思科 NFVI 和编排解决方案使能（日本）乐天云平台

思科杰出工程师 Santanu Dasgupta

最近可能大家注意到了一个激动人心的消息，乐天移动网络公司（此博客中简称为“乐天”）推出了商业友好用户的试用。乐天是日本最新一家移动网络运营商，也是日本的第四家移动运营商。其母公司乐天有限公司是一家在业界享有盛誉的网络和 IT 巨头，拥有例如电子商务、金融科技、富媒体内容和社交媒体应用等多元化的业务。这一点不容忽视，也使得乐天进入电信领域的背景非常不同，与目前业界现有的通信运营商形成了鲜明的对比。乐天有限公司一直在实行内容战略，以适应不断变化的市场动态，并推动非常必要的数字化转型。在乐天发布的新闻稿及其董事长兼首席执行官 Hiroshi（Mickey）Mikitani-san 的博客中，都强调了公司领导层对机遇的看法，提及了新的商业模式和生态圈，并把客户体验作为公司关注的重点。

在技术层面，虽然乐天在无线接入网中首先会采用 LTE-Advanced 技术，但是他们从一开始就为其网络采用了基于 5G 的系统架构。虚拟化和云原生是 5G 系统架构的关键原则，也是乐天技术战略的核心。他们正在利用网络和服务端到端自动化的移动边缘计算，部署从无线接入网（RAN）到核心网的完全虚拟化的网络。对于参与该项目的团队来说，实现无线接入网络的演进是一项令人难以置信的成就，也是 Open vRAN 能够实现多大规模可行性的一个很好的例子。乐天的新闻稿强调了用于托管所有虚拟化应用的通用分布式通信云平台的关键作用。乐天将他们的“通用分布式通讯云”命名为 RCP——即乐天云平台（Rakuten Cloud Platform）的缩写，这个名字最初出现在乐天首席技术官 Tareq Amin 的一篇影响深远的 LinkedIn 帖子中。思科公司从一开始就在这一激动人心的历程中发挥了关键作用，我们对我们的伙伴关系和紧密合作感到无比自豪。RCP 底层由思科的 NFVI 和编排技术提供支持，本博客的目的是分享有关 RCP 设计原则、架构和相关技术的更多详细信息。

我们将说明 RCP 如何涵盖和实现以下术语所体现的设计原则——通信云、通用、分布式和端到端自动化。

通信云——大多数情况下基本上是用于实现网络功能虚拟化（NFV）的私有云，允许将通信应用大规模部署为虚拟网络功能（VNF）。例如乐天的此类通信应用包括：来自多个厂商的 vRAN（vDU 和 vCU）、vMME、vSAEGW、vIMS、vPCRF、vHSS、vDRA、vFW、vCGNAT、vCDN 等。这些网络功能种类繁多，对云平台的要求与传统 IT 应用的差别很大，因此需要专门为此设计的平台。

通信云从架构的角度可以看作是两个主要构建模块的组合：NFVI（NFV 基础设施）和 MANO（管理和编排）。这可以很好地映射到 ETSI NFV 参考架构框架。

在 NFVI 硬件方面，RCP 使用标准的 COTS x86 服务器来实现计算和存储功能。英特尔通过其 Xeon-SP CPU、网卡和 SSD 存储实现了大部分功能，并通过 FPGA 实现 vRAN 工作负载的硬件加速。详情可参见英特尔发布的新闻稿。思科以应用为中心的基础设施（ACI）和 Nexus 9000 系列交换机构成了物理网络交换矩阵，用于互联所有的计算和存储资源，并将它们连接至外部网络。思科 ACI 具备 T 比特级容量、自动化、服务链、遥测和丰富的策略模型，真正为 RCP 提供了 5G 就绪的通信数据中心交换矩阵。

思科虚拟化基础设施管理器（CVIM）是一个完全容器化的 NFVI 软件解决方案，用于创建构成 RCP 核心的云操作环境。CVIM 利用基于 Linux / KVM 的基础设施实现虚拟计算; Ceph 用于虚拟存储; Open vSwitch、fd.io 和 SR-IOV 用于高速网络；docker 及 OpenStack（很快将用 OpenStack 和 Kubernetes 的组合）用于基础设施组件管理和云编排的执行。CVIM 的关键原则是：开放式架构和利用开源技术;因此它嵌入了红帽企业 Linux 发行版（RHEL）和 OpenStack 平台（RHOSP）。大概在 5 年以前，我们决定不再构建自己的 OpenStack 发行版，而是将我们的工程重点放在 OpenStack 构建自动化以及使其更加易于部署和操作、更加安全的工具上。这一策略使我们能够在 CVIM 中开发出许多有价值的功能，包括：完全自动化、零接触的安装配置工具，可以在几个小时内实现从最初的裸机到交付可操作的云;我们开发和打包的丰富的操作工具集;全面的安全加固;完全自动化和支持 CI/CD 的生命周期管理功能，可以进行无间断的维护（包括在线更新和升级）；广泛的性能增强和调优，包括增加了硬实时功能（如果缺少此功能则虚拟 RAN 系统无法工作）；第三方产品支持以及完善的系统级预集成和验证。这些独特的特性为任何准备构建一流通讯云的运营商带来了巨大的价值，并使 CVIM 成为 RCP 明确的选择。

乐天云平台的管理和编排（MANO）层使用带有 NFVO 功能包的思科网络服务编排器（NSO）作为 NFV 编排器，使用思科弹性服务控制器（ESC）作为 VNF 管理器。思科 NSO 是行业领先的模型驱动编排器，通过中间抽象层将服务和设备模型分离，为众多的物理设备和虚拟网络功能提供非常强大的多厂商支持。NSO 在北向和南向都支持多种接口，RCP 选择了使用 ETSI SOL003 接口进行 VNF-M 集成。思科 ESC 将作为所有思科和大多数第三方工作负载的唯一通用 VNF 管理器（G-VNFM），提供丰富的 VNF 生命周期管理功能。在当前的初始阶段还有一个合作伙伴使用自己的 VNF-M，通过基于 SOL003 的接口与 NSO 集成。不过它最终也将迁移到 ESC，从而实现乐天在 RCP 中使用单个 VNFM 和单个 NFVO 的清晰愿景。

下面的图 1 描述了构建乐天云平台基础的通讯云架构：

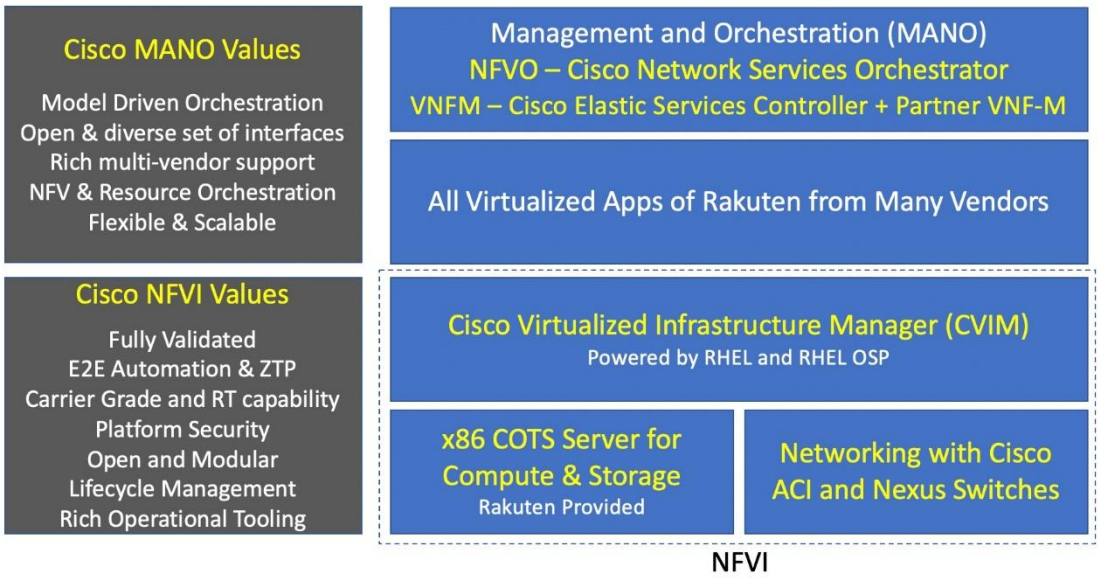


图 1

RCP 在初始阶段作为通信私有云，将支持乐天的所有虚拟网络功能。后续 RCP 将增加对 IT 应用的支持，并扩大其覆盖范围，从而在将来成为混合云平台。

通用——如果我们回溯大约 20 年，当时的运营商有多个网络来支持不同的业务。这些网络采用不同的技术构建，例如 TDM、SONET/SDH、X.25、ATM、帧中继和 IP。运营商最终意识到并行运行多个网络在 TCO 方面不是很经济，从而带来了我们在本世纪前十年所见证的 IP 融合承载趋势。

今天随着各种网络功能的虚拟化，总的来说运营商不想再重复同样的错误：他们不想再构建多个并行的云平台，每个云平台只托管一个或者一部分的目标应用。这些运营商的目标是构建一个通用的云平台，用于托管其网络中涉及的所有厂商的所有虚拟网络功能。这将降低成本并简化操作，同时提供可重用的基础设施。乐天从一开始就决定为所有应用提供通用的云平台，从而朝这个方向迈出了勇敢的一步。当 RCP 上的应用提供商出现兼容性问题时，乐天会促进各方之间的协作，以确保实现所需的结果。

分布式——这是 RCP 非常重要的一个方面。运营商网络需要在网络中不同的“位置”部署各种网络功能。这些位置包括以下网络站点：中央 DC、区域 DC、汇聚/端局、预汇聚/C-RAN 集中点以及接入/基站站点。

如果我们查看典型的运营商中央 DC 和/或区域 DC，我们经常会看到有 IMS 核心网、MME、PCRF、HSS 和 DRA 等网络功能。用于消费者数据服务的分组核心网关（例如 SAEGW）通常也位于这些中央 DC 和/或区域 DC。不过，随着移动边缘计算的出现和 SAEGW 的解耦（控制和用户平面分离），现在可以进一步在网络中的汇聚或者预汇聚位置（更接近消费者的位置）部署 SAEGW 用户平面（SAEGW-U）。这样部署还允许运营商将 SAEGW 用户平面与边缘计算环境中的内容和应用共存，从而更好地支持低延迟通信和/或边缘卸载。

在网络中进一步向下就到了基站站点，即传统的整体式 eNodeB 的部署位置。乐天本着其核心理念和颠覆式天性，正作为首家部署开放虚拟化 RAN 的公司快速进入市场——他们根据解耦的理

念，将无线接入网络层分解成了多个组件。我们与乐天和 Altiostar Networks 共同撰写的白皮书对这种解耦进行了更为详细的讨论。乐天进行解耦的结果是基站精简到只包含远程无线头端（Remote Radio Head，RRH）和天线。这些基站通过裸光纤连接到它们各自的预汇聚站点（在乐天也称为 GC 站点），这些站点上的虚拟化分布式单元（vDU）处理无线协议栈下层功能，虚拟化集中式单元（vCU）处理无线协议栈上层功能。

这种从无线接入网到核心网的完全虚拟化网络（包括移动边缘计算），意味着需要在网络中的不同位置部署各种虚拟网络功能和/或应用。因此，所有这些位置类型都需要存在相同的、可部署虚拟网络功能和/或应用的一致通信云平台。就乐天而言，我们分了三种类型的位置：

1. 中央数据中心——将有 2 个，它们将托管关键应用，例如 vSAEGW-C、vMME、vIMS、vPCRF、vHSS、vOSS、vBSS、虚拟化 Gi 服务以及 MANO 功能等。
2. 区域/汇聚位置的边缘数据中心——大约 50 个或者略少，它们将托管 SAEGW-U、vFW / NAT、vCDN 等关键应用，并将作为乐天的移动边缘计算中心。
3. 预汇聚位置的远端数据中心——将有数千个，它们将托管 vDU 和 vCU 应用，以执行基带/无线处理（虚拟化 RAN）。

下面的图 2 总结了各种位置类型，每种位置类型的关键应用以及规模。

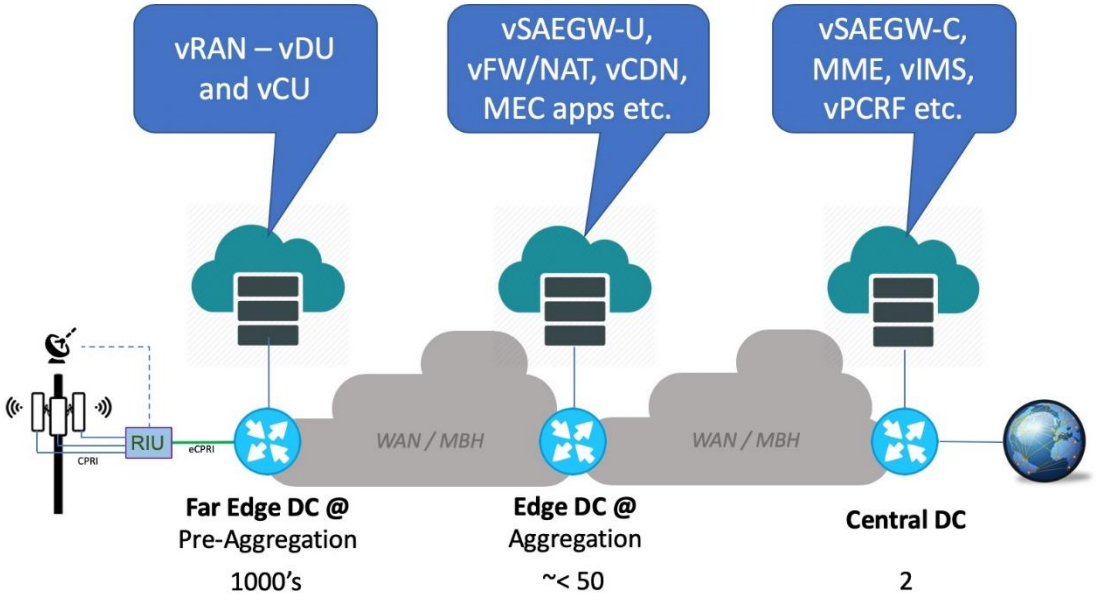


图 2

对于虚拟化而言，这听起来可能是非常大规模的部署，事实上也确实如此。整个部署完成后，将有数千个位置、数万台服务器、超过 100 万个虚拟 CPU、来自所有服务器的接近十万个 25GE 端口，以及几十 PB 级别的存储（虽然大多数网络应用不是存储密集型的）。因此，乐天云平台作为分布式通信云可以看作是网络中数千个位置的各种规模 NFVI 的组合；所有这些都通过集中式的策略、管理和自动化框架整合在一起。下面的图 3 说明了这一点：

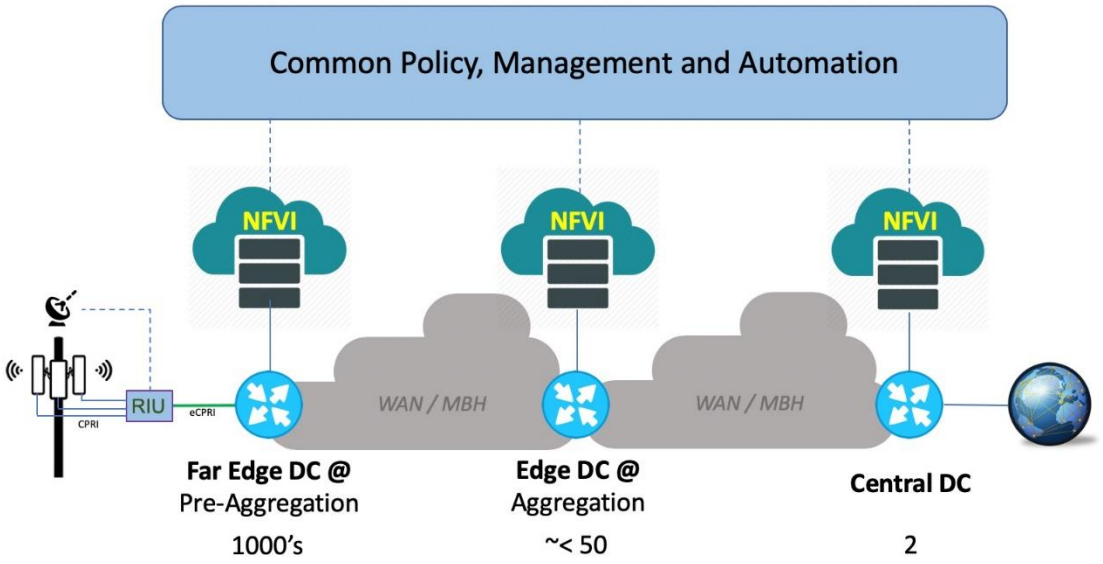


图 3

正如我们之前所看到的，这三种不同位置类型的关键应用各不相同。每种位置类型的扩展性要求也不相同，这取决于每个应用需要多少个实例、每个实例的大小等等。最后，这三种位置类型有不同的物理限制——中央 DC 具有正常的数据中心设施，而远端在空间可用性、每机架功率限制和散热能力、允许的最大设备深度等方面都有很大的局限性。要满足这些要求，需要针对每种位置类型进行不同的 NFVI POD 设计，也需要 CVIM 提供一系列的选项来实现最大的灵活性。

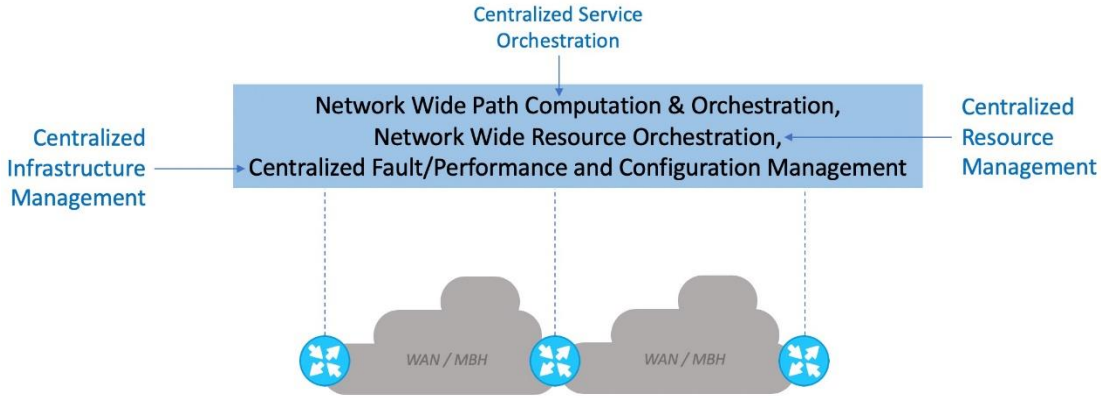
RCP 中央 DC 的每个位置需要 100 多个机架、1000 多台服务器——因此该位置类型的 NFVI POD 设计针对高容量和扩展性进行了优化。为确保操作简单性、易于横向扩展和基础设施可重用性，该位置的所有应用都使用通用的服务器硬件配置（SKU）。在存储方面，根据中央 DC 的应用需求，NFVI 设计中还包括在同一 POD 中同时采用机械硬盘（HDD）和固态硬盘（SSD）的多后端 Ceph 存储。这使得需要高存储 I/O 的应用能够利用基于固态硬盘的后端，而其他应用则利用更经济的基于机械硬盘的后端，从而实现成本和性能之间的良好平衡。

与中央 DC 相比，RCP 的边缘 DC 位置部署服务器机架的规模小得多。因此，这些位置的 NFVI POD 设计需要在占用空间优化、性能和扩展性之间进行平衡。此位置类型也应用了相同的通用 SKU 和存储设计理念。

RCP 的远端 DC 位置则在许多方面都不尽相同。之前已经提到过这种位置在空间、电源、散热和机架深度方面存在限制，不过这种位置类型的占用空间要求也是最低的——因为目标应用只是虚拟化 RAN 的 vDU 和 vCU，每个位置只需要几台服务器。由于 RCP 中将有数千个远端 DC，因此在这些位置推动占用空间优化、最大限度地减少运行多功能云相关的所有开销，是实现合适的成本点的基本要求。为了最大限度地减少开销，业界已经有很多的辩论和争议。有人提出过一种部署模型，包含有集中式的控制和管理平面，在远端站点（在 RCP 中即远端 DC）则只有计算节点，换句话说即是“无头”远端站点模型。对于 RCP 的远端 NFVI 设计来说，可靠性和可预测性至关重要。我们希望即使在远端站点与网络断开连接或者难以与网络的其余部分进行通信的情况下，NFVI 的设计也能够满足这些标准。因此在这里“无头”模型不能成为一种选择，我们采用了不同的方法来设计远端 NFVI POD（本文提出的一个或多个概念包含在一个或多个待批专利申请中）。在这个方法中确立了三个主要原则：

1. 在远端 NFVI POD 中删除存储服务和后端——存储服务在成本、功耗和最终维护方面构成了超小型云平台的巨大开销，因此我们从远端移除了存储后端。这个原则不仅减少了与存储相关的开销（CPU，内存，磁盘），而且还通过删除数千个站点上的大量磁盘驱动器，显著地降低了成本——如果设计目标是每个站点都具有本地存储后端，则需要这些磁盘驱动器。在这个架构中，虚拟 RAN 功能（vDU 和 vCU）的运行时（Runtime）利用了临时存储，而不需要共享存储。对于镜像管理服务，为了确保可靠性和操作简便性，我们在设计中增加了一个通用存储集群，只用几台服务器就可以支持数千个远端 NFVI POD。此通用存储集群也是通过自动化和全面的生命周期管理功能而启用。
2. 在 POD 中的前 3 台服务器上共享控制和计算功能，并使用严格的 CPU 栅栏功能限制控制器所占用的资源——相当长一段时间以来，CVIM 已经支持使用同一服务器实现控制和计算功能的部署模型。在实现这一目标的同时，我们保持了 HA、性能、工具、安全性等相同的功能集。这个模型如果实现得不好，会带来工作负载和控制功能争夺 CPU 资源的风险，我们通过为控制功能分配特定的 CPU 来解决这个问题。任何 Linux 系统都需要一些 CPU 用于操作系统（OS）和硬件。我们在 CVIM 中确保限制操作系统运行于每个 CUP 插槽上的一个内核，以确保在操作系统工作时不会中断工作负载所运行的内核。由于操作系统必须在某处运行，因此这是可能的最小开销。我们将控制功能也运行在操作系统所在内核，并在这些内核和系统的其余部分之间划出“硬边界”，避免在相互之间产生影响。实际上，我们在远端 NFVI POD 上实现了零开销运行本地控制平面，从而使得我们能够最大化 vRAN 工作负载的资源。
3. 管理工具的最终集中化——NFVI POD 在今天拥有本地化的软件管理工具，以通过软件和硬件的维护来推动生命周期管理。最终我们的计划是将这些管理功能集中化成为服务，消除远端 DC 位置的相关开销。

这些关键设计原则确保每个远端 DC NFVI POD 完全自主地拥有自己的控制平面（包括自己的 API 端点）、数据平面和管理平面——所有这些都是在几乎零开销的情况下实现的，以更低的成本满足了我们的可靠性和可预测性目标。这种模式和多年来运营商建设大规模 IP 网络并向 SDN 演进的方式具有相似之处。请看下面的图 4，图中所示 IP 网络在不同的位置部署了路由器。每一台路由器都完全自主，具有自己的控制平面、数据平面和管理平面，能够确保可预测和可靠的操作。但是通过集中式管理以及今天的 SDN，我们就拥有了能够增强 IP 网络的集中式功能，例如网络服务编排（如全局流量工程）、全局资源管理和分布式基础设施管理等。



All routers are autonomous with their own control, data and management plane

图 4

应用与上述相同的原则，我们提出了为乐天云平台提供支持的分布式通信云解决方案，如下面的图 5 所示。我们可以看到，每个位置的每个 NFVI 都是完全自主的，具有自己的控制、数据和管理平面。同时它也通过集中式的服务编排、资源管理和基础设施管理（故障、性能、日志、操作工具）等得到增强，因此该管理框架可以在高度分布式的环境中进行扩展。

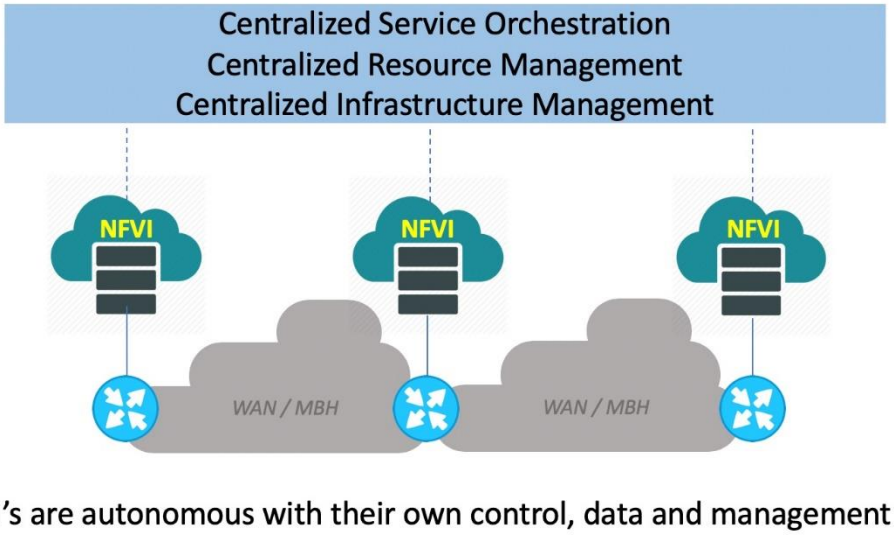


图 5

关于 RCP 远端 DC NFVI 的一个重要方面是 vRAN 应用尤其是 vDU 对性能的严格要求。虚拟化 DU 处理下层无线协议栈，并在虚拟化环境中处理数字化射频信号，要求具有非常高的吞吐和极低的延迟（RAN 前传）。作为类比，想想在虚拟化环境中运行类似 TDM 的应用——那就是 vDU！

从无线的空中接口到 vDU 应用的端到端延迟要求是 250 微秒。在虚拟化环境中，vDU 应用程序的 Layer 1 BBU 进程运行在虚拟机操作系统用户空间，此端到端延迟要求意味着将数据包从网卡移动到此用户空间的延迟最多只能有几十微妙。在高 I/O 吞吐的环境中，对每个数据包都必须完全确保这种低延迟，并且不产生丢包。最重要的是，这一切都必须通过满负载运行的服务器（若干 vDU 占用了所有可用的 CPU 资源）来完成。这需要端到端的硬实时系统，而我们已经在 CVIM 上实现！这是对不能保证此功能的标准 Linux 和 OpenStack/KVM 环境的重大改进。我们还发现虚拟 DU 需要有基于 FPGA 的硬件加速，因为这样可以从通用 CPU 卸载某些计算密集型任务（例如物理层的纠错）。与没有使用 FPGA 相比，使用 FPGA 可以获得更高的扩展性（4 倍以上）。在满足乐天对 COTS 硬件要求的前提下，我们与合作伙伴合作，通过 PCIe 卡增加了 FPGA。CVIM 提供的功能实现了对 FPGA 固件生命周期的支持，所以当 RCP 从支持 LTE（使用 turbo 码）演进到支持 5G NR（使用 LPDC 编码）时，整个系统（包括 FPGA 固件）可以通过标准的编排流程自动升级。CVIM 提供了对开放硬件平台和接口的完整支持，这意味着解决方案中的软件不会被绑定到任何硬件。

端到端自动化——管理虚拟化环境并不简单；当你需要管理的全虚拟化网络涉及 RCP 这样跨越数千个站点的分布式通信云平台时，管理的扩展性将成为首要的挑战，这需要通过适当的设计、实施和运营实践来处理。乐天、思科和参与该项目的其他合作伙伴都非常关注项目的这一方面，从而为乐天的服务和基础设施实现了全自动的网络。在前面的部分中，我们重点介绍了如何使得集中管理框架能够可扩展地管理高度分布式的 NFVI，下面的图 6 则说明了如何为 RCP 实现服务自动化框架。

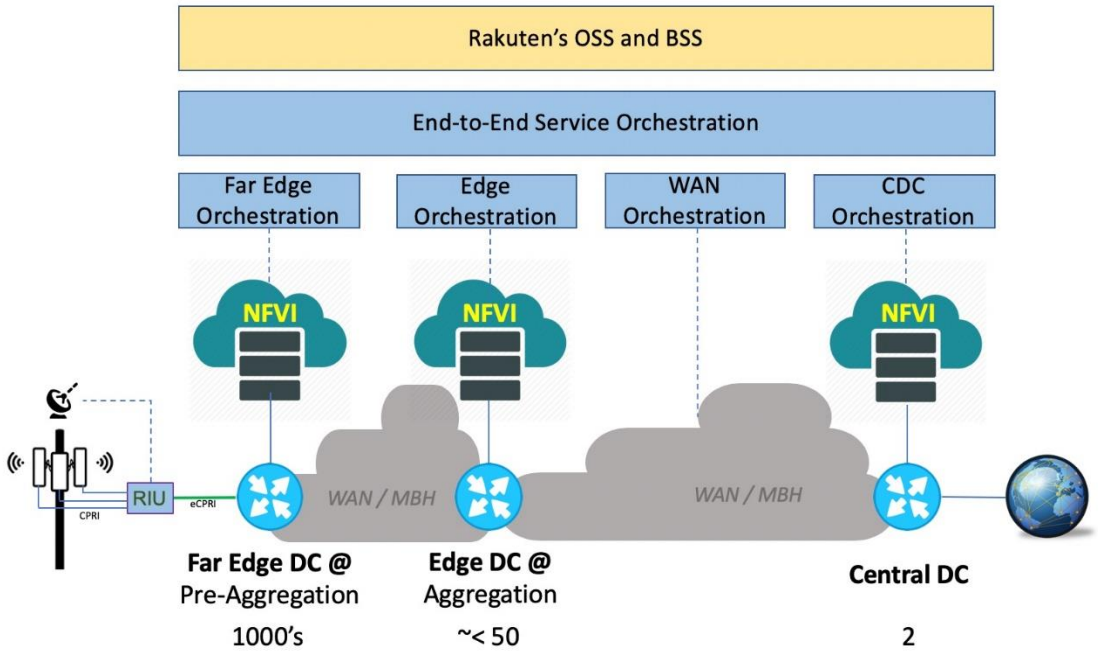


图 6

服务自动化框架基于分层管理和编排的概念，具有四个完全高可用的域级编排系统，可管理中央 DC、WAN、边缘 DC 和远端 DC。域级编排系统由思科 NSO、思科 ESC 和合作伙伴的 VNF-M 构建。这些编排系统在模块化架构框架中整合，由思科 NSO 实现端到端的服务编排。乐天的 OSS/BSS 系统提供了全面的服务生成、服务生命周期管理和操作工作流，服务编排模块则通过北向接口与这些系统交互。

让我们来看一个示例，看看这样的端到端自动化架构在无需人工干预的情况下，如何实现零接触生成 vRAN，生成完全可操作的基站。图 7 显示了零接触部署工作流程的概要，步骤如下：

步骤 1 – 基站的 RIU 启动时向 vRAN EMS 发送通知
步骤 2 – EMS 向 OSS 通知与基站站点相关联的 RIU ID
步骤 3 – OSS 通过 API 向 NSO（E2E）发送通知，请求其激活服务（新站点）
步骤 4 – NSO（E2E）通过 API 通知 NSO（远端）提供 vRAN VNF
步骤 5 – NSO（远端）调用到 ESC 的 API，以在远端 DC 部署 vDU 和 vCU
步骤 6 – ESC 调用目标远端 DC 的 CVIM API，以生成 vDU 和 vCU VNF
步骤 7 – CVIM 生成 vDU 和 vCU
步骤 8 – 生成之后，vCU 查询 EMS 以请求 RAN 配置
步骤 9 – EMS 查询 OSS，获得 RAN 配置和所有的相关参数
步骤 10 – OSS 向 EMS 发送新基站站点的 RAN 配置

步骤 11 – vCU 从 EMS 接收配置
步骤 12 – vDU 从 vCU 接收其配置
步骤 13 – RIU 从 vDU 接收其配置
步骤 14 – 激活所有扇区，基站站点开始运行

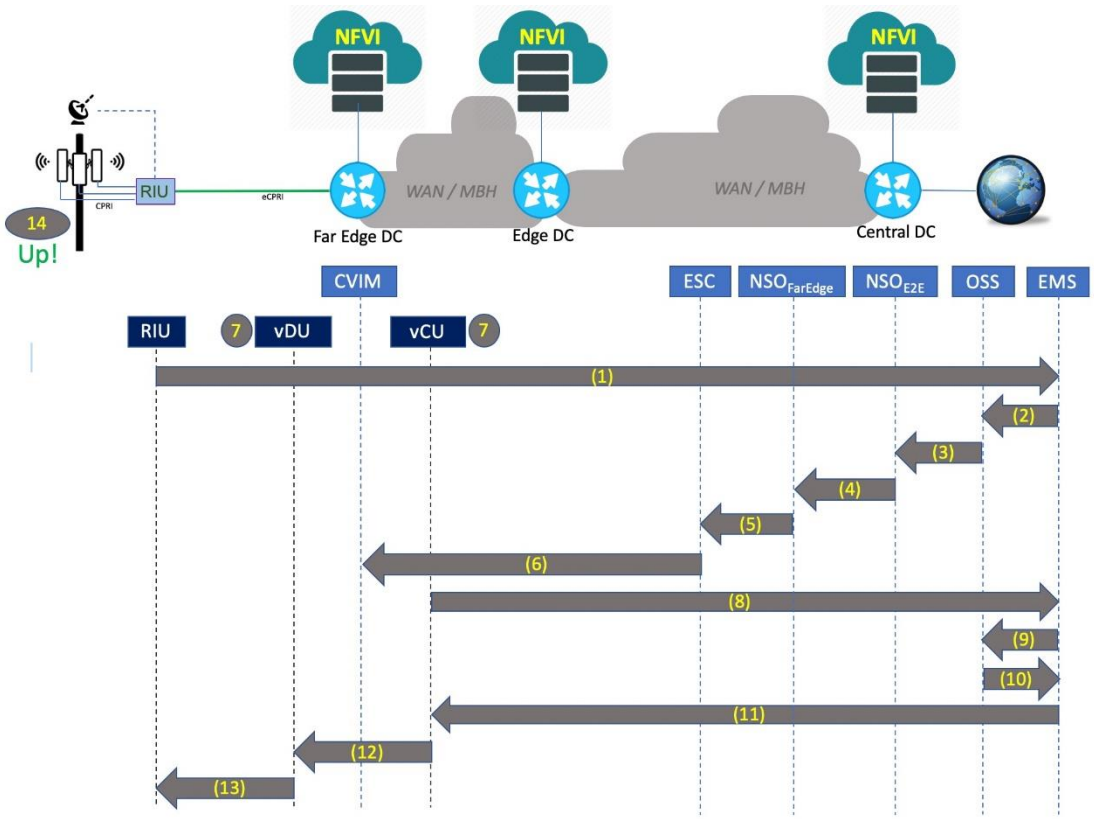


图 7

这是乐天云平台端到端自动化解决方案实现高度差异化功能的一个示例，只需要几分钟即可完成此操作，几乎不需要人工参与。而在典型的移动网络中，相同的操作可能需要数小时到数天，并涉及大量的手动操作，从而增加了运营商的 TCO。

我们希望这篇博文能够为您提供一个有用的视角，帮助了解思科公司和我们的合作伙伴如何在共同创造模式下与乐天密切合作，将他们的愿景转变为业界第一个由端到端自动化的通用分布式通信云 RCP 支持的全虚拟化移动网络，而 RCP 的底层则由思科 NFVI 和编排解决方案提供支持。