MC/ServiceGuard
enterprise HA für Linux

Autor

Kai Dupke
Teamleiter Linux
probusiness AG, Hannover
kdupke@probusiness.de

Abstract

Hochverfügbare Systeme werden im Linux-Umfeld immer häufiger benötigt. Sei es, dass Hilfssysteme produktiven Status erlangen, sei es, dass produktive Systeme zu Lasten von propriäteren Unices auf Linux umgestellt werden. Mit MC/ServiceGuard steht eine enterprise-Lösung bereit, sich den Anforderungen an einen hochverfügbaren Alltag zu stellen. Im folgenden wird diese Lösung dargestellt.

MC/ServiceGuard - enterprise HA für Linux

Hochverfügbare Systeme sind im Bereich unternehmenskritischer Anwendungen Standard. Mittlerweile hat Linux in diesem Bereich verstärkt Einzug gehalten. Somit wächst der Bedarf an geeigneten Hochverfügbarkeitslösungen für Linux. HP hat mit MC/ServiceGuard (MC/SG) eine Lösung auf den Linux-Markt gebracht, die ihre Wurzeln im Enterprise-Umfeld von HP-UX hat.

Konzept

Bei MC/SG handelt es sich um eine Lösung, die einen sogenannten Hochverfügbarkeits-Cluster umsetzt. Hierbei läuft eine Anwendung auf einem einzelnen System und ein oder mehrere andere Systeme dienen dazu, diese Anwendung im Fehlerfall zu übernehmen bzw. von der Clusterlogik zugeordnet zu bekommen. Der Vorgang des Umschaltens wird als Failover bezeichnet bzw. als Handover, wenn er durch den Operator initiiert wurde. In allen Fällen entspricht es einem Neustart der betroffenen Anwendung, im Fehlerfall sogar mit Verlust aller aktuellen im Hauptspeicher gehaltenen Laufzeitdaten.

Das Clusterkonzept von MC/SG setzt auf sogenannten Paketen auf. Ein Paket beinhaltet eine virtuelle IP-Adresse, evtl. benötigten Plattenplatz sowie die zugehörige Anwendungen, Services genannt. In einem Paket können bis zu 900 solcher Services zusammengefasst werden. In einem Cluster wiederum können bis zu 150 Pakete verwaltet werden. Ein Paket ist auf jeweils genau einem Cluster-Knoten aktiv.

Knoten können sowohl auf SCSI basieren, als auch auf FC. Bei Verwendung von SCSI kann ein Cluster aus maximal 2 Knoten bestehen, der Einsatz von FC ermöglicht Systeme mit derzeit bis zu 16 Knoten. In allen Fällen werden die Platten als echtes Shared-Device angesprochen. Hierdurch ist es möglich neben einfachen failover-Clustern auch aktiv-aktiv-Systeme aufzubauen, um die vorhandenen Resourcen effektiver nutzen zu können.

Zur Netzanbindung werden mehrere Interfaces genutzt. MC/SG kann alle angeschlossenen Netzwerke für einen Heartbeat, welcher der Systemkontrolle und internen Kommunikation dient, nutzen. Aus Redundanzgründen heraus müssen mindestens zwei getrennte Netzwerkverbindungen zwischen den Knoten möglich sein. Darüber hinaus kann es noch weitere Netzwerke geben, die evtl. nur den Anwendungen zur Verfügung stehen.

HP-Only

MC/SG ist derzeit leider nur für HP-Hardware erhältlich. Dieses umfasst einige Systeme der Old-HP sowie aktuelle Systeme der New-HP. Auch im Bereich Storage wird nur HP-Hardware unterstützt. Neben den klassischen IA32-Servern bietet HP mit den packaged Clustern eine komplette Lösung für den Einstieg an. Hiermit wird insbesonders die Konfiguration und Integration von 2-Knoten-Clustern deutlich vereinfacht.

Die Ausrichtung auf HP-Produkte ist sicherlich historisch begründet, da die Variante für HP-UX a priori nur für HP-Hardware verfügbar sein musste. Es bleibt zu hoffen, dass die Software-Devision von HP sowohl die Chance, als auch die Bedeutung von hochwertigen HA-Lösungen für Linux mit einer stärkeren heterogenen Unterstützung fördert.

Spannweite

MC/SG stellt sich als echte Enterprise-Lösung dar, wenn es um die Spannweite an möglichen Cluster-Konfigurationen geht.

In der Konfiguration mit zwei Kupfer-SCSI Knoten wird der Einstiegsbereich abgedeckt. Auch hier sind bereits Konfigurationen mit aktiv-aktiv-Betrieb möglich oder die Kombination von mehreren Clustern. Mit FC können bereits Cluster mit bis zu 16 Knoten aufgebaut werden. Auch hier sind mehrere Cluster innerhalb einer Verwaltungsstruktur möglich. Durch den Einsatz von redundanten FC-Komponenten werden Campus-Cluster möglich.

Durch den Einsatz der *HP-Cluster-Extension* werden Metrocluster möglich. Hierbei werden die Daten auf einem zweiten Storagesystem synchron gehalten. Die Synchronisierung wird über MAN/LAN abgewickelt. Am Standort des zweiten Storagesystems wird i.d.R. ein weiterer Cluster aufgebaut. Hierdurch wird MC/SG auch höchsten Anforderungen im Bereich Disaster-Toleranz gerecht.

Heterogen

Eines der wichtigsten Entscheidungskriterien für MC/SG unter Linux ist die Möglichkeit eines heterogenen Einsatzes. So kann MC/SG nicht nur mit den Linux-Distributionen von RedHat und SuSE betrieben werden, sondern auch in Verbindung mit MC/SG unter HP-UX.

MC/SG unter Linux ist derzeit für zwei Distributionen freigegeben. Für den Einsatz mit einem Storage auf SCSI-Basis ist das RedHat in der Version 7.2. Für den Einsatz in einer Umgebung mit FC-Anbindung ist es neben dem RedHat Advanced Server der SuSE-Linux-Enterprise-Server in der Version 8.

Im Mischbetrieb zwischen HP-UX und Linux oder aber bei vorhandenen KnowHow kann MC/SG unter Linux zusätzlich punkten. Die Konfiguration und das Handling des Clusters sind unter Linux und HP-UX nahezu identisch. Lediglich das neu eingeführte Konzept eines Quorum-Servers macht einige kleine Anpassungen in der Konzeption, der Konfiguration und dem Handling notwendig. Ansonsten wird sich ein Administrator in der MC/SG Welt unter Linux sofort wieder zu Hause fühlen.

Noch stärker ist die Integration im Bereich des Operatings. Durch die (java basierte) Console können sowohl Cluster unter HP-UX, als auch Linux, als auch heterogen gemischt verwaltet werden. Somit wird dem Operating eine transparente Clusterlösung bereitgestellt.

Die Übereinsimmungen im Bereich der Administration sowie des Operatings erlauben die Weiternutzung vorhandenen KnowHows sowie eingeführter Verfahrensschemata. Gerade im Bereich Operating fällt die Notwendigkeit zur Einführung neuer Verfahrensweisen weg, so dass alle Chancen auf einen erfolgreichen Einsatz gegeben sind.

Leider ist es derzeit nur in gemischten Umgebungen möglich einen Linux-Cluster zusätzlich zur Überwachung auch per Console zu steuern und z.Bs. Pakete von Hand auf einen anderen Knoten zu verschieben. Derzeit ist das in reinen Linux-Umgebungen nur per Komandozeile möglich.

Installation

Die Installation kann von der Distributions-CD per Skript gestartet werden oder aber händisch per RPM geschehen. In der Dokumentation sind beide Wege beschrieben. Voraussetzung für die Installation ist das Vorhandensein einer der unterstützten Kernel, 2.4.9 oder 2.4.18, inkl. der zur Distribution gehörigen Sourcen. Die Installationsroutine überprüft das System und installiert evtl. notwendiege Pakete, wie z.Bs. LVM nach. Darüber hinaus werden auch neue Treiber für *bonding* und *deadman* übersetzt und installiert.

Der mit MC/SG mitgelieferte Treiber für *bonding* ermöglicht eine HA-Konfiguration der Netzwerkinterfaces. Der bei Linux standardmäßig installierte Treiber ist nur für Channel-Bonding vorgesehen und hat keine HA-Funktionalität.

Plattenplatz kann durch ein externes FC-Subsystem bereitgestellt werden oder durch SCSI-Platten in Verbindung mit aktuellen HP-Controllern bzw. durch Software-Raid auch für ältere HP-Systeme. In allen Fällen übernimmt LVM die lokale Zuordnung und Bereitstellung der Plattenresourcen.

TOC

Deadman ist ein Treiber, der ein Watchdoc-System implementiert. Hierbei wird ein Timer gesetzt, der bei Ablauf das System in einen direkten Reboot, ohne jedwede Datenspeicherung, zwingt. MC/SG setzt diesen Timer regelmässig neu, so dass im Regelfall der Timer keine Auswirkung hat. Im Falle eines Fehlers in MC/SG oder selbst im darunter liegenden Linux-System sorgt dieses Verfahren für einen eindeutigen Status. Vor allem wird verhindert, dass Daten aus dem System-Cache noch geschrieben werden können, was zu einem parallelel Zugriff von mehreren Systemen führen könnte. Ein anderer Knoten kann sodann die Applikation übernehmen. Dieser Vorgang wird als Transfer of Control, kurz TOC, bezeichnet.

Konfiguration

Die Konfiguration eines Clusters erfolgt analog zu der HP-UX-Variante. Mittels des Tools *cmquerycl* wird Kontakt aufgenommen zu allen Knoten im Cluster und die Topologie bestimmt. Zu diesem Zeitpunkt muss die Struktur des Netzwerkes bereits der produktiven Umgebung entsprechen, will man später nicht umfangreiche Anpassungen in den Konfigurationen durchführen. Wichtig ist, dass eine funktionierende Namensauflösung existiert und alle Knoten eine Sicherheitskonfiguration hinterlegt haben, die den gegenseitigen Kontakt ermöglicht. Dieses ist in der beigefügten Dokumentation ausführlich und Schritt für Schritt erläutert. Zum Schluss wird die Konfiguration mittels *cmapplyconf* auf die einzelnen Knoten verteilt.

Script-gesteuert

Die Clusterpakete selber werden durch Skripte generiert. Hierbei erzeugt das Komando *cmmakepkg* ein auf den vorhandenen Cluster angepasstes Template. Dieses Template wird sodann per Editor bearbeitet. Eingetragen werden müssen z.Bs. Plattenbereiche, IP-Adressen, Start- und Stopskripte der Anwendung. Das Skript ist grundsätzlich wohl dokumentiert, wenngleich beim ersten Kontakt eher mit rudimentären Paketen angefangen werden sollte.

Zu einem Paket gehören jeweils auch Skripte, die die Anwendung selber ausführen. Hierbei ist wichtig, dass das Ende eines Skriptes auch immer mit als Failover-Kondition angesehen wird. Bei Anwendungen, die normalerweise als Daemon gestartet werden, ist es also notwendig, eine entsprechende Überwachung mit einzubauen.

Die hinterlegten Skripte sowie die Konfiguration müssen hiernach allen Clusterknoten mitgeteilt werden. Die Skripte müssen hierzu von Hand repliziert werden, die Konfiguration wird wiederum durch *cmapplyconf* automatisch verteilt. Hierbei wird die Konfiguration zusätzlich auch auf Konsistenz geprüft.

Grundsätzlich gilt, dass alle Konfigurationen und Abläufe durch vordokumentierte (!) Shell-Skipte abgewickelt werden. Somit ist die Basis vorhanden, durchaus eigene und nicht standardisierte Konfigurationen zu erzeugen.

Quorum

Beim Clusterstart muss eine Entscheidung getroffen werden, auf welchem Knoten ein Anwendungspaket gestartet wird. Hierzu kann in der Konfiguratin ein preferierter Knoten angegeben werden. Sofern dieser Knoten nicht angegeben oder vorhanden ist, trifft diese Entscheidung das 'Quorum'. Hierzu stehen zwei Mechanismen bereit. Zum einen kann eine Entscheidung durch den Cluster selber getroffen werden, sofern *mehr* als 50% der Knoten betriebsbereit sind. Im Falle eines Clusters mit drei und mehr Knoten ist die Angelegenheit relativ einfach. Bei z.B. vier Knoten bleibt ein Quorum von 75% falls ein Knoten ausfällt und noch drei weitere lauffähig sind. Der Ausfall eines weiteren Knoten lässt das Quorum auf 66% fallen, da noch zwei von drei Knoten funktionsfähig sind. Ein weiterer Ausfall allerdings würde dazu führen, dass das Quorum auf 50% sinkt und damit unterhalb des Eindeutigkeitskriteriums fällt. Dergleichen passiert, wenn der Cluster von Anfang an bereits aus nur zwei Knoten besteht.

Quorum-Server

Sofern das Quorum nicht mehr ausreicht, kommt eine zusätzliche Instanz ins Spiel, der 'Quorum-Server'. Dieses System stellt einen speziellen Dienst bereit, bei dem sich startende Knoten melden. Im Fehler- und Startfall prüft dieser Dienst angemeldete Knoten und kann somit Auskunft darüber geben, welche Knoten betriebsbereit sind. Bei dem Quorum-Server handelt es sich um ein eigenständiges und vom Rest des Clusters unabhängiges System. Als Quorum-Server wird jeder HP/Compaq Server und PC unterstützt, für den die RedHat-Versionen 7.1 bis 7.3 sowie AS 2.1 freigegeben sind. Die Installation erfolgt mittels RPM-Paket und Eintrag in der /etc/inittab.

Unter HP-UX und auch bei anderen Hochverfügbarkeitslösungen ist dieses Quorum häufig als seperate Festplatte oder Einheit auf dem Storagesystem ausgeführt. Allerdings fehlt in diesem Falle die von der Theorie her eigentlich geforderte vollständige Unabhängigkeit des Quorums vom Cluster. Ein Quorum-Server ist in der Lage, bis zu 50 Cluster bzw. 100 Knoten zu verwalten.

SplitBrain

Der Quorum-Server spielt auch eine Rolle im Falle eines sogenannten SplitBrain. Hierbei stellt ein Knoten fest, dass ein überwachtes Paket nicht mehr lauffähig ist. Falls der überwachte Knoten nicht mehr erreichbar ist, kann dieses sowohl darauf beruhen, dass der andere Knoten ausgefallen ist, also eine failover-Situation vorliegt oder es kann sein, dass der Knoten selber vom Netzwerk getrennt wurde. Diese Trennungssituation wird SplitBrain genannt. Im Falle eines fail-overs wird immer zusätzlich die Erreichbarkeit des Quorum-Server überprüft. Sollte dieser ebenfalls nicht erreichbar sein, so geht das System von einem SplitBrain aus und führt mittels TOC einen sofortigen Restart durch, ohne nochmals auf die Plattensysteme zuzugreifen. Hierdurch wird verhindert, Daten auf eine Festplatte geschrieben werden, auf die zeitgleich ein anderer Knoten auch zugreifen könnte - nämlich der Knoten, der das Paket evtl. bereits übernommen hat.

Fazit

Mit MC/SG steht eine ausgereifte HA-Lösung bereit. Sowohl in mit HP-UX gemischten Umgebungen, als auch als standalone eingesetzt präsentiert sich ein in sich geschlossene Konzept. Mit der Bandbreite möglicher Lösungen vom einfachsten Cluster bis zum Metro-Cluster deckt MC/SG das gesamte Spektrum an HA-Lösungen ab. Durch die Historie bedingt steht auch für Linux eine Lösung bereit, die als eingeführt und ausgereift bezeichnet werden kann. Mit HP steht auch ein Hersteller hinter dem Produkt, der über die für Support und Weiterentwicklung notwendigen Resourcen verfügt.

probusiness group

Die probusiness group ist plattformübergreifender IT -Dienstleister und steht ihren Kunden von der Konzeptphase über die Beschaffung bis zum Betrieb zur Seite. Das Leistungsportfolio von probusiness reicht von IT-Consulting über Projektplanung bis zur Erstellung von maßgeschneiderten Lösungen zur Umsetzung neuer strategischer IT -Konzepte.