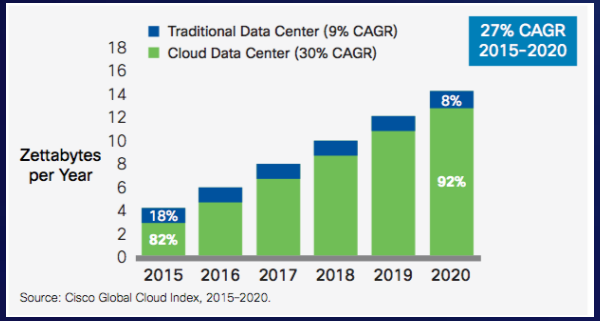
“云汉”- 逢云化雨的技术决策论

UCLOUD云产品的技术实践之路

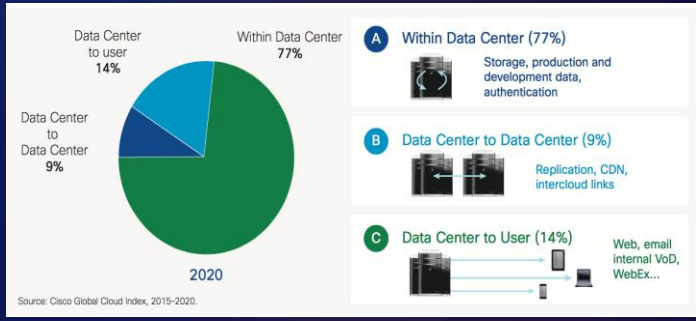
VP OF PRODUCT MARKETING 许杨毅

趋势 1 - 计算规模和存储规模

- “云”入佳境 – 2020年，92%数据在云端

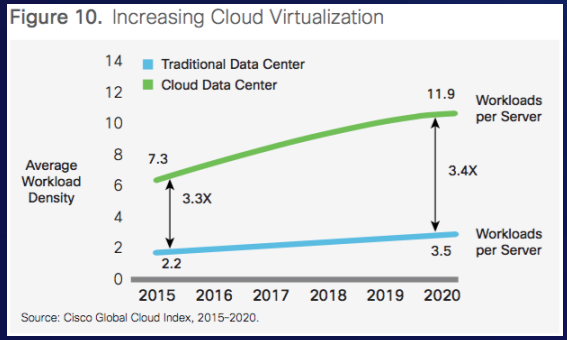


- 全球计算市场趋势 – 2020年，77%的计算在云端

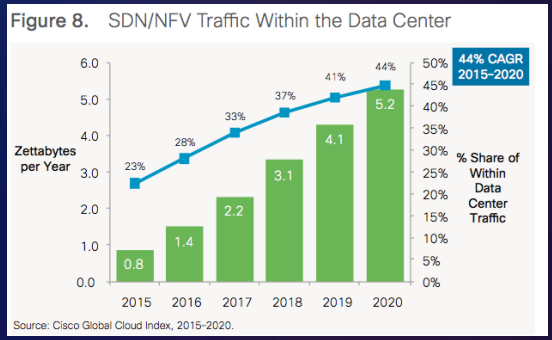


趋势2-计算效率和NFV/SDN

- 云计算基石 – 虚拟化潮流不可阻挡



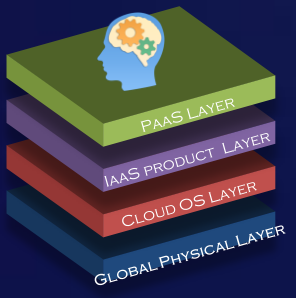
- 云计算的基石 - SDN/NFV的技术成熟度



UCloud云OS操作系统

- UCloud云计算平台内核
- **云计算操作系统**
- **5年研发历程**

- 云大脑
  - 跨域资源管理
  - 复杂任务调度
  - 海量租户隔离
  - 多层应用监控
  - SRE工程体系
  - Boss运营优化



PaaS
UDocker U

基础产品
UHost UDB
SDN VPC ULB
UMS UConnect

云OS
资源管理 资源调度
计费安全 基础运维
SRE体系

全球资源
14地域 19可用区 50 IDC
10万规模主机

运营支撑体系
BOSS系统
SRE工程体系
数据运营系统

计费/租户管理
按需/按秒计费
定制功能
多用户子账号系统

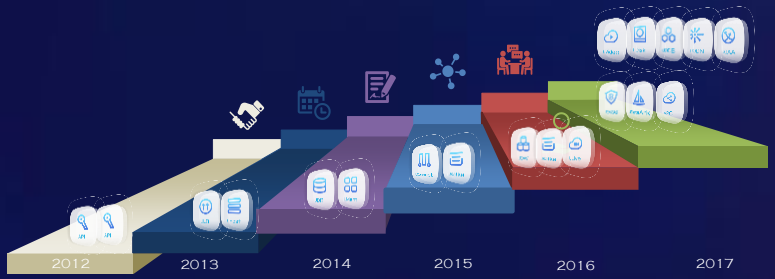
物理资源
14地域
/19可用区
海量设备
跨域连接

全局监控
实时探针
异常检测
告警系统

虚拟化核心技术
计算虚拟化
网络能力虚拟化
存储虚拟化

集群管理系统
资源分配
资源调度
资源隔离.

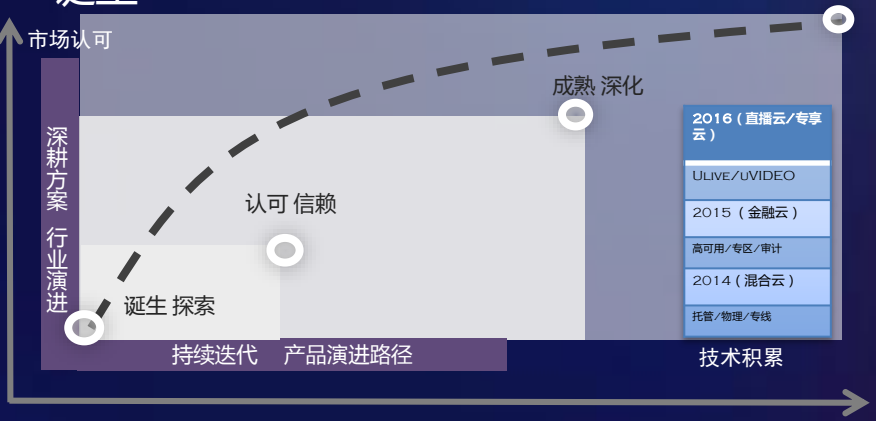
U市场	解决方案		服务支持	合作伙伴	教育培训
基础服务	文曲人工智能解决方案	武曲无服务器应用解决方案	架构师解决方案咨询	服务合作伙伴	原厂培训
研发管理	玉衡金融云解决方案	直播云解决方案	售后技术支持服务	解决方案合作伙伴	官方认证
运维管理	开阳全球服解决方案	启明混合云解决方案	**创业扶持**	渠道合作伙伴	视频资源
安全管理	白驹专有云解决方案	白驹专有云解决方案	创业孵化器		在线实验
行业增值	天罡安全解决方案		UEP企业成长计划		
视频云	**计算2.0**	**弹性计算**	**数据库**	**大数据分析**	**安全&合规&SLA**
云点播UVideo	容器集群UDocker	自动伸缩UAS	云数据库UDB	托管集群	基础网络防护USec
云直播ULive	通用计算UGC	消息队列UMQ	云内存存储UMem	UHadoop	DDoS高防服务UADS
云分发UCDN	自动伸缩UAS		数据传输UDTS	分布式消息队列UKafka	企业应用防火墙UEWAF
开放分发网络ODN	公共镜像Uhub		分布式数据库UDDB	云数据仓库UDW	WEB应用防火墙UWAF
	计算工厂UCF			分布式数据处理UDDP	WEB漏洞扫描UWS
人工智能	**计算**	**网络**	**存储**	**监控**	运维审计系统UHAS
UAI-Service	云主机UHost	基础网络Unet	云硬盘UDisk	监控UMon	数据库审计UDAS
物联网	私有专区UDHost	跨域通道UDPN	数据方舟UDataArk	网络流分析ShockWave	入侵防御UIPS
物联网	物理云主机UPHost	虚拟专有云VPC	对象存储UFile		业务安全UBSEC
	混合云Uhybrid	负载均衡ULB	文件存储UFS		合规
			归档存储Uarchive		SLA



- 主机虚拟化
- 存储虚拟化
- 网络虚拟化
- 计算虚拟化

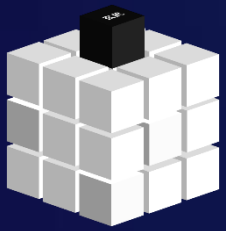
演进 — “云汉”诞生

- 诞生



- “云汉”体系

“玄武” 高可用解决方案



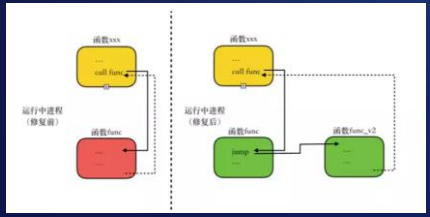
UCLOUD理解，对于所有上云的用户，业务连续和平台的稳定可靠是一切决策的前提和基石。

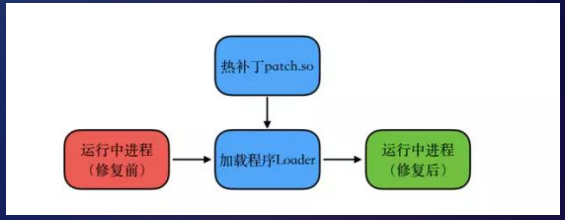
因此玄武整合了所有的高可用产品和特性，作为上层方案的金石之策。

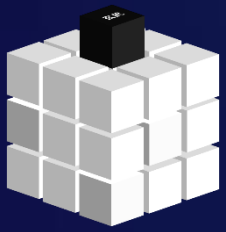
- 业务高可用
- 数据高可靠
- 高可用VIP

- **FLEX- VPC** 技术
- 内核热补丁技术
- 应用程序热补丁技术
- 数据方舟/**UDISK**

“玄武”—高可用基石 内核/应用程序热补丁技术



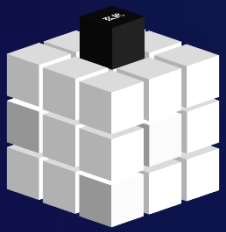




- 热补丁是一种在程序运行时动态修复内存中代码bug的技术。在UCloud，我们使用内核热补丁和应用程序热补丁（也就是进程热补丁）来在线修复核心业务的缺陷和安全漏洞。
- 应用程序热补丁比内核热补丁更加困难和复杂，比如内核对外提供完整的模块加载功能，可以直接加载内核模块形式的热补丁，而应用程序需要通过外部程序通过一系列对其内存和寄存器的复杂操作来注入动态链接库形式的热补丁；应用程序包含多线程；内核在编译时会被限定使用特定的编译方式，而应用程序的编译方式则更加宽泛；内核的二进制相对简单，而应用程序二进制因为需要链接到多种的动态链接库，本身的结构会更复杂。

- 假设我们有热补丁加载程序Loader、目标进程T、热补丁patch.so，目标程序的func函数替换为func_v2。
- **热补丁**
- 编写热补丁源码，编译成动态链接库格式的热补丁patch.so，patch.so中包含func和func_v2的信息。
- 热补丁patch.so在被加载程序Loader加载到目标进程T地址空间的过程中，通过dlsym调用找到func的地址，并将func的入口指令改为可写，同时改变为跳转到func_v2。
- 至此，所有对func的调用都会被重定向到func*v2，func*v2执行完毕后返回，程序继续运行。
- 如图所示

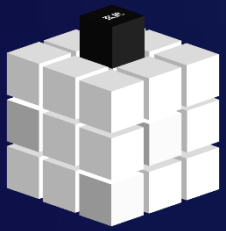
“玄武”—高可用基石 内核/应用程序热补丁技术



```
static void __attribute__((constructor)) init(void)
        {
            int numpages;
            void *old_func_entry, *new_func_entry;

            old_func_entry = dlsym(NULL, "func");
            new_func_entry = dlsym(NULL, "func_v2");

            #define PAGE_SHIFT              12
            #define PAGE_SIZE               (1UL << PAGE_SHIFT)
            #define PAGE_MASK               (~(PAGE_SIZE-1))

            numpages = (PAGE_SIZE - (old_func_entry & ~PAGE_MASK) >= size) ? 1 :
2;
            mprotect((void *)(old_func_entry & PAGE_MASK), numpages * PAGE_SIZE,
PROT_READ|PROT_WRITE|PROT_EXEC);

            /*
             * Translate the following instructions
             *
             * mov $new_func_entry, %rax
             * jmp %rax
             *
             * into machine code
             *
             * 48 b8 xx xx xx xx xx xx xx xx
             * ff e0
             */
            memset(old_func_entry, 0x48, 1);
            memset(old_func_entry + 1, 0xb8, 1);
            memcpy(old_func_entry + 2, &new_func_entry, 8);
            memset(old_func_entry + 10, 0xff, 1);
            memset(old_func_entry + 11, 0xe0, 1);
        }
```

- 此类热补丁适用于动态替换共享链接库中的可见函数，可以修复例如glibc “GHOST漏洞”（CVE-2015-0235）等等，在UCloud我们利用热补丁修复了若干缺陷，在用户没有感知的情况下把bug快速及时的修复。这些热补丁修复程序里，绝大多数代码是通用的，只需少数几行做特殊替换。

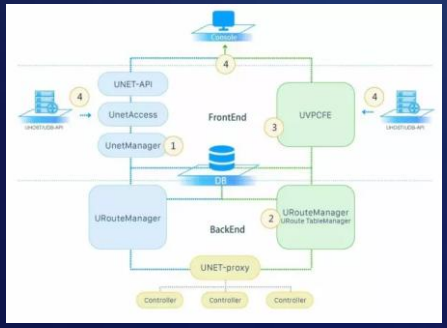
“玄武”—高可用基石 VPC2.0

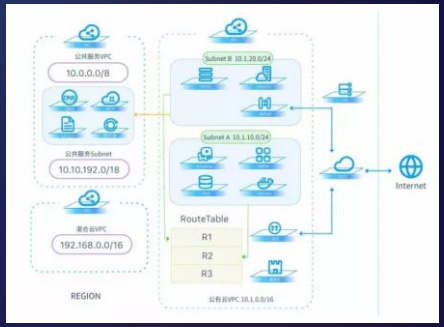


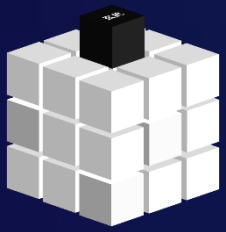
- 很多用户倾向在 VPC 中使用跨可用区的资源部署，因为容灾能力更强。但主机在可用区间迁移时，内网地址肯定会变化。
- 这是怎么回事呢？
- 因为在目前主流的 VPC 技术方案中，对于子网和可用区的分布上存在着限制，即：
- 单一子网只能属于特定的可用区；不同的可用区所拥有的子网段必然不同。
- 举例来说，用户在可用区 A 创建子网1，网段为 10.88.0.0/16。如果接下来在可用区 B 创建第二个子网，其网段必然不同，如 10.99.0.0/16。

- VIP (Virtual IP)是一个可漂移的虚拟IP，通常被配置到多个节点上。当一个节点故障时，VIP 可以漂移到另一个节点，从而实现服务的容灾。可是长期以来，VIP作为一个经典的容灾手段，只能用户自己配置并实现，云厂商并不提供支持。
- 在此之前 VIP 仅限于在单个可用区内使用，VPC2.0发布去除了这个限制。用户可将 VIP 配置在多个可用区的多台主机上，从而搭建覆盖物理云、公有云的跨可用区高可用。相比单一可用区内的高可用，能大大提升业务的健壮性。
- 以下图为例，VIP 被配置在可用区 A 的物理云主机和公有云主机，以及可用区B的物理云主机上，来实现高等级的高可用。。

“玄武”VPC2.0 技术实现 DVR 分布式虚拟路由

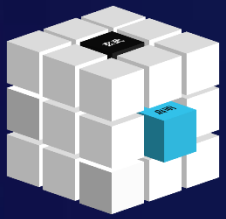






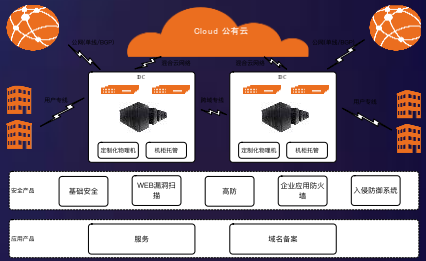
- 现有的VPC网络中，部分产品在使用自定义子网时需要先去创建一个虚拟路由器（VRouter）。
- 路由器具备三个特性：多主机共享出外网、指定端口转发、子网间内网互通。东西向和南北向的流量都要经过虚拟路由器。两个子网互通需要经过VRouter，子网访问外网也需要穿越VRouter，这样不仅物理路径上多了一跳，并且VRouter本身会成为性能瓶颈。

- VPCng重构了路由定义，并抽象为分布式虚拟路由DVR。DVR的载体是分散在各个计算节点上的虚拟交换机，而路由表是其核心。
- 如下图，东西向流量通过虚拟交换机进行分发，实现点对点通信。而NATGW只是作为外网访问的网关设备，提供多子网共享出外网能力。在路由表设计上，支持多种路由类型，包括直连路由、缺省路由、混合云路由、主机路由等等，除此之外还支持策略路由以及定义路由优先级。

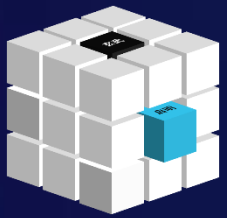


“启明”混合云解决方案

- 先进的混合云的架构
- 全球多数据中心分布
- 全球数据中心专线打通
- 提供单线/BGP多种网络选择
- 自助式可视管理平台、一站式服务，定制化的解决方案



“启明”混合云解决方案 混合云互联之技术实现



- 在传统的混合云方案中，混合云只是作为若干网段接入公有云，用户的诸多日常需求比如路由管理、带宽容量管理、混合云日常运维需要大量人力和沟通成本，亟需优化。
- 混合云抽象为租户的一个独立VPC，其中包含多个子网，子网可以是托管设备网段，也可以是通过光纤、数字链路、VPN连接到公有云POP点的自建IDC网段。混合云VPC的架构图如下：
- 我们设计了与旧混合云控制系统完全解耦的新系统，数据库、控制程序均与旧版本独立。在用户添加路由、修改网段等写动作时，API层利用消息队列对DB进行双写，同时通过DB对账保证数据的一致性。新旧系统分别拥有独立的转发面，分别独立拉取后台配置，确保新旧系统除了双写动作其他操作均解耦。
- 为了支持混合云VPC，我们开发了新版本的混合云网关。通过拉取后台配置，网关执行相关报文的鉴权、转发、封装和解封装等操作。同时，网关集群拥有scale out的能力，可无缝扩容，最高支持高达数百G的流量转发。

混合云网关的无缝切换是升级的核心步骤。我们开发了多个工具以辅助这个关键过程：

（1）基于Netconf的交换机配置API，用于切换及回退交换机侧路由配置。

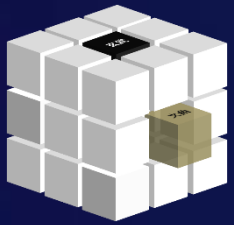
（2）基于flxeVPC的路由切换API，VPC控制后台结合推送与拉取，可以在10s内更新全网三层路由流表。该API用于切换及回退公有云侧的路由配置。

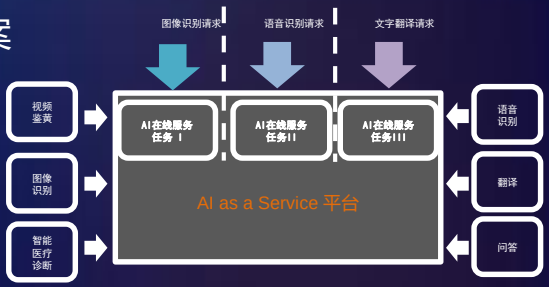
（3）基于OVS PacketOut的连通性检查工具，通过拼接ICMP报文，注入到公有云宿主机的OVS Virtual Interface上，可以模拟用户的业务互通，并基本覆盖用户的连通性黑盒检查场景。

（4）基于交换机统计数据、混合云网关统计数据的流量统计工具，可以从统计角度确认用户切换前后的流量状况。

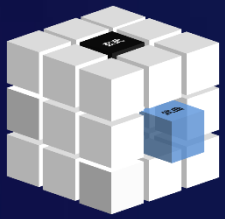
“文曲” 人工智能 AI解决方案

- 在线便捷的使用AI服务
- 节点自动扩容和缩容
- 按需计费
- 支持自定义配置，更加灵活

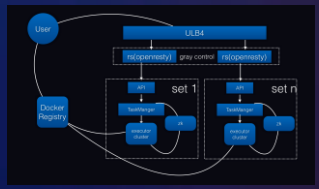


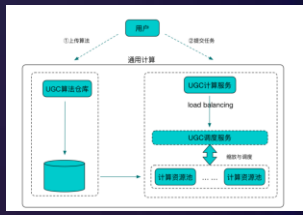


“武曲” SERVERLESS解决方案

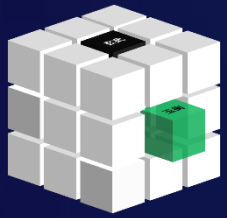


- 在线便捷的使用AI服务
- 节点自动扩容和缩容
- 按需计费
- 支持自定义配置，更加灵活

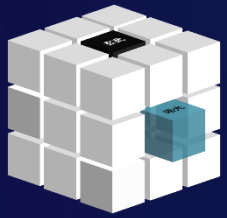




“玉衡” 金融云解决方案

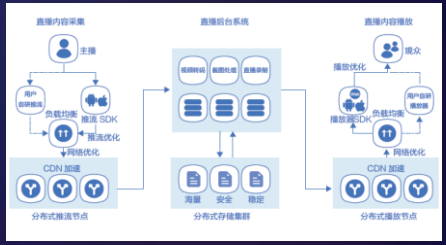


- 监管合规
- 安全防护
- 保障业务连续性
- 混合托管
- 创新业务快速上线
- 快速应变突发业务



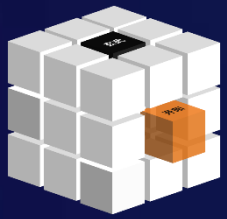
“瑶光”直播云解决方案

- 首屏秒开
- 极低网络延迟
- 多格式、多码率支持
- 低播放卡顿率
- 多终端、多机型、多操作系统
- 一站式服务
- 按需大规模弹性存储／计算／分发能力



“天罡”安全解决方案

- 金牌认证
- 多层纵深防御
- 快速响应
- 按需购买
- 快速接入



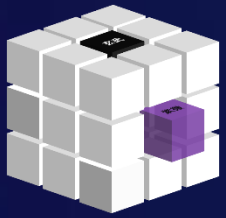
“开阳” 解决方案

- 跨域管理通道UDPN
- 全球网络加速产品UGAA
- 全球运维管理通道
- 网络流分析
- 境外App Store 审核加速方案

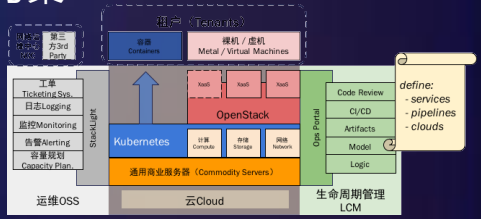
“白驹”专有云解决方案

- 完善的IAAS公有云服务能力
- 稳定的平台
- 优质的客户服务
- 持续革新的产品体系

“紫微” 私有云OPENSTACK解决方案



- 代码开源、自主可控、基础设施无锁定
- 全球第一的OPENSTACK和KUBERNETES发行版
- 业界最广泛的生态系统构建与支持
- 全球最多生产案例（200+）
- 全球最强技术实力
- 开放中立的混合云架构支持



U Define Cloud

演进 ⋈ 云汉