The Case for a Virtualized CDN (vCDN) for Delivering Operator OTT Video
# Table of Contents

Overview 1

Introduction – CDNs and the Evolution of Network Infrastructure for Video Services 1

Emerging Challenges in High-audience Video Delivery 3
  - Capacity Planning – Anticipating Demand 3
  - Agility – Responding to Demand 3
  - Cost – Consolidating Delivery 3
  - Quality – Optimizing Delivery 3

The Evolution of Virtualized CDN Caching 4
  - Step 1 – Physical CDN 5
  - Step 2 – Virtualized CDN 6
  - Step 3 – Elastic vCDN 6
  - Step 4 – Orchestrated vCDN 7
  - Step 5 – Deep Caching vCDN 8

Key Pillars of an Orchestrated Elastic vCDN 9
  - Virtualized Infrastructure 9
  - Policy-driven Service Orchestration 9
  - Software-defined Networking (SDN) Automation 9
  - Deep Analytics 10

Akamai’s Solution – Virtualized Licensed CDN (vLCDN) 10

Looking Forward 11
Overview

Akamai’s Licensed CDN (LCDN) is a cloud-native solution that is seamlessly deployable on various physical infrastructures, virtual infrastructures, or a combination of the two. When instantiated as a virtualized CDN (vCDN), it can operate across a range of virtualized infrastructures such as public clouds, private clouds, conventionally virtualized data centers, as well as all types of network functions virtualization (NFV) deployments at different operator network vantage points. This paper motivates the case for vCDN by addressing specific challenges faced by operators in delivering OTT video, and how such challenges can be solved through a step-wise path to virtualizing network infrastructure.

Introduction – CDNs and the Evolution of Network Infrastructure for Video Services

The ongoing shift from hardware-based, dedicated network functionality to software-based networking built on virtualized infrastructure appears to be in full swing. Many industry observers believe that Content Delivery Networks (CDNs), which play a critical role in the delivery of high-quality video, provide an excellent use case for deployment within these evolving frameworks.

Public CDN servers, such as those operated by Akamai Technologies or other CDN providers, are typically deployed in operator datacenters, Internet peering and/or colocation facilities, where rack space, power, and network connectivity are available. Network operators also sometimes leverage “operator CDNs,” which feature on-net dedicated capacity or equipment, distinct and separate from public CDNs, used to deliver the operator’s own IP video programming. The nature of operator CDNs and their role in video delivery make them particularly appealing in the context of next-generation network and service infrastructure.

In recent years, several factors have played a role in the growth in consumption of online video:

- Emergence of popular OTT content from Hulu, Netflix, Amazon, YouTube, and others
- Increased popularity of live-streamed events such as sporting events and concerts
- Ever increasing premium resolution video offerings — HD, 4K, 8K, 360, VR, and others
- More connected homes and mobile devices, with faster connection speeds
- Operators embracing delivery of first-screen content via IP rather than traditional QAM
These trends are helping drive the shift to virtualized deployment for CDNs. Accommodating new technology and satisfying customer demand for video, however, are merely subtexts to the larger set of benefits promised with virtualized infrastructure. Among them:

- A more CapEx-friendly use and reuse of available server resources, in which the operator can leverage the same servers for CDN and/or other services and applications when and where needed
- Simplified provisioning, failure management, and other operational tasks, targeted at reducing OpEx
- Reduced time to market for new services or programming, intended to enhance revenues and customer satisfaction

Despite the benefits, however, there are and will be challenges that will only be solved by intelligent blends of technology, vendors, and products. That is the primary focus of this paper.

Akamai Technologies has led the way in developing and collaborating with network operators on orchestrated virtualized (vCDN) solutions that leverage four pillars — virtualized infrastructure, dynamic service orchestration, software-defined network (SDN) automation, and deep analytics. What’s notable about this is not Akamai’s ability to instantiate CDN software functionality on VMs; getting software to work on a virtual machine is relatively straightforward. What’s notable is the elasticity element, i.e., the ability to dynamically and efficiently instantiate, then eventually tear down, that function based on the combination of policy-driven factors such as demand and available processing capacity. This is a much more in-depth and complex operational undertaking made all the more important by the need to deliver large-scale video events at high (1k) or higher (4k, 8k, etc.) resolution, or quickly recover from network or server failure, to cite just a couple of examples. This capability is sometimes referred to more simply as “cloud native,” and it has become a benchmark of sorts for anyone developing products that provide service-enabling network functions.
Emerging Challenges in High-audience Video Delivery

In an increasingly competitive and diverse market, operators must constantly drive down costs, even as they offer subscribers an ever-increasing choice of video programming that must be delivered reliably and with the highest possible picture quality. High-audience events exist for both Video on Demand (VoD) and live linear television. For VoD, binge watching of new episodes (e.g., “House of Cards”) or the on-demand consumption of blockbusters (e.g., “Game of Thrones”), has become a well-established trend. Live linear television, still an important component of the overall service mix, presents challenges as well as opportunities. Even beyond that, subscriber behavior, which may include transitioning back and forth between live, time-shifted, recorded, and on-demand content — sometimes across devices connected to different networks — presents additional challenges. Operators looking to overcome these challenges must focus on four key areas: capacity planning, agility, cost, and quality.

Capacity Planning – Anticipating Demand

It’s not a question as to whether popular events will drive larger traffic volumes, it’s a question of how much. Historically, CDN capacity has been allocated based on variables such as the amount of provisioned bandwidth, the number of viewers in a specific geography, and the aggregated byte volume of delivered content. Additional capacity is pre-provisioned to accommodate spikes in consumption. Still, it’s difficult for operators to anticipate whether event-driven traffic will spike 3x or 10x beyond normal volumes. Over-provisioning leads to wasted resources. Under-provisioning leads to poor quality, and even potentially more damaging, service outages. The challenge, then, is to remove the guesswork when anticipating demand.

Agility – Responding to Demand

Between the ordering process for new servers, the amount of time needed to build and deliver servers, and the effort required to install and configure those servers and software, the process from start to finish in building out CDN capacity can take weeks or even months. Though many high-audience events provide ample preparation time, events that trigger unanticipated and instantly high viewership — for instance, popular sporting events, concerts, the sudden death of a national figure, natural or man-made disasters, or important local or world events — can significantly strain network and CDN resources. If demand for these events exceeds available resources, there simply isn’t enough time to plan for, deploy, configure, and activate additional CDN capacity.

Cost – Consolidating Delivery

As noted previously, it can be wasteful to over-spend on capacity and potentially harmful to under-spend. Additionally, the build-out of separate and distinct physical infrastructure for each type of service can be overly burdensome and costly to network operators. This ties into the agility factor mentioned previously in that it’s not only difficult to time certain events, it’s sometimes difficult to anticipate their requirements. Purpose-built servers, i.e., systems designed specifically to provide functionality for one type of service (e.g., VoD), may perform well in certain use cases but may be limited in other ways (e.g., for live events), including ability to scale. In short, purpose-built solutions are not the answer given the forecasted growth of IP video. Even general-purpose servers, when dedicated to one service, are not as efficient as they could be. The challenge, therefore, will be for operators to, as much as possible, consolidate service delivery to a single, flexible, highly scalable multi-purpose platform.

Quality – Optimizing Delivery

There are any number of reasons why traditional CDN servers, and even some virtualized CDN deployments, are deployed in larger centralized and regional data centers as opposed to very small points of presence (PoPs). Large content libraries used for delivering VoD, for example, are well suited for a more centralized approach. For high-audience events such as popular live linear content or VoD releases, however, it’s more efficient to deliver from CDN resources located in small PoPs deeper in the network. Generally, operators should identify vendors able to help them provide sustained high quality as well as better granularity and flexibility as the growth of IP video of all types continues. Cost-wise and from a software/operational complexity perspective, however, it may be challenging to deploy and manage lots of widely distributed servers. The challenge will be for operators to find a way to cost-effectively deploy CDN resources “closer to the eyeballs” as well as in more centrally located facilities.
The Evolution of Virtualized CDN Caching

Despite the challenges, a key strength for network operators is that they own networks and facilities, and already use them to deliver video as well as other services. From the perspective of OTT-style video delivery moving forward, one key to success will be how they use their existing facilities for CDN-enabled delivery.

Operator networks and facilities are typically structured as a hierarchy, beginning with core network and large centralized data centers, extending to regional aggregation networks and data centers, and finally reaching “last mile” access networks and small PoPs located close to subscribers.

Because larger facilities have more space for equipment and personnel, today they tend to house more IT servers and other elements used to deliver video and other services, while small, low-level PoPs usually house fiber or copper termination equipment, access routers, some switching gear, and not much else. In many of these tiny PoPs, there is barely room for a desk, let alone personnel to manage the equipment that resides there. Because these PoPs can literally operate with the lights turned off, they are sometimes called “lights out PoPs.” Edge locations, such as these lights out PoPs, may hold the key to next-generation video delivery at scale.
To address the operational complexity of leveraging each of these network vantage points, the ultimate solution will be an orchestrated, elastic vCDN. Arriving at this ultimate solution is and will be an evolutionary progression as shown below.

**Step 1 – Physical CDN**

While public CDN servers that deliver content from a wide range of websites and content owners are often deployed at network operator peering locations, private CDN servers, which are dedicated to delivering the operator’s content, are often placed closer to subscribers, depending on the operator’s size and footprint. Either way, while a physical operator CDN delivers video with far better quality and reliability than video delivered directly from the origin, an inherent problem with physical CDN delivery is that it is based on dedicated, physical servers, so does not readily address capacity planning, agility, or cost challenges.
Step 2 – Virtualized CDN

The step beyond physical CDN delivery is virtualized CDN delivery, which helps reduce cost by moving away from dedicated, purpose-built servers towards shared, multi-purpose servers. With the ability of today’s processors to support multiple VMs, a CDN function used for delivering video can run in one VM while other services can operate in different VMs on the same physical server. Because this model lets operators better utilize server resources, a virtualized CDN can be more cost effective than a CDN running on dedicated infrastructure. But while the vCDN addresses cost, it still does not solve challenges related to capacity planning, agility, or quality.

Step 3 – Elastic vCDN

Added to a virtualized CDN, elasticity allows for rapid scaling of CDN caching and bandwidth capacity relative to demand. In other words, an elastic vCDN can better handle unanticipated spikes in viewership. As such, it lets an operator pursue revenue opportunities — e.g., high-audience events ranging from sports to popular VoD title releases — for which one historically needed to overprovision with dedicated CDN capacity weeks or months in advance. Elasticity thereby enables the minimization of total cost of server ownership while supporting peak capacity events.

The key to the true effectiveness of elasticity, however, lies in how resources are expanded and contracted. Today’s elastic vCDN solutions often operate using manual intervention or basic calendaring. Global resource management and the necessary orchestration to properly utilize shared server infrastructure remain relatively nascent. For that reason, today’s elastic vCDNs often are limited to deployment in higher-level data centers. While it solves capacity planning, cost, and, to an extent, agility challenges, simply adding elasticity to a vCDN in this fashion still falls short of the end goal. What is necessary is a policy-driven orchestration solution that can make better use of the shared IT servers comprising the virtualized infrastructure.
Step 4 – Orchestrated vCDN

Orchestration is part of a larger overriding set of management functions, known as Management and Orchestration (MANO). While there are many MANO-like paradigms, in the context of virtualized infrastructures, one can loosely capture them as three components working together: Service Order Management (SOM), Resource Management (RM), and Virtual Infrastructure Manager (VIM).

These three “orchestration” components work together to execute the business request, policy rules, and instantiation mechanisms that result in the corresponding lifecycle operations of services operating on a virtualized infrastructure.

The RM typically interacts with a service-specific Element Manager (EM) that configures the underlying service relative to the allocation and deallocation of service-specific virtual machine instances. This is shown in the functional diagram below.

The net results are that the operator achieves two important goals:

1. **Increased flexibility** in terms of how services, and the compute, storage, and network resources needed for those services, are allocated in a fully automatable manner

2. **Decreased operational costs**, with a virtualized, all-IP delivery infrastructure, better suited to meeting the anytime/anywhere demands of subscribers in a fully automatable manner

Orchestrated vCDNs thereby address the first three challenges identified earlier — capacity planning, agility, and cost — that operators view as barriers to competing more effectively, and expanding and sustaining their subscriber and revenue bases.

Note that the orchestrated vCDN also must consider business factors over and above the elastic CDN. Where the elastic CDN is solely driven by customer requests placed on the CDN, orchestrated virtualized infrastructure is, effectively, a bus onto which vCDN as a function boards and de-boards as necessary, along with other bus “passengers.” For that reason, operators must factor in policies that govern and respond to, among other things:

- **Bandwidth usage and allocation** during peak and non-peak usage
- **Non-CDN technologies** that also impact network utilization and operations
- **Demands for compute, storage, and network resources** from other services/functions
Step 5 – Deep Caching vCDN

With the SDN/NFV network transformation underway, there are various efforts to re-architect data centers, or PoPs, located much closer to viewers. Examples of such transformation efforts include open source and standards projects such as Central Office Re-architected as a Datacenter (CORD), Mobile Edge Computing (MEC), myriad variants led by operators themselves, and initiatives such as the Edge Computing working group part of Facebook’s Telecom Infrastructure Project (TIP). The PoPs might be Central Offices for telco operators, or head-ends for cable operators, either of which can serve as anchor points for fixed and/or mobile broadband offerings. Since these PoPs tend to be unlit, numerous, and widely distributed, lifecycle automation via a policy-driven orchestrated vCDN solution eliminates the complexity of technicians manually assigning CDN resources across a variably provisioned virtual infrastructure deployed across many physical locations.

The net results are that the operator achieves several important goals:

1. **More traffic offload**, particularly through the “middle mile,” or the network that connects centralized data centers to small PoPs located at the network edge. At this level in the network, the unicast CDN effectively has network efficiency akin to multicast, as often it is at this point where multicast fans out into unicast streams anyway.

2. **More scale**, i.e., more capacity, performance, and network throughput available to deliver video at high quality on behalf of high-audience events

3. **Better QoE**, since video is delivered from closer proximity to the end user, high-speed broadband offers plentiful bandwidth, even for ultra-HD, and there is less resource contention for video delivery in a more highly distributed CDN

This further addresses the final two challenges: optimizing cost and quality.
Key Pillars of an Orchestrated Elastic vCDN

Understanding the evolution to an elastic vCDN with deep caching, as well as its implications and benefits, requires a basic overview of its four foundational components: virtualized infrastructure, policy-driven orchestration, software-defined networks, and deep analytics.

Virtualized Infrastructure

Virtualized infrastructure, in contrast to physical infrastructure, refers to a platform in which the required functionality for a given service or application operates in virtual machines (VMs) distributed across a network of servers. Virtualized infrastructure consists of:

- **Distributed servers running x86 hypervisors.** Hypervisors, or virtual machine managers, are programs that let multiple guest virtual machines share a single physical host. Examples of hypervisors include VMware and KVM.

- **Virtual infrastructure managers (VIM),** which expose APIs for assigning VM instances to the appropriate underlying compute resources onto which “virtualized” functions (CDN cache, request router, object storage, analytics, etc.) can be deployed.

Policy-driven Service Orchestration

As discussed previously, orchestration will play a critical role in the ultimate success of implementing an elastic vCDN solution. Orchestration not only provides the dynamic control needed to rapidly assign or remove functional elements to and from the virtualized infrastructure, it also provides the operator the ability to control how functions from different types of services share that virtualized infrastructure. Ideally, service orchestration will be driven by business/operational policies, or a set of rules, that govern why and when virtual resources get allocated for a given service. In the case of high-audience video delivery, for example, the orchestration platform may, based on a pre-provisioned policy rule, allocate enough CDN resources at a given date and time to effectively stream the important football match, but when the score remains close late in the match, and demand begins to exceed pre-allocated resources, a second policy rule based on a given number of HTTP “request-URI” commands may trigger the allocation of additional CDN resources so that the availability and quality of that event remain high. The orchestrator would thus include the following capabilities:

- **Ability to take actions to manage CDN components** in response to business/operational policies, e.g., elastic scaling out/in of CDN components to achieve a QoE objective or to save costs, scheduled downtime/maintenance workflows, etc.

- **Ability to interact downstream** with VIMs (OpenStack, vSphere, etc.), SDN (if separate API), and a service’s element manager

- **Responsible for overseeing** similar aspects of multiple virtualized services

Software-defined Networking (SDN) Automation

Software-defined network (SDN)-based solutions have emerged to offer templated network automation, VLAN-like isolation, and service chaining. Just as virtualized infrastructure decouples service or application-level functions from physical compute and storage resources, at a minimum SDNs decouple network resources (IP address assignment, MAC assignment, etc.) from physical elements, so that virtual network resources can quickly and logically be allocated to specific services or applications where needed. While an SDN is not strictly required solely for an elastic vCDN solution, it certainly simplifies the at-scale deployment of cache instances in increasingly more complex network environments involving other NFV elements.
Deep Analytics

An at-times overlooked aspect of an orchestrated vCDN, perhaps because it is not explicitly identified in some of the popular architectural diagrams, is real-time analytics. This analytic data is from the underlying virtual infrastructure, CDN-related data, as well as data pertaining to other virtualized functions. The combination of analytics can offer sufficiently deep insight to guide when and where to act, which can aid operators to craft appropriate policy-driven orchestration workflows to do traffic engineering, failure mitigation, and dynamic elastic scaling. For example, if a lot of traffic comes from a region where no vCDN caches are deployed, then relative to provision policies the orchestration solution can either scale out existing cache sites or deploy caches closer to the source of the traffic demand.

Akamai’s Solution – Virtualized Licensed CDN (vLCDN)

As the move to orchestrated, elastic, deep vCDNs for high-audience video evolves, Akamai has already established leadership, with a virtualized version of its Aura Licensed CDN (LCDN). Distinct from Akamai’s public CDN, the Akamai Intelligent Platform™, Aura LCDN is licensed software, deployed and managed by network operators primarily to deliver their own content. Note the following:

- In 2012, Aura LCDN was the industry’s first virtualized CDN to be included in virtualization proof of concept testing prior to the ETSI-launched NFV industry specification group. The testing was conducted by the same team that penned the first NFV white paper, which in turn informed the CDN use case described in this paper.
- Since 2013, Aura vLCDN has been deployed in a Tier-1 telecommunications operator’s commercial offering, running on top of VMware ESX. Akamai demonstrated peak network I/O performance to support heavy video traffic workloads per cache.
- In 2014, Akamai demonstrated elasticity for its vLCDN at the SDN World Congress, together with a layer-3, overlay SDN solution on an OpenStack-based platform.
- In 2015, Akamai completed successful proof-of-concept testing of its vLCDN as an instance of an orchestrated vCDN offer with a major European operator and demonstrated that solution at the 2016 Mobile World Congress event in Barcelona, Spain.
- In 2016, Akamai’s Aura vLCDN was incorporated as a Deep vCDN proof of concept of the Central Office Redesigned as a Datacenter (CORD) project run by Open Networking Lab (ON.Lab) and demonstrated at the 2016 Open Network Summit (ONS) and Broadband Forum (BBF).
- As of 2017, Akamai’s Aura vLCDN is in production trial with a Tier 1 operator who is using a commercial orchestration solution together with OpenStack, CEPH, and SDN.

Aura vLCDN’s core service components, HyperCache and Request Router, both operate on servers in virtual machine (VM) environments. In addition to the deployments noted previously, Akamai has demonstrated vLCDN’s readiness with VMware ESXi and KVM hypervisors, vSphere and OpenStack Virtualized Infrastructure Managers (VIM), OpenFlow-based underlay SDN solutions and several overlay SDN solutions, decoupled block storage solutions such as CEPH and iSCSI, and several orchestration solutions.
Looking Forward

As the competition for viewers among network operators continues to heat up, and business and technical challenges arise, scaling delivery of high-audience, high-quality video is becoming, and will continue to become more difficult and more expensive for network operators.

For most operators, it will be suboptimal to deploy dedicated servers for CDN caching in lower level data centers, which are lights-out facilities that are space/power/cooling constrained. Furthermore, the process of bringing up and configuring applications such as CDN caches to support such services, along with potentially complex network interconnection requirements, will not scale given the limited personnel available to perform these operations.

The migration towards virtualized infrastructures has the potential to address these challenges. With the vCDN evolution outlined in this white paper, it will become possible to deploy cache capacity on demand, both centrally and deeper towards the network edge, to meet the needs of peak events.

The net result will be an increased ability to scale out and scale in the CDN resources needed to optimize video delivery, and possibly other content, in a more granular and efficient fashion, with reduced infrastructure costs, and the significant reduction in the time required to do so.

Akamai, with its technical experience and expertise, financial stability, vision, partnerships, and — most importantly — the right solutions, is looking forward to enabling broader adoption of vCDN by network operators to address their growing challenges of delivering high-quality video services.