



Geteilte Last – halbe Last

Geschäftskritische Anwendungen fordern die permanente Verfügbarkeit sowie kürzeste Reaktionszeiten der entsprechenden Server-Dienste. Bei TCP/IP-basierten Netzwerkdienste wie Web-, News-, FTP- und Mail-Server bieten Loadbalancer (Lastverteilsysteme) diesbezüglich wertvolle Unterstützung.

Sei es der Trading-Server eines Internet-Brokers, die Services eines ISP oder die Web-Dienste einer grösseren Unternehmung – die ungenügende Performance oder gar der Ausfall von Server-basierten Diensten verursachen oft schwerwiegende Schäden. Nebst dem folgenschweren Imageverlust sind vom betroffenen Unternehmen oft auch verlorene Kunden und Aufträge zu beklagen. Um derartigen Szenarien vorzubeugen, sind entsprechende Vorkehrungen ein Gebot der Stunde. Es gilt, Server-Dienste sowohl hochverfügbar als auch performant auszulegen. Einen wichtigen Beitrag in dieser Bestrebung leisten Server-Farmen. Diese bestehen aus zwei oder mehreren Servern, die dieselben Dienste zur Verfügung stellen und gespiegelte Datenbestände halten. Server-Farmen beinhalten oft mehrere Cluster mit unterschiedlichen Diensten und sind zum Schutz vor Katastrophen wie Feuer, Wasser und Leitungsausfall auf unterschiedliche Standorte verteilt.

Virtual Resource Management (VRM)

Allerdings reicht eine auf Verfügbarkeit getrimmten Topologie nicht aus, performante Services zu garantieren. So zwingen beispielsweise Lastspitzen schwach dimensionierte Server in die Knie – eine Problematik, die durch den Ausfall einzelner Server noch akzentuiert wird. Diesbezügliche Abhilfe schaffen Load-Balancing- beziehungsweise Virtual Resource Management (VRM) Systeme. Sie verteilen die vorhandene Last an die einzelnen Server mit dem Ziel, eine gleichmässige Auslastung der Systeme zu erreichen. Beim Ausfall oder bei der Wartung eines Servers sind Lastverteiler zudem in der Lage, Client-Anfragen an ein anderes verfügbares System zu leiten und so wesentlich zur hohen Verfügbarkeit der angebotenen Dienste beizutragen. Ferner ermöglichen Load-Balancing-Systeme die einfache Skalierbarkeit von TCP/IP-basierten Netzwerkdiensten und schaffen eine standortübergreifende Fehlertoleranz von Server-Verbindungen.

Gegenüber den Usern sind Load-Balancing-Systeme komplett transparent. Sie besitzen lediglich eine virtuelle Adresse, der die einzelnen Server zugeordnet sind (NAT-Proxi). Dabei spricht ein Client beziehungsweise Web-Browser stets die virtuelle Adresse (z.B. www.boll.ch) an – der Loadbalancer selbst ist dann dafür besorgt, dass die Anfrage nach vordefinierten Regeln an einen der zur Verfügung stehenden Server durchgereicht wird.

Intelligenz bei der Lastverteilung

Die Verteilung der Anfragen an den Server-Pool erfolgt nach benutzerdefinierten Kriterien, wobei anhand intelligenter Algorithmen dafür gesorgt werden soll, dass jede Anfrage an den best geeigneten Server geleitet wird. Abhängig von der vorhandenen Infrastruktur sowie den gestellten Anforderungen anbieten sich unterschiedliche Verfahren

Round Robin (RR): Die eingehenden Requests werden nacheinander und regelmässig den einzelnen Server zugewiesen. Dieses Verfahren eignet sich nur, wenn alle Server identisch ausgerüstet sind. Es kann die unterschiedliche Auslastung der Server jedoch nicht verhindern.

Weighted Round Robin: Bei diesem Verfahren wird der Leistungsfähigkeit der einzelnen Server Rechnung getragen. Schwächere Systeme werden bei der Verteilung sporadisch übersprungen und Server mit einem höheren Gewicht werden gelegentlich zweimal ausgewählt.

Least Connections: Die Vergabe einer neuen Verbindungen erfolgt an den Server, der die geringste Zahl an offenen Verbindungen aufweist. Vor dem Hintergrund, dass nicht jede Session dieselbe Last erzeugt, kann es folglich trotzdem zur Überlast einzelner Server kommen.

Weighted Least Connections: Bei diesem Verfahren werden die offenen Verbindungen über eine Gewichtung normalisiert. Leistungsfähigere Server erhalten folglich mehr Verbindungen zugewiesen als Server mit einer geringeren Kapazität.

Server Latency: Über die Messung der Server-Antwortzeit (TCP Connect oder Application Response Time) wird die Vergabe einer neuen Session an den aktuell schnellsten Server möglich.

Weighted Server Latency: Der unterschiedlichen Leistungsfähigkeit der einzelnen Server wird durch eine Normalisierung der Antwortzeit Rechnung getragen. Dadurch wird verhindert, dass schwächeren Servern keine Sessions zugewiesen werden.

Server Direct Measurement: Die wohl intelligenteste Verteilung der Sessions erfolgt auf Basis der tatsächlichen momentanen Last der einzelnen Server. Dazu wird jeder Server mit einem Softwarepaket bestückt, das dem Balancer die momentane Last übermittelt.

Persistenz der Anwendersitzung

Auf Grund der Tatsache, dass die meisten Web-basierenden Dienste interaktiv sind und aus mehreren Einzelsessions bestehen (Aufbau einer neuen Session pro Webseite), müssen Load-Balancer anwendungsspezifisch dafür sorgen, dass der einzelne User innerhalb einer Anwendersitzung immer mit demselben Server kommuniziert. Diese Forderung kommt unter anderem bei SSL-gesicherten Transaktionen, beim Mail-Austausch oder bei Session-ID protokollierten Transaktionen zum Tragen. So ist die mit „Session persistence“ bezeichneten Funktion beispielsweise bei E-Shops notwendig. Dabei ist von zentraler Bedeutung, dass der Warenkorb und die daraus resultierende Bestellung zwingend über denselben Server bedient werden, um Unterbrüche und Inkonsistenzen zu vermeiden. Vorgängig jedoch – beispielsweise bei der Betrachtung einzelner Produkte – ist noch keine Session persistence notwendig, weshalb bei jeder neu aufgerufenen Seite auf den geeignetsten Server zugegriffen werden kann.

Um die Aufgabe der Sitzungs-Konsistenz zu lösen, stehen primär drei Varianten zur Verfügung. In der trivialsten Form erfolgt die Überprüfung und Zuteilung der einzelnen Sessions aufgrund der übermittelten Absenderadresse. Nachteil dieser Lösung ist jedoch die Tatsache, dass viele Anbieter (z.B. AOL) aus Sicherheitsgründen die IP-Adresse des Absenders im Laufe einer Sitzung ändern und folglich eine konsistente Behandlung der Sitzung verhindern. Als alternative Variante anbietet sich die Nutzung von Cookies, mit deren Hilfe der Load-Balancer in die Lage versetzt wird, den Client nicht mehr aus den Augen zu verlieren. Allerdings besteht diese Möglichkeit nur dann, wenn Cookies auf dem Kundensystem zugelassen sind.

Um der Problematik der beiden erstgenannten Lösungsvarianten zu begegnen, setzen neue Load-Balancer wie etwa die Modelle E350 und E450 von Coyote Point auf die Überprüfung der SSL-ID. Diese auf der Layer 7 Ebene kommunizierten Information garantiert die konsistente Behandlung der Anwendersitzung.

Globaler Einsatz schafft Sicherheit

Ausgereifte Load-Balancer sind nicht auf den lokalen Einsatz limitiert. Vielmehr ermöglichen sie den Aufbau dezentraler, weltweit verteilter Systeme. Dadurch wird einerseits die Verfügbarkeit von Diensten beim Ausfall eines ganzen Standorts erhöht. Andererseits lassen sich Anfragen an den nächstgelegenen Server routen, was die Verzögerungszeiten zwischen Anfrage und Antwort reduziert.

Bei dieser mit GEO-Load-Balancing bezeichneten Struktur wird den einzelnen Loadbalancern der identische Host-Name im DNS zugeteilt, wobei jedes System eine eigene IP-Adresse besitzt. Um eine standortübergreifende Lastverteilung zu gewähren, pflegen die Loadbalancer der einzelnen Standorte einen regen Informationsaustausch.

Um zu verhindern, dass ein Loadbalancer selbst zum Single Point of Failure wird, ist die Integration eines zweiten Systems als Backup im Hot-Standby empfohlen.

Vereinfachtes Management von Server-Farmen

Der Einsatz lastverteilender Systeme lohnt sich nicht nur aufgrund der erhöhten Verfügbarkeit der angebotenen Dienstleistungen. Ins Auge stechen ebenfalls die vielfältigen Erleichterungen bezüglich Unterhalt und Betrieb von Server-Farmen. Diese lassen sich im laufenden Betrieb administrieren, erweitern und warten. Wird ein Server vom Netz genommen, wird die Last automatisch auf die verbleibenden Systeme verteilt. Zusätzlich integrierte Server andererseits werden „on the fly“ zum integralen Bestandteil des Server-Pools.

Ein weiteres Augenmerk gilt der vereinfachten Verwaltung von SSL-Zertifikaten. Erweist sich die Installation oder die Pflege der SSL-Verschlüsselung auf den einzelnen Servern als zu aufwendig, kann diese zentral auf dem Load-Balancer erfolgen. Dieser wird dadurch zu einem eigentlichen zentralen Zertifikatsverwaltungssystem.

Zentralisierter Komfort

Damit die durch den Load-Balancer gewonnene Systemverfügbarkeit uneingeschränkt zum Tragen kommt, werden an die Verwaltung des Lastverteilers selbst hohe Anforderungen gestellt. Moderne Systeme wie die Equalizer-Serie von Coyote Point bedienen sich einer intuitiven Weboberfläche, die der zentralen Verwaltung, Konfiguration und Überwachung sämtlicher vernetzter Load-Balancer und Cluster dient. Das einfache Hinzufügen und Entfernen von Servern im laufenden Betrieb gehört ebenso dazu wie das Umstellen der Regeln (Algorithmen) für das Traffic-Management.

Wertvoll erweist sich auch das grafische Monitoring der einzelnen Server. Es informiert auf einen Blick sowie mit einem hohen Detaillierungsgrad über die Last-Situation jedes einzelnen Systems.

Vorzüge durch den Einsatz von Load-Balancern

- Permanente Verfügbarkeit der angebotenen Dienste
- Optimale Lastverteilung auf die Server innerhalb eines Clusters
- Einfache Wartung der Server im laufenden Betrieb (ohne Unterbruch der Dienstleistung)
- Flexibilität bezüglich Gestaltung der Server-Infrastruktur (Auf und Ausbau mit unterschiedlich leistungsfähigen Servern)
- Gewinn an Sicherheit durch dezentralen Aufbau (GEO Load-Balancing)
- Zentrale Zertifikatsverwaltung
- Intuitive Verwaltung und Konfiguration der Load-Balancing-Systeme sowie der einzelnen Server und Cluster