KEMP | LOAD BALANCERS

High Availability and Performance for Microsoft SharePoint

Technical White Paper

By Jason S. Dover, Director of Technical Product Marketing

Introduction

Microsoft SharePoint provides a collaborative workspace to share ideas, publish content, build productivity applications, track projects and keep your team connected in the modern workplace. According to Microsoft market research, 78% of Fortune 500 companies use SharePoint and 20,000 new unique users have been added every day since the year 2006. Because of this, a growing number of IT teams are tasked with ensuring successful deployment and continuity for SharePoint infrastructures. Microsoft recommends that when planning a SharePoint deployment, high availability and disaster recovery are of the highest priority since other important aspects such as performance and capacity are negated if server farms are not available or cannot be recovered after an unexpected failure event. In addition to these considerations, a plan for fostering adoption early in the project lifecycle is key to recoup the costs of the initial investment. Along with a clear understanding of individual stakeholder requirements, key enablers for this include:

- Comprehensive and accessible training
- Ease of use
- Long uptimes
- Low number of helpdesk tickets (contributed to by a low number of unexpected outages)

It's also important to agree on acceptable organizational expectations regarding uptime for the SharePoint environment based on SLAs (service level agreements) and RTOs (recovery time objectives) for internal and external clients. As shown in the adjacent table, even minor diversions can make big differences.

Availability %	Downtime per year	Downtime per month	Downtime per week
99%	3.65 days	7.20 hours	1.68 hours
99.9%	8.76 hours	43.2 mins	10.1 mins
99.99%	52.56 mins	4.32 mins	1.01 mins
99.999%	5.26 mins	25.9 secs	6.05 secs
99.9999%	31.5 secs	2.59 secs	0.605 secs

As with most applications, to facilitate availability, resilience and an overall sound architecture that can withstand unexpected anomalies, a lot of planning and consideration is required. Since enterprise SharePoint deployments are typically architected with multiple tiers and distributed across multiple sites they can be very complex. Fortunately, Microsoft has integrated a number of intelligent availability features in recent versions of SharePoint and extended the supported environments. This, combined with application delivery technology that enables multi-site distribution and traffic acceleration makes SharePoint more reliable, available and scalable. Support for Microsoft Azure also enables hybrid deployments and flexible DRaaS (disaster recovery as a service) options.

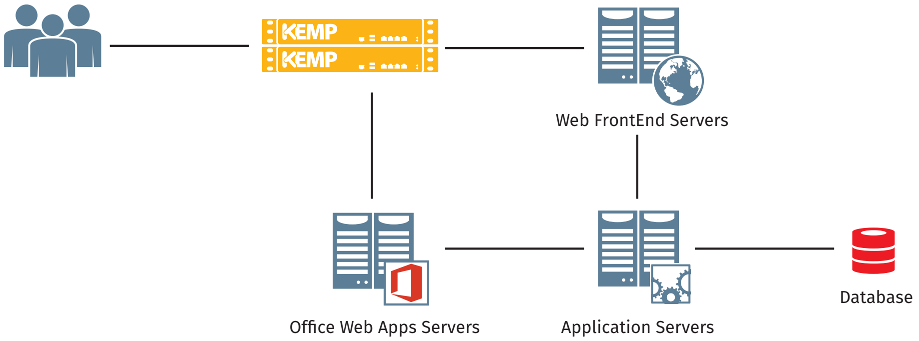


Figure 1 -Typical HA multi-tier SharePoint topology

High Availability in the Cloud

From its original introduction, Microsoft Azure has rapidly matured to be a premier cloud platform for enterprise applications. With official support extended for Microsoft SharePoint it's a prime destination for new and augmented deployments, equipping customers with the benefits of a cloud resource consumption model and simplified administration and provisioning. Based on ease of deployment and an attractive cost model for transient environments, Azure is often used for SharePoint development and test. However, the ability to reduce costs, increase resource flexibility, lower data center commitment and provide on-demand scalability has made Azure the production platform of choice for many enterprises as well.

By placing redundant SharePoint farms across different Azure regions, built in site resilience is gained without the need to maintain on premise secondary and tertiary data centers. Cold, warm and hot standby methods provide options to balance initial and ongoing investment with complexity, cost and recovery time. Evaluation of organizational RTOs and RPOs (recovery point objectives) require consideration to choose the right model.

DR Model	Description	Time to Recovery	Pros	Cons
Cold	Fully built farm with VMs stopped requiring periodic starting to update and maintain	Hours to days	Cheapest to maintain	Slowest recovery
Warm	Running smaller version of farm requiring additional provisioning and database steps at recovery	Minutes to hours	Inexpensive to recover	Expensive and time consuming to maintain
Hot	Fully provisioned, updated and running	Seconds to minutes	Quickest to recover	Expensive to configure and maintain

The use of Azure Availability Sets enables redundant virtual machine deployment in each SharePoint tier across distinct fault domains (FD) and update domains (UD). Virtual machines within the same availability set operate on different underlying physical hardware and utilize different power sources and network switches to withstand planned Azure infrastructure maintenance events and unplanned outages. With the use of properly configured Availability Sets, at least one virtual machine in a set will always be available meeting Azure's 99.95% SLA.

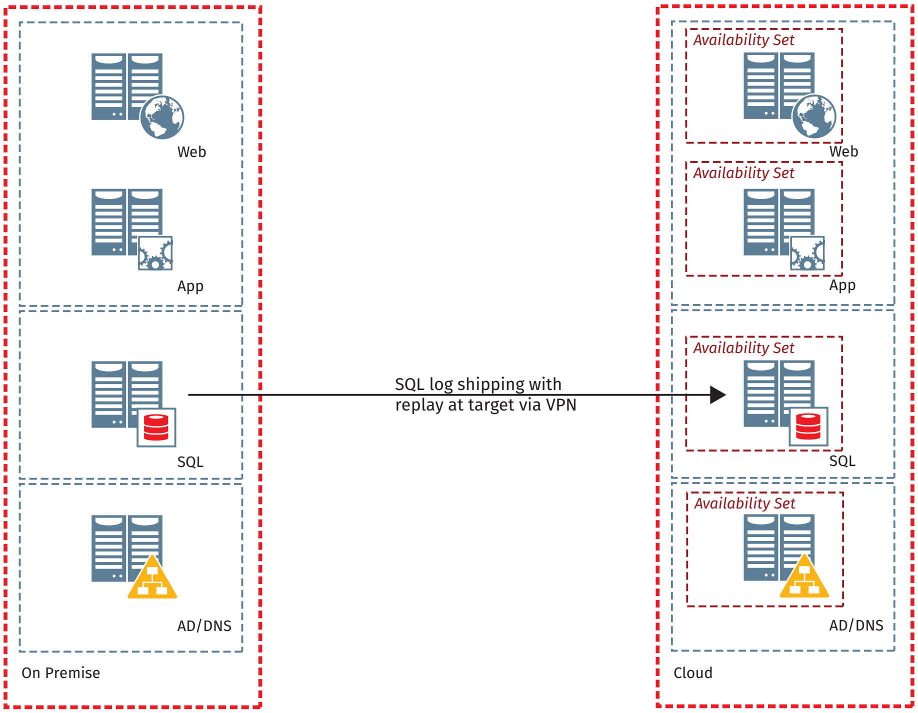


Figure 2 - Warm standby scenario with on premise and Azure deployment

Database Tier Availability (SQL)

The database tier of a SharePoint environment utilizes SQL and contains all configuration and content data that allows the deployment to function. Microsoft has engineered several HA features into the modern versions of SQL making it easy to plan and implement availability.

Log Shipping, which operates at the database level, can be used to maintain multiple warm standby or secondary databases for a single production or primary database to handle failures and facilitate minimally intrusive maintenance windows. With SQL AlwaysOn Failover Cluster Instances (SQL Failover Clustering prior to SQL Server 2012), high availability is possible at the SQL server instance level as well. A failover cluster instance is a single instance of SQL server installed across a set of Windows Server Failover Clustering (WSFC) nodes appearing as a single instance while providing redundant failover capabilities.

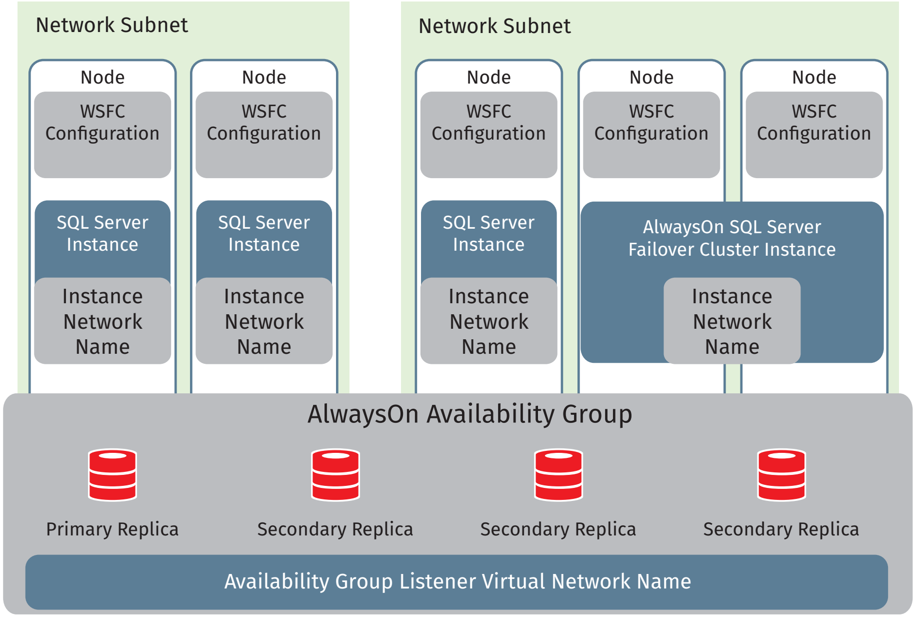


Figure 3 - SQL and AlwaysOn Availability Groups

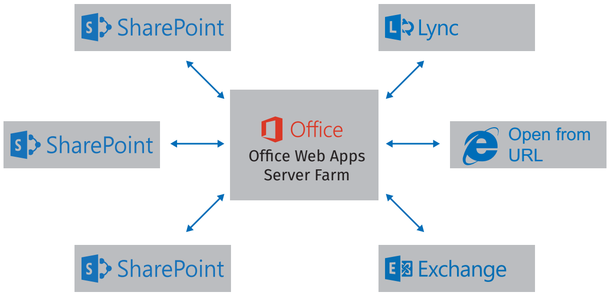
Database mirroring was previously leveraged to maintain a single standby database, or mirror database, for a corresponding production database also known as the principal database. Despite the effectiveness that this HA methodology has served in SQL environments for several versions, it should be noted that it's been deprecated in SQL 2014. Because of this, Microsoft's current guidance is to use AlwaysOn Availability Groups instead.

Application Tier Availability

Farm servers in the application tier typically host the Central Administration website and other services that require dedicated resources or separation from the web tier such as crawl components, query components and profile pages. By design, SharePoint allows for scaling at the Application tier by simply adding additional redundant servers for each deployed service. SharePoint provides built-in load balancing capabilities that round-robins requests to web service applications. A service application proxy generally requests an endpoint for a given connected service application from the software component load balancer which maintains a list of available endpoints and returns the next available one based on its round robin algorithm. For service applications local to the SharePoint server farm, the load balancer looks into the configuration database and for applications operating in remote server farms, the load balancer leverages the Application Discovery services to locate the endpoint. The topology web service on the remote federated farm copies endpoint location information into the local configuration database on an intervallic basis by means of the Application Addresses Refresh timer job. When endpoint failures occur, the failing endpoint is taken out of the round-robin rotation for the period specified by the "BadListPeriod" parameter of the Set-SPTopologyServiceApplicationProxy cmdlet. Additionally, SharePoint health score (X-SharePointHealthScore) allows for automatic protection of web applications and corresponding servers by enabling throttling when performance counters breach acceptable levels. This allows for high availability and optimal performance at the application tier of SharePoint.

Office Web Apps

Even though Office Web Apps is no longer tightly integrated into SharePoint as of SharePoint 2013, it still plays a key role in deployments by delivering web-based versions of Office products such as Word, PowerPoint and Excel. It allows users to access and edit files using these services from SharePoint as well as Lync, Exchange and URL access. The new standalone farm deployment model introduced in 2013 allows for easier scalability as resource requirements change for the organization as well as simplifies management since it's hosted on dedicated virtual or physical systems and the SharePoint deployment itself no longer has to be optimized in order to host Office Web Apps.



As with all direct and related components for SharePoint deployments, it's important to plan for system failures. High availability for Office Web Apps is accomplished by placing multiple servers in a farm and creating a load balanced VS (virtual service a.k.a virtual IP) on port 443 using a load balancer. The load balancer monitors the availability of the individual nodes by issuing health checks. One method involves sending HEAD requests to the '/hosting/discovery' directory. Assuming there are no network or access issues, failure to successfully complete this attempt is a clear indicator that the server is likely not functioning properly. The load balancer will also maintain affinity or stickiness to the appropriate node to prevent dual-authentication and other session-related problems. This is often accomplished using cookies.

Web Tier Availability and Performance

Servers at the web tier, known as Web Frontends (WFE) handle user requests, serve web pages to clients as well as host web services and Web Parts. Frontends direct requests to the appropriate servers in the Application Tier, which then serve the requested response, back to the Web Tier. They can also be leveraged to host-dedicated query, crawl and other service components. High availability and secure publishing is achieved at this tier by adding multiple WFE servers to a farm and placing them behind a load balancer or load balancing service that both distributes traffic and acts as a reverse proxy. While there are a number of options to achieve this (Windows Network Load Balancer, Application Request Routing, Web Application Proxy) many customers opt to use external load balancers because of their ability to provide core application publishing and load balancing along with intelligent complementary application frontend services, application awareness, global site and hybrid cloud traffic distribution as well as integrated security services. KEMP Technologies LoadMaster application delivery controller (ADC) provides these services along with other

Layer 7 application traffic capabilities, which help optimize flows and client UX for SharePoint environments. Following are a few examples.

- Data Compression

LoadMaster data compression reduces the amount of data that has to be transferred for many object types by utilizing GNU zip (gzip) compression. Leveraging Lempel-Ziv (LZ) compression and HTTP/1.1 gzip content encoding, bandwidth utilization is reduced for high compression files that are transmitted between clients and WFEs such as HTML, CSS, and JavaScript. LoadMaster compresses the application payload in each request, reducing network bandwidth consumption without degrading content quality or negatively impacting response time.

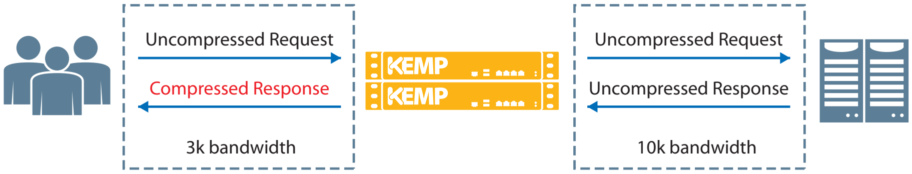


Figure 4 - Application delivery controller data compression

- Static Content Caching

LoadMaster's advanced caching engine saves valuable WFE server processing power as well as farm-side bandwidth that can be re-allocated to other critical business application logic by caching static content in memory. Chatty protocols such as HTTP(s) require frequent creation and closure of connections for fetching static content. By caching this information in memory, user response time is quicker and only requests for new or dynamic content ever gets directed to the workload servers, helping to save resources and improve SharePoint server performance.

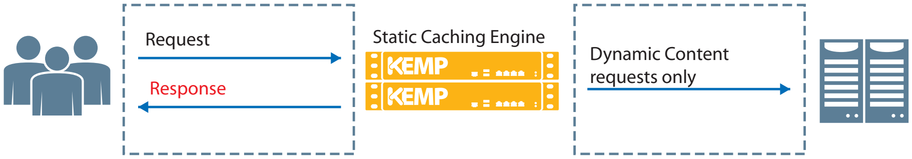


Figure 5 - Application delivery controller content caching

- TLS (SSL) Termination

With TLS (SSL) overlay services enabled, client TLS sessions are terminated at the LoadMaster allowing for intelligent Layer 7 processing such as content switching and advanced session persistence. This is a requirement for any complex traffic manipulation since headers must be read and this can't be done without decryption. Some services in SharePoint require the traffic to reach the target service encrypted and for those, LoadMaster supports re-encryption for terminated traffic streams to ensure secure end-to-end flows. For those that don't, KEMP's LoadMaster provides a security overlay for applications, even which may have not been originally developed to leverage TLS (SSL) sessions, to improve infrastructure security without adding extra processing burden to the workload servers. LoadMaster further extends security capabilities for administrators to restrict ciphers that can be used to access protected SharePoint services.

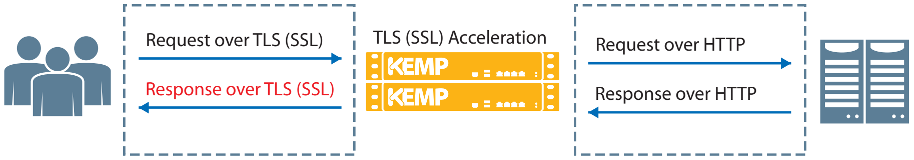


Figure 6 - Application delivery controller TLS (SSL) acceleration

- Content Switching

Content switching enables requests to be directed to subsets of servers in a given pool based on content requested, method used, header properties and other heuristics of incoming traffic. This enables clients to be directed to the best target for their individual session and also allows for requests to one service or resource to be redirected to another (e.g. redirecting port 80 requests to 443). As an example, an enterprise may have developed a custom application that entails the use of fat clients. When the time for an upgrade comes, there may be a need for updates on both the client side and server side for the configuration to continue operating. During this interim period of such a project, LoadMaster can be configured to detect the version of the client issuing the request and direct them to the appropriate server based on revision. Content switching also enables simplified load balancing configuration by allowing different services using the same port set to operate under a single virtual service with sub virtual services – content switching directs individual requests to the appropriate directory based on host header.

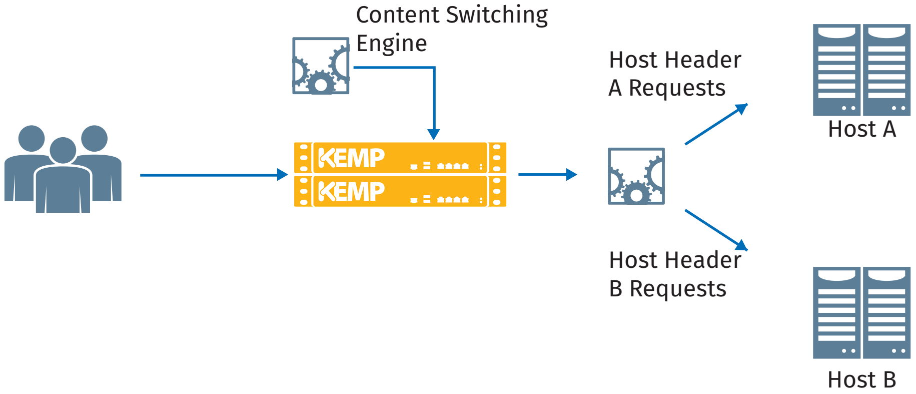


Figure 7 - Host header-based content switching

- Web Application Firewalling (WAF)

Because SharePoint deployments are often published both internally and externally, multiple channels exist for risks to be exploited. Since web-facing applications often harbor vulnerabilities, serious consideration should be given to the security strategy supporting the SharePoint deployment. There are a number of areas that should be addressed to ensure that risk to the environment is minimized. As an example, maintaining clear matrices of who has access to which data will go a long way in preventing data from accidentally getting into the hands of unauthorized parties as well as make it easier for administrators to provide needed records when audits are conducted. Multi-factor authentication and validation of access to resources before allowing requests to reach application servers can also limit the possibility of hackers gaining unauthorized access or executing DDoS attacks.

Another important area is that of security focused at the actual application. SharePoint has no native web application firewalling (WAF) capabilities built in and because of this an external solution is required. While network firewalls and other network appliances contribute to a holistic security strategy, web application firewalls play a key role in securing SharePoint environments since they are able to fill a critical gap based on their operation at the upper layers of the network. For instance, SQL injection, cross-site scripting and cookie tampering attacks would all go unnoticed by a network firewall which is primarily focused on ports and sockets but would be successfully prevented by a WAF solution such as KEMP Technologies' Application Firewall Pack (AFP) which is able to perform deep inspection of http flows and payloads even when they are encrypted. In addition, data leak prevention capabilities ensure that the SharePoint infrastructure isn't intentionally or unintentionally used to leak sensitive information on published webpages or otherwise.

Additional LoadMaster capabilities including authentication services, DDoS mitigation, intrusion prevention (IPS) and intrusion detection (IDS) all work together to support an enterprise defense-in-depth strategy.

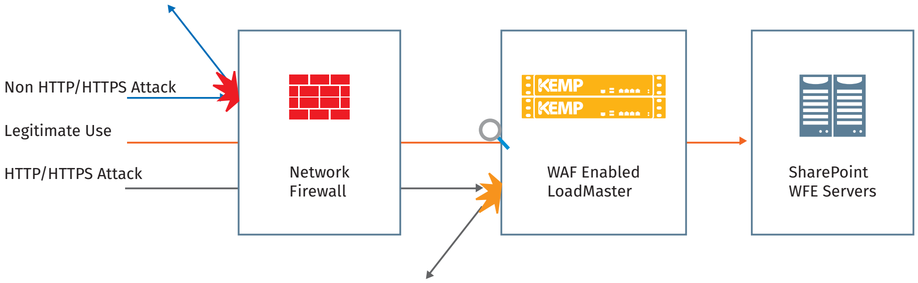


Figure 8 - WAF working with traditional network firewall to secure SharePoint

- Global Site Load Balancing (GSLB / Hybrid)

A core part of site resilience involves having services distributed across multiple locations. This may encompass on premise data centers in different regions, utilization of multiple regions in a public cloud or a configuration stretched across on premise data centers and public cloud for a hybrid deployment. One key enabler for this type of model includes GSLB (global site load balancing) technology. GSLB makes it possible to delegate the responsibility of a distributed namespace to an intelligent device that monitors the system health, utilization and responsiveness of each involved location and distribute requests based on pre-defined business logic. This may include distribution based on proximity and response time in the case of an active/active deployment or the use of a timed delay before failover from primary to secondary in the case of an active/standby deployment. KEMP's GSLB technology is known as GEO and enables flexible configuration models to help customers scale and protect their SharePoint deployments across data center and cloud boundaries.

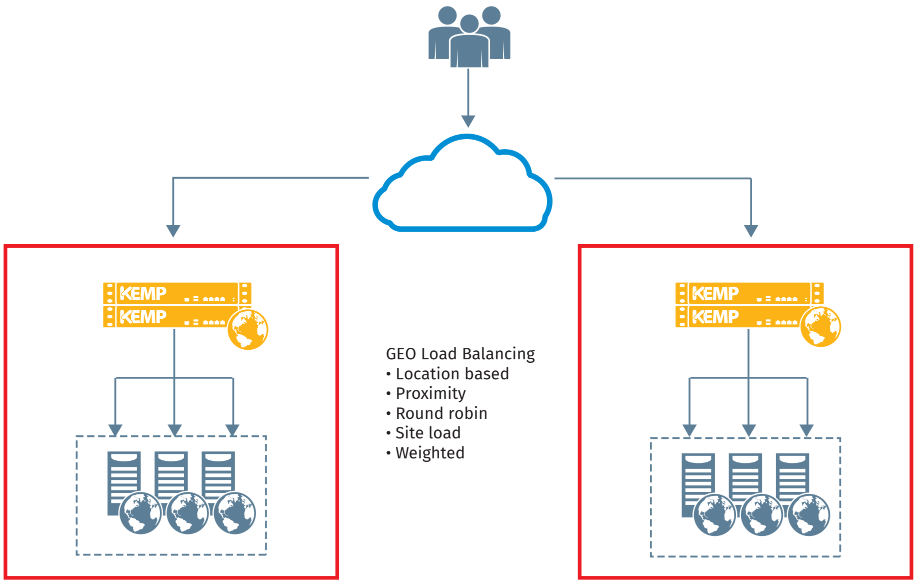


Figure 9 - Typical GSLB distribution across hybrid infrastructure

Application Delivery Controller Service Configuration for SharePoint

SharePoint uses multiple ports for internal and external farm communication. The default ports that generally need to be configured on the load balancing service are those necessary for communication between clients and WFEs and Central Administration (80, 443, 8080 and 8443). A number of additional ports for east-west and backend database communication are needed for SharePoint to function properly. The majority of these ports can be modified from the out of the box configuration in order to meet the needs of individual environments.

Port	Usage	Description
80	Web Front-End (http)	Web Application / Site Access
443	Web Front-End (https)	Web Application / Site Access
8080	SPCA (http)	Central Administration
8443	SPCA (https)	Central Administration

As noted earlier, a virtual service is used on a load balancer to provide access to servers behind it. Such a service consists of an IP address, allowed connection ports and a number of other settings that dictate how client traffic is processed and routed. Three core components of load balancing are scheduling, health checking and session persistence.

- Scheduling Methods

Since a load balancer serves as a single aggregation point for all incoming traffic to a SharePoint farm, an intelligent mechanism is required to decide how traffic is distributed among available servers. KEMP's adaptive scheduling makes distribution decisions based on availability of resources on individual member servers. In other cases, the service can be configured to send requests to the server with the least connections or simply alternate requests in a round-robin fashion. The particulars of the deployment and configuration of the farm will determine the best method to use.

- Health Checking

To ensure high availability it's required to confirm whether or not a server in a pool of a distributed application is healthy enough to receive client requests before forwarding these requests to the target. Intelligent server and application health checking mechanisms such as TCP port checks or attempted execution of HTTP methods allow a LoadMaster to recognize whether or not a given WFE is healthy or not. If it isn't, it's automatically removed from rotation and existing connections are re-routed to an alternate target until it begins functioning properly again.

- Session Persistence

As the name implies session persistence ensures that clients remain connected to the server that they initially start their conversation with for the duration of their session. Depending upon configuration, it will also ensure that clients are reconnected to the same system if their connection is temporarily interrupted because of an IP change on a mobile device or a network issue. This contributes to an improved user experience by preventing login requests mid-session and allowing clients to pick up where they left off in their work stream if they step away from their device momentarily. To determine a user's identity a variety of individual and combined values can be used including IP address, cookies, URL hash, a unique fingerprint based on headers and client version, etc.

Summary

A great deal of planning and administration must go into the architecture and maintenance of a SharePoint deployment in order for it to succeed. Two key objectives for IT teams tasked with SharePoint projects are to ensure that they perform optimally and can withstand unexpected anomalies that lead to downtime and lost productivity. SharePoint has been engineered with a wealth of features that enable administrators to meet these goals. KEMP's LoadMaster helps to further satisfy these requirements by delivering complementary advanced Layer 7 security along with hybrid traffic distribution and content manipulation capabilities. By combining these technologies, taking advantage of a hybrid cloud model using Azure and leveraging the various availability features that Microsoft has implemented, enterprises can gain the benefits of a highly available and optimally performing SharePoint infrastructure.