Scaling Content Delivery for IPTV and WebTV

http://www.convergedigest.com/bp/bp1.asp?ID=484

Introduction

New challenges are underway in serving and storing digital content for on-demand viewing via IPTV or WebTV. As the amount of digital content accumulates, key issues arise in not only storing the data, but making it easily instantly accessible to a large number of viewers.

Audiences today want the best of both worlds...a huge library of titles to choose from and instant access. Content-centric companies and media outlets strive to meet those needs but typically it is easy to either serve a small amount of data quickly, or make a large amount of data accessible with slower access times, not both. One alternative has been to outsource part of the distribution to content delivery networks, replicating data across dispersed Internet locations and creating a high added cost.

But now solutions based on centralized storage caching give content businesses the best of all worlds, instant access to the most popular titles, a vast library archive to fit broad preferences, and the ability to effectively serve content from company owned data centers.

Background

Content-centric businesses seek to maximize the exchange between the large libraries and the highest number of users. Business models rely on both measurements increasing in order to reach significant profitability, as shown in Figure 1.

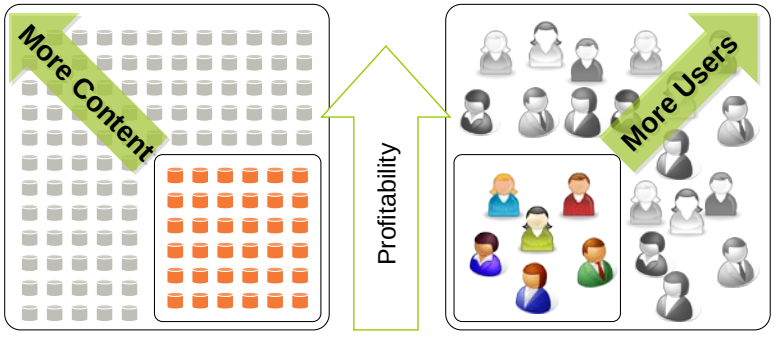

Figure 1: Content-centric businesses aim to maximize content and users

But scaling this type of compute and storage infrastructure presents its own challenges. On the storage side, capacity and performance of the content library, such as a conventional disk array, are at odds with each other. The larger the library, the harder it is to serve content quickly due to disk contention and the need to manager multiple client connections. This limited capability hampers the ability to reach a large audience.

As a result, content-centric companies typically outsource part of the distribution effort to content delivery networks (or CDNs) that replicate content across geographic areas to guarantee instant access to users. With the successful adoption of key content and services, the cost of relying on a CDN can skyrocket.

Architectural Models for Content Delivery

The base model for content delivery involves two main components, a user/Internet component residing outside the data center and the origin or "owned" infrastructure component residing inside the data center. The origin infrastructure typically comprises three tiers starting with load balancers and web servers reside at the data center egress. That tier is supported by a set of application servers, which in turn access a central storage or database repository.

With moderate scale, this model works very well in that the origin infrastructure is relatively simple to maintain, and there is only some reliance on CDN which does not require excessive content replication or cost.

To date, two models have been used to scale content delivery beyond the base implementation, the CDN model and the over-provisioned model. A third, new model is emerging based on centralized storage caching. These are outlined below and shown in more detail in Figure 2.

CDN Model

When demand spikes and the origin infrastructure can no longer handle response time requirements, content-owners often turn to outsourced content deliver networks. These providers offer the ability to host and/or replicate content to a distributed network serving users from multiple points-of-presence. This model is very effective at meeting user demand quickly and can be implemented with relative ease.

However, large scale content distribution via a CDN, particularly with heavily-requested videos or music, can become costly. And while there will always be a need for edge-focused distribution, it makes sense to build origin infrastructures that can absorb additional peak load without sending a corresponding peak to the bill!

Over-provisioning Model

Fortifying an origin infrastructure in the data center to scale can also be costly, requiring added expenditures and excessive manual steps. Meeting high volume demand typically involves replicating content across individual application or web servers. For very small data sets this is possible, but it is not an option with larger libraries. Maintaining multiple content copies requires a tedious update processes and adds significantly to the capital expenditure at the application and web server layers.

Another tactic to scale data center performance involves over-provisioning at the storage layer, replicating the content for performance needs. This too can add significantly to capital expenditures and software costs to maintain multiple consistent copies.

Centralized Storage Caching Model

A centralized storage caching approach to scaling content delivery reduces the high variable cost of a content delivery network and eliminates data center over-provisioning. By caching content in a high-speed, high-capacity centralized pool of memory resources, lightweight application and web servers have instant access to a single content repository, with headroom to scale.

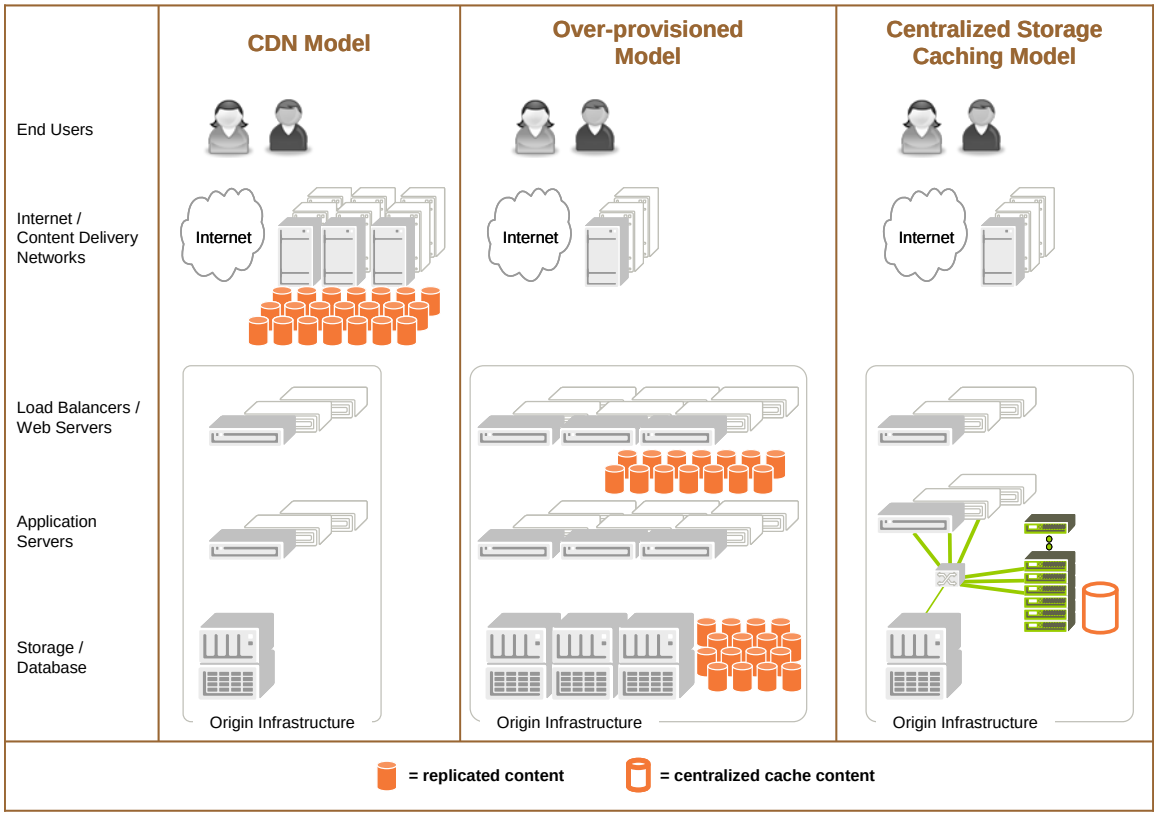

Figure 2: Options for scaling content delivery

Details of Centralized Storage Caching for Web Scale Distribution

This approach uses scalable caching appliances with access to a single storage repository, and relies on the caching appliance to deliver the speed and performance necessary for the application and web servers. As shown in Figure 3, the underlying storage/database component holds a single instance of the original content. As application servers make requests, they are served from the scalable caching appliance.

Caching by its very nature is dynamic. As requests are made, they are served from cache. If a request is made that does not reside in cahce (known as a cache miss), the caching appliance simply retrieves that content object from the persistent storage device, such as a NAS server. Future requests for that object are now served from cache.

In the event that the cache becomes full, the least recently used content object will be removed from cache, but still reside in the persistent storage device, allowing a dynamic cycle to take place with little to no management required.

By retaining frequently accessed data in a centralized pool of high-speed, high-capacity memory, the caching appliance can serve thousands of application servers simultaneously without sacrificing performance. This enables the use of lightweight application servers that have less internal memory and disk than what would be required without the caching appliance. As the user based grows and demand increases, the scaling centers around maintaining appropriate caching resources and adding lightweight and cost-effective servers.

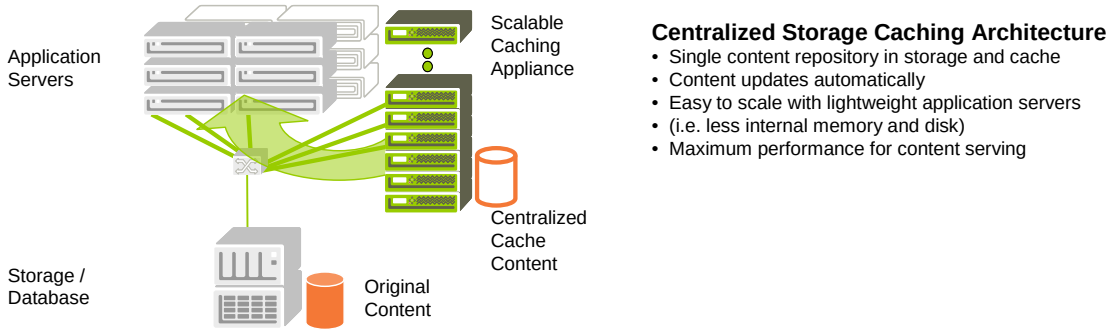

Figure 3: Details of A Centralized Storage Caching Approach

Conclusions

With the rise of Internet-based distribution for music, videos, and other types of content, the ability to scale both the content store and the user base helps ensure profitability and success. To date, distribution models relying too heavily on content delivery networks and those relying on over-provisioned infrastructure have proved costly.

Now options combining centralized storage caching deliver the performance and economics for scalable, web-based content distribution that is easily managed, maintained, and dynamically adjusts to meet user demand.