信息化金融是更高层次的信息化建设，对传统的运营和服务进行改造或重构，在金融服务方面取得质的提升。P7

ACI的核心价值在于确定什么东西是重要的，并将其动态地转换成底层设备行为，通过网络基础设备交付服务。P12

思科不仅将国外最先进的技术和解决方案介绍到国内，更重要的是将发达国家金融行业信息化建设的宝贵经验引进中国。P35

新一代智能数据中心

智能 · 敏捷 · 安全 · 高效

挑战带来机遇，发展依靠创新。面对挑战和机遇，国内金融行业必须要变革和创新，创新是金融业生存和发展的根本。金融信息化的发展趋势是随着金融发展趋势递进的，而金融发展趋势又是在金融与信息技术的相互作用中发生的。未来的发展趋势将证明，信息科技是实现金融机构业务创新的决定性力量。


CISCO

新一代智能数据中心
智能·敏捷·安全·高效

CONTENTS 目录

CONTENTS 目录

解决方案

附录

前言

金融机构转型
从新一代智能数据中心开始

》今天，国内金融行业特别是占据主要地位的银行业面临着诸多严峻挑战。

利率市场化的推进、各种直接融资渠道的快速发展已经成为商业银行转型的巨大推动力。商业银行必须改变传统的依赖净息差的盈利模式，进行金融服务创新，积极开发新产品和新的业务模式。

发展综合化经营帮助银行改善收入结构、应对金融脱媒以及利率市场化等的挑战，满足客户金融需求多样化的需要。部分商业银行已经加快综合化经营步伐，积极介入基金、保险、证券、金融租赁等领域。

银行业的准入和退出机制将不断得到完善。首批民营银行试点将在2014年展开。同时银监会正在酝酿加快推出银行破产条例，建立存款保险制度。因此各银行间的竞争加剧，尤其是中小型银行将面临更大的压力。

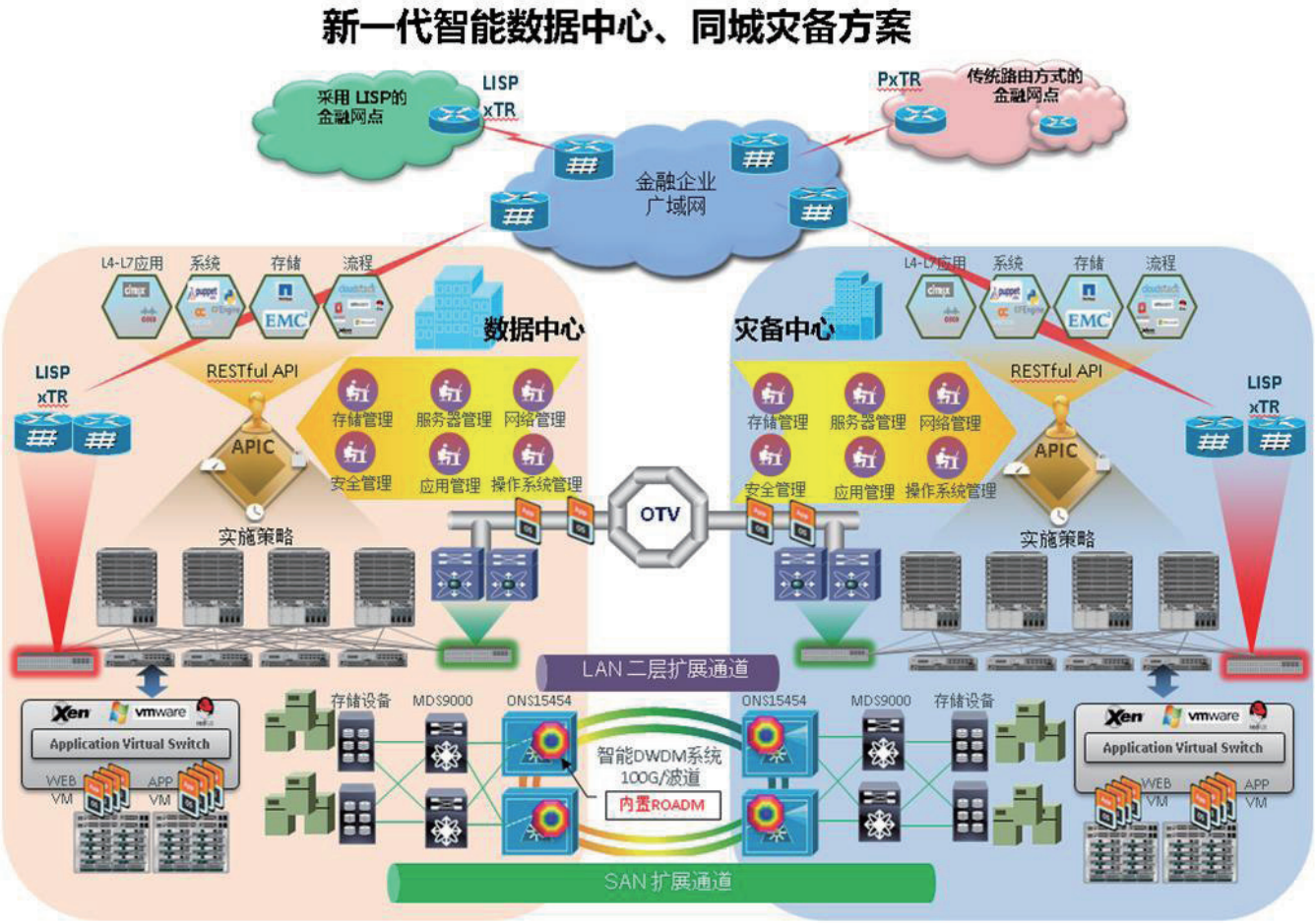
人民币国际化进程将一步加快。根据SWIFT统计，人民币已经成为世界第八大交易货币。以亚太和欧洲地区为核心的人民币离岸市场体系将逐步建立起来，人民币在全球支付交易规模比重将会提升。国内的商业银行将继续大力拓展海外市场，扩大经营机构，实施全球化经营。

同业业务监管加强，加速金融体系去杠杆化。有利于商业银行加快业务转型，构建更加“差异化、多元化、专业化”的经营模式。

国际金融体系监管改革和《巴塞尔协议III》对中国金融机构资本充足率管理体系提出了更高要求。新资本管理办法的实施将加大商业银行的资本压力，各银行资本工具创新的动力逐步加强。将对国内银行业的经营模式转型、风险管理、资本管理、信息披露机制、人才队伍以及预期效益等带来长期的战略性影响。

互联网金融逐渐影响和改变着传统金融服务业的运行模式。依托开放式平台、大数据和云计算的广泛应用，互联网企业快速向金融业态发展与渗透。“线上金融”的快速发展将对依靠物理网点扩张的传统业务模式带来巨大挑战。银行需要借鉴互联网金融成功的经营模式，再造全新的网络银行，更加重视用户体验，拓展金融服务的广度和深度，推动金融产品与服务发生革命性变革。

前言



思科公司一直致力于与国内客户分享国外发达国家金融机构信息化建设的经验，并带来该领域最新的技术成就。思科公司关于新一代智能数据中心的建设方案将带领您超越同行，在日益激烈的市场竞争中保持领先。

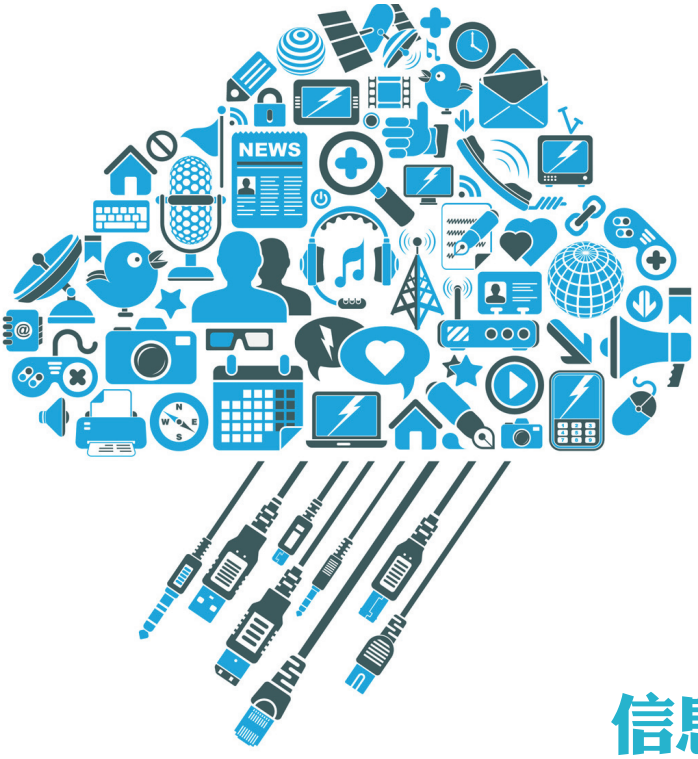
挑战带来机遇，发展依靠创新。面对挑战和机遇，国内金融行业必须要变革和创新，创新是金融业生存和发展的根本。

国外发达国家金融业的金融创新已成为体现金融企业核心竞争力的主要因素，而95%的金融创新都极度依赖信息技术，信息技术可以帮助分析复杂金融产品的定价并进行风险管理，使这些产品的交易成为可能。

中国金融业信息化十二五发展规划中明确提出了信息化支持金融机构转型，提高特色化、综合化、国际化经营能力。金融信息化的发展趋势是随着金融发展趋势递进的，而金融发展趋势又是在金融与信息技术相互作用中发生的。未来的发展趋势将证明信息科技将成为实现金融机构业务创新的决定性力量，功能单一、服务手段落后的传统金融机构最终将退出历史舞台。

新一代智能数据中心

一、综述



信息化金融
-----信息技术引领金融行业的变革和创新

在过去的一年，人民银行提出了中国的金融业信息化正经历着向信息化金融的转变，信息化金融已逐渐成为中国金融业的发展方向。传统的金融信息化是通过信息技术投入，硬件设施升级等基础性信息化建设，实现工作效率的极大提升。而信息化金融则是更高层次的信息化建设，对传统的运营流程、服务产品进行改造或重构，在金融服务方面取得质的提升，获得更高效快捷的金融服务。

通俗的讲，过去的金融信息化，IT是后台，被认为仅仅是金融业务的载体而已。未来的信息化金融，IT不仅将直接参与到产品的创新过程中，而且也将对金融机构前后台运营方式的变革产生重大影响。

毋庸讳言，对于金融机构来说，数据中心是建设信息化金融的核心所在。应用上收是金融机构目前和今后的总体趋势，随着“线上金融”的发展这种趋势会进一步加快。

创新产品的成功要依靠相关应用系统的开发和运行支撑，应用系统从开发、测试到上线运行的整个周期越短，新产品就能越早问世而抢先占领市场。

目前金融机构业务承载的核心——数据中心中通常部署了包括布线系统、IP网络、存储网络、计算系统、安全控制系统、网络服务系统、应用系统等在内的诸多IT组件。在目前的运营方式中，这些组件位于各个独立的管理“孤岛”中，相关资源的调度、配置、使用都很低效。试想一个新的产品上线，让各个处于“孤岛”的组件运行起来，它的周期会有多长，上线后它的运行维护会有多复杂，人为风险也会增加。一个复杂且缺乏灵活性的后台运营模式如何支撑前台产品的竞争和服务的竞争？ 现而易见，一个不仅要在功能组成上，而且在运营模式上达到灵活智能的数据中心是未来金融机构产品持续创新和保持竞争力的关键所在。本文将从多个角度深入探讨如何建设金融机构新一代智能数据中心。

理想中的智能的数据中心应该具有以下功能：

- **根据应用的需求来灵活的调度各种IT组件，实现敏捷的业务快速部署，加速新产品的推出。**
- **面向应用级的安全控制而又不丧失业务部署的灵活度和传输性能，安全监管的粒度更加精细。**
- **提供云计算和虚拟化的能力来高效的调度资源和最大限度地使用资源，为金融机构节约投资成本、提高利润产出。**
- **能够满足各种新应用所带来的计算特征需求，如大数据的运用对于吞吐量和时延都提出了很高的要求，为金融机构推出个性化的产品提供支撑。**
- **支撑多个数据中心之间的协同工作，提供金融机构的业务连续性，同时也带来资源的有效使用。**
- **自动化的运维和监控，降低人为的IT运行风险。**

当前数据中心的挑战

降低成本、提高效率的挑战

很多数据中心已经面临成本危机：一方面，能源成本高昂，并且没有足够的电力和冷却能力，无法满足新一代高密度服务器和存储设备的需要；另一方面，IT基础设施的容量增长受到场地、空间的严重制约。伴随数据中心能力不足的是资源浪费严重：大部分数据中心中的服务器和网络设备的利用率仅在24%~30%之间，有的CPU利用率、硬盘利用率都在在10%以下。绿色数据中心不是简单的每台设备的能耗降低，合理的架构和规划是关键。

业务连续性和灾难恢复的挑战

局部的突发性灾难事件，如地震、洪水、飓风、火灾或者恐怖活动等，都可能对金融企业的业务产生重大影响，导致收入减少，利润下降甚至失去客户。而重 大灾难事件则很可能导致金融企业一蹶不振乃至倒闭。而当前许多数据中心不能正常应对内外部的许多安全性挑战和威胁、满足业务连续性和可用性的要求，往往由于IT故障和各种灾难使得金融企业停止提供服务，造成很大的损失。

加快应变速度的挑战

互联网金融对传统金融行业的冲击目前看还主要是观念上的，让我们思考如何提升企业的业务应变能力。目前金融机构业务变革的速度正在日益提升，一方面变革产生的各种风险随之增加，因而IT系统以更快的响应速度和更有效的应对措施，来降低这类风险也就变得愈加重要。另一方面，变革速度的加快给企业数据中心带来时间上的更大压力，这也迫使企业IT系统提高响应速度。

实现自动化部署，迎接云计算的挑战

要提升数据中心的效率，实现业务的快速部署，云计算是目前的最佳解决方案。而资源整合、虚拟化和自动化是迈向云计算的必由之路。实现数据中心的自动化不是简单的管理工具的堆叠，而是从根本上实现服务器、网络、存储和管理的统一业务编排的自动化，就好比交响乐团的表演需要指挥来进行乐曲演奏的编排，而不是乐器演奏家进行各自的表演那样。

业务数量暴涨对运维的挑战

随着金融行业应用数量的爆发式增长，要求数据中心支

持应用的迅速上线，同时要精细化管理，实现每个应用的可视可控，对传统的运维工具和运维模式提出了巨大的挑战。

传统数据中心基础架构普遍存在的问题

计算虚拟化和资源池建设

数据中心的虚拟化技术涵盖了 网络、计算、存储、管理等多个层面。各个层面的虚拟化意味着应该能够为最终用户提供更低成本的计算、管理和监督，在降低 IT 费用的同时大幅提高效率和敏捷性。

这几个层面的虚拟化并不是各自进行，互不干扰的，以计算虚拟化为例，它会带来对网络、存储、运维管理、外部监管的影响。

1、虚拟机流量的可见度与安全问题

在传统的网络中，可以通过使用端口镜像SPAN捕获数据包，以相当简单的方式进行分析。在虚拟化环境中，数据可能永远不会通过一个物理交换机或网络，而是留在同一个物理主机，使得监测变得困难。如果没有虚拟机级的可视化，安全部门就无法对恶意攻击进行检测和调查，采取纠正措施并预防未来的恶意攻击。如果没有这种可视化，服务器和网络部门就无法了解在任何给定的时间内网络的表现。

2、网络边界的消失给运维和管理带来的问题

计算虚拟化之后，虚拟机的网络配置由系统（服务器）部门使用虚拟机管理器进行设定，网络部门对于虚拟机用了哪些VLAN，有哪些端口级别的安全防护一无所知。网络边界的消失带来了可怕的管理“黑洞”，模糊了金融机构目前各部门的责任界限。

3、X86 SAN Boot与基础设施的矛盾

实施了虚拟化的X86服务器如果想要做到SAN启动，就需要建设一个庞大的SAN存储网络，x86服务器上购买HBA卡是一笔不菲的投资，更重要的是机房可能就没有足够的光纤资源组建到延伸到标准机架的SAN网络。

4、虚拟机实时迁移与网络的联动

如果虚拟机从一台物理机迁移到另外一台物理机，网络参数配置和策略却不一致，则可能导致虚拟机无法提供服务。

5、虚拟化的效率问题

在实施虚拟化之后，有时会发现虚拟机其实还有不少空闲的资源，但从业务响应速度上，看到的却是比传统物理服务器更慢的效果。虽然每个虚拟机处理器占用率很低，但经常会出现多台虚拟机并发网络访问的情况，此时的网络I/O吞吐量就成为了制约虚拟机性能的瓶颈。

6、安全监管与大资源池的矛盾

经常听到金融机构的系统部门抱怨网络上为什么要划分那么多的功能分区，限制了计算资源共享和虚拟机实时迁移、虚拟机集群的部署。但是网络部门和安全部门认为要满足监管的要求，觉得数据中心功能区已经无法再精简了。因此如何实现在满足监管的前提下，跨分区的大的计算资源池共享是设计智能数据中心架构需要重点考虑的技术要点之一。

从上面的对比表可以看出，虽然针对这些问题思科相比业界而言都有不错的解决方案，但是思科认为单一解决问题的方式不能够满足数据中心未来的发展要求，进而提出以应用为中心的基础设施（ACI）的新一代智能数据中心解决方案。

大数据带来的问题

实时的数据分析能力对于金融机构实现个性化至关重要，大数据在金融行业的采用将日趋普遍。然而大数据给服务器、网络和存储带来了巨大冲击，服务器的可靠性和性能要求相对降低，更多地将采用x86和低成本，存储向本地化发展，并且从面向块存储Block转向采用面向对象或文件的存储，网络要求高密度和大容量，要能支持按需增长和横向扩展。

传统数据中心更多的是为了南北向的客户端到服务器访问的流量模型而设计，对于大数据应用中的东西向服务器到服务器的大流量模型就存在很多问题。比如多层架构的网络设计导致收敛比过大，服务器互访的实际可用带宽低，延迟

针对上述相应问题的解决方案对比表

数据中心遇到的问题	业界的解决方案	思科目前的解决方案
虚拟机流量的可见度与安全问题	依赖于Hypervisor厂商，如vShield，开销大	利用N1000v和VSG，可提供运营简单性、部署灵活性、增强的性能、以及集中化管理能力等优势
网络边界的消失给运维和管理带来问题	虚拟接入层的网络配置，只能由系统人员去设置，存在管理黑洞。	VM-FEX, 将每个虚拟机的网卡连接到交换机的虚拟接口上，实现了一对一的连接。网络部门和服务器部门分工明确。
x86 SAN Boot与基础设施的矛盾	投资存储网络，或者使用简单的FCoE。	部署思科聚合式以太网CNA卡可以减少50%的前期成本。支持Unified Port可以将交换机端口设成1G/10G以太或8G FC，支持多跳FCoE。
虚拟机实时迁移与网络的联动	需要二层网络，或者传统网络上配置叠加的逻辑网络，开销大而且低效难管理。	可使用VXLAN、NVGRE等实现跨三层的虚拟机迁移，是支持物理机与虚拟机并存情况下的VXLAN部署的技术领先者。使用DFA动态矩阵架构技术的网络配置Network Profile实现二层内的虚拟机迁移时，通过将ARP广播、DHCP等限制在末端叶子节点内，避免了二层故障域过大的问题。使用ACI以应用为中心的基础架构中应用网络配置模板ANP实现跟随虚拟机迁移而实时修改交换机的策略。
虚拟化的效率问题	I/O吞吐能力低，而40G的骨干网络中每个端口需要4对光纤,数据中心原有基础设施无法满足。	1微秒低延迟交换机，单对光纤的40G BiDi技术，usNIC支持应用支持调用网卡，支持HPC高性能计算和高频交易。
安全监管与大资源池的矛盾	缺乏解决手段。	使用新一代交换矩阵架构跨区互联，在保障安全的前提下实现大资源池共享。

大，扩展性差等。

思科智能数据中心采用使用新一代交换矩阵架构，服务器之间互访路径可达16条，未来可扩展到64条路径，每条路径可由16根40G的链路捆绑而成。横向扩展能力极强而且按需扩容简单，数据中心接入端口不足就增加机架和接入层叶子节点（Leaf）交换机的数量，性能不足就增加骨干（Spine）节点交换机的数量。

私有云带来的问题

随着金融机构的综合化发展，金融集团或控股公司要求建设统一的数据中心、一致性的金融数据，协同的应用服务。传统的数据中心设计很多还是采用竖井式的基于应用的封闭模式资源使用，无法适应多租户的独立管理和资源共享需求。

思科新一代交换矩阵架构，支持内置的逻辑分段Segment，在一个统一的资源共享平台上虚拟出众多的逻辑的数据中心供各个子公司或分行使用。每个租户可以使用门户页面进行自助式服务，独立管理属于自己的服务器、网络交换机、软件防火墙，而数据中心内部进行自动化的资源分配、管理和回收，同时提供每个租户的服务质量协议（SLA）保障。传统的数据中心迈向云计算是一项复杂的工程，思科智能的数据中心系统注重开放性和多厂商跨平台的异构环境的统一管理能力，努力构建适合云计算的业界生态系统。

新一代智能数据中心

二、智能数据中心

2.1 建设新一代智能数据中心

目前的数据中心内部分区大多采用基于POD(Point of Delivery)的模块化设计（图1)。这种架构存在明显不足：首先，POD是为客户机到服务器的南北向访问而设计的，无法满足云计算和大数据背景下服务器到服务器的东西向访问大流量低延迟要求。其次，防火墙和负载均衡器等网络服务组件无法对POD内部的应用按照重要程度进行优先等级的服务，只是先进先出FIFO式的顺序服务。而且，只能限于POD内部实现虚拟化所需的资源共享。每个POD都有自己专用的网络服务组件，但是对于南北向访问，这些网络服务组件一般利用率不足，对于东西向访问，它们又成为制约可用带宽和增加延时的性能瓶颈。此外，POD互联的核心无法实现横向的按需增长，只能升级核心设备或者改成折叠式CLOS结构。

总的来说，目前数据中心的架构是僵化不灵活的，服务器的安全策略和网络配置决定于它所放置的位置，如果一个应用要得到这样的网络服务就必须要固定在这个位置。

2.1.1 新一代数据中心总体架构

思科新一代智能数据中心（图2)采用以应用为中心的基础架构（ACI)，ACI价值的核心在于确定什么东西对于应用是重要的，并将其动态地转换成底层设备的行为，通过网络基础设备交付该应用所需要的服务。应用所需要的各种服务可以按照任意顺序组合，不存在传统模式的边界瓶颈问题。在ACI的环境里，资源共享不存在POD边界，做到整个数据中心范围内在物理和虚拟环境任意扩展。

ACI由应用策略基础设施控制器(APIC)、Nexus 9000产品搭建的NG Fabric和NX-OS操作系统的增强版组成。ACI以应用为中心，将物理机和虚拟机按照IP地址、端口号、VLAN、VMware port group、DNS、VXLAN segment ID等信息编入终端端口组（EPG)，应用之间的访问抽象为EPG互访，由Contract来定义网络转发策略、安全隔离策略、QoS服务质量策略、负载均衡等高可用策略等。因此说，ACI天生就支持在一个Fabric上实现多个逻辑网络（图3)。

传统数据中心想要实现云计算所需的多租户，只能通过VRF或者VLAN进行隔离，管理复杂并且难以实现资源共享。新一代数据中心ACI的多租户隔离是基于APIC控制器的策略，是白名单式的隔离，因此安全性更高，并且非常灵活。与纯软件的SDN方案相比，ACI工作效率高，可实现硬件和软件并存环境下的高度扩展，单个ACI的Fabric可以支持6

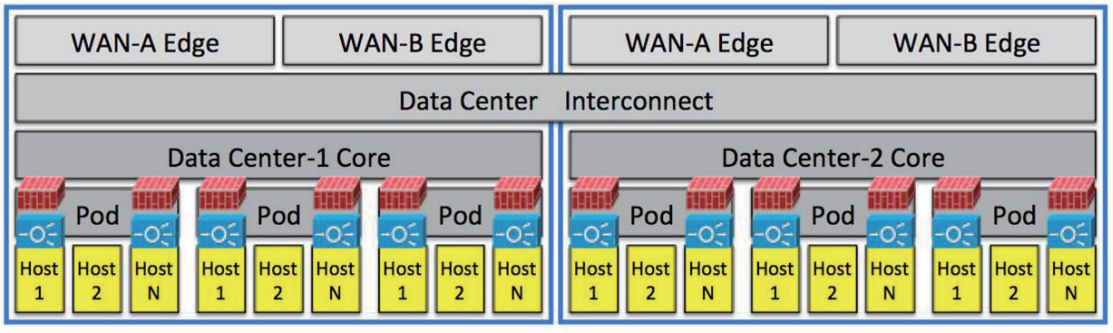


图1

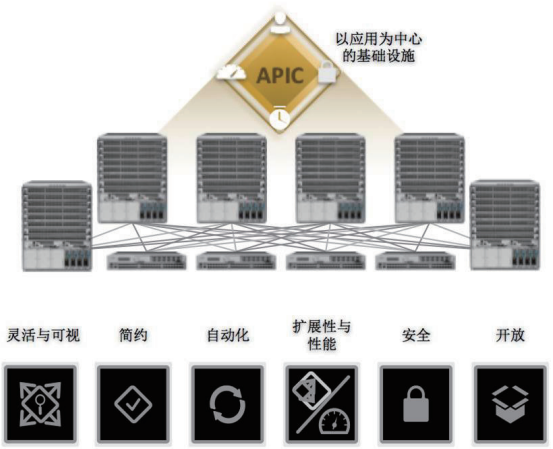


图2

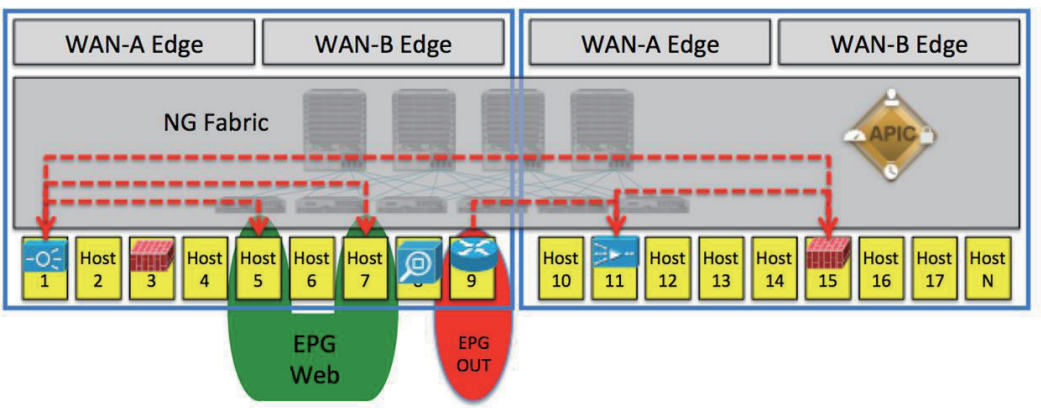


图3

万4千以上个租户、1百万个终端、20万个万兆物理端口。

ACI将软硬件系统和专用集成电路芯片(ASIC)中的创新与基于开放式API的动态应用感知网络策略模型完美结合在一起，APIC控制器的南向接口驱动交换机硬件，北向接口连接各类管理工具。对比在传统网络上进行逻辑叠加(Overlay)的SDN纯软件解决方案，例如VMWare NSX则忽略了应用对底层硬件网络的要求，代价就是：1、每台服务器和虚拟机的开销增加。2、管理节点繁多。3、扩展性差，如NSX目前只支持5000台主机和5个控制器。4、无法感知应用，而虚拟机上应用的信息被NSX增加的包头所掩盖，物理网络的QoS无法生效。5、数据中心非虚拟化的服务器与其交互复杂。6、无法实现应用可视，不利于故障的检测、隔离和排除。7、安全性差，缺乏防火墙、IPS、SSL等，也无法使用外部的硬件安全服务组件，而自身的OVS软件交换机的每个接口仅支持30个访问控制表项。

因此，纯软件SDN下的应用性能却依然得不到保证。这也是思科ACI理念得到业界热烈反响的重要原因之一。

2.1.2 业务的敏捷性和快速部署

要实现业务的敏捷性和快速部署，关键在于包括计算资源、网络、存储在内的数据中心基础设施的无状态化。而且，无状态化在虚拟机的应用很好理解，虚拟机不必绑定在某台特定物理机上，使得实时迁移成为可能。

思科UCS统一计算系统引入了服务配置模板Service Profile的概念，目的就是无状态化计算，将物理机特有的个性参数，如MAC地址、UUID等抽离出来，使得作为计算资源的主机不必绑死在某台特定物理机上，使得硬件服务器的快速上线成为可能。

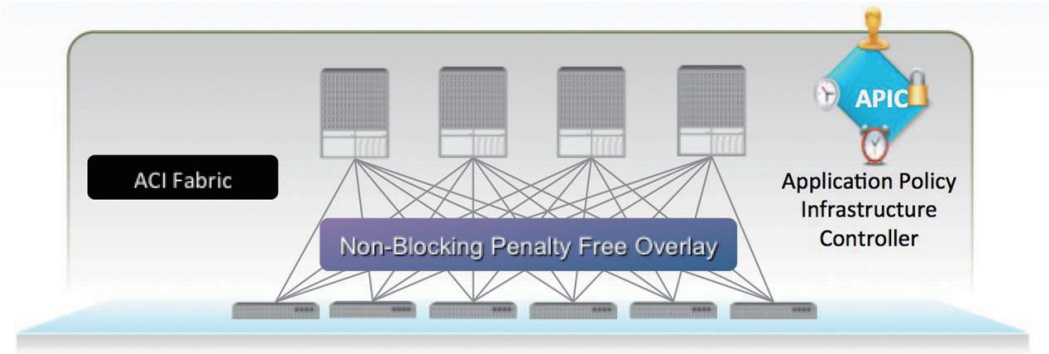


图4

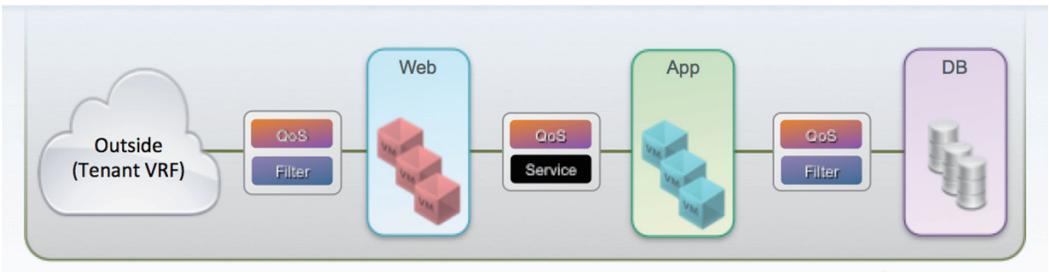


图5

那如何理解网络的无状态化呢？我们知道传统的局域网交换机的端口是通常是静态设置VLAN的，而对比园区网用户接入时采用802.1x或者Web认证的方式，交换机端口的VLAN实际上是按照用户认证情况而动态设定的。显然最终使用者不必打电话给网络运维人员，让他们赶紧修改端口VLAN，这就是无状态化带来的灵活方便。

很遗憾，在今天的数据中心，交换机的端口还是网络人员静态指定VLAN。遗憾在哪里？一台服务器一旦被划分到某个VLAN，那就只能用于某种用途，就必须受制于该VLAN代表的安全策略、负载均衡策略。事实上的逻辑应该是，这一台服务器目前正在用于某种应用，网络就应该配置成能够提供该应用所需要的安全服务和负载均衡服务（图4）。

在新一代交换矩阵架构中，交换机的端口配置会根据所连接的应用而动态变化，整个网络就是无状态的，依赖于应用策略基础设施控制器（APIC）的指令，实现逻辑网络上的配置，不同的服务器被动态地划分到Web组、应用组、数据库组（图5）。

2.1.3 量化的网络性能监控与应用可视化

在数据中心如何有效监控网络性能，一直是个难题。使用Netflow也只能基于统计地进行异常模式分析，使用SPAN进行数据包镜像抓取，又难以分析应用级的性能，也无法实现跨设备的端到端性能分析，至于根据网络性能而动态调整网络策略那就更是无从谈起。

思科新一代智能数据中心使用的Nexus9000系列交换机支持内置的硬件原子计数器，可以比较两个叶子节点交换机之间通讯时，不同路径上入口节点发送数据包与出口节点接收到数据包数量的差异，并且根据策略动态地自动选择性能好的路径（图6）。

不过设备的硬件资源总是有限的，为每个应用分配独立的硬件计数器显然不现实。思科特有的Telemetry遥测技术使用软件定义网络SDN的概念，支持按需测量，从APIC控制器发送指令，让交换机只计数特定的数据流，除了可以是传统的IP地址、TCP/UDP端口、VPN VRF指定外，还可以根据EPG（End Point Group)应用组来指定数据流。交换机将计数结果再返回APIC控制器，为APIC控制器下发新的策略提供了依据，如调整该数据流使用路径或提升该数据流优先等级。

反应网络性能的当然不只是丢包这一个参数，实时测量不同接入交换机之间端到端通讯的延迟，以及反应延迟差异的抖动参数也在思科新一代智能数据中心中得到实现。跨交换机的

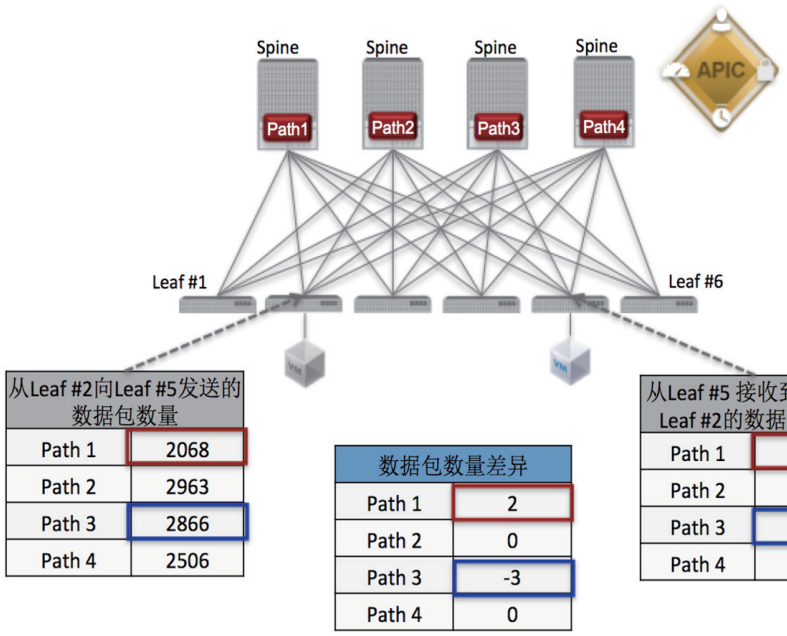


图6

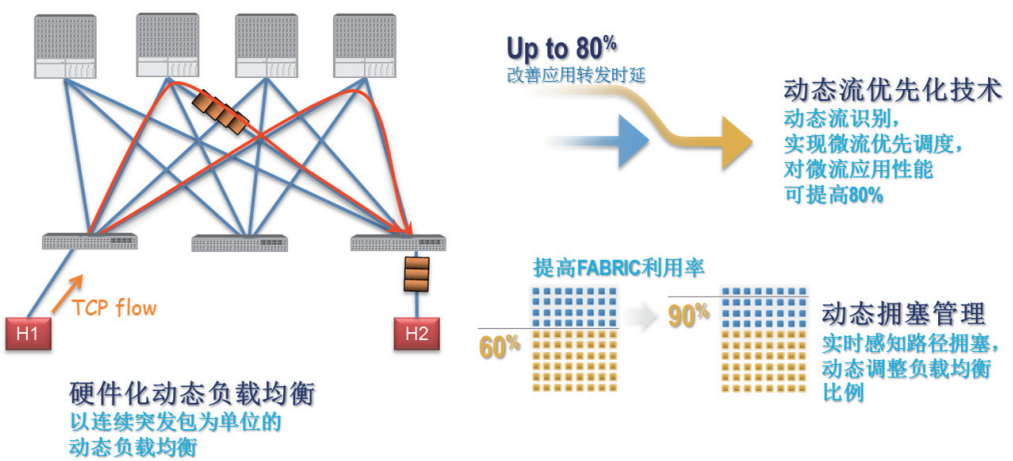


图7

时间同步采用标准的IEEE 1588协议，即“网络测量和控制系统的精密时钟同步协议标准”，思科的核心交换机引擎中配有专门的接口，可以连接专业的外部高精密时钟源。

思科以应用为中心的基础架构采用硬件动态负载均衡技术，基于交换机微流识别和链路拥塞检测的结果，单一会话流量以连续突发包为单位动态地在多条等价路径间转发，打破了原有等价转发路径间通过HASH算法按流负载均衡，单一会话流量无法突破单一链路带宽的限制，从而优化数据流在Fabric内转发的时延，提高Fabric的利用率，适应更为严苛的应用环境（图7）。

2.1.4 应用级自动化

在当前的环境中，IT非常复杂且缺乏灵活性，这阻碍了业务的增长。同时，由于当前的数据中心技术无法支持资源共享架构模式，IT专业人员被迫在各自低效的孤岛中工作，无法在单一的视图中查看影响应用性能的所有软硬件组件，导致IT组件配置困难、故障排除复杂、变更繁琐。

要实现数据中心的高效节能和快速部署，这些孤岛就必须

被打破，将网络、存储、计算、网络服务、应用和安全保护等所有IT组件统一起来，作为单一的动态实体进行管理，同时丝毫无损性能，而这正是以应用为中心的基础设施（ACI）的优势所在。

思科ACI使用完备开放的RESTful API，与开源社区密切合作，包括OpenStack、Open Daylight、OVS虚拟交换机和VXLAN等，推动形成一个广泛的ACI生态系统，包括管理、协调、控制、虚拟化、网络服务和存储合作伙伴等。数据中心开发团队能够通过RESTful API和Puppet、Chef、CFEngine、Python脚本编程工具等，利用ACI的自动化优势实现更紧密的集成（图8）。

思科ACI为Microsoft Hyper-V、RedHat KVM、VMware vSphere和其他虚拟化平台提供了自动化虚拟网络管理与遥测技术。以VMWare为例，APIC控制器管理员创建应用网络配置模板（ANP），服务器管理员将虚拟机的网卡匹配到ANP对应的port group，APIC控制器就自动将ANP的配置策略推送到网络设备（图9）。

APIC控制器为自动化和管理ACI矩阵、进行策略编程和监控状态提供了一个统一平台。它能够优化性能，支持将应用部署在任意位置，将防火墙和负载均衡器等网络服务部署在任意位置，将服务器网关部署在任意位置，并统一物理和虚拟基础设施的管理工作。与传统SDN控制器不同，它不受交换机数据和控制平面的影响，因此即使APIC离线，网络也能够对服务器和虚拟机的变化做出响应。通过集中管理、应用网络配置文件、L4-7网络服务自动化和开放式API，ACI可显著提高业务灵活性，将应用部署时间从数天缩短至仅仅几分钟。相比纯软件的SDN方式实现网络虚拟化和无状态化，ACI能够消除每个虚拟机的开销，并可充分利用现有的布线投资，可节省75%的总体拥有成本。

通过推出专注于业务应用的网络交换矩阵，思科将能够帮助企业实现基于策略的自动化，改进性能，提高可视性和简化运营，从而获得更高的工作效率和灵活性，同时大幅节省成本。

2.1.5 网络服务的分布式部署

在数据中心网络架构中，除了提供连通性需求外，各类网络服务的部署也至关重要，比较常用的数据中心内部网络服务有状态化包过滤，入侵防护，服务器负载均衡、应用防火墙及应用可视化分析等。

在传统的数据中心架构设计中（图10），网络服务节点通常部署在汇聚层，形成一个网络服务子层，通过网关配置、策略路由、路由策略、xflow或者流量镜像等方法完成网络服务的按需部署。在设计部署规划时首先把需求转化为数据流量模型，然后通过网络配置实现流量转发模型，在服务节点上部署服务策略最终实现网络服务部署。总的来说，网络服务的部署遵从：网络服务需求——>网络流量模型——>策略部署 的步

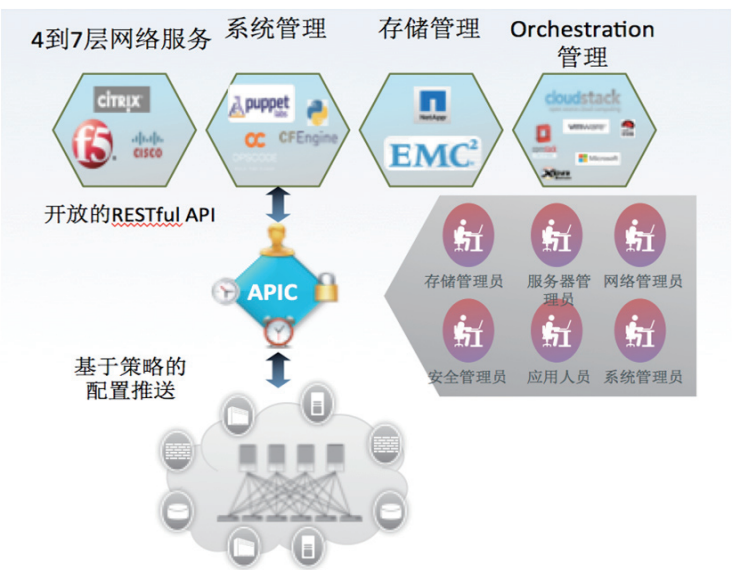


图8

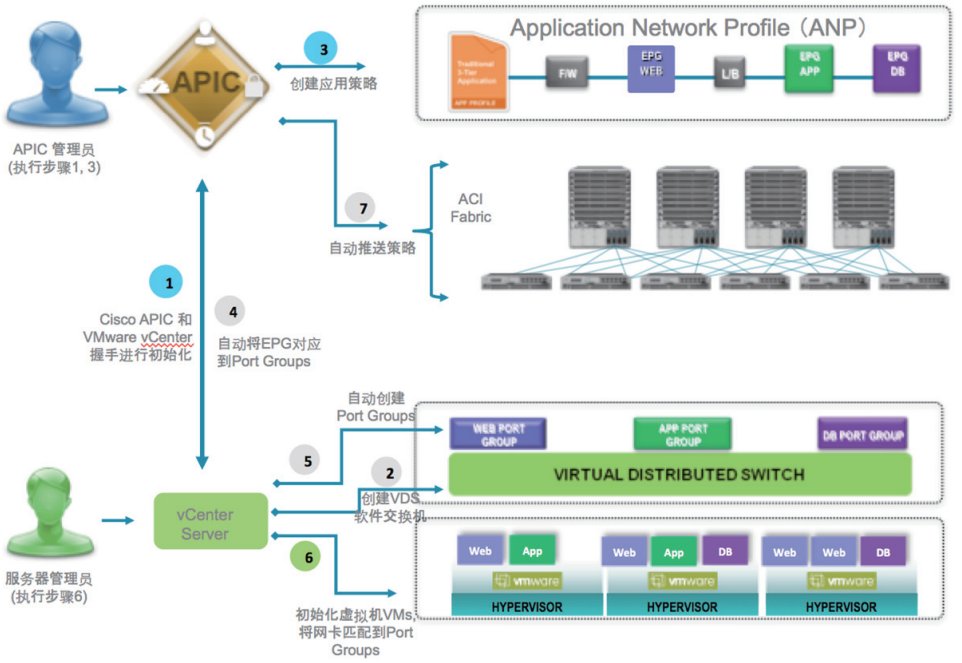


图9

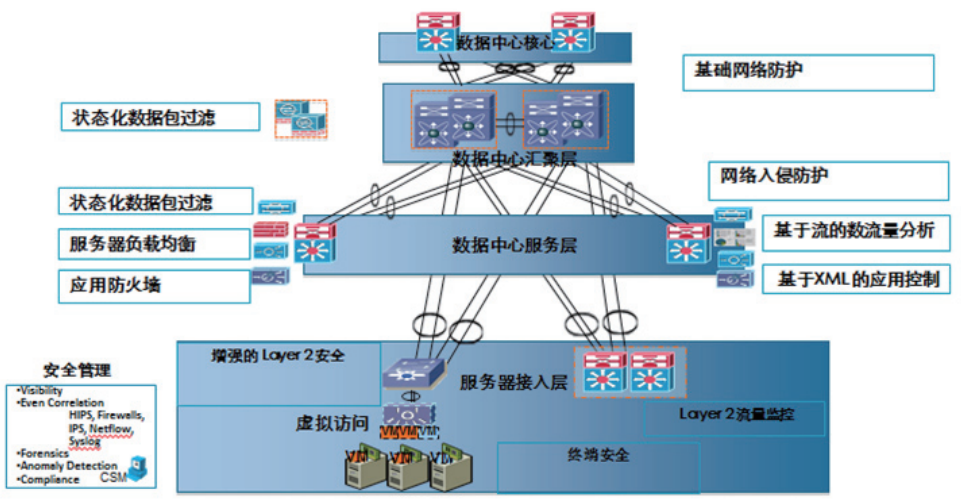


图10

骤，将需求转化为专业化的网络概念和配置后得以实现。

在以应用为中心的架构中，网络服务的部署模式更为简化，直接面向应用需求进行网络服务的部署；同时对于网络服务节点的部署的最佳实践也从传统的汇聚层集中式部署转变为叶子节点分布式部署。

Application Network Profile

Application Network Profile（图11）是以应用为中心的基础架构的一个显著特点，其中不仅包含基本的访问连通性要素，同样也包含了网络服务要素。对于某一具体应用，在访问应用的哪个阶段（从客户端到WEB，WEB到APP，APP到DB等），需要在哪个网络服务节点（FW IP，SLB VIP等）进行何种网络服务（安全，负载均衡，应用控制等）均可在Application Network Profile中以合约（contract）形式通过图形化UI直观地配置和显示。这样对于每个应用的每种访问关系（EPG之间的访问逻辑）所需要的任何类型的网络服务都

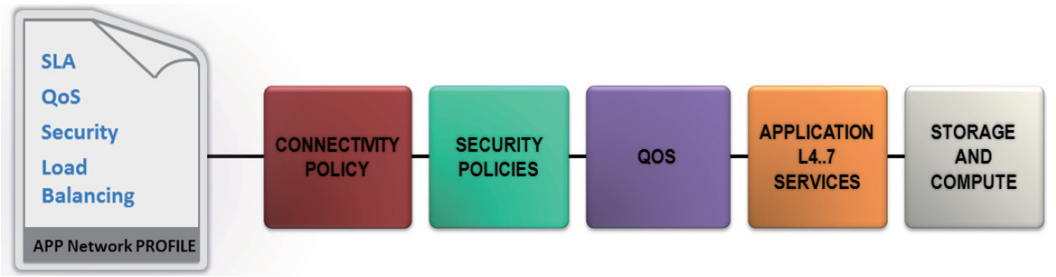


图11

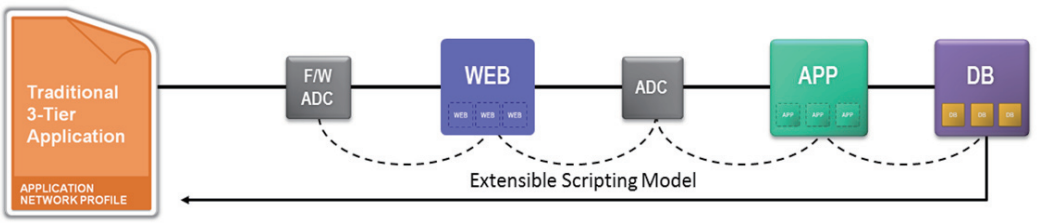


图12

可以根据应用的需求直接制定，而不需要像在传统数据中心的部署方法中通过调整VLAN，SVI部署位置，路由策略和策略路由等方式，依靠复杂网络逻辑来定义（图12）。Application Network Profile制定完成后会通过APIC翻译为设备语言下发给每个网络节点执行。

网络服务分布式部署

下一代数据中心架构中，Fabric成为网络规划的基本架构，实现了低耦合、高可靠、高性能、高扩展的网络架构。基于Fabric的网络架构，网络服务节点的部署不再有局限性，可以部署在任意一个Leaf节点上，可以根据需求进行分布式的网络服务部署，当然如果想依然保持集中式的部署模式也可以直接迁移；部署在Leaf节点上的网络服务节点可以采用HA和Cluster等方式保持网络服务节点本身的高可靠，HA和Cluster内的不同节点可以部署在不同的Leaf上从而保障整体高可靠；对于不同应用可以通过Application network profile的配置，轻松地配置成通过不同的网络服务节点提供网络服务，如网银通过SLB1进行负载均衡，而手机银行通过

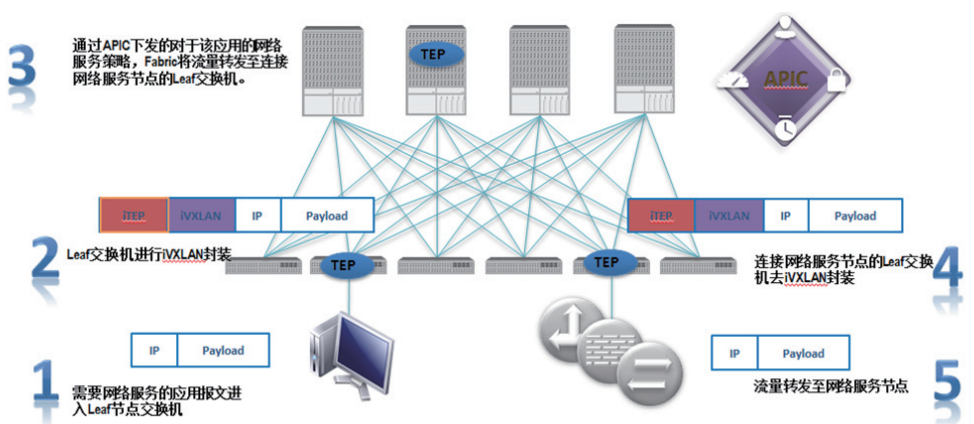


图13

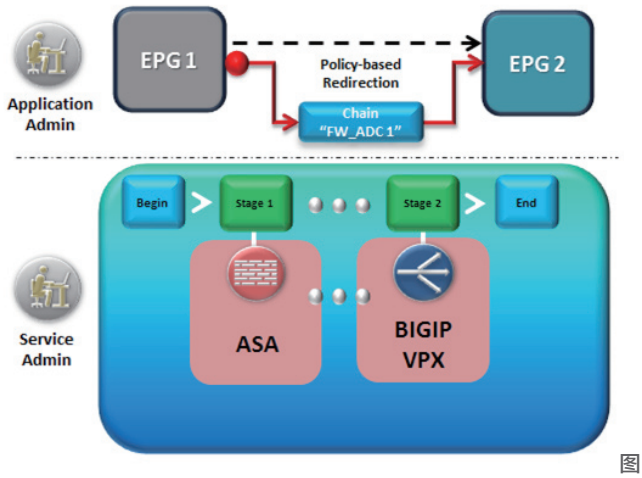


图14

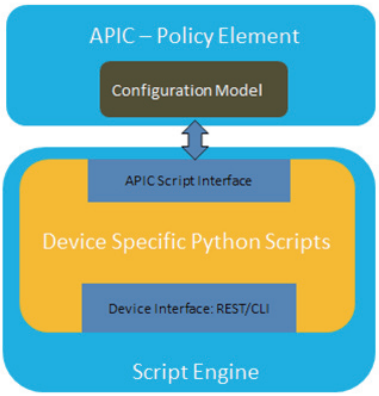


图15

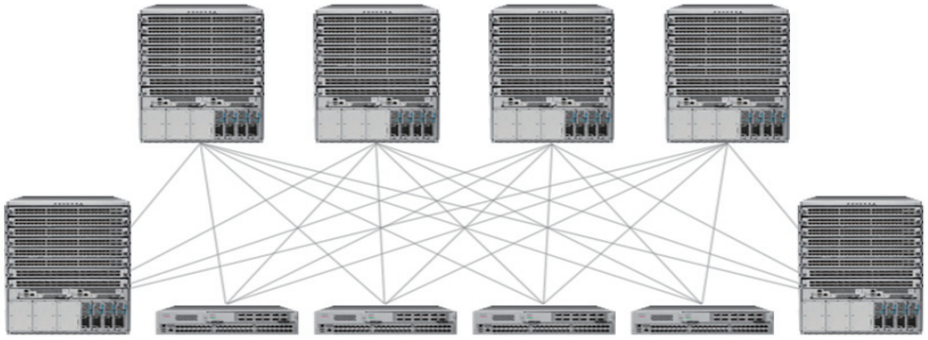


图16

SLB2进行负载均衡；Fabric中各Leaf通过学习APIC上配置的服务策略或者通过NSH，将需要网络服务的应用流量转发至连接网络服务节点的Leaf（图13）。

网络服务策略部署

网络服务节点的策略可以通过两种方式部署，一是依然保持现有的独立部署的模式，通过网络服务节点自身的UI界面或者CLI进行网络服务策略配置（安全策略，负载均衡策略等）（图14）；二是APIC控制器将自动化生成的网络服务配置自动加载到网络服务节点（图15）。通过这两种方式，可以适应各种网络服务要求并同各种类型的网络服务设备兼容。

2.1.6 高性能骨干与集约化布线

近年来互联网金融业务的兴起进一步促进了国内金融行业以业务创新为驱动的转型，传统业务越来越多的被植入互联网的基因，数据分析类应用的使用越来越普遍，体现在数据中心网络架构上，主要的特点就是东西向流量增加迅速，在数据中心内部流量占比不断提高。另一方面，随着万兆网卡成本进一步下降同千兆网卡基本趋同，虚拟化应用的普及使得服务器网卡利用率不断提升，数据中心内网络流量增长迅速，数据中心服务器万兆接入的时代已经到来，要求数据中心骨干网络架构拥有更高的性能和更好的扩展性。

Fabric骨干架构

下一代数据中心采用Fabric骨干网络架构（图16），任意两个Leaf节点间有大量等价转发链路同时转发，可以通过简单的扩展SPINE和Leaf平滑扩展Fabric性能，不会影响到已有的流量，任意链路的中断和设备故障达到毫秒级收敛。

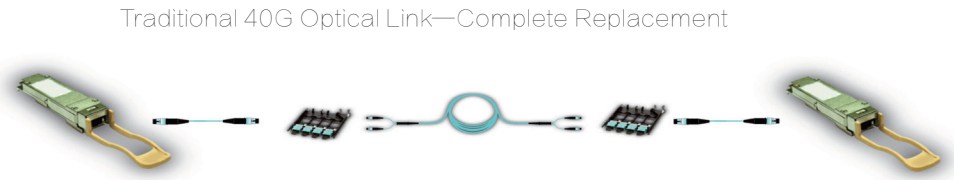


图17

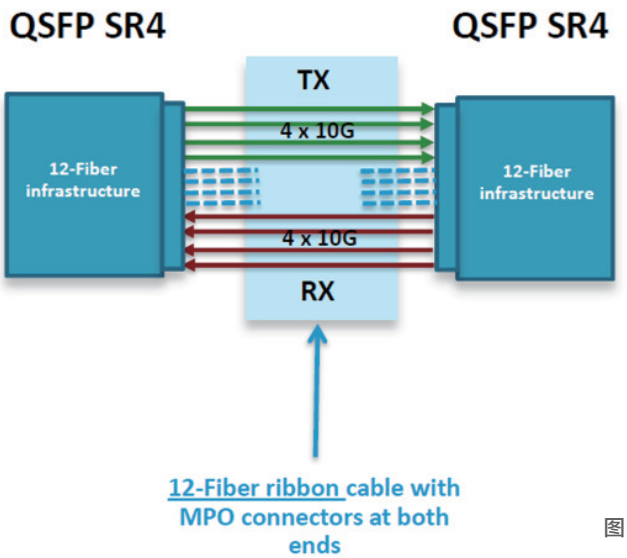


图18

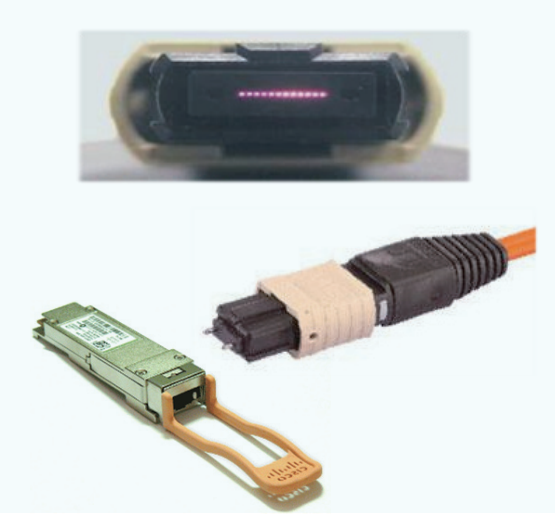


图19

40G集约化布线

除了通过Fabric架构的海量等价路径扩展性能，在服务器大规模万兆接入的情况下，采用40G以太网技术提高单根链路的带宽是提高数据中心骨干网络性能最为重要的方法。40G以太网技术已经商用较长的一段时间，目前40G以太网光模块的价格已经大幅下降，但是传统40G的布线系统费用相比较于10G骨干布线系统成本而言却是大幅度增加，阻碍了40G骨干网络的普及。

传统的40G布线模式

传统多模40G布线系统（图17）包含设备上40G多模光模块，光配架MPO面板，机柜间40G互联Trunk光缆。传统40G多模布线系统（图18）中网络设备上采用QSFP+ 40G多模光模块（图19），与10G多模光模块不同的是40G多模光模块采用MTP/MPO接口，需要8根纤芯完成一个40G传输，每根纤芯上承载一个单向10G，一般单个40G需要使用1根12芯光缆。

还有一些其他的布线组件：12芯LC到MTP转换器，3个40G 的8芯转2个12芯MTP转换器等。因此采用传统40G系统无法兼容原有的10G布线系统（接口形式不同，骨干布线数量不够），不能从10G骨干网络平滑的升级到40G骨干网络，这也是国内金融数据中心40G骨干网络还没有大规模应用的一个重要原因。

思科Bidi技术40G布线模式

思科40G Bidi技术弥补了传统40G的不足，使用LC光纤接口（图20），采用波分技术（图21），每根纤芯上双向各传输20G带宽，这样2根纤芯即可完成40G连接，采用OM3光纤传输距离为100m，OM4光纤传输距离为125m。在采用40G Bidi技术后，使用原有的10G布线系统可以平滑升级到

图20

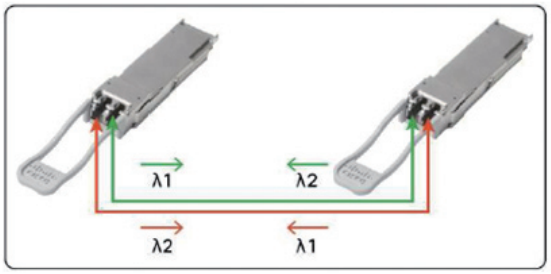


图21

Fiber Cable Cost*	30m	60m	100m
QSFP SR4 (288 x 12-fiber connectors) (US$)	$32,058	$53,562	$83,412
Cisco QSFP BiDi (288 x 2-fiber connectors) (US$)	$7,884	$12,966	$19,647
Savings (US$)	$24,174	$40,599	$63,765
Savings per 40-Gbps port (US$)	$84	$141	$221
Percentage cost reduction	75%	76%	77%

图22

40G，且Bidi光模块价格较传统光模块更低，使得部署40G骨干的总体成本和部署复杂度大幅度降低。

传统40G布线和采用思科BiDi技术的40G布线系统的成本对比如下，基于思科40G Bidi技术的布线系统能够节约传统40G布线系统75%以上的成本（图22）。

2.1.7 数据中心安全设计

下一代数据中心采用了Fabric的架构，安全网络服务可以分布式的部署在任意Leaf节点。结合数据中心的虚拟化部署，以应用为中心的基础架构中数据中心的安全设计涉及到多方面的考量。

Fabric内部的访问安全控制

以应用为中心的Fabric架构中，EPG（End Point Group）间的访问关系是由合约（Contract）来定义和约束的。在没有定义Contract的EPG间默认不存在访问关系，是不能相互通信的。这种白名单性质的Contract在网络连通性层面保障了各EPG间的安全访问关系。

EPG间的Contract包含Filter、Action和Label，可以实现EPG间特定端口的通信，从而在Fabric内部过滤不允许的访问。Filter用于定义特定通信，类似于Access List的作用，Action包含Permit、Deny、Redirect、Copy、Log和Mark等动作，可以实现对于指定流量的各种转发动作，Label用于标记该Contract（图23）。

网络安全服务

除了Fabric内部的Contract，下一代数据中心安全设计中还可以通过分布式的网络安全服务对通信安全进行进一步的保障。由于Fabric架构本身是一个高可靠、可平滑扩展的高性能架构，在传统金融网络安全部署中我们常用的防火墙HA方式由于其性能不可灵活扩展，当使用高性能设备替换低性能设备时需要一定割接窗口时间成为网络扩展时的瓶颈。因此思科于2012年推出了ASA Cluster技术，可把多达8台ASA防火墙（图24）以集群的方式在单一网络安全服务节点部署，同一Cluster内的ASA执行相同的安全策略，安全策略仅需配置一次即可在集群范围内自动同步。基于ASA Cluster技术（ASA可以提供状态化访问控制、应用防火墙、IPS、VPN等各类网络安全服务），我们可以根据性能需求部署网络安全服务，当需要扩展性能时，直接在Cluster中增加ASA的数量即

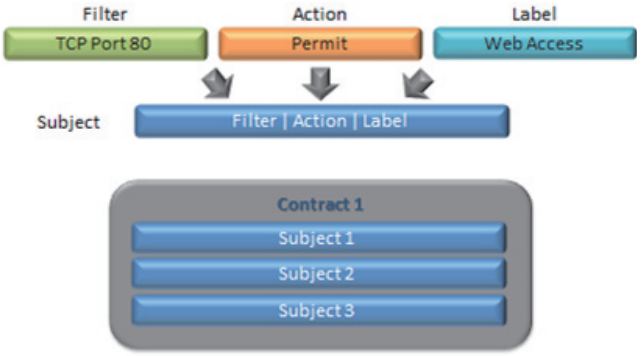


图23

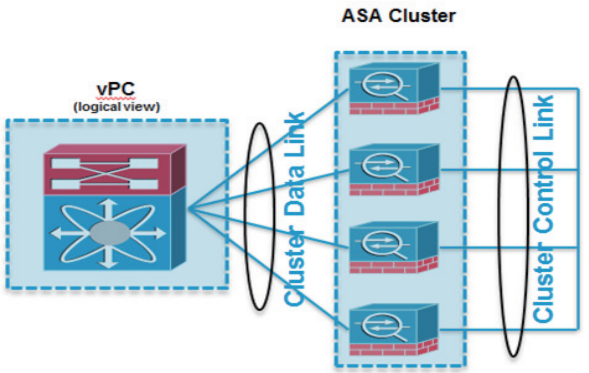


图24

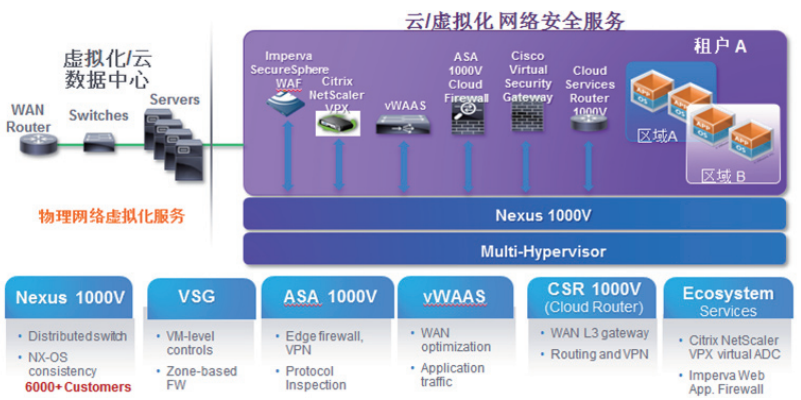


图25

可完成性能扩容，实现网络安全服务按需扩容、简化管理和保护投资的三重目的。2013年思科进一步推出了数据中心互联（DCI）验证的ASA Cluster版本，使得同城数据中心间的网络安全服务可以有效融合，配合网络层面Fabric的扩展技术，在真正意义上把同城数据中心网络架构打造成为地理位置分散的而业务连续的高可靠双活数据中心网络及安全服务。

虚拟化环境下的网络安全

在国内金融数据中心，大部分生产类应用部署在非X86平台（大机和小机）上，但随着近年来X86平台性能的发展，虚拟机的使用越来越广泛，同时X86的成本优势越来越明显，所以国内金融行业尝试向X86平台迁移的努力已经显现，很多WEB应用已经迁移到X86平台之上，因此X86虚拟机环境下的网络安全也是下一代数据中心安全设计中的一个重要组成。思科提供完善的虚拟化数据中心安全解决方案，满足数据中心层次化安全控制的需求，包括虚拟机到虚拟机、虚拟机到物理机、物理机到物理机各种类型访问的安全控制（图25）。

- **在虚拟机和虚拟机之间由VSG进行访问控制，VSG和vPath技术配合，将虚机间通信的流量引导至VSG进行安全检查，首包通过后该会话的其余数据包直接通信，提高安全控制效率**
- **在虚拟机边界部署VASA完成虚拟机和外部的安全控制。**
- **不同区域或网段间的安全控制可通过物理ASA（ASA Cluster）部署完成**
- **除此之外，思科及合作伙伴还可提供虚拟化环境下的负载均衡，WAAS和虚拟路由等多种虚拟化网络服务。**

2.2 数据中心交换矩阵技术的比较和选择

所有新技术的出现都是依赖于业务发展需求的驱动，而非凭空产生，对于构建智能数据中心的核心技术—交换矩阵技术来讲也是这样。

随着金融机构产品创新的速度加快，作为数据中心建设的核心——交换矩阵技术必须适应不断变化的业务和应用的需要，提高应用部署的敏捷性；同时实现精细化的安全控制以满足安全合规的要求；提供应用的可视化及传输优化，以改善应用响应时间；整个系统应具备良好的扩展性，支持不断扩大的网络规模。从根本上简化、优化和加速应用系统上线的进程。

另一方面，虚拟化技术的蓬勃发展和广泛应用，云计算和大数据等新的计算部署模式的出现，用户之前按照所谓“生产”、“管理信息”等业务区分的网络分区边界已经限制了应用的快速部署和数据交换性能，分区内服务器的规模也在不断增加，要求网络平滑地扩展以及设计上满足物理机或者虚拟机位置的灵活部署以及计算资源调度的便利性。

服务器的虚拟化使得传统的服务器和物理交换机的边界变得模糊，网络需要知道虚拟交换机下连接的主机情况。对于大量虚拟机的出现，相应地基于业务上线的自动化配置和编排甚至业务处理流程的优化也被提到议事日程上来。

此外，整个交换矩阵的可编程能力以及对外提供开放的应用程序接口的能力，对于方便用户的使用和部署也成为考虑的要点之一。

面对不断涌动的需求，思科提出了多种处理技术和解决方案，每个方案都有针对性地解决了数据中心发展所面临的不同阶段和不同层面的问题，下面我们做一简要的概述和比较。

2.2.1 FabricPATH 交换矩阵技术

传统的数据中心交换网络架构设计，每个业务分区内部通常是一个二层交换域，二层交换网络的特点是配置简单，服务器的接入类似“即插即用”，同时很好地支持虚拟机的资源调度，包括VM迁移。但是对于分区之间的虚拟机迁移和VLAN延伸的需求（二层网络的灵活接入），要求新的解决方案来解决（图1）。

可能你会说，可以将开放平台业务分区（金融机构中业务扩展最快的区域）合并成一个大的业务分区，并采用传统的分区结构设计。不错，但是随之而来的问题是可靠性、效率和扩展性问题。传统的生成树协议收敛慢、带宽利用率低、管理负担重（调整最优路径）、局部出现的问题会扩散到整个二层广播域、设备的MAC地址表的容量问题等等，显然这是不能够为大规模的二层交换网络部署带来可靠性和稳

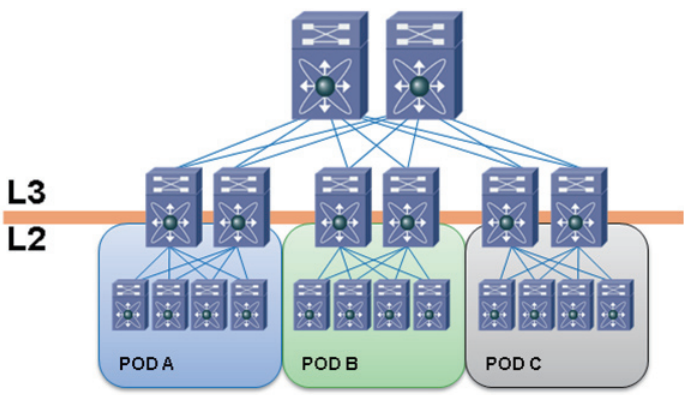


图1

定性的保障。而采用跨交换机的链路捆绑技术（例如VPC技术），虽然解决了传统生成树技术所遇到的绝大部分问题，但是配置相对复杂、后台仍然需要运行生成树以避免非VPC接入设备造成的可能的环路，进而对拓扑设计方面的灵活性产生一定约束，缺乏扩展性（当网络规模需要多于两台VPC使能的汇聚层设备时）。

Fabric PATH技术的产生正是要解决上述的问题。

拓扑稳定性：Fabric PATH彻底告别了生成树协议，采用路由技术来构建交换机设备之间的拓扑结构，收敛快，可自动计算最佳路径。

性能提升：支持多达16路等价路径（东西向流量的带宽利用率得到极大提升，提高吞吐量、降低时延）。Fabric PATH中的Spine（骨干）节点可以配置到16个，并且支持Anycast HSRP网关（提升南北向流量通道带宽）。

灵活性和扩展性：拓扑结构设计灵活，不必担心环路的问题。整个交换矩阵支持拓扑结构的横向扩展。Fabric PATH交换机不会学习泛洪报文中的MAC地址，只有在接收的单播帧中的目的MAC地址为本地MAC地址的条件下才会学习新的MAC地址，即会话式MAC 地址学习功能，该功能减少了交换矩阵中对FP边缘交换机的MAC地址表容量的压力 。

配置简便：配置工作量相比传统的跨交换机的链路捆绑技术减少约90% 。（图2）

综上所述，Fabric PATH技术方案 增强了二层交换网络的拓扑结构稳定性、大幅提升了性能（包括东西向和南北向）、具备强大的结构扩展性和组网灵活性以及配置便利性，是取代传统网络分区中二层交换域结构的最佳选择。因此可以说该方案不仅适用于新建二层网络交换域结构，也适合传统二层交换域网络结构的改造。

下面，我们看一下如何将传统网络分区结构迁移到Fabric PATH架构（图3）。

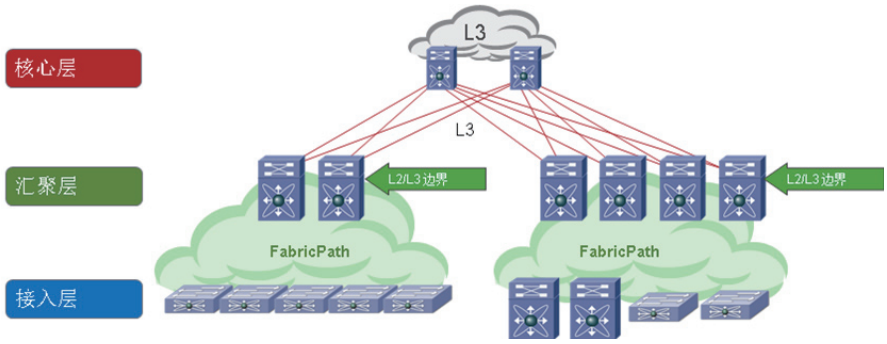


图2

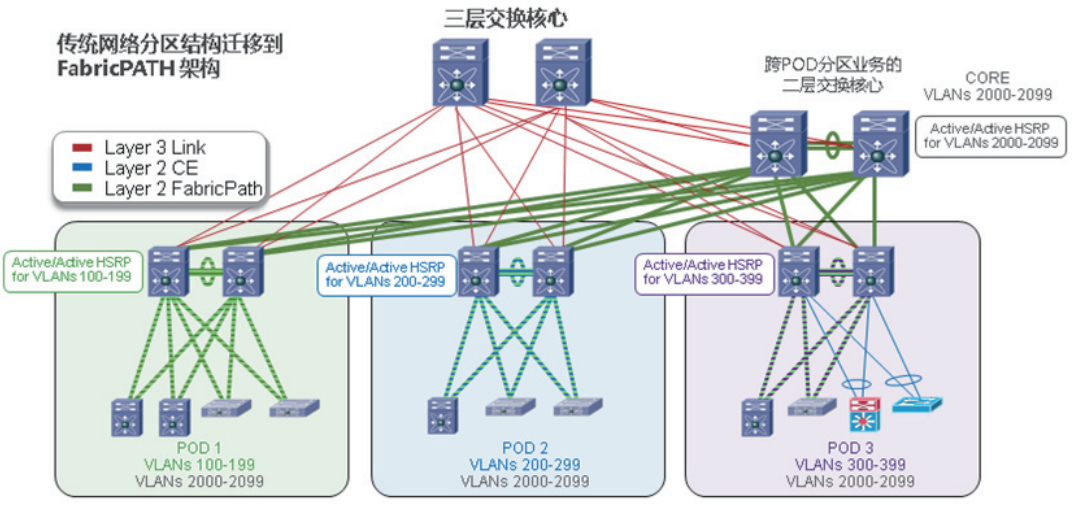


图3

图例中POD1，POD2，POD3为传统业务分区，分区内二层交换网络架构升级到Fabric PATH，每个分区内的Fabric PATH运行两个独立的路由拓扑，一个是本地POD的拓扑，另外一个是跨POD分区的拓扑（缺省拓扑）。本地Fabric PATH拓扑只包含本地VLAN，如例子中POD1 的本地Fabric PATH拓扑只包含VLAN100-199，跨POD分区的拓扑（缺省拓扑）包含跨越POD1-3的VLAN 2000-2099。本地VLAN终结在本地POD的HSRP网关设备上，跨POD分区的VLAN2000-2099终结在新建的二层交换核心设备上。

这里有两点需要说明：一个是每个POD分区内的VLAN是可以重用的，例如POD1,2,3 可能原先都是使用VLAN100-199（这种情况在很多客户的实际网络中是存在的），配置到Fabric PATH时采用各自独立的拓扑即可避免冲突。另一个是跨POD分区的拓扑（缺省拓扑）中不需要运行所有的VLAN（如VLAN100-199，VLAN200-299，VLAN300-399）只需要配置跨POD分区的VLAN2000-2099。本例中共配置了四个Fabric PATH拓扑。

从这个例子我们可以看出Fabric PATH组网的灵活性。目前，思科的交换机支持多达8个Fabric PATH拓扑处理能力，可以满足目前金融机构各业务分区迁移到Fabric PATH架构的需求。

2.2.2 TRILL协议和Fabric PATH的关系

TRILL（Transparent Interconnection of Lots of Links，多链路透明互联）的设计目标与Fabric PATH相似，目前只发布了几个“建议标准”类型（Proposed Standard）的RFC（“建议标准”类型的RFC不要求具备实际部署和使用经验。注意：不是所有RFC都是标准，只有标准轨迹类型的RFC才能成为各厂家在实现相关技术时所必须遵循的标准），根据IETF组织的规定“建议标准”类型RFC只有经过广泛使用后才可升级为互联网正式标准（将被正式赋予STD编号）。思科参与在IETF的TRILL工作组，并一直力图完善其各项功能。尽管目前思科的数据中心交换机在硬件上完全支持TRILL封装和处理，但是由于目前看TRILL存在诸多尚未解决的问题，距离实际部署还有相当长的距离。实际上TRILL还在不断的标准化过程中，依旧有大量的草案在讨论中。有兴趣的读者可参考http://datatracker.ietf.org/wg/trill/。

我们大致看一下目前TRILL存在的一些尚未解决的实际问题：

- **TRILL不支持多拓扑能力，不具备FabricPATH的灵活性，限制了用户网络的迁移和使用。**
- **TRILL不具备FabricPATH中会话式MAC地址学习功能，导致TRILL设备中保存不必要的MAC地址信息，产生MAC地址表容量可能溢出的问题。**
- **TRILL不支持Active-active FHRP，导致南北向流量出现瓶颈。**
- **TRILL不支持客户端CE做跨设备的多链路捆绑，导致客户端接入不能实现多路径和快速收敛。**
- **TRILL设计的封装报文具备不必要的额外开销，相比FabricPATH封装传输效率低。**

2.2.3 动态交换矩阵自动化架构 —— Dynamic Fabric Automation

尽管FabricPATH为用户构建稳定、灵活可扩展的二层网络提供了相比以前最好的解决方案，但是随着用户网络中虚拟机的大量运用以及多租户需求的出现，要求网络具备自动化配置能力来进一步提高灵活性，简化物理服务器与虚拟机部署工作，并支持其在交换矩阵中快速移动。同时对于交换矩阵自身的可靠性和稳定性也提出了更高的要求。显然需要另外一种全新的解决方案来满足上述业务发展的需求。

思科提出的动态交换矩阵自动化架构DFA解决方案主要提供四大功能：交换矩阵集中管理、工作负载的自动编排、优化的交换矩阵架构、虚拟交换矩阵（多租户）。

2.2.3.1 交换矩阵集中管理

优势在于提供单一的交换矩阵集中管理点(CPOM)，可减少管理接触点，缩短排障时间，实施交换矩阵整体监控。此外它还包括如下重要的功能：

提供交换矩阵内所有网络设备的自动配置能力（POAP）：

思科的管理软件DCNM已经集成了POAP引擎，可以实现配置模板的自动下发，资产管理。

互联线路管理和一致性检查：DCNM实施拓扑管理，同时可检测线路异常（自动发现设备之间错误的连线并进行处理），检测可针对DFA和非DFA设备。

虚拟交换矩阵（多租户网络）和主机的可视化定位（仅限于DFA）（图4）：可在DCNM上将查询的主机和租户网络定位在设备节点上，极大地方便了运行维护工作。

2.2.3.2 工作负载的自动编排

当业务编排管理员配置虚拟服务器和物理服务器时，该功能可自动将创建的网络策略应用于DFA叶子交换机和虚拟交换机中。当虚拟服务器在DFA交换矩阵中移动时，网络策略将被自动迁移到所连接的DFA叶子交换机和虚拟交换机中。大大简化基础设施部署工作，为虚拟机（VM）上线提供基础设施动态配置。可提供开放的API，与云管理平台实现出色的集成。

支持完全自动化的网络配置、半自动化的网络配置、手工配置三种配置模式。

全自动化的网络配置：集成了计算和存储的自动化编排能力、网络和网络服务的自动化编排能力。业务编排管理员定义逻辑组织网络（包括管理Segment-ID资源）并和自动配置文件的名称相映射，云管理编排软件（如UCS Director，OpenStack，vCloud Director等）直接和vSwitch交互（包括Segment-ID信息）。网络管理员在DCNM（CPOM）准备自动配置文件模板。一旦虚拟机上线，MAC地址学习或者VDP信令触发配置下发动作，包括动态VLAN ID，SVI，VRF以及动态VLAN ID和Segment-ID映射等配置信息从DCNM（CPOM）推送到DFA的叶子交换机，进而VLAN ID信息通过VDP协议传递到vSwitch。

半自动化的网络配置：没有云管理编排软件的参与，网络管理员需要手工完成虚拟交换机网络信息和DCNM（CPOM）自动配置文件模板中信息的映射关系。例如在不支持VDP的环境下，网络管理员在叶子交换机配置移动域，在DCNM（CPOM）配置自动配置文件模板信息（包括虚拟机静态VLAN ID和移动域的映射），手工配置vSwitch 的Port-Profiles及Port-Groups，一旦虚拟机上线，MAC学习触发DFA叶子交换机在DCNM的LDAP库检索后下载网络自动配置 文件，包括VLAN ID，SVI，VRF以 及VLAN ID和Segment-ID映 射 等配置信息。如果是在VDP环境下，网络管理员在叶子交换机配置动态VLAN ID范围，在DCNM（CPOM）配置自动配置文件模板信息（包括动态VLAN ID和Segment-ID的映射），一旦vSwitch通过VDP触发了Segment-ID信息，DFA叶子交换机在DCNM的LDAP库检索后下载网络自动配置文件。

手工配置（在小型的网络中或者用户重点考虑DFA其它功能时，可采用该模式。它更适合传统网络向DFA网络过渡时使用）：遵循传统的网络管理模式，网络管理员在DFA叶子

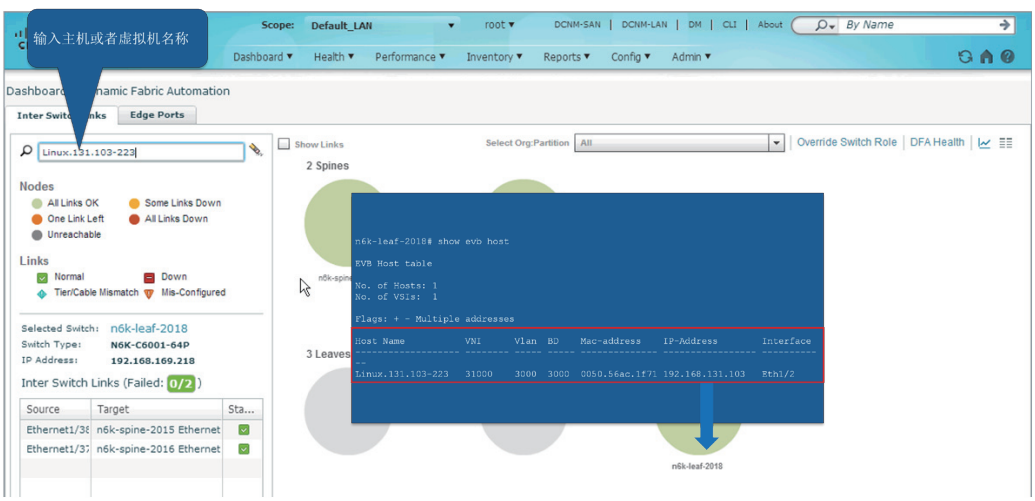


图4

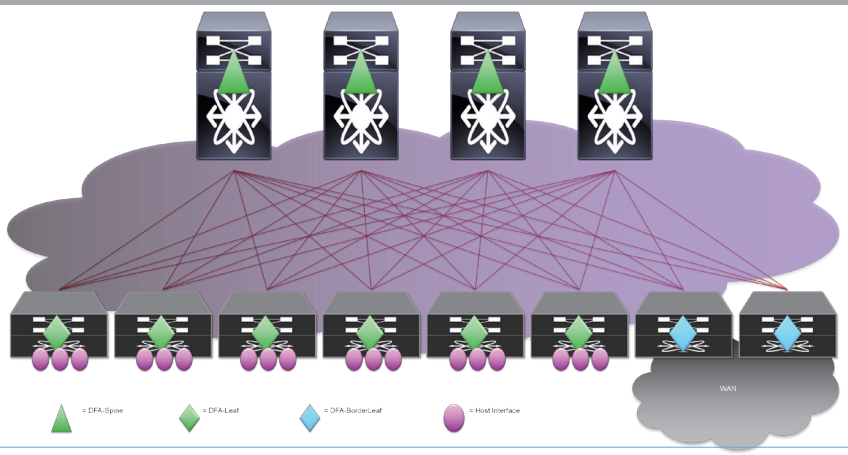


图5

交换机上手工配置网络VLAN、SVI、转发模式以及VLAN到Segment-ID的映射。

2.2.3.3 优化的交换矩阵架构

我们回顾一下在Fabric PATH架构中，尽管在交换机互联的拓扑上用路由方式完全取代了传统的生成树协议，实现了拓扑结构的稳定性，优化了二层网络，但是客户端报文的转发没有独立的控制平面来控制（TRILL存在同样的问题），客户端报文的转发决定是依靠交换机数据驱动（MAC地址的学习）得到的，这样当客户端发送诸如ARP等广播报文时，将穿越整个交换矩阵，所在VLAN中的所有终端将接收到该报文，如果出现问题，排查的故障范围过大（图5）。

在DFA交换矩阵架构中，增加了客户端报文转发的控制层面，将故障域隔离到叶子节点交换机下联的局部范围内，大大提升了整个交换矩阵架构的稳定性和可靠性。该控制层面通过注入客户端主机路由方式实现，路由的宣告、分发和控制采用业界成熟的技术MP-BGP，并做了适当增强来实现。采用分布式网关技术、将网关服务下沉到每个叶子节点上，这样客户端以及网络服务设备（如防火墙）的接入位置可以非常灵活地部署，同时BGP已经被证明可以很好地支持百万级路由处理，所以DFA底层的交换矩阵架构设计具备极好的可靠性、稳定性、灵活性和扩展性。

简单描述一下DFA交换矩阵架构的底层设计：包括三个控制平面，两个转发平面。

DFA 交换矩阵控制平面——网络拓扑的控制平面（图6）：与FabricPATH完全一样，仍然采用ISIS作为设备之间拓扑

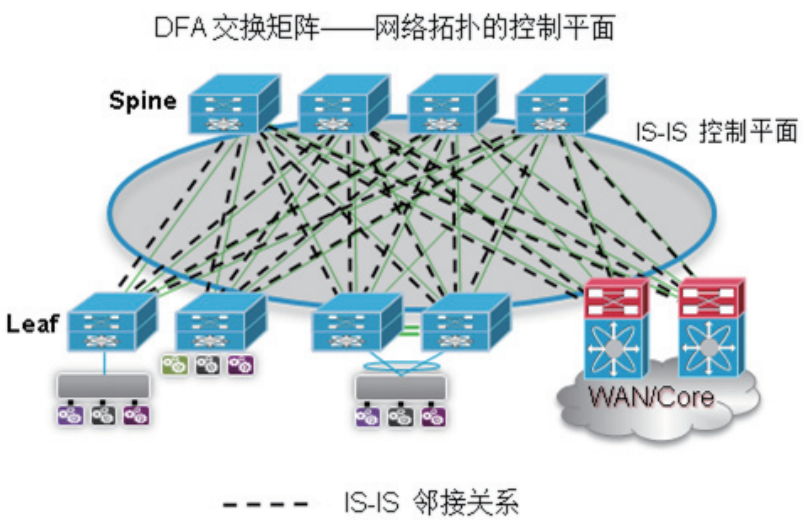


图6

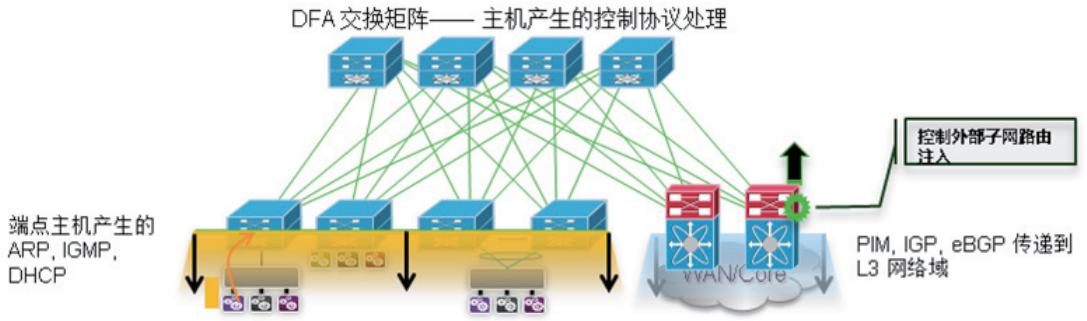


图7

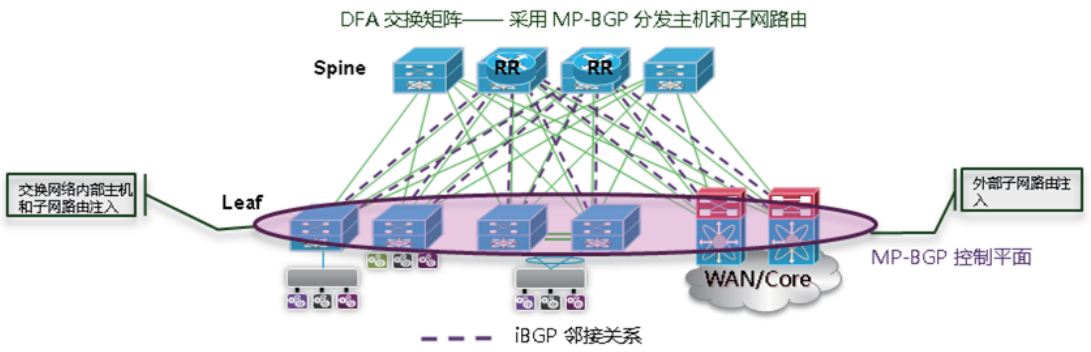


图8

建立和选路的协议。

DFA 交换矩阵控制层面—— 主机产生的控制协议处理（图7）：与FabricPATH不同，主机产生的ARP, IGMP, DHCP等消息中止在DFA的叶子节点。 控制泛洪和故障域。来自外部网络的 PIM, IGP, eBGP 等路由消息中止在DFA的边界叶子节点。

DFA 交换矩阵控制平面—— 采用 MP-BGP 分发主机和子网路由（图8）：主机路由分发工作从DFA链路状态协议（ISIS）中分离，叶子节点采用MP-BGP分布内部主机/子网路由和外部可达性信息。

可选择两台Spine节点设备作为BGP的路由反射器，以增强扩展性。为了宣告主机路由的可达性信息，叶子节点必须首先发现本地连接的设备。本地主机的检测，基于 VDP或者ARP/DHCP。远程主机的检测，收到MP-BGP通知消息，安装路由到Unicast RIB表。（当配置L3 会话式学习时，只是选择性安装路由）。

DFA 交换矩阵转发层面——代理网关模式和任意播网关模式：DFA在叶子节点采用分布式网关设计，所有叶子节点上共享所有客户端主机子网的网关IP和MAC(非HSRP操作)。ARP报文中止在叶子节点，在叶子节点外没有任何泛洪。帮助实现虚拟机迁移，工作负荷分布，组建任意的工作集群系统，在物理主机和虚拟机之间提供无缝的二层或者三层通信。针对叶子节点下面的每个VLAN，存在如下两种转发模式配置：

Proxy-Gateway 代理网关模式（缺省配置模式）——利用 proxy-ARP 确保两个主机之间的通信永远是被路由的（不论是同一子网内部还是不同子网之间的通信），叶子节点总是执行L3 查询操作。

Anycast-Gateway 任意播网关模式（仅用在存在静默主机的VLAN）——确保分布式缺省网关部署在每个叶子节点但是针对同一子网内通信转发采用传统的FabricPath行为（例如：ARP是被泛洪到整个交换矩阵，叶子节点执行L2查

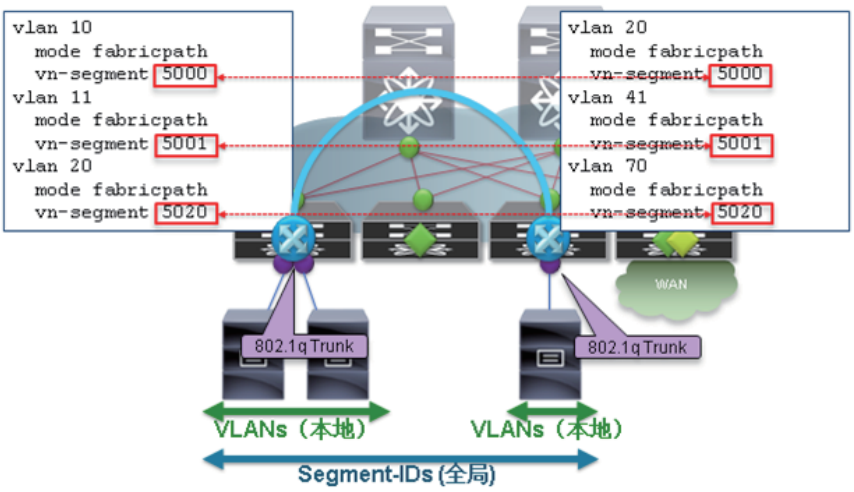


图9

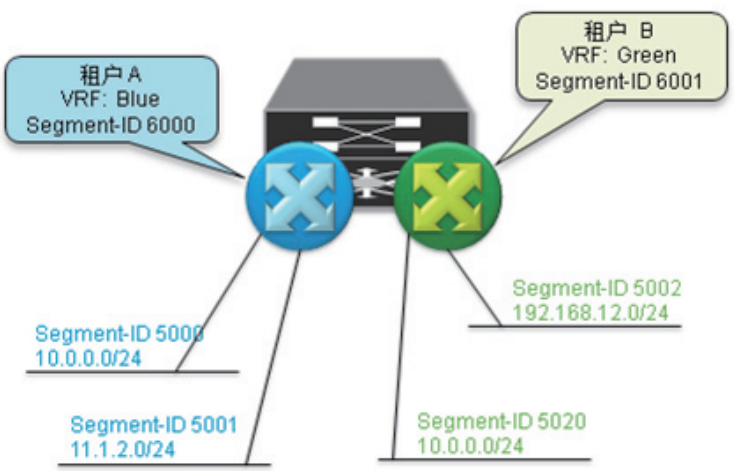


图10

询，数据平面针对终端的MAC地址基于会话式学习），不同子网的转发是按照三层转发操作，这个和Proxy-Gateway模式的操作行为一致。

2.2.3.4 虚拟交换矩阵（多租户）

Fabric PATH（包括Trill）在设计上无法支持多租户的部署场景，而在DFA交换矩阵中使用Segment-ID来实现VLAN ID的重用和区分多租户网络（图9）。传统的标识二层交换域唯一性的VLAN ID 的空间最大为4096个，而DFA交换矩阵内部使用Segment-ID（24位）使二层交换域在理论上扩大到1600万个（实际数字需要看Nexus交换机的硬件处理能力），在叶子交换机上执行VLAN ID到Segment-ID的1:1映射，Segment-ID是全局唯一的，而VLAN ID仅仅在本地叶子节点连接的区域有效。同时DFA还使用专用的Segment-ID来标识每一个VRF，VRF用来映射到租户Tenant（图10）。

DFA交换矩阵架构不仅提供了支撑云计算平台的能力，支持虚拟机业务上线的自动化网络配置、多租户能力，提供了统一管理和监控、统一配置和自动下发。更重要的是从底层交换矩阵设计上将可能出现的故障域压缩到最小范围，进一步提升了全网架构的稳定性和可靠性，支持物理机和虚拟机的部署，支持第三方云计算自动化业务编排软件工具。全分布式网关的设计保证各种类型业务（服务器、网络服务设备等）部署与网络连接的物理位置无关。强大的控制平面和转发平面的弹性设计，使得网络部署规模的适用范围很广，不仅包括大型网络，也

可以是很小的业务分区。

2.2.4 以应用为中心的交换矩阵技术–ACI

随着金融行业之间的竞争加剧，新的应用产品不断涌现，应用架构的整体趋势正变得日趋复杂和庞大，基础架构在支撑业务的快速部署、资源调配方面面临着巨大的挑战。

挑战一：传统的网络设计将应用之间的逻辑实现和访问与网络元素（IP和VLAN等）绑定，将应用访问的安全策略和服务质量控制与网络元素绑定，根据网络元素（而不是应用）来识别和调度流量，这种紧耦合的设计方法造成应用上线所导致的网络配置和管理的复杂性（图11）。

不仅在业务上线过程中，而且在日常的运行维护管理中，管理人员需要将大量的网络配置翻译成应用之间的逻辑关系，因为他们面对的是一堆网络元素（IP和VLAN等）。防火墙、负载均衡器和网络设备上也存在大量本不应存在的访问控制。这种复杂性造成了应用维护和迁移的困难（图12）。

由于当今的网络架构对于运行其上的应用没有感知，因此还创造了额外的解决方案去实现应用的感知，例如深度包检测技术。

挑战二：安全防护和控制问题永远是金融机构的首要考虑因素之一。传统的网络设计中，业务分区划分的重要考虑因素便是安全控制。很多业务流需要跨越多个业务分区，甚至整个网络。一个完全的应用访问可能需要经过多个安全控制点，但是这些控制点的设备是分散的，缺乏一个针对应用访问路径的

图11

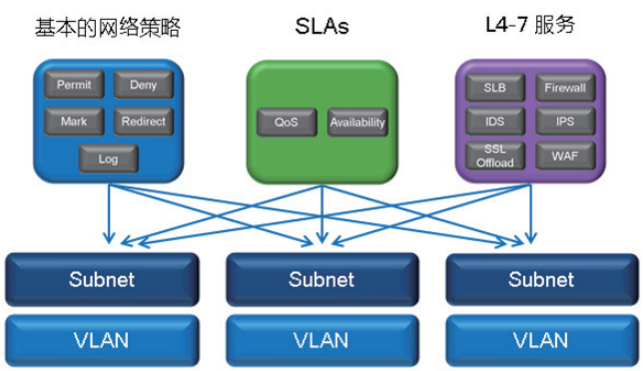


图12

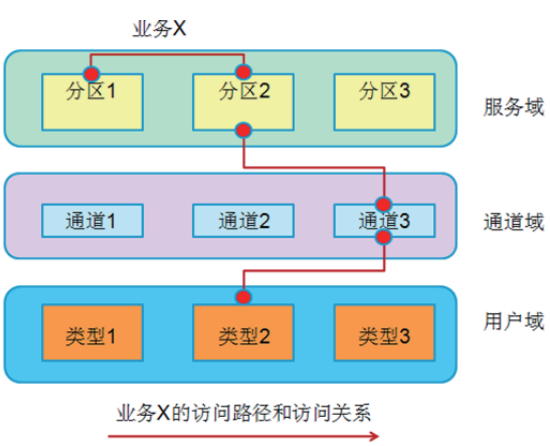


图13

图14

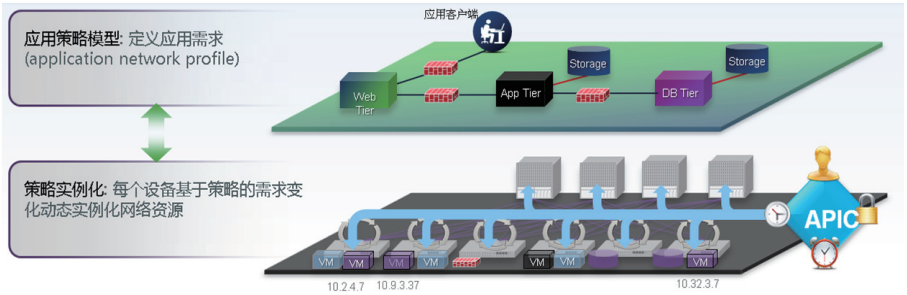


图15

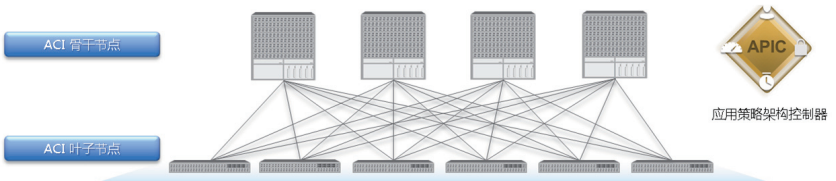


图16

全局性通道控制和整体的策略描述（图13）。

众多的安全控制点降低了业务部署的灵活性，资源往往被限制在某个区域内，同时安全控制区域之间的访问性能也受到极大的影响。

挑战三:考虑到开发自动化配置管理工具的负荷，金融客户希望通过最简单的方式来实现多租户网络的虚拟化能力，并且能够提供图形化的控制管理面板。同时提供云计算平台，具备自动化业务编排能力，可以高效的调度计算和存储资源和最大限度地使用计算和存储资源，为金融机构节约投资成本、提高利润产出。

挑战四:对于交换矩阵技术来说，用户最担心无法有效监控应用流在矩阵内部多条等价链路和设备上的传输问题，包括性能监控，延时和丢包情况。

面向应用为中心的交换矩阵技术ACI，采用了全新的设计理念以应对上述挑战，为智能数据中心的构建提供了理想的解决方案。

ACI在应用传输和底层的网络元素之间定义了新的抽象层ANP（应用网络模版），该层独立于外部网络，不会改变传统的网络协议传输方式，而在ACI内部，ANP基于应用逻辑之间的合约 来定义应用服务提供者和使用者之间的访问关系（图14）。而将网络实体元素按需映射到不同的EPG（Endpoint Groups)中，增加抽象层的好处显而易见，它可以使得面向应用的网络服务策略保持独立性，网络实体元素即使发生变化，比如IP地址变化，而应用的访问关系之间的策略依然保持不变，并且该策略可以被其它抽象层重复调用。应用服务端点的标识、位置和关联的策略解耦合，所有这些独立于底层网络（图15）。

ANP中定义了各层级应用逻辑之间的访问策略，进而将这些策略下发到交换矩阵的转发层面，形成一个端到端的转发通道。由于采用了类似DFA的分布式网关设计，消除了主机IP控制层面（ARP, GARP)的泛洪需要，主机与交换矩阵之间的连接与物理位置无关，IP 地址可以携带迁移到矩阵中任意位置，极大地方便了实际部署。进入到交换矩阵的数据报文（包括使用不同封装格式的报文，如802.1Q VLAN, IETF VXLAN, IETF NVGRE 等)将被用ACI定义的eVxLAN封装统一化处理，进而在矩阵内部传送（图16）。

ACI交换矩阵可以提供自动服务插入和重定向。安全控制和转发操作完全从物理或者虚拟网络实体元素属性中解耦合出来。

下面是一个采用基于策略转发模型的设计举例（图17）。被定义为“Users”的EPG发起对目标为”Files”的EPG的应用访问，数据流到达交换矩阵的物理端口，将根据已经从APIC下发到端口硬件的策略实施转发，该策略来自于ANP定义的EPG之间的访问合约。合约中依据安全管理员定义的策略需要将该应用流重定向到防火墙处理，则该数据流报文被交换矩阵重定向到安全设备处理完毕后，转发到目标的EPG”Files”。

ANP构建了应用访问的全程逻辑隔离的通道，卸载了部

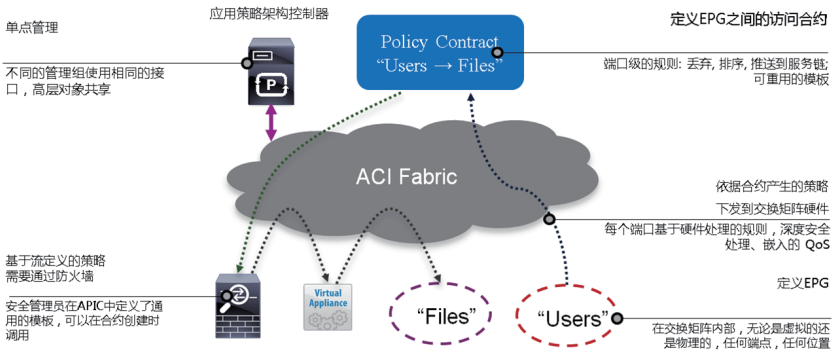


图17

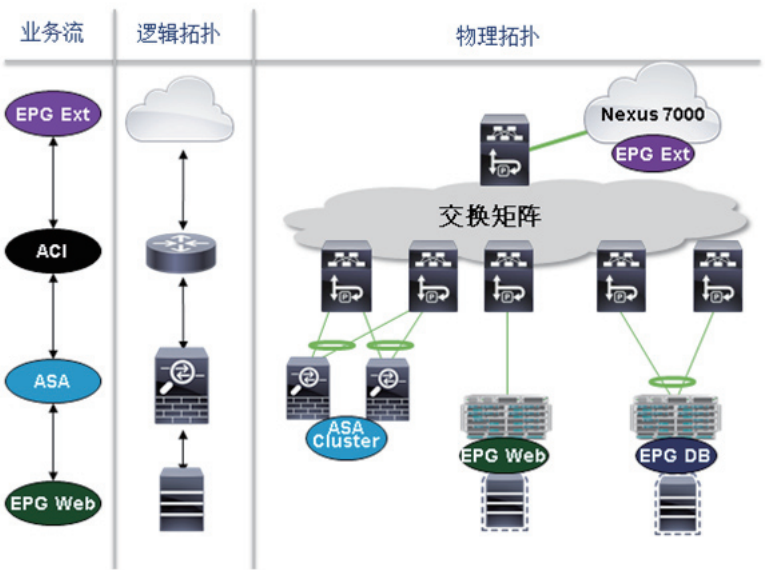


图18

分防火墙等安全设备的处理负担，提供了业务流的全程处理视图，为智能数据中心的构建提供了弹性扩展的基础支撑。

对于一个应用访问的业务流，APIC给出了完整的逻辑拓扑，将安全精细化处理到每个应用处理的端到端控制（图18）。

相对于其它交换矩阵技术实现多租户网络的配置，APIC采用集中式GUI配置，简便易操作，支持高达64000多个虚拟的网络定义。把整个ACI交换矩阵看做一个交换机，每个租户网络相当于在上面构建了多个单跳的虚拟叠加网络。相比传统的SDN解决方案，这些运行在虚拟叠加网络之上的应用系统在交换矩阵内部传输时，是完全可视化的（图19）。

应用可视化的实现依靠ACI交换矩阵提供的原子计数、内部高精度延时测量的实时遥测技术，ACI交换矩阵可以获得应用传输延时和线路拥塞的状态，并且基于Flowlet交换算法，可以实现应用传输流从拥塞线路到轻载线路的动态转移，动态调整应用流传输的排队优先级，实现动态负载均衡。

ACI交换矩阵提供针对云计算业务编排软件的北向接口，可实现自动化的资源调配。首先业务编排软件按照业务需求，确定所需资源配置；定义相关的虚拟网络服务；从计算资源池调配服务器配置虚拟计算机，完成计算资源配备；划分虚拟存储区，定义虚拟存储网络服务，连接虚拟存储区，完成存储配备；指定操作系统并启动服务器，应用系统启动准备及上线；

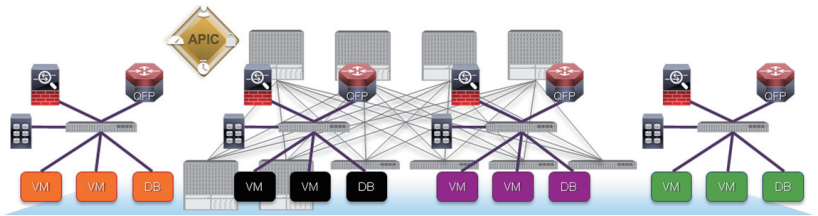


图19

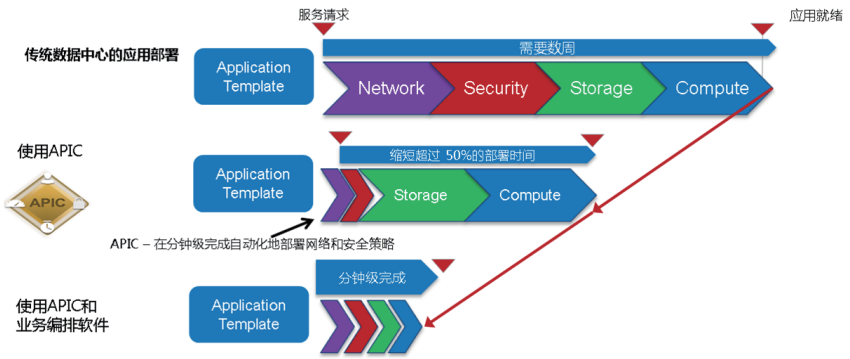


图20

对应的数据流被导向到新建的逻辑服务区。（图20）

传统的数据中心实现应用的部署，往往需要数周时间，而单独采用APIC应用策略基础架构控制器后，可在分钟级内完成网络架构和安全策略的部署，如果和云计算业务编排软件配合使用，可实现包括计算和存储资源调度在内的分钟级部署。

ACI交换矩阵架构提供了业务部署的最大的灵活性（包括多租户网络的建立）和便利性、应用传输和监控的可视化和高性能、以及提供精细化的面向应用级的安全控制。它也是一个开放的系统，提供丰富的南向和北向接口。

2.2.5 总结

针对不同的应用场合和需求，用户可采用不同的交换矩阵解决方案。

场景一：期望构建非常稳定的二层交换网络的拓扑，矩阵内支持VLAN内主机任意物理位置的部署和迁移，配置要简单明了。在大型网络中需要考虑解决东西向和南北向扩展需要。推荐采用FabricPATH交换矩阵技术。

场景二：大规模的数据中心网络部署，不仅需要网络拓扑结构稳定，还希望将本地接入侧的故障域压缩到最小范围，矩阵内支持主机任意物理位置的部署和迁移（不局限于在VLAN内部，而是在整个IP层面），支持多租户网络部署，支持虚拟机业务上线的自动化网络配置，集中化的图形配置管理工具，提供与第三方云计算自动化业务编排软件工具的北向处理接口。推荐采用动态交换矩阵自动化架构 —— Dynamic Fabric Automation。

场景三：适用于任何规模的数据中心网络，期望通过简便的手段能够非常清楚地了解和控制数据中心内部端到端的应用访问过程，实现面向应用的全局视角的精细化地安全控制。应用的部署和迁移过程非常地快速和方便，东西向和南北向的流量处理不会出现瓶颈，提供交换矩阵内部各个应用流传输的性能优化（丢包、延时、负载均衡）以及可视化监控，矩阵内支持主机在任意物理位置的部署和迁移（不局限于在VLAN内部，而是在整个IP层面），非常方便地支持多租户网络部署，支持虚拟机业务上线的自动化网络配置，集中化的图形配置管理工具，提供与第三方云计算自动化业务编排软件工具的北向处理接口，提供交换矩阵内各服务设备的南向处理接口。毋庸置疑，建设未来智能的数据中心推荐采用以应用为中心的ACI交换矩阵解决方案。

新一代智能数据中心

三、思科解决方案

3.1 利用网络虚拟化技术（VDC，VPC和FEX等）构建新型数据中心

思科不仅将国外最先进的技术和解决方案介绍到国内，更重要的是将国外发达国家金融行业信息化建设的宝贵经验引进中国，为中国的金融信息化建设提供强有力的技术支撑和可靠保障。现将思科在国内数据中心建设的经验分享如下：

客户面临的主要问题和挑战

- **数据中心空间、电力、空调制冷等资源严重不足，扩展数据中心时遇到挑战；**
- **传统纯物理式网络设备的部署架构提高了网络设备采购成本，增加了管理维护节点，导致运维成本居高不下；**
- **传统EOR布线方式增加了数据中心项目实施和日常维护的难度系数，浪费了大量的人力、物力和时间；**
- **HSRP等交换机冗余技术已经不能满足金融行业对业务稳定性的高要求，亟需找到一种在HA部署模式下，当有一台网络设备故障时备用设备能将网络快速收敛并恢复正常的技术；**
- **如何废除STP技术在二层网络中的应用，以避免广播风暴且提高传输带宽；**
- **如何充分利用网络设备的性能和端口，避免投资浪费。**

1）8台Nexus 7000每两台成一组共四组：其中一组为核心交换机，作为整个数据中心的核心骨干；其他三组为汇聚交换机，分别为基础服务器区、业务服务器区和测试服务器区这三个服务器汇聚区的汇聚交换机，保证了极高的可靠性和非常高的传输性能。Nexus 5000与Nexus 2000一起构成了数据中心的接入层。

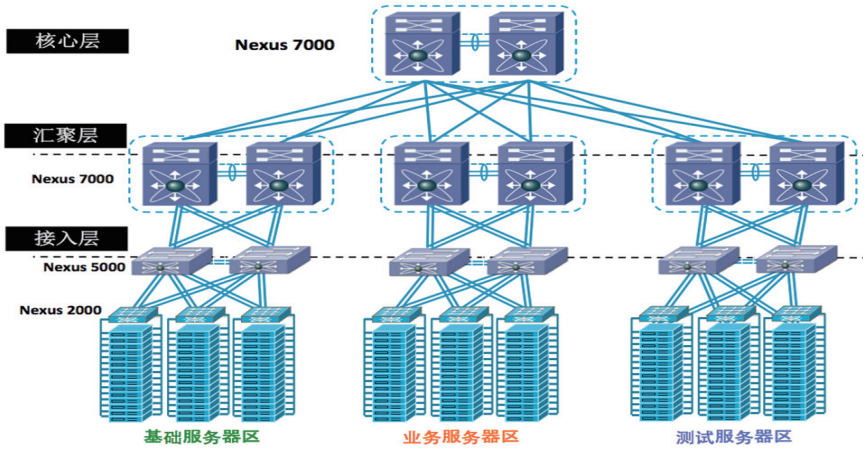


图1

2）接入层通过Nexus 5000使用FEX技术链接其远程扩展板卡Nexus 2000来实现。Nexus 2000作为柜顶交换机使用，但是该交换机的配置、管理、维护都是通过Nexus 5000实现的，因此既简化了布线，也减少了管理节点，同时还使得VLAN在不同机架之间扩展更容易。（图2）

3）使用Nexus 7000和Nexus 5000都支持的vPC技术，使得底层交换机双上连至上层两台交换机时，实现了跨机箱的以太网链路捆绑，逻辑上消除了环路，从而废除了STP并提高了传输带宽。（图3）

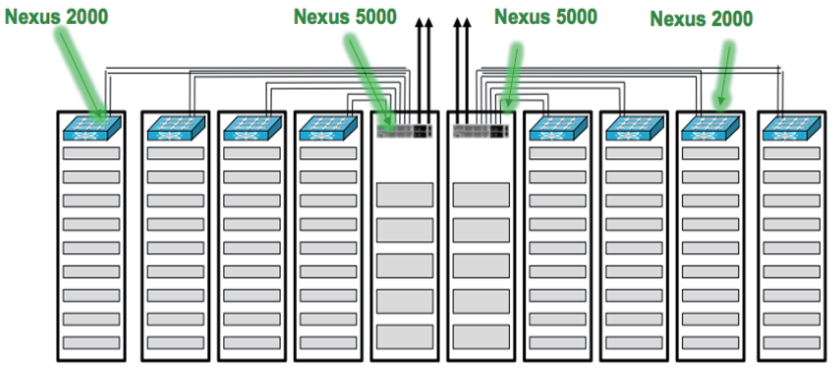
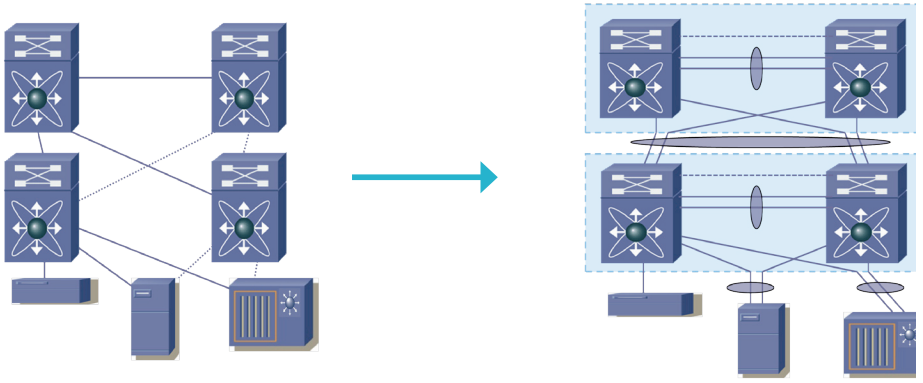


图2



没有vPC时的逻辑拓扑　　使用vPC的逻辑拓扑

图3

中小银行——思科7/2解决方案

1）两台Nexus 7000使用VDC技术共虚拟出8台交换机，非物理的两台配对成一组共四组：其中一组为核心交换机，作为整个数据中心的核心骨干；其他三组为汇聚交换机，分别为基础服务器区、业务服务器区和测试服务器区这三个服务器汇聚区的汇聚交换机（Nexus 7000目前可虚拟出8+1台虚拟交换机，其中1台作为管理交换机使用，而根据中小银行数据中心规模，每台Nexus 7000虚拟出4台交换机基本上可以满足要求）。同时为了简化数据中心架构、降低成本、减少转发延迟，Nexus 7000与Nexus 2000也一起构成了数据中心

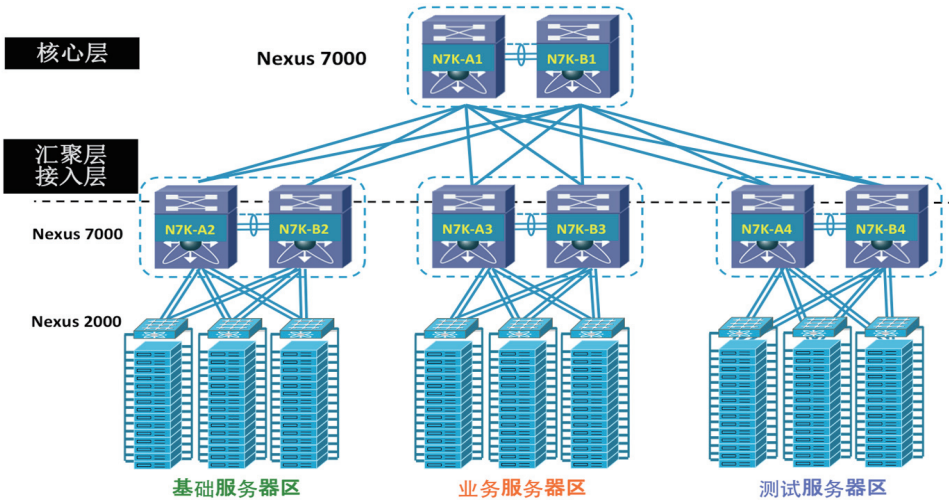


图4

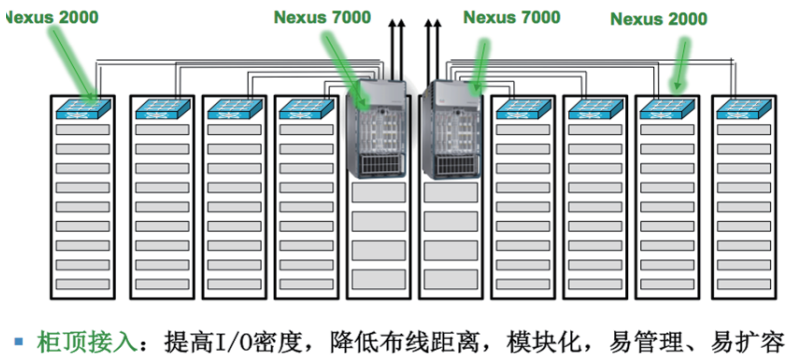


图5

的接入层（图4）。

2）接入层直接通过Nexus 7000使用FEX技术链接其远程扩展板卡Nexus 2000来实现。Nexus 2000作为柜顶换机使用，但是该交换机的配置、管理、维护都是通过Nexus 7000实现的，因此既简化了布线，也减少了管理节点，同时还使得VLAN在不同机架之间扩展更容易（图5）。

问题的解决和客户的受益

- 通过VDC技术实现了网络虚拟化，即交换机虚拟化。这样可以减少投资，降低成本，减少电能消耗，节省机柜和空间；
- 通过VDC技术，可以实现交换板卡上的端口灵活划分，从而避免了投资浪费；
- 通过vPC技术，交换机与交换机之间不再使用STP，逻辑上消除了环路，从消除STP阻塞端口、使用所有可用的上联带宽、双上联服务器可以运行在双活模式、在链路和设备故障时提供“零秒级”快速收敛；
- 通过FEX技术，实现了TOR部署方案，优化了布线结构，从而提高了数据中心项目实施进度，降低了后期运维的难度。由于所有Nexus 2000都不用单独管理，这就减少了管理节点；另外Nexus 2000还具有即插即用无需配置的功能，因此可以实现故障快速恢复，满足金融行业对稳定性和可靠性的要求；
- VDC、vPC、FEX结合模块化设计方案，使得数据中心的扩展更安全、更方便、更迅速，加快了新业务的部署进度。

3.2 存储融合的统一网络架构

客户面临的主要问题和挑战

- 业务增长带来的数据存储管理问题：传统分散的存储区域网络(SAN)中每个SAN服务于分离的应用，SAN不断零散增长给SAN管理提出了巨大的挑战；
- SAN兼容性要求：随着数据存储的增长，客户数据中心中存储设备的品牌也会相应增加，SAN面临多存储厂商兼容性问题；
- SAN网络设计需要给整个系统提供稳定和可预测的交换时延。

思科的解决方案：

思科通过 MDS 9000系列多层控制器(Director)和光纤通道交换机提供更高的端口密度、交换带宽、性能、多协议功能和可靠性，用于建设数据中心综合性存储网络。思科MDS 9000系列产品可以充当一个集中系统，提供SAN网络互联和高级服务。思科MDS 9000系列产品都是模块化的系统，针对很高的端口密度和数据中心应用的性能进行了优化。

思科MDS 9000系列多层控制器(Director)和光纤通道交换机还可以提供多种功能和服务，例如虚拟SAN、高级ISL链路汇集、LUN分区、故障通告(Call Home)、高可用性和不中断软件升级。MDS 9000系列产品还包括一个强大、内嵌的矩阵管理器应用，它可以配置、监控和诊断存储网络（图6）。

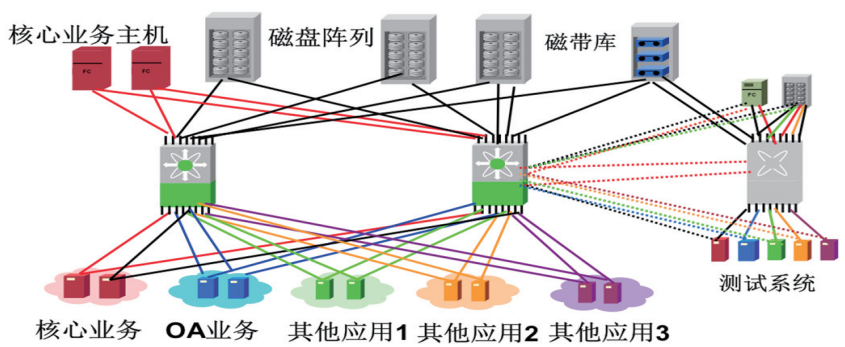


图6

问题的解决和客户的受益

- 多协议存储网络：通过集成 FC、FICON、FCoE、iSCSI 及 FCIP 等不同的协议，实现更低的总拥有成本 (TCO)、更高的灵活性；
- 企业级存储连接：支持虚拟化程度更高的工作负载，以提高可用性、扩展性和性能；
- 以服务为导向的存储区域网络：将所有网络服务扩展至任何设备，而不受协议、速度、供应商或位置的限制；
- 思科的SAN导向器可以加速存储I/O处理，进一步降低时延，同样的灾备距离下实现更低的I/O延迟或者同样时延容限下实现更远距离的灾备能力，与上层应用和下层存储都无关，实现存储的透明化和异构化；
- 统一操作系统和管理工具：降低运营开支、简化操作、实现无缝互操作并提供稳定一致的功能。

3.3 数据中心统一计算平台建设

客户面临的主要问题和挑战

- 现有计算资源不足，服务器性能偏低、数量不足；
- 资源利用率不高，且系统资源分散，难以实现集中管理，各业务系统之间无法实现资源调剂、统一调度；
- 现有的服务器系统管理模式下，随着服务器数量的增多，管理非常不便；
- 对基础设施的可用性要求越来越高，否则业务连续性难以保障；
- 技术人员人手不够，而IT系统越来越复杂、规模也越来越大，因此更加需要简便的技术管理手段；
- 现有机房空间有限，供电消耗巨大，新产品需要具有节能环保、节省空间等绿色技术。

思科的解决方案：

Cisco 统一计算系统 (UCS) 是适用于刀片式服务器计算的革命性新架构。Cisco UCS 是新一代数据中心平台，将计算、网络、存储访问和虚拟化功能整合到一个聚合型系统中，旨在降低总体拥有成本 (TCO) 和提高业务灵活性。

UCS系统集成了低延迟、无损万兆以太网统一网络结构与企业级 x86 架构服务器。此系统是一个集成的可扩展多机箱平台，系统中的所有资源都参与统一管理域。

Cisco 统一计算系统能够提高可扩展性，但不会提高复杂程度，无论系统内只有 1 个服务器，还是有 320 个服务器（带数千个虚拟机），都能将它们作为一个系统来进行管理。通过端到端配置和对虚拟化系统及非虚拟化系统的迁移支持，Cisco 统一计算系统能够简单、可靠和安全地加速提供新服务（图7）。

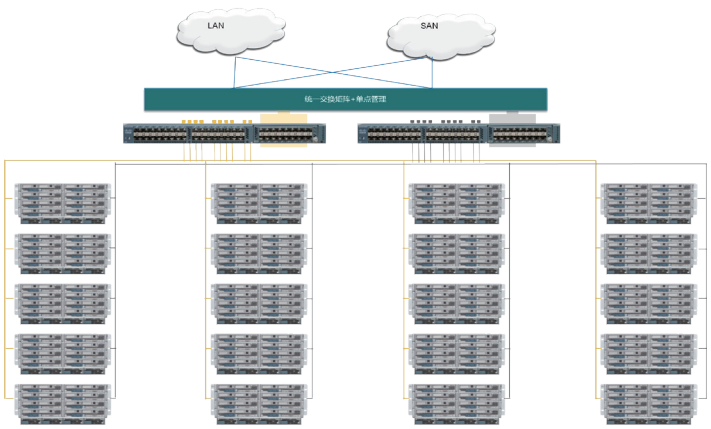


图7

问题的解决和客户的受益

- 统一的计算和网络平台；
- I/O虚拟化及高速吞吐能力；
- 简化的网络环境；
- 无状态的计算能力；
- 机架和刀片可以统一管理；
- 虚拟化的数据中心在大幅提高硬件资源使用效率的同时，还增强了系统的可用性；
- 保证了虚拟化网络的可视性和性能，增强了对虚拟化环境的管理和控制；
- UCS的一体化结构、无状态计算等特性使得技术人员对服务器的维护管理工作大为简化；
- 融合架构的基础设施简化了管理，减少了布线、制冷、机房空间等间接成本；
- 绿色节能的新一代数据中心，在数据中心不断扩容的情况下减少能源功耗。

3.4 商业银行灾备DWDM方案

客户面临的主要问题和挑战

- 同城灾备是实现商业银行业务连续性要求的重要环节，而成熟、稳定、高效、多业务支持的数据中心网络互联解决方案是亟须解决的问题；
- 同城灾备系统往往采用同步数据复制的技术保证两地数据完全同步，数据零丢失。因而要求跨越两地的存储通道具备高吞吐量、超低时延（以微妙计算）和极高的可靠性；
- 高可靠性：存储网络通道的高可靠性设计是同城灾备系统保障两地数据完整性和一致性的基石；
- 兼容性：具备系统的互操作证明保证集成最稳定，用户放心使用。 需要考虑与主流存储设备厂家的互操作证明，否则在实施时可能遇到互操作兼容性问题；
- 多业务支持：能够支持IP网络、以太网、SAN网络。

思科的解决方案：

ONS15454 DWDM是Cisco公司推出的新一代波分复用系统，是这个业界最先进的DWDM系统之一。该系统将传统长途DWDM的稳定性、大容量等特点和OADM波分系统的灵活性融为一体，为用户组网带来了全新理念（图8）。

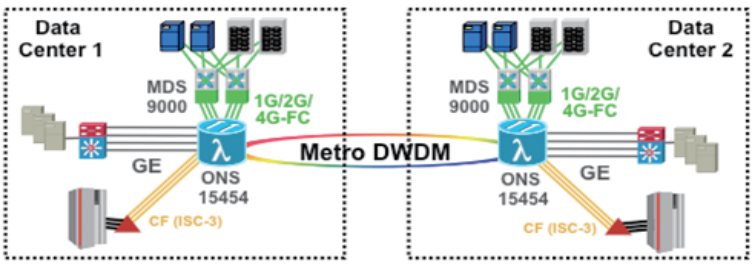


图8

问题的解决和客户的受益

- 行业标准组件：思科是智能光网络的创始者和积极推动者。思科全面主导相关标准化组织（如IETF,OIF,ITU-T等）关于智能光网络标准的起草和制定。思科DWDM是经过 MEF、IBM、NEBS、Telcordia 和其他机构认证的服务与平台；
- 设备稳定可靠，广泛应用于金融业和电信业光传输网络： 思科的光传输产品运行稳定可靠，具有电信级系统可靠性，可靠性高达99.999%，网络恢复时间低于50ms，因此众多金融行业客户、运营商、大企业选择作为光传输网络平台；
- 多业务功能：思科DWDM设备可以有效地汇聚和集中IP、以太网、语音和存储服务，并且在整个城域光传输网络中进行优化传输；
- 集成化的、灵活的光传输网络：同一个机箱可以支持多种光传输速度，并且可以在不影响正常业务的情况下进行系统升级，以满足不断增长的网络需求；
- 超低时延：满足同城灾备中存储系统在同步数据复制方面的苛刻要求；
- 自动化网络设计：无需人工计算光纤衰减、色散和光功率级别等单调而容易出错的数据；
- 自动发现网络服务：提供准确的线路卡和服务视图，包括自动配置、安装和维护；
- 智能化配置和管理：动态灵活地实现通道业务的创建和变更、通道均衡、通道监测， 操作和维护全部图形化，减少运维成本。

3.5 采用思科OTV技术助力双活数据中心部署

当金融机构建设同城双活数据中心时，可以通过在两个数据中心之间部署OTV 二层通道来灵活构建远程应用集群系统所需的内部互联子网和对外服务的VIP子网（图9）。

除此之外，OTV技术也可以支撑双活数据中心之间计算资源的移动，可实现在三层网络中逻辑分割的多个物理分区之间进行无中断地计算移动，组成云计算资源池（图10）。

在主数据中心利用OTV技术将位于两个不同的三层分隔的业务分区A1和分区A2中实现某个VLAN子网的连通。或者在主中心和同城灾备中心之间实现某个VLAN子网的连通，支撑例如VMWARE的虚拟机的迁移或者二层网络中资源池的调度。

图9

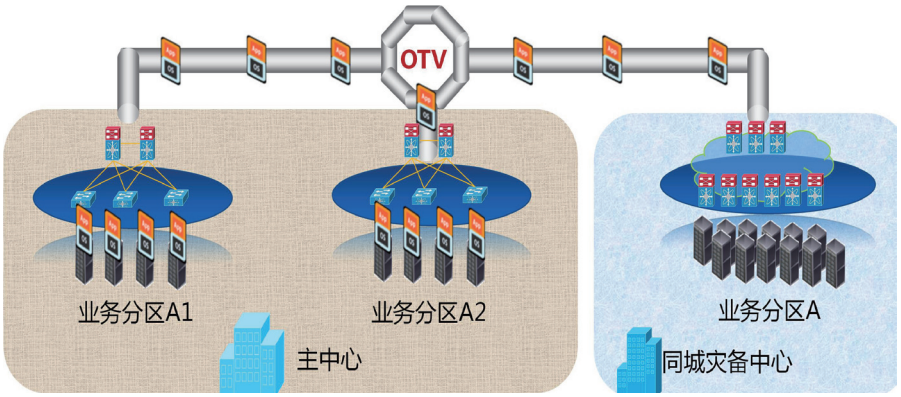


图10

客户面临的主要问题和挑战

- 维护站点的独立性：在多个数据中心之间扩展第 2 层域时会导致协议和故障扩散，即各数据中心通过 IP 网络互联时，那些通常会被隔离的协议和故障会蔓延至其他数据中心。这些故障会在开放的第 2 层泛洪域中自由传播。这就需要一个既能建立第 2 层连接又能限制故障蔓延至泛洪域的解决方案，以遏制故障并保留通过使用多个数据中心获得的弹性；
- 传输独立性：数据中心之间的传输性质会因数据中心的位置和不同区域服务的可用性和成本而有所不同。经济划算的数据中心互联解决方案在传输方式方面不能有局限，而且要能使网络设计人员根据企业和运营偏好，灵活选择数据中心之间的传输方式。支持 IP 的传输是最通用的传输方式，它可以提供灵活性并实现长距离连接。通常认为，可以使用 IP 传输的解决方案灵活性最好；
- 多宿主和端到端环路预防：LAN 扩展技术应提供高度的弹性，因此需要在 VPN 上的第 2 层站点实现多宿主。必须提供一种机制，能在连接多宿主桥接网络时防止环路产生；
- 兼具复制、负载均衡和路径分集的带宽利用率：跨数据中心扩展第 2 层域时，必须对数据中心之间可用带宽的使用进行优化，以最低的成本获得最佳连接。在数据中心和传输网络之间建立弹性连接的同时平衡所有可用路径上的负载，需要提供比传统以太网交换和第 2 层 VPN 更高的智能。应该对组播和广播流量进行最佳复制，以减少带宽消耗；
- 可扩展性和拓扑独立性：由于数据中心中部署了 LAN 扩展，因此提供不影响网络设计、能够在拓扑中任意节点进行部署的解决方案至关重要。这种灵活性通常需要 LAN 扩展解决方案具有较高的可扩展性，能够随着数据中心接入功能的增加以及边缘设备数量的增加而扩展；
- VLAN 和 MAC 地址可扩展性：数据中心之间的 LAN 扩展需要多个 VLAN 同时扩展。此外，在某些应用中，会出现重复的 VLAN ID，它们即使在同一个 LAN 扩展中，也必须分别传送。因为站点是互联的，所涉及的 MAC 地址数量也将增加，而 MAC 地址空间不能汇总；如果不能正确处理，这会导致问题发生，并会限制解决方案的覆盖范围；
- 复杂运营：第 2 层 VPN 可在数据中心之间扩展第 2 层连接。但是，通常会涉及复杂协议、分布式调配和运营密集型分层扩展模式。包含内置功能和点到云调配的简单重叠协议对降低建立该连接的成本至关重要。

思科的解决方案：

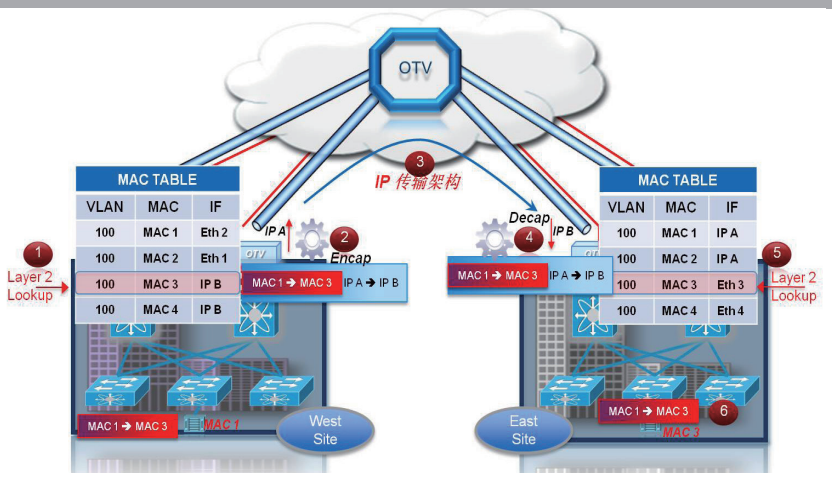
OTV是一种“将MAC地址封装在IP包里”的技术，支持2层VLAN在任何传输链路上的扩展。传输链路可以是基于包交换的、基于标签交换的或者任何支持IP数据包的其他类型的协议。通过使用MAC进行路由的原则，OTV提供了基于2层链路的覆盖链接，同时保证了两个互联的2层域是隔离的，也就是保证了两个互联的数据中心2层域独立从而确保了故障域的有效隔离、灵活自如的扩展和负责均衡。

OTV的核心工作原理是通过控制协议来通告MAC地址的可达信息（而不是通过数据平面学习）并且使用IP封装的包交换技术处理并转发2层流量（而非使用电路交换）。这些是OTV区别于其他传统的2层VPN技术的重要特征。传统的2层VPN技术不但严重限制了2层VPN网络的扩展，同时，由于缺少控制平面的管控也制约的LAN在不同数据中心之间的延伸。

OTV采用控制协议定义了MAC地址和经过网络核心可达的下一跳IP地址之间的映射关系。OTV就像是通过MAC进行路由一样：目的地址是MAC地址、下一跳是IP地址，业务流量被封装进了IP包，这样业务流量就可以流过IP核心网络，通过MAC路由而到达下一跳。因此，介于源主机MAC地址和目的主机MAC地址之间的流量通过覆盖封装就转变成了相关边界设备之间IP源地址和IP目的地址的IP流量。这些数据采用封装处理的方式进行传输而不是通过隧道的方式进行传输，这

是因为封装是可以动态变化的。由于这些业务流量借助IP转发，因此OTV在IP核心网里就变得十分高效，不但可以优化负载均衡流量、复制组播流量，而且还能够快速地实现故障恢复。下图清晰地描述了OTV动态封装机制（图11）。

数据转发表里的MAC地址与下一跳IP地址之间的映射关系由控制协议通告，故而消除了未知的单播数据包必须穿越数据中心间互联链路的要求。控制协议具有可扩展性，除了必要的具体MAC地址信息外，还包括VLAN、站点ID和关联IP地址。这些丰富的信息，如果仅靠数据泛洪学习方式其中大部分是无法获得的，但是对于实现多宿主、负载均衡、抑制环路、本地FHRP及本地ARP转发等OTV必备功能来说，却是至关重要的信息（图12）。



MAC1（站点1）与MAC2（站点2）之间通讯　图11

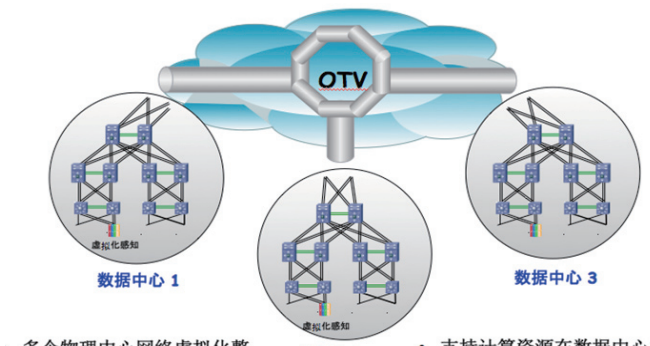


图12

问题的解决和客户的受益

- OTV具有内置过滤功能，可以对最常见的链路本地网络协议（生成树协议、VLAN中继协议、HSRP等）进行本地化，并防止它们穿过互联区域。此功能可防止协议故障在不同数据中心站点之间传播。第一跳弹性协议（HSRP、VRRP等）的本地化既能隔离故障，又能帮助确保最佳路由；
- OTV 的覆盖性质使它能够在任何传输中运行，唯一前提是此传输链路可以转发IP包。满足不同用户使用多种传输链路进行数据中心互联的要求；
- 作为OTV控制协议的一部分，多宿主自动配置和检测也包含在内。此功能无需其他配置或协议，便可实现站点的多宿主。提高了数据中心互联的可靠性；
- OTV可对跨全主用多宿主部署中的多个边缘设备的流量执行有效的负载均衡。负载均衡基于OTV控制协议提供的信息遵循等价多路径(ECMP)规则。从而最大程度地利用了数据中心之间宝贵的链路资源；
- OTV 可扩展密度相对较大的边缘设备且可支持数量较多的服务器资源，为了发挥该功能的最大效果，在数据中心内部选择什么位置部署该功能就变得至关重要。数据中心网络的汇聚层是部署所需边缘设备的一个便利位置。这种部署可在必要时将现有的第 2 层域直接放入 OTV 中，从而简化网络设计和运营。在聚合层定位边缘设备需要使用一个能够支持大量边缘设备的解决方案，而 OTV 提供了所需的可扩展性；
- OTV 提供了一个协议来满足 LAN 扩展的不同要求。OTV 提供的自动发现机制内置于单个协议中，新站点或者新服务器的加入，不会对现有站点产生任何影响。

3.6 FabricPATH交换矩阵设计

客户面临的主要问题和挑战

- 随着商业银行业务的不断发展，新的应用不断增加，原有网络数据交换能力不能适应新一代系统数据处理要求；
- 因传统的竖井分区限制，网络安全区域划分不够合理和细致，网络基础架构不够弹性；
- 网络如何实现对业务的灵活支持，减少服务器系统部门和网络部门之间的协调工作量；
- 服务器虚拟化环境下的网络部署、安全问题。服务器虚拟化的部署，需要网络为虚拟机资源迁移提供最大的灵活性；
- 传统数据中心网络令人头痛的二层物理环路/二层故障问题。

思科的解决方案：

采用思科FabricPATH技术组网，Spine节点采用Nexus7000系列，Leaf节点采用Nexus5000系列，下联Nexus2000系列交换机。消除生成树协议，收敛快速，构建非常稳定的二层交换网络的拓扑，矩阵内支持VLAN内主机任意物理位置的部署和迁移，配置非常简单，真正实现即插即用（图13）。

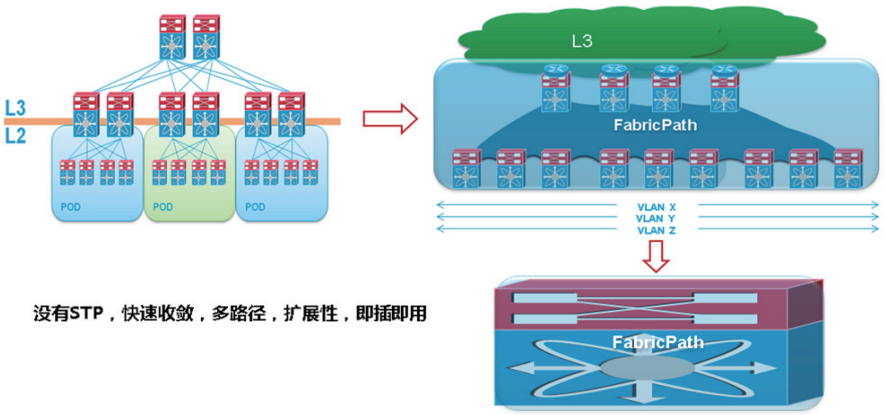


图13

问题的解决和客户的受益

- 网络架构简洁。整个交换矩阵系统就像一台单一的交换机，相对上层用户系统，网络具备一定的独立性；
- 网络运行高效： 减少了交换机数量，提供了更高的端口密度和更低的收敛比，支持高效访问和使用资源；
- 具有高可扩展性： 网络规模非常容易扩展；
- 服务器部署便捷：理论上没有服务器部署的物理位置限制，可实现快速弹性的部署和移动性；
- 网络运维简单： 交换矩阵能够在线升级和重新配置，维护和排障如同管理三层路由网络一样方便。

新一代智能数据中心

附录：成功案例（部分）

6.1 利用网络虚拟化技术（VDC，VPC和FEX等）构建新型数据中心

Nexus7000高端交换机自2008年问世以来，全球已经超过8000多个客户部署，销售超过40000多台设备，而以Nexus5000为代表的固定端口的数据中心级交换机更是超过15000多个客户部署。国内金融行业客户为中国工商银行、中国农业银行、中国建设银行、中国银行、中国交通银行、中国邮政储蓄银行、人行清算中心、华夏银行、招商银行、南京银行、大连银行、晋商银行、锦州银行、铁岭银行、丹东银行、赣州银行、郑州银行、江苏银行、上海银行、青海银行、齐鲁银行、湖北农信、上海农商银行、江阴农商行、云南省农信社、无锡农商行、江苏省农信社、东莞农商银行、成都农商行、国泰君安、上海期货交易所、大连商品交易所、中国金融期货交易所、东吴证券、永安期货、广发证券、西南证券、深圳证券通信有限公司、中国人寿、上海农商银行、国泰君安证券、民生银行、泰康人寿、光大银行、恒丰银行、兴业银行、渤海银行、长江证券、广发银行等。

6.2 存储融合的统一网络架构

中国交通银行、中国建设银行总行、中国农业银行总行、中行江苏分行 、兴业银行、广发银行、人行清算中心、福建海峡银行、徽商银行、郑州银行、国泰君安证券、中宏保险、中国人民保险集团股份有限公司、中国银联数据中心、汉口银行等。

6.3 数据中心统一计算平台建设

全球有超过32000个用户部署UCS系统，位列全球X86刀片服务器市场的第二名。国内金融机构有国家开发银行、人行清算中心、山东省城市商业银行合作联盟有限公司、华融湘江银行、深圳证券通信有限公司、广发证券、中国证券登记结算有限公司、中宏人寿保险公司、中国太平保险集团、长沙银行等。

6.4 商业银行灾备DWDM方案

中国银行总行、中行广东分行、建设银行南数据中心、中行浙江分行、中行湖南分行 、中行江西分行、包商银行、江苏银行、成都银行、重庆银行、贵阳银行、南京银行、江苏农信省联社、广州农商银行、东莞农商银行、四川农信、上海证券交易所、中国证券登记结算公司、上海农商银行、上海期货交易所、中国金融期货交易所、太平洋保险、交通银行、浦发银行、华融湘江银行、郑州银行、人行清算中心等。

6.5 采用思科OTV技术助力双活数据中心部署

全球有800多客户成功部署，包括著名的大型金融机构，如富国银行、苏格兰皇家银行、瑞信银行、美国证券交易所等，国内包括中国工商银行、中国农业银行总行、中国交通银行、人行清算中心、中国银行、华夏银行、招商银行、南京银行、大连银行、锦州银行、铁岭银行、丹东银行、赣州银行、郑州银行、江苏银行、青海银行、齐鲁银行、上海农商银行、江阴农商行、云南省农信社、无锡农商行、江苏省农信社、东莞农商银行、成都农商行、国泰君安、上海期货交易所、大连商品交易所、中国金融期货交易所、东吴证券、永安期货、西南证券、深圳证券通信有限公司、中国人寿、湖北农信、广发银行等。

6.6 FabricPATH交换矩阵设计

全球已经有超过2000多用户购买了 13,000多 FabricPath许可证书，其中超过250个用户组建了较大规模的 FabricPath网络。国外著名的金融机构如美国银行(BANK OF AMERICA)、瑞士瑞信银行、德意志联邦银行、GE Money 银行、VISA公司以及国内的 四川农信数据中心、人行清算中心等。


CISCO