# KPI Mapping for Virtual Infrastructure Scaling for a Realistic Video Streaming Service Deployment

Rufael Mekuria[+], Michael McGrath[*],Christos Tselios[-]
*Unified Streaming[+], Intel Labs Europe[*]*
*Amsterdam, The Netherlands[+], Leixslip, Ireland [*]*

Dirk Griffioen[+], George Tsolis[-], Shahar Beiser[x]
*Citrix Greece[-], Nokia Bell Labs Israel[x]*
*Patras, Greece[-], Tel Aviv, Israël[x]*

***Abstract* –Current advances in cloud computing architectures and software defined networking are enabling virtualized server and network infrastructures. This allows on-the-go scaling of network and compute resources, when supporting some higher level service. This is useful when deploying a service with varying resource demands in cost effective manner. In such cases scaling infrastructure on-the-go can be of great help to achieve Quality of Experience (QoE) without over-provisioning. In practice, QoE driven scaling of virtualized infrastructure is challenging. Firstly, it requires detailed knowledge of the application specific Key Performance Indicators (KPI's) related to the underlying virtualized compute and network infrastructure. Secondly, it requires a toolset for collecting measurement data related to these KPI's, i.e. telemetry. Thirdly, appropriate analytics are required to transform the telemetry into actionable insights. Finally an application specific strategy and toolset for resource scaling decisions needs to be deployed, i.e. orchestration. In this paper we present a deployment of adaptive infrastructure scaling for a popular large scale video streaming application. We detail the KPI's identified and present the toolsets for telemetry and orchestration. This work paves the way to scale compute and network resources on the fly for large scale video streaming services, reducing current overprovisioning practice.**

***Index Terms – Video Streaming, Virtualization, Key Performance Indicators, Quality of Experience***

## I. Introduction

Video Streaming on Demand (VoD) services are popular on the Internet. Contrary to linear broadcast TV, it gives users more freedom to watch what they want, how they want and when they want. Video traffic is already a major source of traffic on the Internet today, expected to reach up to 80% of all Internet traffic by 2019 [1].

The deployment of video streaming services requires costly provisioning of bandwidth, compute and storage resources in datacenters. Resource allocation is further complicated by the varying resource demands, especially during peak hours, or when specific videos become popular leading to traffic bursts. Due to technical difficulties related to changing the physical infrastructure when a service is operational, resources are often overprovisioned.

In video streaming practice, adaptation instead often happens at the client. For example, in busy periods clients will adapt and switch to lower bandwidth/quality streams. Such techniques called Adaptive Bit-Rate Streaming (ABR) are popular and widely deployed. It enables video streaming systems to operate, but will often lead to lower quality video being received by users in busy periods.

In addition, client side ABR fails to identify root causes of its reception quality. Clients often have a very limited view of the overall network and service. They can't detect the cause of a problem and cannot make the correct decisions all the time. This often results in unstable behavior and quality fluctuations. Advances in virtualized compute and network infrastructure, introduce the possibility of infrastructure adaptation. By scaling the underlying virtualized infrastructure, aforementioned problems related to client-side ABR could be alleviated, while also limiting resource overprovisioning.

In this work, we explore scaling of virtualized network and compute infrastructure for a popular realistic large-scale video streaming service. This extends beyond introducing QoE management in cloud applications, as presented in [2]. We chose the Unified Origin (UO) [3] video streaming platform for this study. UO is a commonly deployed video streaming system worldwide. We present the KPI's of this service when deployed on a virtualized network/compute infrastructure. Further, we detail the tools used for telemetry and orchestration to enable on-the-go scaling. This paves the way for QoE-driven scaling when using the KPI mapping to establish a good quality of user experience.

Section II overviews relevant virtualization and cloud technologies. Section III presents UO and its KPI mapping. Section IV presents tools for telemetry and orchestration.

## II. Compute and Network Virtualization

*Public Clouds* such as the Amazon EC2, Microsoft Azure etc. allow you to run virtual machines in their datacenters spread across the world. In *public clouds* you often pay per time unit that a VM is used, and you can rapidly instantiate new virtual machines and increase resources on existing VM's such as memory and/or processor. Beyond such Infrastructure as a Service, public clouds can also provide ready applications and platforms (AaaS), (PaaS) etc.

*Private Clouds* can be deployed by any entity running a private datacenter. An important open source platform to deploy, manage and run private clouds is OpenStack [4]. OpenStack has projects for compute virtualization (Nova), networking (Neutron) and object storage (Swift) and image loading (Glance). These four projects enable you to run a cloud in your datacenter where you can start, allocate and scale Virtual Machines (VM's), object storage, network resources. Besides the core OpenStack projects, there are a large number of complementary projects. This modular approach allows

efficient private cloud deployment in a relatively straightforward manner.

*Network Virtualization* aims to introduce the principles of hardware virtualization to the underlying network infrastructure. The two dominant technologies to achieve this are Software Defined Networking (SDN) [5] and Network Function Virtualization (NFV) [6]. SDN focuses on control plane and forwarding plane decoupling. In the SDN paradigm, the controller controls the forwarding behavior of all the underlying IP links. With SDN, network resources can be allocated for end-to-end network paths/flows. This can enable the virtualization of network resources, akin to compute virtualization in cloud environments. NFV allows network functions (i.e. NAT, routing, firewall) to become operational on minimalistic VM's running on commodity hardware inside the overall network. This approach facilitates rapid deployment and testing of network functions in real-time.

We expect these technologies to be key building blocks of the next generation of mobile and wired networks.

## III. KPI Mapping for Unified Origin

Unified Origin (UO) [3] is a media streaming software platform for large scale deployment that typically runs inside a web server instance. UO implements adaptive bit-rate streaming (ABR) based on protocols such as MPEG-DASH, Apple HLS etc. The architecture is shown in Fig.1. Streams from a live encoder or stored file (on backend server or the same server) are retrieved by UO. UO then packages this data into media segment and manifest files for each of the protocols. The segments are then returned to clients, possibly through a CDN. The on-the-fly packaging in UO reduces the storage needs and makes content preparation very easy for content owners. Content owners can keep the simple MP4 and fragmented MP4 formats, not having to worry about protocol specific manifests and segments. Deployment of this architecture poses requirements on the datacenter. In Table I we capture them as measurable KPI's. The server side KPI's CPU, memory, cache and Disk I/O, can be monitored and provide an indication for upscaling. The client and player side KPI's are hard to measure in a realistic streaming scenario. Therefore, in experiments with workload generators we will find the relation between client and server side KPI's first. We will use this mapping to tune a scaling strategy based on server side KPI's only, that leads to satisfactory client side KPI's in practice. Fig 2. Illustrates the setup where we use Hammer, a tool for generating video streaming requests. This tool also measures client side KPI's enabling us to empirically optimize the analytics and scaling heuristics based on reported results.

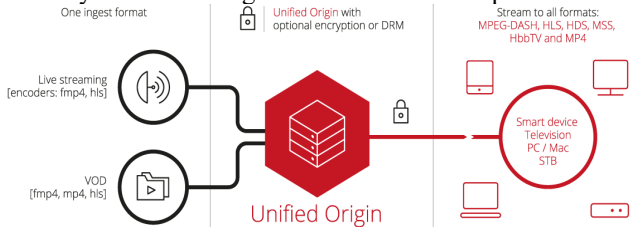


Fig. 1 Unified Origin implementing just-in-time packaging [3]

TABLE I KPI Mapping for Unified Origin

| KPI | Explanation | level |
|---|---|---|
| Throughput per client | Received bytes per client (MB/s) | Client |
| Latency per user request | Delay to first byte received | Client |
| Transaction Failure rate | Number of requests resulting in an error | Client |
| Requests / second | requests handled per second per user | Client |
| CPU usage | CPU at the VM running origin | O.Server |
| Memory usage | RAM usage at origin | O.Server |
| Disk I/O | Disk I/O ops (if files at origin MB/s) | O.Server |
| Cache Usage | Cache usage in the origin MB | O.Server |
| Backend Throughput | Data throughput from backend MB/s | B.Server |
| Requests to Storage | Number of requests to the storage | B.Server |
| QoS Storage - Origin | Bandwidth, delay. packet loss, jitter between storage and origin | Network |
| QoS Origin – Client | Bandwidth, delay. packet loss, jitter between origin and client | Network |

## IV. Telemetry and Orchestration

The purpose of orchestration is to automatically coordinate and manage resources and services running in a datacenter. The initial resource allocation (virtual CPU, network etc.) is often done via a template. For online orchestration, such as scaling, telemetry is needed. In our framework this is based on the SNAP framework [7].

SNAP can measure and report the metrics corresponding to the server side KPI's in Table I. Metrics reported by SNAP can be used by an analytics module such as developed in [8]. For the orchestration, we envision the use of Nokia Cloudband.

This work is supported by the European Union (H2020 RIA, GA No. 671566).

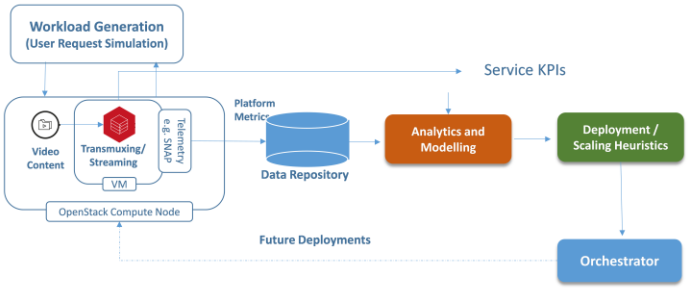


Fig. 2 Infrastructure scaling of UO with telemetry, analytics and orchestrator

## References

[1] CISCO, "Cisco Visual Networking Index," 2015 http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[2] T. Hossfeld, R. Schatz, M. Varela, C. Timmerer, "Challenges of QoE Management for Cloud Applications," IEEE Comm.Mag., April 2012.

[3] Unified Streaming "Unified Origin" ," 5 3 2016. [Online]. Available: http://www.unified-streaming.com/products/unified-origin

[4] OpenStack foundation, "OpenStack," 5 3 2016. [Online]. Available: https://www.openstack.org/.

[5] B. A. A. Nunes et al. "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks" in *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1617-1634,

[6] J. Martins, M. Ahmed, C. Raiciu, V. Olteanu, M. Honda, R. Bifulco, and F. Huici. "ClickOS and the art of network function virtualization" In Proceedings of the *11th USENIX (NSDI'14)*. 459-473.

[7] Intel SNAP https://nickapedia.com/2015/12/02/what-if-collecting-data-center-telemetry-was-a-snap/

[8] T. Metsch, et al. 2015. Apex Lake: A Framework for Enabling Smart Orchestration. In Proceedings 16th International Middleware Conference (Middleware Industry '15). ACM, New York, NY, USA.