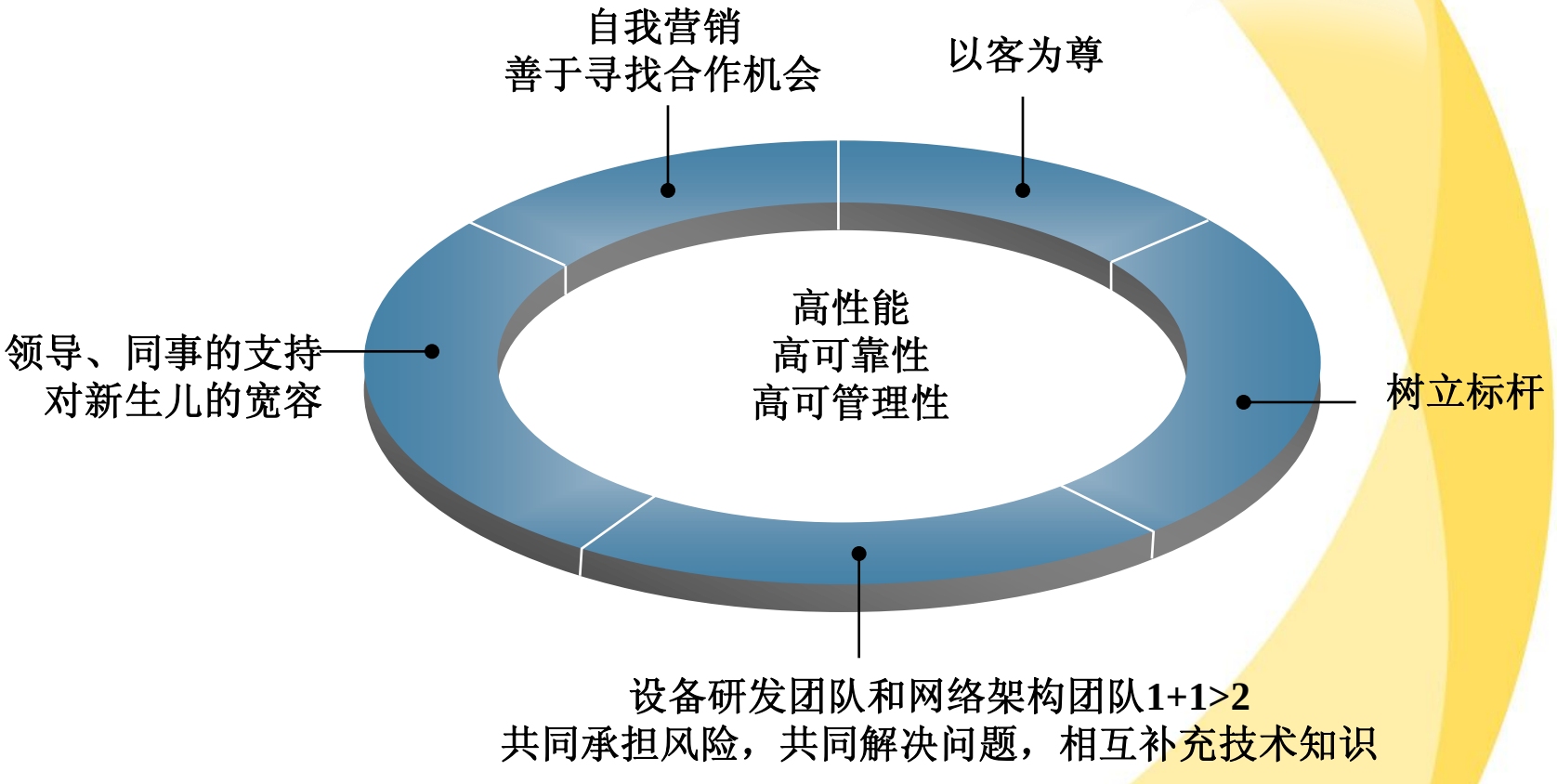
•负载分析：

•服务配置：

解决命令行方式配置时的繁琐、易出错问题

省钱大比拼！

新浪软件负载均衡系统-成功的秘密2



火焰山新传

俺老孙再也不用找那铁扇公主了

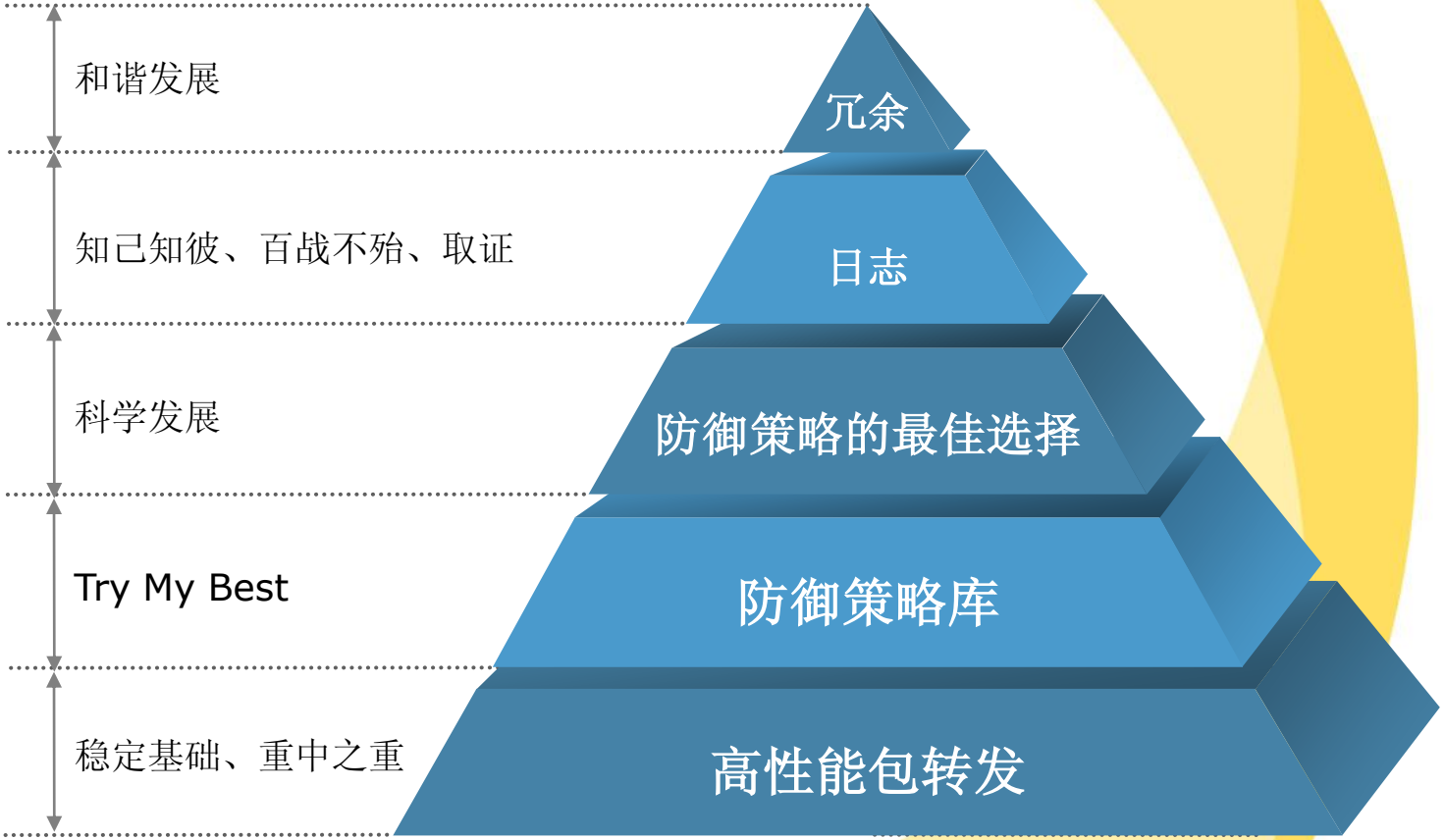
有了新浪防火墙

火焰山新传

新浪DDOS防火墙简介

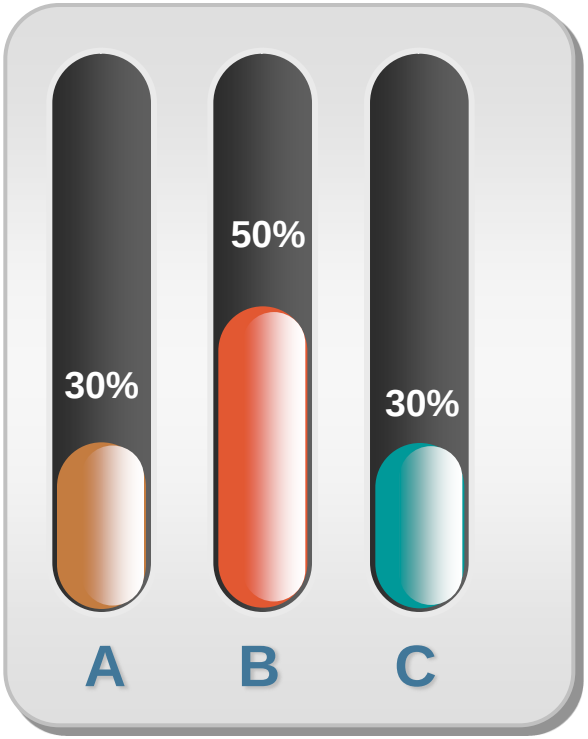
- 09年架构部-基础技术团队拳头产品
- 核心功能由新浪自主研发
- 融入了多项创新技术

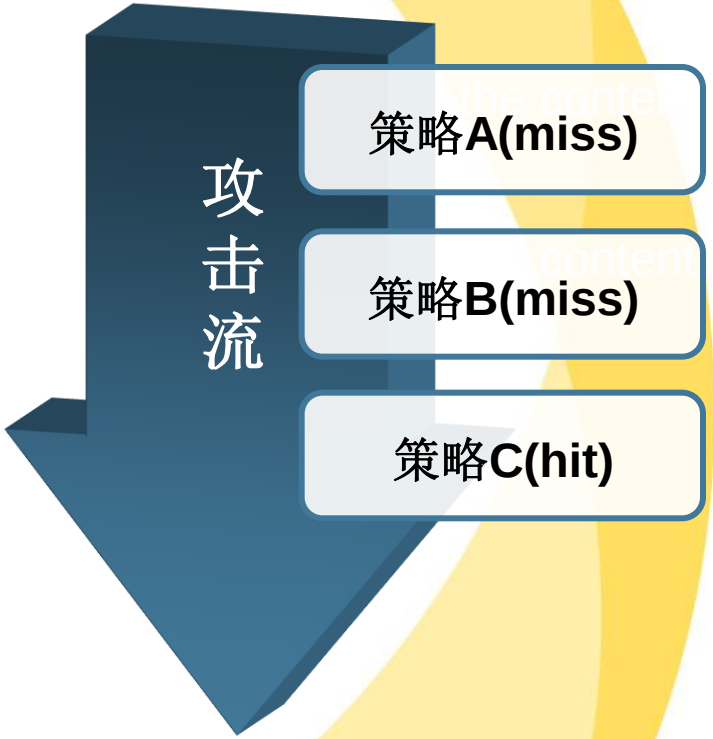
DDOS防火墙宝塔图



火焰山新传

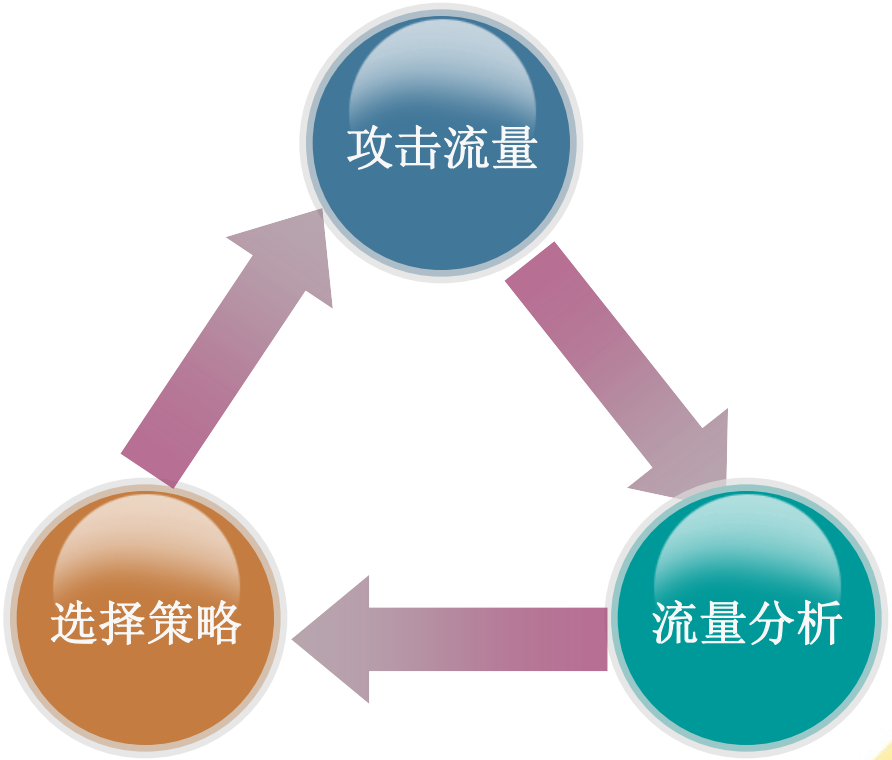
针对攻击流选取最佳防御策略----必要性





各防御策略的CPU占用率

针对攻击流选取最佳防御策略----常规解决方法

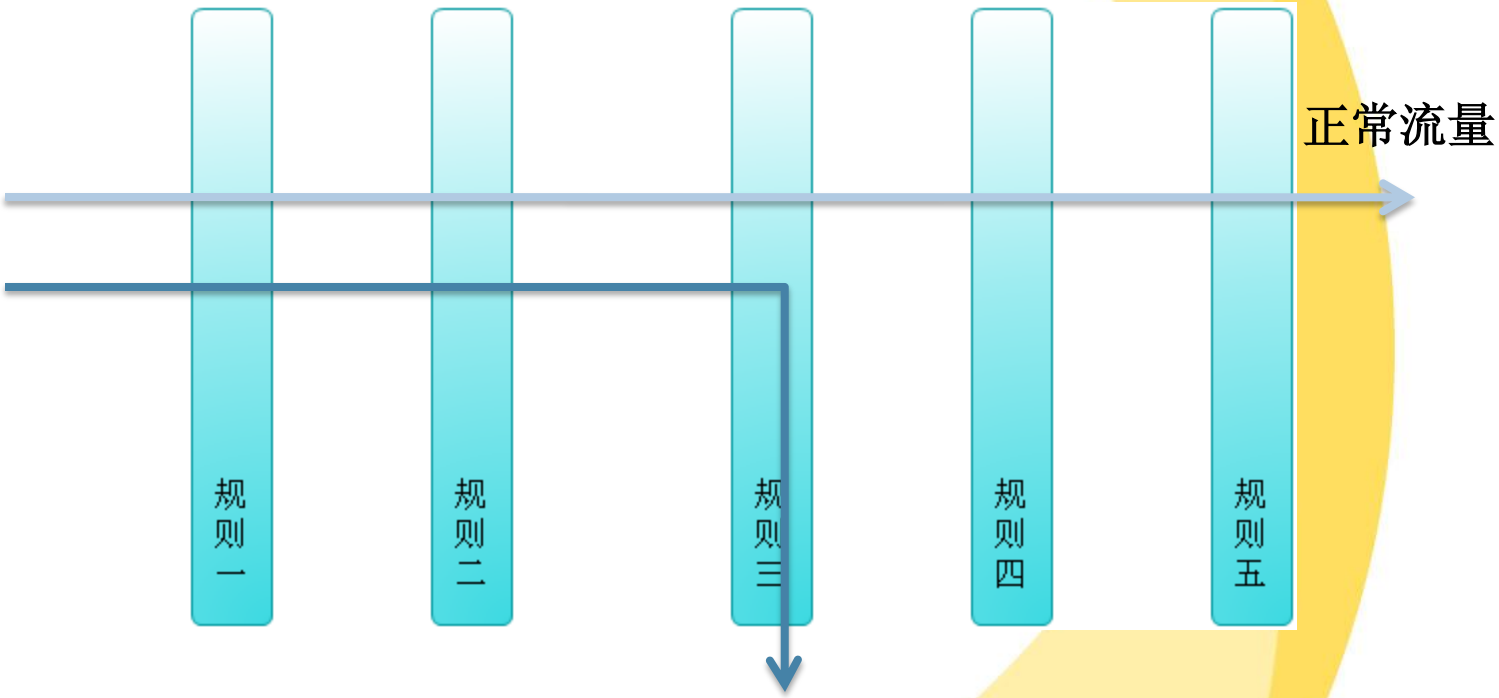


实际效果未必好！

实现的成本较高！

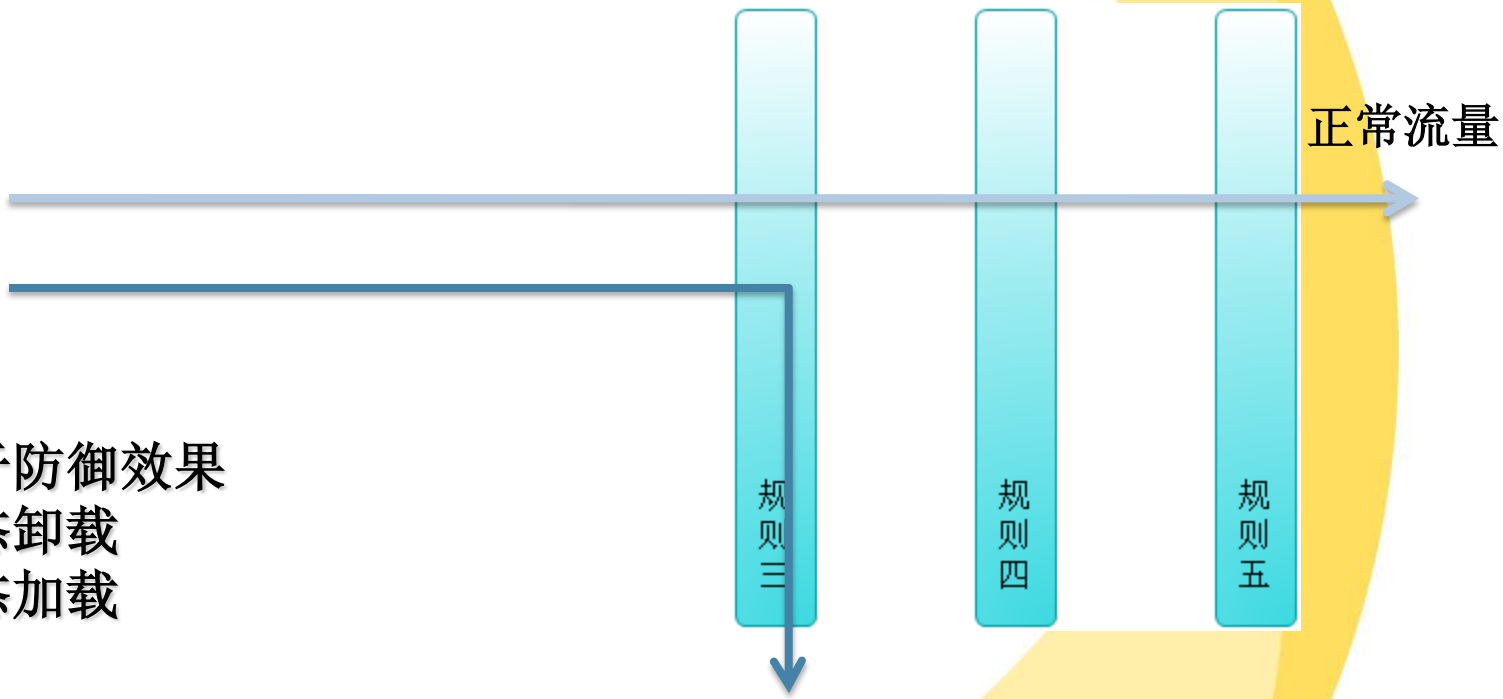
火焰山新传

针对攻击流选取最佳防御策略----新浪解决之道



火焰山新传

针对攻击流选取最佳防御策略----新浪解决之道



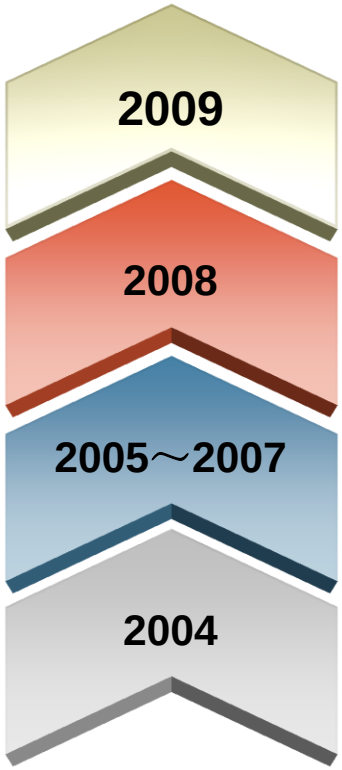
基于防御效果
动态卸载
动态加载

侦破70码！

监控录像去哪里了？

侦破70码！

我的流量分析历程



- 2009：•提出“流量快速分析”理念
- 2008：• sniffer(零拷贝网卡、pfring)
- 2005～2007：• netflow、sflow、SNMP、sniffer(pfring)
- 2004：• MRTG、RRDTool、 ntop

侦破70码！

流量分析过程中的一些困惑

- SNMP、netflow无法看到包的Payload
- IDC数量、带宽的猛增给全网sniffer分析带来很大的成本和性能挑战
- 用户分析需求的不确定性
- 流量分析是手段而不是目的，用户更关注的是否能实现目标
- 实时sniffer分析结果的使用率问题
- Tcpdump、wireshark等工具使用不方便

。 。 。 。 。

侦破70码！

新浪流量快速分析系统

抓包
- tcpdump
- Wireshark

上传pcap包
- Rsync
- Web方式

选择分析插件
- 源目标IP数据包数
- tcp 标志位统计
- Tcp握手时间统计
- ……

输出分析结果
- Web方式

流量快速分析系统

- **实际效果：**
在多次异常流量分析中发挥了重要作用

- **低成本、高灵活性**

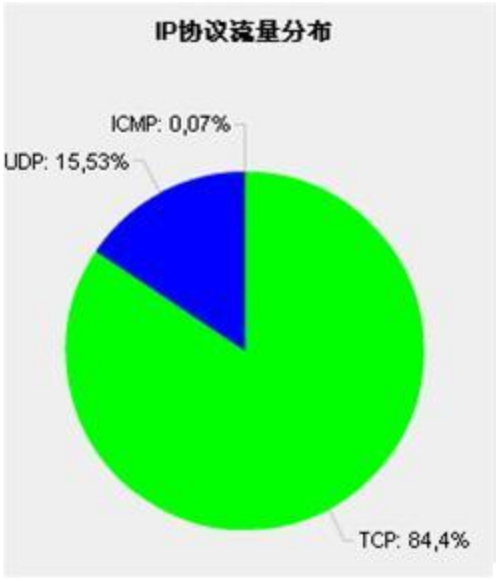
侦破70码！

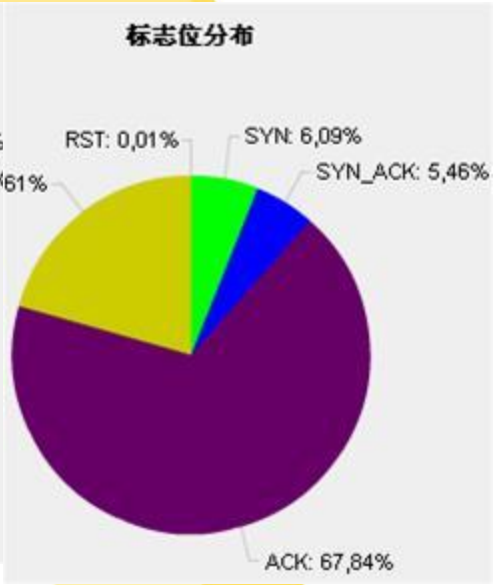
- pcap包上传与插件选择

侦破70码！

- guiyin-all_20090311191011
  - 按插件分析
    - 握手分析
    - 流量分析
    - 访问统计
    - 源地址统计
    - TCP明细
    - UDP明细
    - ARP明细
    - ICMP明细
    - ICMP概况
    - UDP概况
    - ARP概况
    - TCP概况
  - 按协议分析
    - IP分析
      - TCP
        - HTTP
          - 响应时间
          - 请求时间
          - 返回值
        - 通用
          - 握手时间
          - 标志位
      - ICMP
        - 应答

•流量分析结果：





源地址统计 TOP20		
源地址	流量（M）	包数（个）
[illegible]24	717.80	67751
[illegible]	469.93	40841
[illegible]227	391.00	97878
[illegible]	371.66	34890
[illegible]	342.30	248526

侦破70码！

新想法－－－面向症状的流量分析

选择症状

- 交换机CPU有些高
- Ping某台服务器的RTT值特别大,且有丢包
- 出口流量出现异常波动

……

结果

- 发现ARP Flood
- 正遭受DDOS攻击，攻击特征为……
- IP地址*.*.*.* 占用了500Mb的带宽

……

茴香豆的奥秘

茴的几种写法

茴香豆的奥秘

精彩继续进行

2009.4~2009.7

巧用iptables获得TCP各端口流量统计
iptables规则的变通配置法

2009.4

服务器双网口利用方案：
等价路由 PK Port-channel

2008.7

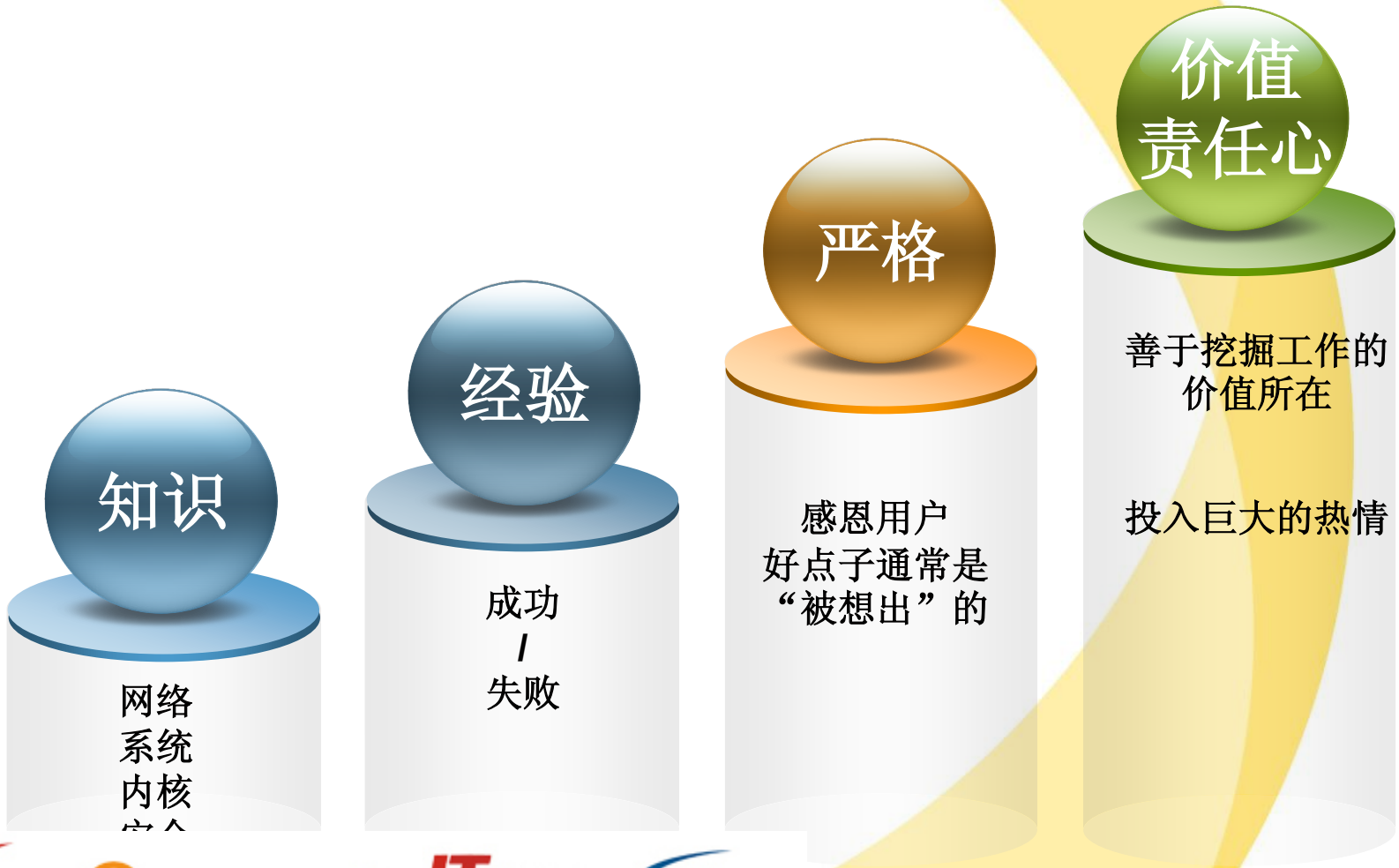
• LVS特殊单臂解决方案
“0”硬件成本

2008.3

• 利用交换机Port-channel做负载均衡

总结

第六感小宇宙如何被“激发”？



You are the one

一切由你开始

Q & A

感谢您 Thank you!