Web Caching in the Cloud

Cloud computing is emerging as the preferred way for users, organizations and service providers to access, use and manage computing resources. Web caching, and in particular caching based on open-source Memcached software, has emerged as a key technology for web site operators and developers to economically scale their sites. But cloud architectures and Memcached do not seamlessly coexist.

Gear6 has developed offerings compatible with both cloud architectures and Memcached, allowing site operators to simultaneously exploit the benefits of both.

In recent years, the phenomenon of cloud computing has gained great momentum, and is quickly becoming the preferred way for users, organizations and service providers to access, use and manage computing resources. In the cloud model, computing details are abstracted from the users who do not need or want any knowledge of, expertise in, or control over the technology infrastructure. Cloud computing represents a new supplement, consumption, and delivery model for IT services based on the Internet, and it typically involves the provision of dynamically scalable and often virtualized resources as a service over the Internet.

The advantages of this model are numerous. Organizations can decide that the design, construction and management of physical data centers is not a core competence, and through cloud computing those tasks can be outsourced to third-party cloud service providers who are experts in these areas. Significant economies of scale can be realized, as the varying needs of many organizations can be sourced and managed together.

And most cloud service providers can allocate additional computing resources to a customer almost instantly, allowing companies to react to changing demands, and eliminating the need to provision physical equipment for peak loads that is not needed at other times. Cloud computing provides a “pay as you go” model for computing, slashing capital expenditure (“capex”) costs and incurring operational expense (“opex”) only when computing resources are required.
Solution Brief: Web Caching in the Cloud

Commercial offerings:
Amazon’s EC2

Amazon.com is a pioneer and leading provider of cloud-based computing services, through its Amazon Elastic Compute Cloud (EC2) offering. EC2 allows Amazon customers to rent computers at Amazon facilities, on which they can run their own computer applications. EC2 allows scalable deployment of applications by providing a web service through which a user can boot an Amazon Machine Image (or AMI) to create a virtual machine instance containing any software desired. Many enterprises and web service providers have moved, or are considering moving, some or all of their computing infrastructure to the EC2 cloud environment.

Web Caching

A separate and generally unrelated technology that has gained traction in recent years is web caching. Web site operators have realized that by caching frequently-accessed data in high-speed memory, instead of in databases, they can dramatically reduce response times, economically scale their sites, and substantially reduce the load placed on their databases, file servers, and other information sources.

The open-source distributed cache system Memcached has quickly become the de facto standard for web caching, particularly for operators of dynamic, high-growth web sites. Memcached stores frequently-used data in DRAM instead of in databases, providing response time improvements of 100x or more.

Commercial offering:
Gear6 Web Cache Server

As Memcached has grown in importance, a number of its shortcomings have become evident. Its use of memory is not particularly efficient, it lacks high-availability features that are a must-have for many web-based businesses, and if offers very little visibility into its inner workings for management, troubleshooting and planning purposes. And as open-source software, it lacks the backing of a support organization that can help users to operate it properly. Gear6, based in Mountain View, California, has addressed these shortcomings with Gear6 Web Cache, an enhanced Memcached distribution with substantial advantages in memory efficiency, cost, fault tolerance, management tools and support. More information on Gear6 Web Cache is available at www.gear6.com.

Cloud computing and Memcached

It is not surprising that many large organizations are interested in both cloud computing and Memcached web caching, since they offer complementary advantages in cost savings, flexibility and ease of management. But they too have evolved in different communities, largely in isolation from one another, and they do not operate together as seamlessly as might be desired.

As an example, the ability to rapidly scale applications up or down in size is a key advantage of cloud computing, but standard Memcached does not accommodate it very well. Changing the size of a Memcached cache tier causes the cache data to be flushed and then rebuilt, this typically causes degradation in site performance, and causes spikes in the demands on source databases and file servers. Rebuilding the contents of a large cache tier can take hours, and site performance may not be restored to normal until the end of the rebuilding process.

Another issue with the joint use of cloud computing and Memcached is that they need to be integrated properly to provide the best performance. Many users of cloud services lack the time or expertise to fine-tune Memcached to the Amazon EC2 architecture, or to that of another cloud service provider.
In addition, each of the shortcomings of Memcached in the data center environment can be seen when Memcached is deployed in the cloud: the inefficient memory utilization, the lack of fault tolerance, the limited visibility and management capabilities, and the lack of support.

**Gear6 Web Cache Server for Amazon EC2**

For these reasons, Gear6 now offers a version of Gear6 Web Cache for Amazon EC2 cloud-based computing. Called Gear6 Web Cache Server, it begins to equip cloud architectures with the same cache tier advantages that Gear6 Web Cache provides in data center deployments.

**Memory efficiency**

Despite all the other advantages of standard Memcached, there are some significant shortcomings in the way it manages memory:

- Memcached sets an object size upper limit of 1 megabyte
- Memcached places object in buckets of pre-set sizes, resulting in significant amounts of wasted memory
- Memcached employs a slab architecture that does not respond well to changes to the mix of object sizes
- Memcached uses a least-recently-used (LRU) eviction algorithm that does not assess the frequency of use or creation cost of objects before evicting them

Gear6 Web Cache Server, like Gear6 Web Cache, replaces the Memcached memory system with an elegant, dynamic, block-based “cache without borders” system that treats all available memory as a unified pool, allocating memory as needed without restriction or waste. Please see the Gear6 white paper “Cache Tier Memory Efficiency with Gear6 Web Cache” for a complete overview of Gear6 memory management.

In addition, Gear6 Web Cache Server for Amazon EC2 is integrated with Amazon’s block-based memory architecture, applying to it the same techniques that are used with flash memory in Gear6 data center offerings. This results in the highest possible utilization of available memory, with the lowest development and support burden on web site operators.

**Dynamic provisioning of Memcached resources**

For applications requiring an exceptional degree of flexibility, Gear6 provides an optional, complementary offering called Gear6 Web Cache Aggregator. Used together with Gear6 Web Cache Server, it allows cache nodes to be created and deleted at will, consistent with the flexibility offered by other cloud-based offerings. This capability is highly valued by Gear6 customers whose web demands have seasonal or event-triggered spikes that are much larger than those seen under normal
Fault-tolerant web caching

As mentioned previously, fault tolerance becomes important in web environments with very large cache tiers. The failure of any component in a standard Memcached deployment can conceivably bring down the entire cache tier. Even under best circumstances, such a failure can result in performance degradations that will persist until the failure is rectified and cache data is rebuilt, a process that can take several hours. Gear6 Web Cache, for data centers, incorporates a sophisticated clustering technology to address these issues. The same capabilities are currently being developed for Gear6 Web Cache Server for cloud environments, and are scheduled for release in the second half of calendar 2010. Please contact Gear6 if this is of particular interest to you.

Manageability features

Gear6 Web Cache Server for Amazon EC2 also includes a web-based management interface, and a Representational State Transfer (REST) interface that is well-suited to cloud deployments. These interfaces provide a wealth of historical and real-time cache tier data that can be used for troubleshooting and planning purposes.

Summary

Operators of high-growth, complex web sites now have the ability to exploit both cloud-based computing and Memcached web caching, through the use of Gear6 Web Cache Server for Amazon EC2. They can move parts, or all, of their computing infrastructure to the cloud, and still provide the substantial performance and other benefits of Memcached and Gear6 Web Cache.

About Gear6

Gear6 is the leading provider of scalable Memcached solutions. Focused on the emerging need for a dedicated web caching tier, the company enables high-growth web sites to deliver and scale dynamic applications and content. Gear6 customers include leaders in media, social networking and content aggregation. The company is privately held, and is headquartered in Mountain View, California. For more information on Gear6 and its offerings, visit www.gear6.com.