SACC
2012中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2012
架构设计 · 自动化运维 · 云计算

搜狗VPS架构演进与运维实践

■ 搜狗运维部系统组 裴彤

■ 2012-09-13

自我介绍

本科：北师大物理系，物理学

石家庄二中，高中物理教师

硕士研究生：中科院国家天文台，天文技术

搜狗运维部系统组，运维工程师

邮箱：peitong@sogou-inc.com

QQ：39397166

主要内容

Ø VPS：what & why

Ø 搜狗 VPS 概况

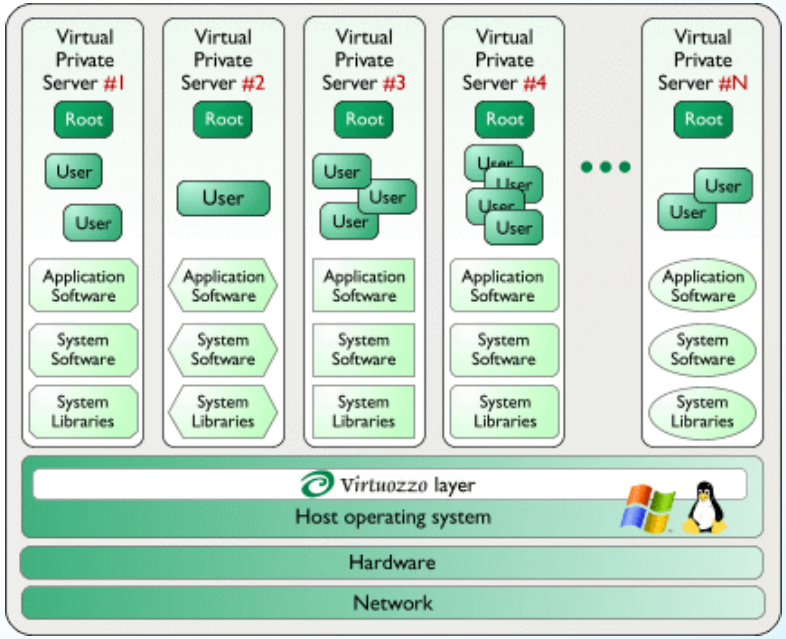
Ø 搜狗 VPS 架构演进

Ø 一些经验&问题&设想

What - 什么是VPS？

- VPS全称“Virtual Private Server”，一般译为“虚拟专用服务器”，指的是利用虚拟化软件在一台物理服务器上创建的多个相互隔离的小服务器。

- 每个VPS都可分配独立公/私网IP、独立操作系统、独立存储空间、独立内存、独立CPU资源、独立执行程序和独立系统配置等。
- 常见的虚拟化方案：VMware，xen，kvm，virtualbox 等。



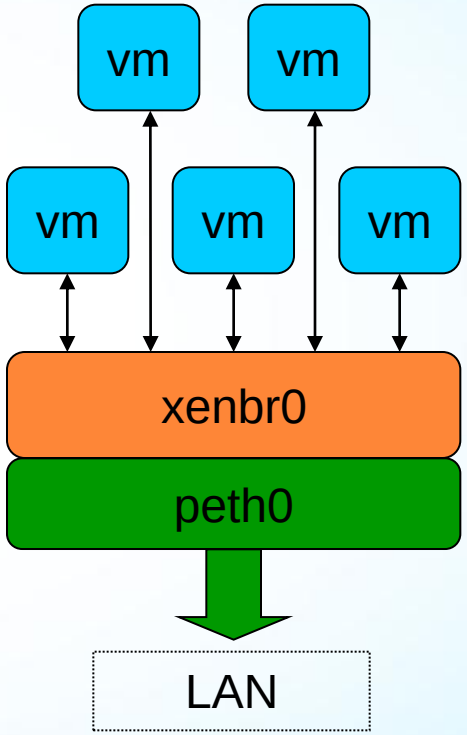
Why - 为什么要发展VPS？

- 初级阶段
  - 用较低甚至极低的成本获得独立主机，用作个人开发/测试机、低负载线上服务器等
  - 老旧物理服务器P2V，实现冗余，腾出机架
  - 灵活分配物理服务器硬件资源，提高资源利用率

- 高级阶段
  - 提高业务连续性，如通过动态迁移，做到硬件故障、升级、搬迁时无需停机
  - 物理服务器间负载均衡
  - 改善灾难恢复工作，如硬盘数据恢复等
  - 桌面虚拟化

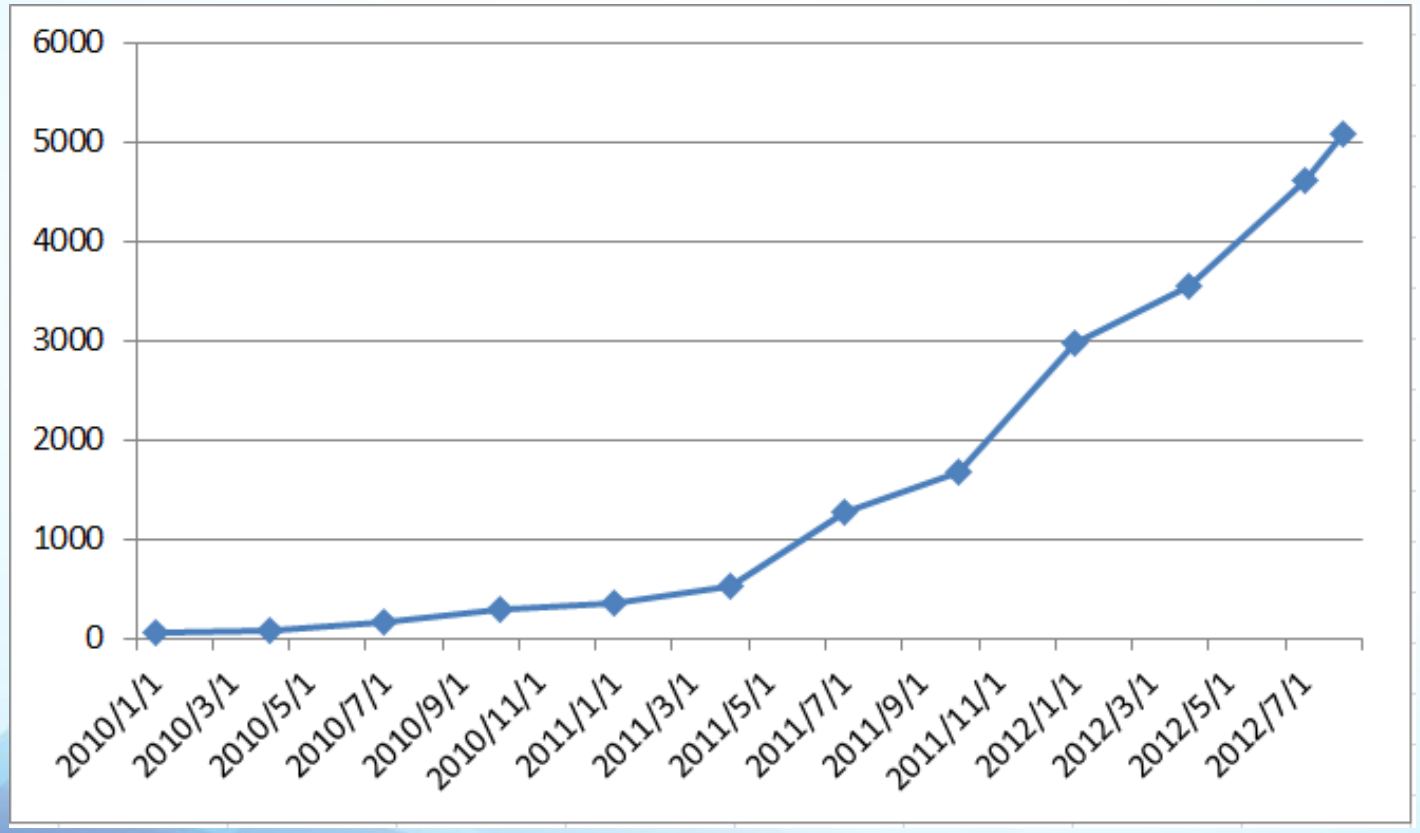
搜狗 VPS 概况

- 2009年开始至今，从无到有，宿主集群从几台扩到几十台乃至数百台，虚机增加到5000+，运维人手始终 2 * 0.5 个
- 使用本地存储，每台宿主可提供90G内存、3.5T硬盘
- 一个公用宿主集群，若干个专用宿主群
- 多为 CentOS 5 + XEN 平台，少量 CentOS 6 + kvm
- 虚机网络采用桥接方式，同宿主虚机共享4块千兆网卡



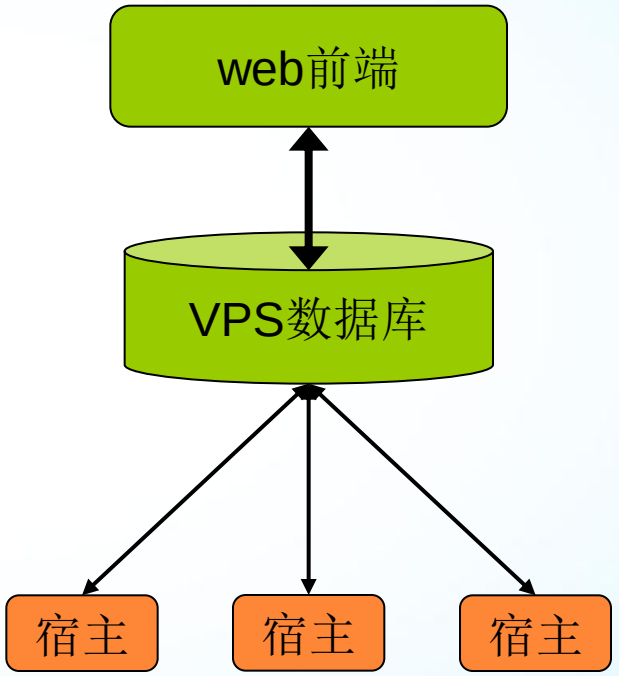
搜狗 VPS 概况

- 虚机增长曲线



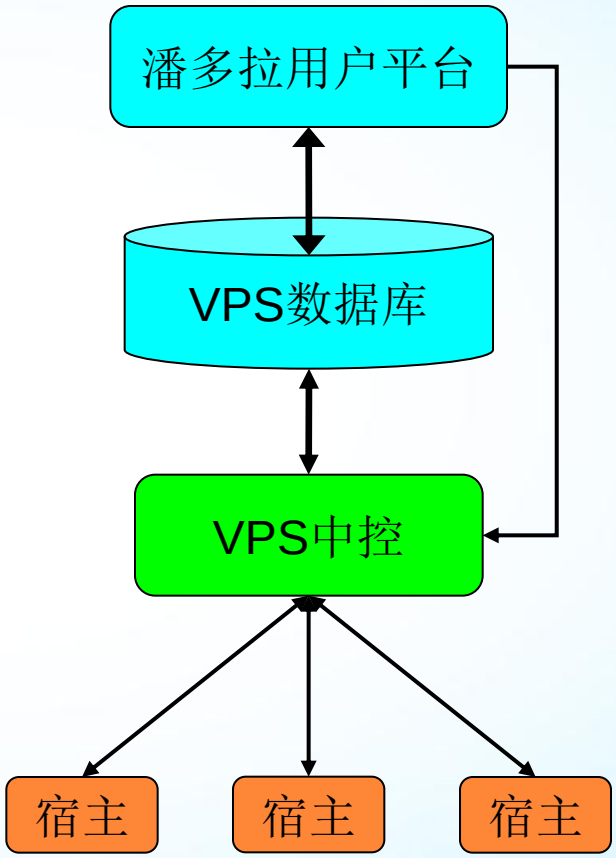
搜狗 VPS 架构演进（规范化->平台化->自动化）

- 很久很久以前（数台宿主）
  - 只有几台宿主，用 wiki 记录虚机信息。
- 后来（数十台宿主）
  - 使用数据库，开发了简单的 web 界面，包括虚机申请、虚机列表、宿主列表、可用 ip 等
  - 人工 ssh 登录宿主，执行脚本创建、删除虚机
  - 虚机申请、交付有邮件提醒
  - 每台宿主定期运行脚本，将宿主/虚机信息收集入库（更新）



搜狗 VPS 架构演进（规范化->平台化->自动化）

- 现在（数百台宿主）
  - 设置VPS中控，宿主上部署 sogou-vps-agent，接收并响应来自中控 sogou-vps-manager 的指令
  - 宿主不再直接读写数据库，只与VPS中控交互，中控与再数据库交互
  - 所有操作（虚机创建/删除/修改/重启/重装/分ip/收集信息...）都在中控上进行
  - 中控增加web界面，鼠标操作
  - 用户界面升级，迁入“潘多拉”平台



界面截图

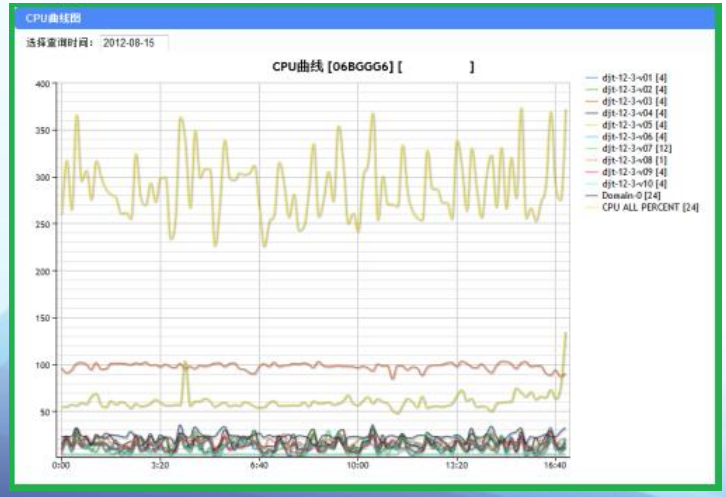
- 用户可在“潘多拉”平台上查询自己的虚机，查看详细信息，进行重启/申请下线/续约等操作

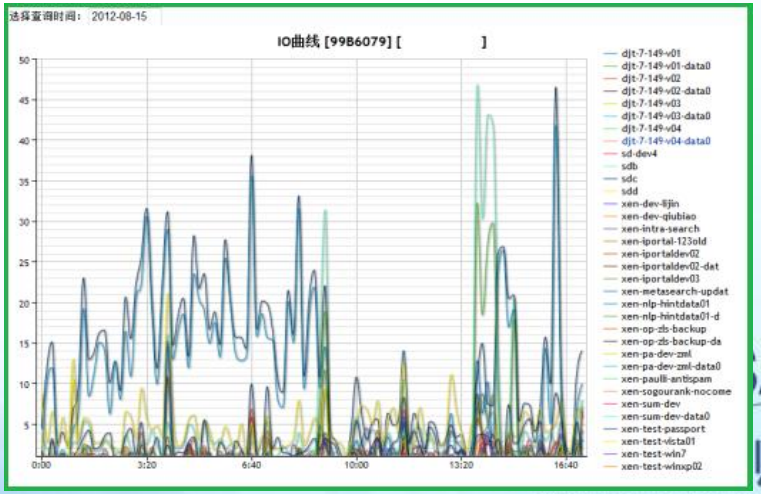
界面截图

- 各种统计：资源总览，cpu曲线，io 曲线，各事业部某段时间内消耗的资源，等等

当前位置：先知 >> 私有云管理 >> 云总览

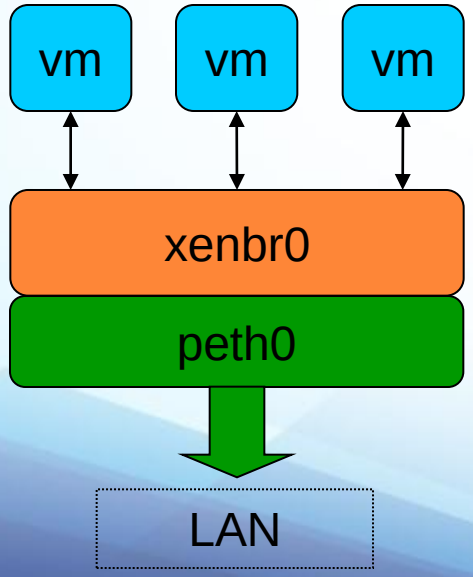
服务名称	宿主数	虚机数	vcpu_total	vcpu_free	vcpu_usage	mem_total	mem_free	mem_usa
宿主机	36	202	3816	2653	30.5%	2183.6	-182.6	108.4%
VPS_	217	2441	31696	9430	70.2%	18446.7	2255.8	87.8%
vps	5	5	560	554	1.1%	209.9	199.9	4.8%
VPS_O	207	2426	30704	8482	72.4%	18045.7	1898.8	89.5%
kvm	3	17	432	200	53.7%	218.1	28.1	87.1%
main_d	44	524	6016	1257	79.1%	3957.5	524.3	86.8%

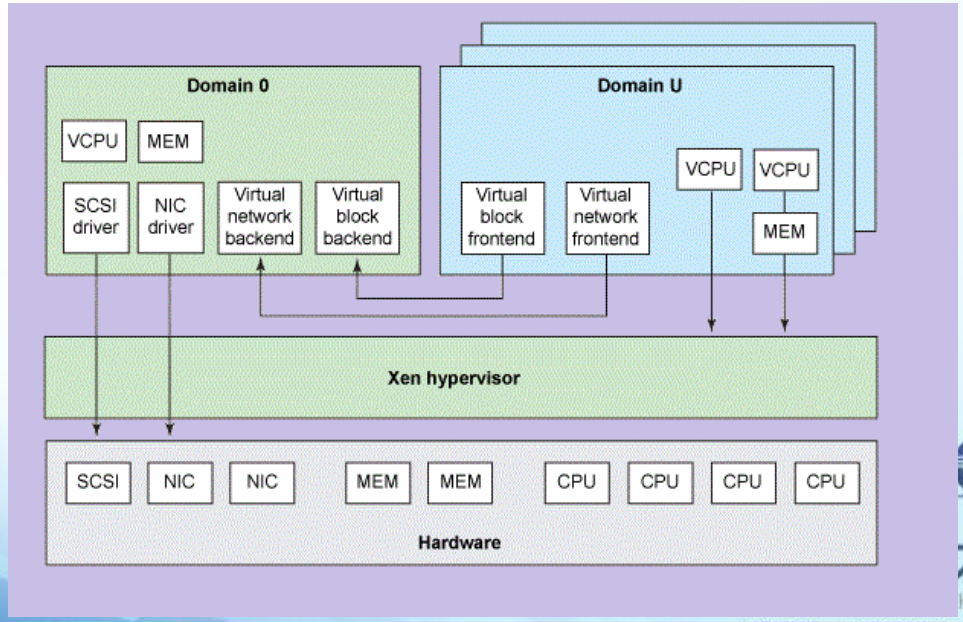




一些问题及解决方案

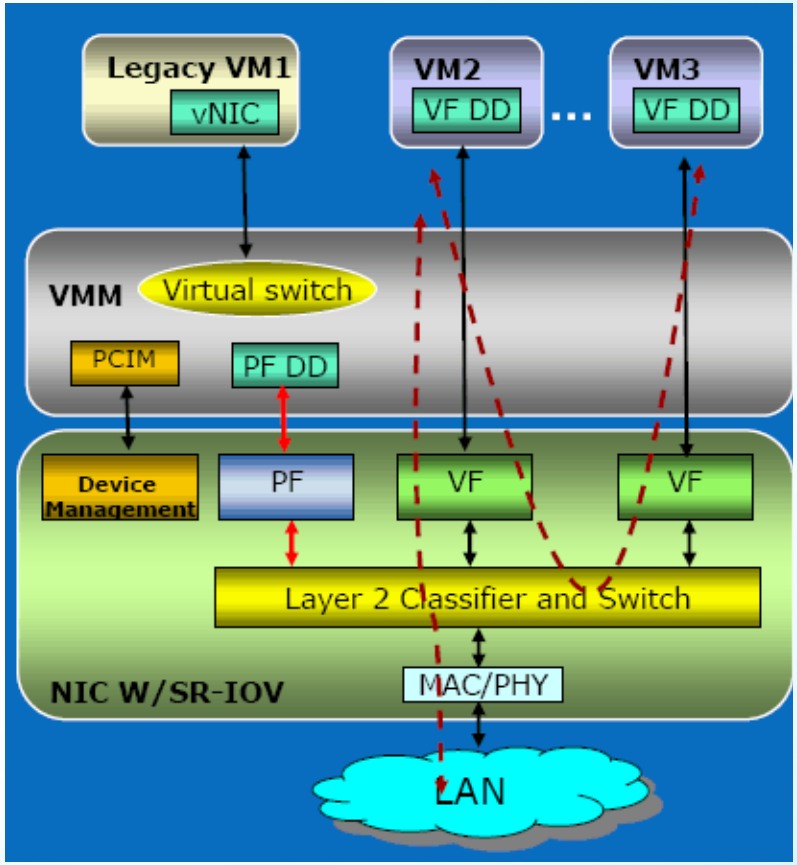
- 虚拟网桥性能不理想
  - 宿主繁忙时，丢包严重
    - 分析原因：DomU 的网络、磁盘IO 均由 Dom0 处理
    - 解决办法：为 dom0 保留独享CPU核心和内存





一些问题及解决方案

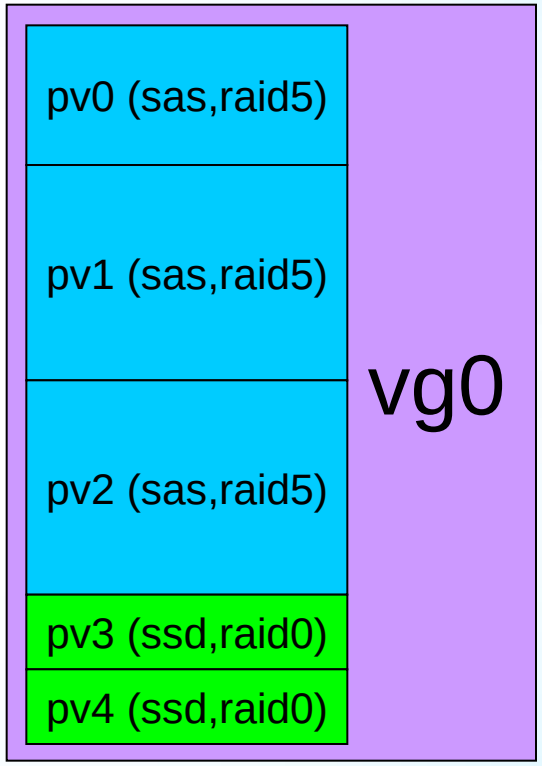
- 虚拟网桥性能不理想
  - 极限性能只能达到物理网卡的60%~70%
    - 虚网桥交换性能存在瓶颈
    - 可选方案：SR-IOV 等硬件方案



一些问题及解决方案

■ 本地存储

- 优点：架构简单，性能较好，可做存储分级
- 缺点1：IO不易隔离
  - 初步解决：分 pv ，隔离 io；同 vg，便于 pvmove
  - 后续：CentOS 6，Control Groups (cgroups)
- 缺点2：扩容受限，不便做快照、快速部署；虚机迁移困难
  - 设想解决方案：外部存储



一些问题及解决方案

- 不同类型虚机之间的协调（分类，隔离，资源复用）
  - 某些业务有专用宿主群
  - 虚机分为 dev 和 srv ，每台宿主限制 srv 虚机数量
  - 资源权重调节
    - CPU：sched-credit weight/cap
    - 磁盘和网络IO还不能很好控制（考虑kvm，cgroups）

未**来目标**

- 虚机性能更好，可靠性更高，支持动态漂移；大规模宿主集群，物理资源更加灵活地调配。
  - 基础架构充分优化，网络硬件针对虚拟化专门设计
    - 大二层，跨DC，流控等等
  - 使用统一的外部存储（网络存储集群）
  - 更专业的平台（用户/管理员界面，资源监控，预警，报表 等等）
    - openstack

Sogou搜狗

谢谢！ Q&A. . .