SMPTE

# MOTION IMAGING JOURNAL

*Covering Emerging Technologies for the Global Media Community*

# Broadcast Media Creation as a Service: Using Infrastructure-as-Code and the Public Cloud to Power On-Air Media Creation Platforms

**Introdução:**

O artigo a seguir é uma leitura mais que obrigatória para entendermos o futuro à nossa frente, futuro onde a nuvem e suas infinitas possibilidades não param de surpreender-nos (positivamente). Os autores propõem a criação de mídia como um serviço e utilizam, para isso, a combinação de múltiplas máquinas virtuais (VMs) baseadas em nuvem habilitadas para GPU, escaláveis dinamicamente e com armazenamento nativo em nuvem de alto desempenho aliado juntamente com tecnologias disruptivas como IaC (infraestrutura como código) ,que finalmente prometem tornar a criação de mídia em nuvem pública uma opção viável. **Como já disse:** Leitura mais que obrigatória para o futuro que se descortina à nossa frente!

**Tom Jones Moreira**


## Abstract

*Traditionally, the infrastructure required for broadcast media creation has consisted of specialized graphics workstations, high-performance storage arrays, complex networking, and video wiring, all of which are capital-intensive and expensive to maintain. Over the past two years, NBCUniversal has developed a cloud infrastructure model to provide on-air media creation as a service to its creative teams. Utilizing GPU-enabled cloud-based virtual machines, dynamically scalable high-performance cloud-native storage, and infrastructure-as-code methodologies, we are now able to provide full end-to-end media production workflows in the public cloud. With an automation-forward approach, we know exactly what is going to be applied, how it will propagate through the infrastructure, and what dependencies are involved. Elastic scalability of compute and storage eliminates over-provisioning (and its related capital investments) and provides the ability to dynamically add/remove resources as necessary. This paper aims to highlight the technical details of our programmatically reproducible solution as well as the challenges and benefits of media creation in the cloud.*



## Keywords

*Automation, broadcast, cloud, configuration management, graphics processing unit (GPU), infrastructure-as-code (IaC), media creation, virtual desktop infrastructure (VDI)*



**Media creation is an integral component of broadcast television production. On any given day, thousands of graphic assets are created for news broadcasts, late-night entertainment shows, websites, digital streaming platforms, and mobile applications. Across its diverse brands and platforms, NBCUniversal hosts numerous creative teams of graphic artists and video editors. In an effort to address traditional infrastructure challenges and limitations, NBCUniversal has developed a fully cloud-native media creation platform that is nimble, automated, and elastically scalable.**


## Introduction

Media creation is an integral component of broadcast television production. On any given day, thousands of graphic assets are created for news broadcasts, late-night entertainment shows, websites, digital streaming platforms, and mobile applications. Across its diverse brands and platforms, NBCUniversal hosts numerous creative teams of graphic artists and video editors.

In an effort to address traditional infrastructure challenges and limitations, NBCUniversal has developed a fully cloud-native media creation platform that is nimble, automated, and elastically scalable.

Before diving into our solution, let us first discuss why we went down this road in the first place, namely the issues with traditional broadcast media creation, and how we went about solving them.

## Traditional Broadcast Media Creation Infrastructure

Tight deadlines and the overwhelming demands of live television production require creative teams to be equipped with specialized graphic workstations, high-bandwidth networking, and high-performance shared storage. Traditionally, these requirements force creative teams to be centrally located with physical hardware and cabling.

A creative user's desk typically consists of a Mac or PC workstation, dual-desktop computer monitors, video router control panels, and a broadcast video monitor.

*Digital Object Identifier 10.5594/JMI.2020.3024028*
*Date of publication: 11 November 2020*

This hardware requires a large amount of power, video wiring, and network cabling to be run to each desk.

In addition to this desk-side infrastructure, there are numerous back-end servers that support media creation workflows, including directory services, font management servers, plugin license servers, databases, and file shares. As these services directly impact data-driven graphics for on-air broadcasts, they are run on enterprise-grade redundant hardware with high-bandwidth network connections.

High-performance storage is a necessity in any media creation environment. Artists need fast, reliable storage that can handle large media files (video, audio, and images) and allow for collaboration. Creative users need a file share that allows them to access and work on the same content simultaneously. These files ultimately get delivered to production control rooms for on-air graphics playout.

Across NBCUniversal, these creative environments take up a large physical footprint and are expensive to both commission and maintain.

High performance, low latency, and data reliability come at a premium price.

### *Limitations of Physical Hardware Infrastructure*

Physical hardware is both expensive and capital-intensive. Creative workstations generally cost over $10k each, and to utilize the latest media software and applications, workstations must be replaced with newer upgraded models every three years.

Storage systems also follow a three-to-five-year refresh cycle. Storage capacity grows exponentially as more content is created and higher resolutions are used [high-definition (HD), 4K, and 8K], creating the need for more storage tiers and archive solutions.

Scalability is also a major challenge. Wiring and configuring workstations is a slow and expensive process. As systems cannot be commissioned quickly, extra units are often purchased and provisioned in case of system failure or growth in headcount. This model plans for the worst-case scenario, while during off-hours many systems sit idle. This is an example of the "locked-island problem," where compute resources are tied up in unused workstations. Most applications running on bare-metal servers cannot take full advantage of their over-provisioned central processing unit (CPU), random access memory (RAM), or storage.

The physical hardware and cabling at each desk also require artists to report to the same seat every day. There is no flexibility to work from another office or from home. Companies routinely spend thousands of dollars each year just for moving creative workstations to different locations within a facility.

All these limitations boil down to a common theme—lack of flexibility. Let us say we were to onboard a new creative team—we would have to spend capital, provision equipment, and have the necessary manpower to commission it all. And that is how we have always done it—but what if the group doubled in size over the next six months, and how would we deal with that? What if they need to relocate or add remote workers? What about short-run productions that only last a few weeks?

NBCUniversal needs an agile infrastructure, one that can scale and provide these services on demand. And that is exactly where our journey to a cloud-native platform began.

## Migrating Creative Workflows to the Public Cloud

### *Step 1: Server Virtualization*

Server virtualization is a technology that allows for a single physical server to be divided into multiple virtual machines (VMs). A hypervisor extracts the physical server's resources (CPU, RAM, and storage), allowing them to be consumed by VMs without contending for resources. VMs can run different operating systems (Windows or Linux) on the same host server.

Utilizing server virtualization, we were able to virtualize the back-end systems (directory services, font management servers, plugin license servers, databases, etc.) and host them on a small number of servers. This consolidation helped reduce rack space and the power consumed.

### *Step 2: Virtual Desktop Infrastructure*

Virtual desktop infrastructure (VDI) enables desktop operating systems to be virtualized.

VDI architecture provides the flexibility for end-users to work remotely. Unlike the traditional model with physical hardware and cabling, VDI allows users to connect to the desktop from any location with network connectivity.

#### GPU Virtualization

The biggest hurdle in bringing VDI to the broadcast industry was the need for graphics processing unit (GPU). The GPU is a specialized card in creative workstations that processes images and renders graphics, essential to the software used by creative users.

Recently, a specialized virtual GPU solution was introduced that integrates with a hypervisor to allow a physical graphics card to be divided up and allocated to VMs. The hypervisor treats the GPU (and its graphics memory) the same way it treats CPU, RAM, and storage. Guest OS desktops and applications are presented with a graphics profile and view the allocated GPU memory as a physical card connected to the system. This breakthrough in technology allows graphics-intensive applications to utilize the graphics card as it would in a physical hardware workstation.

#### VDI Delivery Protocols

Enhancements have also been made to virtual desktop delivery protocols (and associated agents) to provide a faster, more interactive experience for media creation artists and editors. Depending on the network

bandwidth, artists can now experience lossless video streaming from their VMs to their endpoint.

New features include higher desktop resolutions [4K/ultrahigh-definition (UHD)], GPU acceleration, enhanced color representation, increased color depth, lossless compression, increased text clarity, and distortion-free graphics. This new feature set now allows for a truly interactive experience. Artists who connect to GPU-enabled VMs over this protocol share the same experience as those who connect directly to high-powered physical machines.

With these improvements, VDI is now a viable option for media creation workloads. This model provides the same benefits as traditional server virtualization—centralized resources, ease of management, and increased security.

### *Step 3: Storage Virtualization*

Storage virtualization allows for physical storage from multiple devices to be pooled into a single virtual storage capacity that can be accessed by VMs. Recently, the introduction of hyperconverged infrastructure (HCI) has simplified this process. HCI combines storage virtualization, compute virtualization, and management into a single system. This solution uses software and ×86 servers, eliminating the need for purpose-built hardware. The virtualization software layer abstracts storage and compute resources, and dynamically allocates them to VMs.

### *Step 4: Automation*

By utilizing the technologies listed above, we can now effectively virtualize all the major components of media creation infrastructure. Back-end servers and VDI workstations can be replaced with VMs running on hypervisors, with storage dynamically allocated as needed.

Although this solution provides many tangible benefits, it is an example of the "Lift and Shift" model, where infrastructure has simply been virtualized but not rearchitected. Physical servers must still be purchased to handle the maximum number of users. These servers (although fewer in quantity) are expensive and are bound by hardware limitations.

Physical systems are replaced by their virtualized counterparts, but the overall infrastructure still requires the same level of maintenance. Instead of managing hundreds of individual physical workstations, teams now manage hundreds of VMs. The technology has advanced but the ease of management has not.

NBCUniversal requires a solution that provides elastic scalability of compute and storage resources to eliminate over-provisioning as well as a platform that can dynamically allocate resources on demand. We needed an automated approach to simplify management and maintenance.

Infrastructure-as-code (IaC) is the process of managing and provisioning data centers through scripts or declarative definitions, rather than manual processes or tools. This methodology treats compute, storage, database, and network resources as software. By using this approach, whole environments can be dynamically provisioned and configured to seamlessly accept and propagate change.

Above all, IaC is basically a design philosophy that prioritizes automation, efficiency, versioning, and reusability. IaC applies changes to an environment programmatically by leveraging application programming interfaces (APIs) and other similar methods, contrary to manual configuration.

Instead of manually creating VMs by clicking through a GUI, IaC tools allow the process to be automated. By declaring the end-state in code, teams can be sure that a specific VM always comes up in the same state every time. This becomes especially powerful when provisioning hundreds or thousands of machines.

The same can be said for storage and network resources. IaC tools allow programmers to automate the creation of all infrastructure components, declaring precisely what should be provisioned, with what resources, and at what time.

Closely aligned, configuration management picks up where IaC leaves off. Configuration management is the process of automating the deployment and configuration of settings and software for both physical and virtual machines. The process declares exactly what settings will be applied, what dependencies are involved, and how it will propagate through the infrastructure. This automation-forward approach can be used to deploy software or configurations to many systems simultaneously; the same effort goes into provisioning 1 or 1,000 systems.

With a declarative approach, programmers define what (as opposed to how) infrastructure should be provisioned. The code always specifies the end state, meaning programmers can always see what systems are currently deployed and how they are configured. The code is the documentation.

These automation tools allow us to manage the full lifecycle of our broadcast systems. We have now virtualized and automated our entire technology stack, but we are still managing our own equipment, networking, power, and cooling … enter the public cloud.

### *Step 5: Public Cloud*

Public cloud is the logical evolution of virtualization, providing a platform for compute services hosted and sold on-demand by third-party providers. Customers access these services over the public Internet and pay for only the CPU cycles, storage, and bandwidth they consume.

Unlike on-premises virtualization, public clouds offer a pay-as-you-go model. Instead of making large capital investments in physical hardware, it utilizes an operational expensive model, charging only for the resources needed, with costs directly related to the services consumed. This eliminates the need to purchase, manage, and maintain on-premises hardware and the associated infrastructure (power, cooling, networking, rent, etc.).

Some public cloud-native services include serverless compute, identity management, directory services, code repositories, databases, and many more. Administration services including billing, monitoring, backup, and replication are also provided. The cloud service provider is responsible for all management and maintenance of the provided infrastructure.

One benefit of public cloud services is that systems can be provisioned on-demand and are elastically scalable. Cloud architecture scales dynamically based on the number of resources needed at any given time. This eliminates over-provisioning and provides appropriately sized systems. It also allows for seemingly infinite resources. Service consumers can now make intelligent decisions about scale and resource allocation on the fly.

Public cloud services are also policy-driven, enabling us to interact with them via API and automation tools. Hosted services allow programmers to focus on value-add solutions instead of day-to-day maintenance. Developers and owners no longer need to worry about hardware infrastructure or where it resides; instead, they can focus on media creation workflows.

Public clouds also offer GPU-enabled VMs, providing the ability to host graphics workstations in the cloud. Storage solutions can now deliver cloud-native high-performance (SAN-like) storage. Optimized for media workloads, these file systems can outperform most on-premises storage systems, and unlike on-premises storage solutions where systems need to be (over) provisioned for worst-case scenarios, these file systems can scale elastically up and down as needed.

Cloud-native architecture provides high availability, reliability, and redundancy. Systems can span multiple geographic availability zones to provide true disaster resiliency and ensure that data remain accessible regardless of its location.

## NBCUniversal Solution

Over the past two years, NBCUniversal has combined these technologies into a working solution. Utilizing GPU-enabled cloud-based VMs, dynamically scalable high-performance cloud-native storage, and IaC methodologies, we are now able to provide full end-to-end media production workflows in the public cloud.

A specialized team within NBCUniversal was formed to lead this cloud-forward effort. To avoid the "Lift and Shift" model, the project was treated as greenfield, with no constraints or dependencies to legacy infrastructure. Using an automation-forward approach, all infrastructure resources (compute, storage, database, network resources, etc.) were created through IaC tools.

Where possible, public cloud-native services have replaced back-office servers (directory services, DNS, account management, etc.), providing reliability and ease of management. Third-party cloud software-as-a-service (SaaS) solutions have replaced plugin and fonts management servers, even further simplifying the back-end.

When a new show or production is needed, creative workstations and their associated shared storage are created on-demand. A web input form triggers the automation stack to provision the environment, creating back-end infrastructure, virtual networking, storage, and VDI desktops **(Fig. 1)**.

Using declarative IaC tools, a new isolated cloud environment is created on demand with network rules to communicate with native cloud provider services. Workstations are provisioned from scratch using the latest OS images. Once booted, configuration management tools install the necessary graphics and edit software. Newly created workstations are automatically added to directory services and the VDI connection broker.

Once this automation process is complete, users are presented a dashboard with detailed information about

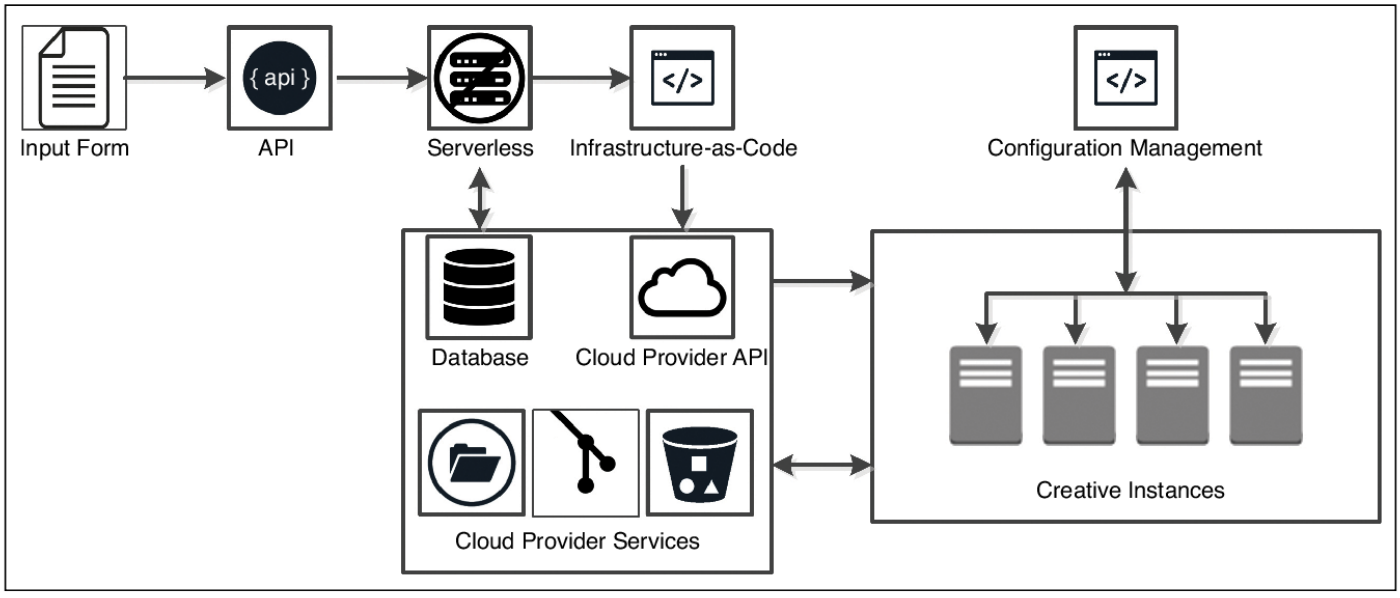


FIGURE 1. Automated provisioning of media creation cloud resources.

the newly provisioned infrastructure. Here, they can view host names and IP addresses of the systems, edit system specifications, add users/groups, access policies, and manage configuration of the high-performance storage cluster.

By design, all provisioned workstation desktops are nonpersistent, meaning they are destroyed when not in use and recreated when needed. NBCUniversal always uses system images with the latest OS patches applied. The nonpersistent desktop approach ensures all systems have the most up-to-date patches. All user data (desktop, documents, application settings, etc.) live on shared storage. Folder redirection allows artists to access their personal data from any provisioned workstation. All media assets live on the high-performance storage.

Once shows/productions are complete and resources are no longer need, users have the ability to "collapse" their environment. When a show is collapsed from the dashboard, all compute resources (workstations) are destroyed and all media data are archived off to cloud object storage. If necessary, shows can be "rehydrated" at a later date. All data are retrieved from the archive back to the high-performance storage tier. Compute resources are also recreated on demand.

NBCUniversal utilizes the latest VDI desktop protocols to deliver a secure, high-definition, and highly responsive computing experience to its creative teams. Artists and editors working in an NBC facility connect to their workstations via a zero client. These facilities are wired with high-bandwidth direct network connections to the public cloud provider for the highest quality experience.

Satellite creative workers utilize a virtual private network (VPN) connection to access their workstations in a secure manner from anywhere over the public Internet. With most of NBCUniversal's creative teams working from home due to COVID-19, this has become the preferred connection method. The ability to access the same desktop resources from anywhere allows creative teams to transition to a work-from-home model without missing a beat.

Collaboration between geographically diverse teams is also now a viable option with this platform. Creative users from New York and Los Angeles can now work on the same shared storage. We no longer have to build identical tech-stacks in multiple locations and worry about syncing content. Hosting this infrastructure in the public cloud has enabled users from anywhere to access the production environment on-demand. Over the past six months, our cloud-platform userbase has quadrupled in size.

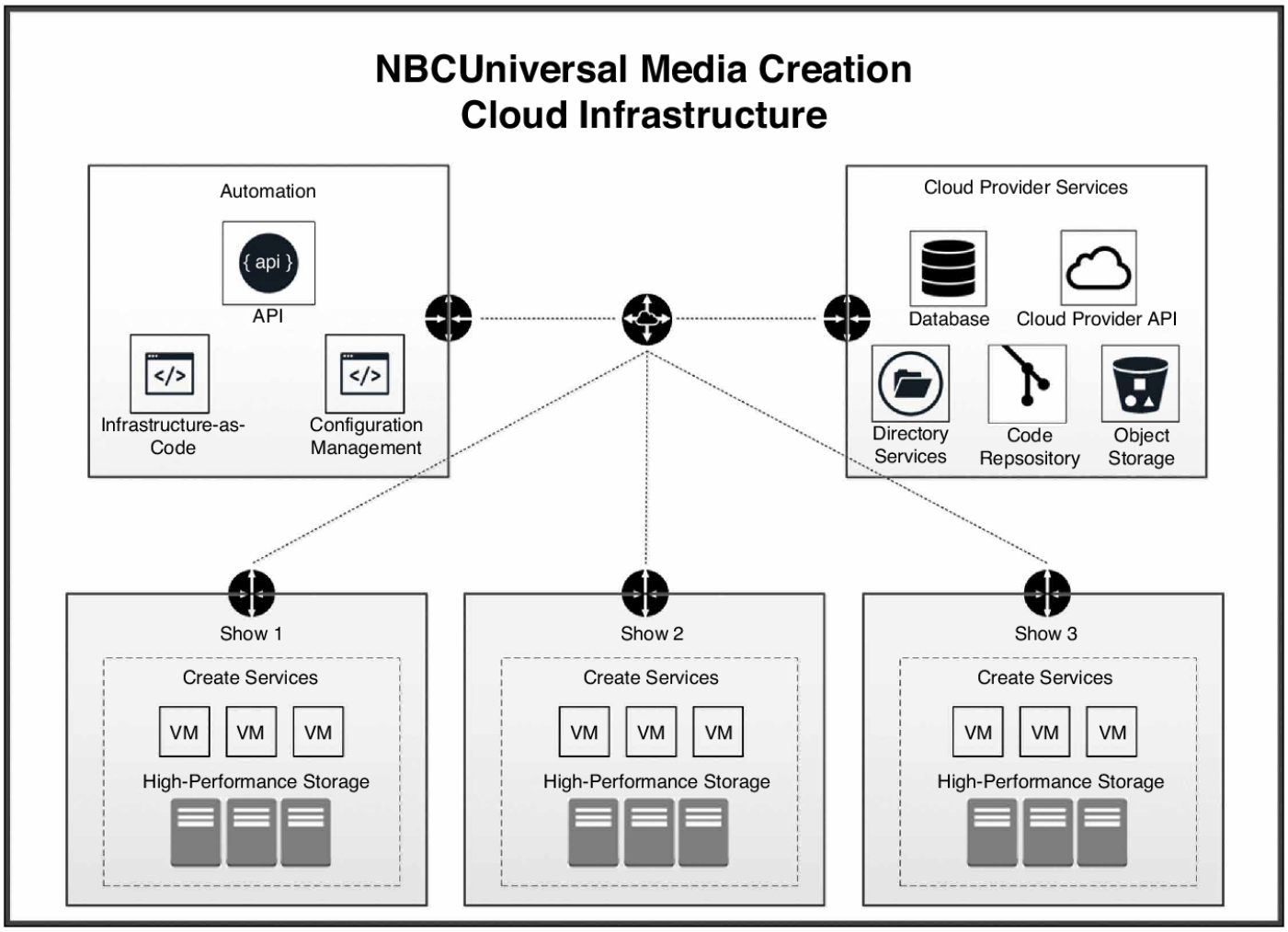


FIGURE 2. NBCUniversal media creation cloud infrastructure.

One major benefit to this platform is scalability. Cloud architecture scales dynamically based on the number of systems needed. This eliminates over-provisioning and allows for the addition/removal of resources as necessary. Instead of running hundreds of physical workstations round the clock, we only provision (and pay for) what is needed at any given time.

Automation tools further enable elastic scalability, allowing the infrastructure to grow and shrink to meet production needs. (For example, a creative team needs to onboard 25 new editors to meet a show launch deadline. Through automation, we can easily provision an additional 25 edit systems to be ready in less than an hour. Once the project is complete, we simply destroy these systems and right-size the environment for the current workload.)

This rapid-scale functionality also enables small-run productions to have fully functioning creative environments without the need for any capital purchases. Our infrastructure is now truly ephemeral, which is in stark contrast to the traditional physical hardware model **(Fig. 2)**.

Automation tools provide the flexibility to easily scale up VM specs; instance types and storage capacity can easily be adjusted through code as needed. Software versions can also be upgraded/modified on demand. A new software package can be deployed to all systems in the environment via a simple code change.

This is truly the power of automation. Once a system end state is validated and defined in our code-base, we know that every subsequent system will be deployed with the exact same configuration, severely reducing management overhead and system drift.

### *Solution Challenges/Benefits*

A successful migration to a fully virtualized cloud platform should be seamless to the end user. If implemented the right way, creative teams (artists and editors) should not notice any difference in their day-to-day operations; the experience should be the same no matter where the infrastructure resides. Support and maintenance teams, however, will need to adjust. Instead of maintaining physical hardware and wiring, the focus of these teams must shift to managing the infrastructure and application services. Most of the application-layer support tasks should remain the same, but the management of virtual infrastructure (and its automation tools) will require a change in skillset. For NBCUniversal, this new model changes both the skillset and structure of its maintenance teams.

Security is always a concern when migrating workflows to the public cloud. Proprietary company data are now hosted in datacenters which are hundreds of miles away. Additional security measures must be put in place including encryption, hashing, VM hardening, identity, and access management.

Source material needs to be ingested to cloud workstations for creative manipulation, and final assets need to be delivered back to on-premises control rooms for on-air playout. Network bandwidth and security considerations must be made, and media-specific file delivery options must be addressed. Granular network rules (security groups) should be put in place to microsegment applications and only open communication between servers on necessary ports. Cyber security tools should be put in place to monitor and log network traffic. A vulnerability assessment by a third-party company is recommended to ensure security of company data.

Aside from the technical challenges, one of the biggest hurdles in migrating to a cloud-native platform is the financial implications. As mentioned, public cloud costs follow a pay-as-you-go operating expense model. Teams only pay for the resources they need, with costs directly related to the services they consume. When evaluating a potential migration to the public cloud, companies must factor in data egress costs in addition to compute and storage. Cloud costs may initially seem more expensive than on-premises solutions, but on-premises total cost of ownership (TCO) should also be considered. TCO includes infrastructure costs associated with power, cooling, and rent, expenses that do not exist in the public cloud model. Companies should also factor in enterprise volume agreements, which may also help reduce cost.

The size of the team being migrated to the cloud is also a factor. High-performance storage systems are costly, making the solution more cost-effective for larger teams. At the time of this writing, the breakeven point for NBC is around 40–50 creative cloud users.

The pay-as-you-go model is only cost-effective if systems are properly managed. Public cloud charges are incurred only when systems are in use, so it is imperative to make sure systems are hibernated when idle. Leaving idle machines running overnight will significantly impact monthly bills. Configuration management tools can provide a powerful solution to this problem. Policies can be set to shutdown systems on user-logout or after a set period of idle time.

Another way to avoid exorbitant cloud compute costs is to ensure systems are properly sized. With seemingly unlimited resources in the cloud, teams may be tempted to over-provision systems as they did in the physical hardware model. This is unnecessary and can lead to excessive costs; higher system specifications equal higher rates. It is important to "right-size" systems when they are provisioned. With automation tools, it is easy to redeploy with higher specifications if necessary down the line.

Media creative teams should also be wary of burst costs when in production. Unlimited cloud resources provide great flexibility, especially for cloud rendering solutions. Users can choose to spin up as many compute resources as possible to complete a render job, but this comes with a price. These tasks should be weighed with a time versus cost model to determine the best practice for each production.

Storage utilization is another big factor in cost-efficiency. It is important to "right-size" build high-performance cloud storage. This hot tier should remain a small fixed capacity.

Policies should be put in place to automatically tier colder data off to lower-cost cloud object and archive storage. In this model, the high cost storage capacity (and price) remains constant, while the lower cost storage can grow over time. This not only reduces cost but provides unlimited data retention, eliminating the need for separate archival solutions.

## Conclusion

Although most of the underlying technologies have been around for years, applying these tools to broadcast media creation has been a major challenge. The combination of GPU-enabled cloud-based VMs, dynamically scalable high-performance cloud-native storage, and IaC methodologies have finally made public cloud media creation a viable option. NBCUniversal has pioneered this effort and will continue to develop its platform, migrating even more production workflows to the public cloud.

## About the Authors

***Kevin Fornito*** is the director of production infrastructure at NBCUniversal, responsible for the provisioning and automation of core infrastructure supporting live broadcast production. He has spent the past 12 years at NBCUniversal and is currently focused on the migration of production workflows to the public cloud, utilizing cloud-native, and micro-service architectures.

***Chris Zembower*** is a senior systems architect at NBCUniversal, focusing on hybrid cloud, automation, and orchestration for the production engineering group supporting live studio operations at 30 Rock. With more than 15 years of experience in media and technology roles, he has sought to introduce modern technology principles and DevOps-oriented tooling to broadcast and media workflows.

***Steve Sneddon*** is the senior vice president of production engineering at NBCUniversal, where he oversees the design, build, and support of the technical facilities and infrastructure across NBCU's live production locations in New York for MSNBC and the NBC Network, Telemundo Center in Miami, FL, and Universal City Studios in Los Angeles, CA. He has nearly 20 years of experience in large-scale broadcasting technology and has spent the last 15 years at NBCUniversal.

*Presented at the SMPTE 2019 Annual Technical Conference & Exhibition, Los Angeles, CA, 21–24 October 2019.* 