Hochverfügbare Virtualisierung mit KVM, DRBD und ArchivistaVM

Ein Cluster-Konzept für KMU-Unternehmen und Private

Contents

1 Einleitung **2**

2 Einichten des Clusters **3**
- 2.1 Darum genügen 10 bis 15 Minuten nicht 3
- 2.2 Cluster-Aufbau von Hand (Stand 7.11.2011) 3

3 Hardware für preisgünstige Cluster **5**
- 3.1 SwissRocket-Cluster 5
- 3.2 SwissRocket-Budget-Konzept 7
- 3.3 Raid-Kontroller und Racks . . 8
- 3.4 Abenteuer SSD 8

4 Software und mehr **10**
- 4.1 Begriff Hochverfügbarkeit . . 10
- 4.2 Virtualisierung mit KVM . . 11
- 4.3 Plattenspiegelung mit DRBD 12
- 4.4 Management mit ArchivistaVM 12

5 Details zur Technik **14**
- 5.1 So funktioniert die Demo . . 14
- 5.2 Netzwerkarten und MAC-Adressen 14
- 5.3 Crash-Kurs drbdadm 15
- 5.4 Switch-Over 15
- 5.5 Fail-Over 16

6 Abschliessende Bemerkungen **17**
- 6.1 Ausblick für ArchivistaVM . . 17
- 6.2 Über den Autor und die ArchivistaBox 17
- 6.3 Copyright-Hinweise 18

1 Einleitung

Der Vortrag stellt die Open Source Lösung Archivista SwissRocket vor, mit der hochverfügbare Virtualisierung in 10 bis 15 Minuten aufgebaut werden kann. Dabei werden die Komponenten KVM, DRDB und ArchivistaVM vorgestellt. Der Vortrag ist in zwei Teile unterteilt. Zunächst wird ein Cluster aus der Sicht des Anwenders aufgebaut. Im zweiten Teil erfolgt ein Blick hinter die Kulissen der Komponenten. Gemachte Erfahrungen (positive wie negative) bei realisierten Projekten runden den Vortrag ab.

Soweit der eingereichte Text für den diesjährigen Vortrag. Die schelchte Nachricht gleich vorweg, die Lösung ist leider nicht fertig geworden. Die noch schlechtere Nachricht lautet, die Lösung wird wohl nie fertiggestellt sein. Aber, wie ja allgemein bekannt ist, dies trifft wohl auf alle Software-Projekte zu.

Und nun zur besseren bzw. guten Nachricht. Der Titel des Vortrages ist nicht mehr aktuell. Doch alles schön der Reihe nach... Zunächst soll aufgezeigt werden, wie ein Cluster eingerichtet wird. Dafür benötigen wir ArchivistaVM in einem aktuellen Release. Für all jene, die nicht am Vortrag teilnehmen, hier die Demo dazu:

www.archivista.ch/avcluster.gif

Die Demo zeigt auf, wie ein Cluster (3 Instanzen) komplett automatisiert in ca. 100 Sekunden aufgebaut werden kann. So gesehen darf von einem Cluster in 100 Sekunden und nicht in 10 bis 15 Minuten die Rede sein.

2 Einichten des Clusters

Dass für die Demo eine virtualisierte Umgebung verwendet wurde, sei an dieser Stelle nicht verschwiegen. Dies einmal, weil ansonsten drei Rechner hätten mitgeschleppt werden müssen, und weiter weil es relativ schwierig ist, auf einem Screen das Aufsetzen dreier Rechner zu präsentieren.

Verwendet wurde eine ArchivistaVM-Box-Universal mit ca. 300 MByte-Durchsatz (6 Festplatten mit Raid10). Dies entspriecht etwa dem Druchsatz dreier ArchivistaVM-Budget-Rechnern bzw. dem weiter unten vorgestellten Budget-Cluster-Konzept für den Eigenbau.

2.1 Darum genügen 10 bis 15 Minuten nicht

Beim Einreichen des Vortrages bestanden bei Kunden und bei uns einige ArchivistaVM-Cluster. Nachdem der erste produktive Verbund von Hand ca. 8 Stunden Zeit in Anspruch nahm, wurden Skripte realisiert, mit der ein Cluster in 10 bis 15 Minuten aufgesetzt werden konnte. Der ursprüngliche Plan für diesen Vortrag bestand darin, diesen Vorgang anlässlich eines Vortrages durchzuspielen und dabei die Technologie dahinter vorzustellen.

An sich sind 10 bis 15 Minuten nicht viel Zeit für das Aufsetzen dreier Server. Immehin werden dabei die Festplatten auf jeweils einen zweiten Rechner gespiegelt. Derartige Installation gelten ganz klar nicht als 08-15-Job eines jeden Informatikers, geschweige denn eines/r Anweders/in.

Als der Vortragstermin näher rückte, stellte sich die Frage, ob neben den 15 Minuten für das Aufsetzen des Clusters genügend Zeit für die technischen Erläuterungen vorhanden sein würde? Die Problematik bei den Skripten war, dass sie ohne fundierte Kenntnisse nicht ablaufen, oder noch schlimmer, bestehende Festplatten zerstören, sofern die falschen Parameter mitangegeben werden. Natürlich hätte genau dies im Rahmen eines Vortrages erklärt werden können.

2.2 Cluster-Aufbau von Hand (Stand 7.11.2011)

Der Vollständigkeit halber sei hier diese Kurzanleitung angeführt, mit der von Hand ein ArchivistaVM-Cluster aufgesetzt werden kann:

```
a) Annahme Rechner 192.168.0.141, 192.168.0.142, 192.168.0.143
b) Rechner 192.168.0.141:
   pveca -c
c) Rechner 192.168.0.142 und 192.168.0.143:
```

```
   pveca -a -h 192.168.0.141
d) Netzwerk einrichten (alle drei Rechner):
   perl /home/cvs/archivista/jobs/cluster.pl
e) Zweite Festplatte einrichten (alle Rechner):
   perl /home/cvs/archivista/jobs/clustermd.pl 1 1 b
f) 1. Verbund (Primary): 192.168.0.141:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r1 1
g) 1. Verbund (Secondary): 192.168.0.143:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r1
h) 2. Verbund (Primary): 192.168.0.142:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r2 1
i) 2. Verbund (Secondary): 192.168.0.141:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r2
k) 3. Verbund( Primary): 192.168.0.143:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r3 1
l) 3. Verbund (Secondary): 192.168.0.142:
   perl /home/cvs/archivista/jobs/clusterdrbd.pl r3
```

Diese Anleitung hat ein Kunde von uns am 7. November 2011 erhalten; für den Fall eines Falles, dass der ausgelieferte Cluster erneut aufgesetzt werden muss. Dabei gilt es folgendes zu beachten. Die Anleitung setzt voraus, dass zuvor die drei Rechner korrekt (von Hand) aufgesetzt werden. Danach ist auf einer root-Konsole auf dem jeweils richtigen Rechner der richtige Befehl abzusetzen.

Der Autor bemerkte Mitte November, dass (sofern ausführliche Erklärungen beim Aufsetzen erfolgen sollten) das Ziel in 45 Minuten wohl zu ambitioniert sein würde. Entweder musste der Vortrag technisch betrachtet sehr kurz ausfallen, oder dann musste das Einrichten des Clusters einfacher werden. Das Resultat der Demo zeigt, dass die letztere Variante gewählt wurde. Bevor die Lösung detaillierter vorgestellt wird, zunächst einige Ausführungen zur Hardware.

3 Hardware für preisgünstige Cluster

An sich ist ein Server ein Server und damit ein Computer. Dennoch sind die Anforderungen an die Hardware nicht zu unterschätzen. Natürlich könnnen mit der Virtualisierung CPUs und die Festplatten geteilt werden; dies gilt aber auch im umgekehrten (eher negativen) Sinne. Pro aufgesetzter Instanz 'knabbert' die Virtualisierung ganz schön die Festplatten an. Bei vier Instanzen kriegt jeder Gast nur noch einen Viertel, wenn alle gleichzeitg auf die Platte zugreifen möchten. Und doch, konzeptionall kann bei der Wahl der Hardware sehr viel herausgeholt werden, wie das nachfolgende SwissRocket-Cluster aufzeigt.

3.1 SwissRocket-Cluster

Die Vorteile des SwissRocket-Cluster-Konzeptes sind ausführlich beschrieben unter:

www.archivista.ch/de/pages/aktuell-blog/swissrocket-cluster.php

In einem Satz gesagt geht es darum, dass bei kleineren Cluster-Verbünden (2 bis 7 Server) einerseits die Festplatten über die Maschinen gespiegelt werden und weiter die Switches nicht redundant ausgelegt werden, indem direkte Crossover-Kabel zwischen allen Knoten aufgebaut werden.

Die linke Abbildung zeigt das in der Praxis meistens realisierte Konzept in KMU-Unternehmungen, die rechte Abbildung das Konzept von ArchivistaVM SwissRocket. Beim Cluster-Konzept von ArchivistaVM wird vermieden, dass die Switches redundant aufgebaut werden müssen (ansonsten beim Ausfall des Switches alle Rechner stillstehen würden). Dies gilt es zu vermeiden. Einmal ist der Aufbau von Clustern, die mit redundanten Switches arbeiten, nicht ganz einfach, weiter kosten 10-GBit-Switches auch heute noch gut und gerne einen fünfstelligen Betrag pro Switch, d.h. ca. 20000 bis 30000 Euro oder Franken (so weit liegen die ja nicht mehr auseinander) sind alleine für redundante Switches zu kalkulieren.

Denn eines muss für schnelle Server für die Virtualisierung beachtet werden: Neben guten CPUs sind schnelle Disk-Verbünde und mindestens gleichschnelle Netzwerkkarten aufzusetzen, ansonsten Cluster-Verbünde wenig Sinn ergeben. Dazu ein Beispiel, bei sechs Platten mit RAID10 können ca. 300 MByte Durchsatz pro Sekunde erreicht werden. Würde nun mit 1-GBit-Karten gearbeitet, so könnte der Inhalt der Festplatten nur mit ca. 100 MByte auf die Festplatte des zweiten Rechners übertragen werden. Damit würde der 300-MByte Festplattenverbund mit maximal 100 MByte gemächlich vor sich hindümpeln. Aus diesem Grunde müssen 10-GBit-Netzwerkkarten verwendet werden, erst damit lassen sich (bis ca. 900 MByte pro Sekunde) an Daten auf einen zweiten Rechner übertragen.

Dass dies an dieser Stelle explizit ausgeführt wird, hat seine Gründe. Der Autor hat bei Kunden viele hochpreisige Lösungen namhafter Hersteller angetoffen. Alle Verbünde 'krankten' daran, dass die 10-GBit-Karten nicht verbaut wurden. Und damit wir uns richtig verstehen, die Lösungen lagen teilweise durchaus bereits im sechsstelligen Franken/Euro-Bereich. Die entsprechenden Cluster-Landschaften bestanden im übrigen allesamt aus Verbünden zwischen drei und fünf Servern (bis 48 CPUs).

Die hier aufgezeigte Lösung beinhaltet zwischen zwei und sieben Servern. Das mag zunächst als wenig erscheinen. Da aber nicht sowohl redundante Maschinen für die Instanzen (RAM) und Speicher (Festplatte) aufgebaut werden, muss maximal 1 Maschine für die Ausfallsicherheit 'geopfert' werden. Auch dazu ein Beispiel: Bei 7 Servern mit je 24 CPUs können von 166 Prozessoren immerhin deren 144 verwendet werden.

Und falls dies noch immer nicht reicht, eine Verdoppelung auf 288 Cores könnte mit 4-Sockel-Boards (anstelle der verwendeten 2-Sockel-Boards) erreicht werden. Allerdings dürften in dieser Grössenordnung zwei redundante 10-GBit-Switch nicht mehr einen ununberwindbaren Budget-Posten darstellen.

3.2 SwissRocket-Budget-Konzept

Ein SwissRocket-Cluster bei drei Knoten schlägt mit ca. 12000 Franken zu Buche. Dabei stehen 36 CPUs (24 einsetzbar) zur Verfügung. Im Rahmen eines Kunden-Projektes fragte ein Kunde nach, ob es nicht kostengünstiger ginge.

Der Kunde wollte einerseits zwar Redundanz bei den Festplatten haben, nicht aber derart tief in die Tasche greifen müssen. Daraus ist die Idee entstanden, preisgünstige Cluster ohne 10-GBit-Netzwerkkarten aufzubauen. Ein solcher Budget-Cluster dürfte in der Reichweite eines jeden Budgets (selbst für Private) liegen. Wer bereit ist, die benötigen Komponenten selber zusammenzutragen, kann für sehr wenig Geld in die Cluster-Landschaft einsteigen. Im einzelnen werden benötigt.

- 3xGehäuse mit Platz für zwei Platten. Vier Platten bringen nichts, da die 1-GBit-Netzwerkkarten den Durchsatz nicht erbringen werden.
- 3xBoard (mATX oder ATX). Benötigt werden minimal zwei Steckplättze PCI-x1. PCI-x4 oder höher ginge auch, nicht aber (weil zu langsam) PCI.
- 3xCPU. QuadCore oder höher. Günstige AMD-CPUs reichen durchaus. Weiter kann bei AMD 'fehlertolerantes' ECC-Ram (unbuffered bzw. unregistered) verwendet werden.
- 6 Intel-1-GBit-Netzwerkkarten. Intel-PCIx-Karten sind nicht extrem teuer, bieten aber einen Durchsatz nahe bei 100 MByte/Sekunde, bei mITX-Boards müssten weit teurere Dual/Quad-Port 1GBit-Karten verwendet werden.
- 6 SATA-Festplatten. Auch wenn die Plattenpreise im Moment hoch angesiedelt sind, es sollte dennoch zu SATA-Platten mit 7200-Umdrehungen/Minute gegriffen werden.

Hinzugefügt seien zwei Punkte: Erstens ist hot-swap nicht zwingend notwendig; das Konzept sieht ja vor, dass ein Knoten (Rechner) 'abrauchen' darf und zweitens möchte der Autor an dieser Stelle gesagt haben, dass wer nach detaillierter Hardware nachfragen wird, über kurz oder lang eine Offerte für die ArchivistaVM-Budget-Server in den Händen halten wird. Nicht bewährt hat sich das Herausgeben der detaillierten Komponenten innerhalb eines Beratungsgespräches. Auch kann im Rahmen eines Open Source Projektes keine kostenfreie Hotline für das Zusammenbauen allfälliger erwähnter Komponenten betrieben werden. Dies sollte an sich selbstverständlich sein, die Praxis hat gelehrt, dass dem leider nicht so ist.

3.3 Raid-Kontroller und Racks

Wer frühere Vorträge zu ArchivistaVM durchstöbert, der findet einerseits die Empfehlung des Autors, dass Racks nicht verwendet werden sollten und weiter wird der Einsatz von Hardware-Raid-Kontrollern mit Batterie-Modul empfohlen.

Nun gibt der Autor gerade die umgekehrte Empfehlung ab. Einerseitz können mit ArchivistaVM mittlerweile bis zu 26 Platten ohne Hardware-Raid-Kontroller verwendet werden (in einem Desktop-Gehäuse geht dies schwerlich) und weiter ergibt der Einsatz von Racks ab einer gewissen Cluster-Grössenordnung aufgrund der Möglichkeit, die Maschinen entsprechend ausbauen zu können (insbesondere beim RAM), durchaus seinen Sinn.

Hat der Autor nun 'Kreide' gefressen? Sagen wir es so, Rack-Server müssen nicht partout einen hohen Stromverbrauch zur Folge haben. Viele Server-CPUs sind mittlerweile äusserst sparsam. Dies gilt aber nach wie vor nicht bei SAS-Festplatten (auch nicht bei SATA-Platten jenseits von 7200 Umdrehungen/Minute), die üblicherweise in Server-Gehäuse verbaut werden.

Die Empfehlung zu Hardware-Raid-Kontrollern lag darin begründet, dass ArchivistaVM bzw. die ArchivistaBox ganz einfach nicht in der Lage war, Software-RAID überhaupt anzusprechen und dass daher keine Lösungen über 2 TByte ohne Hardware-Raid-Kontroller machbar waren. Und anstelle von Betterie-Modulen für die Kontroller können/sollten USV-Geräte (Batterien für die Server) eingesetzt werden; damit ein Verbund (bzw. zumindest die Gast-Systeme) bei einem drohenden Stromunterbruch sauber heruntergefahren werden kann.

3.4 Abenteuer SSD

Dass ArchivistaVM mittlerweile bis zu 26 Platten per Software-Raid ansprechen kann, liegt daran, dass der Autor der Werbung für schnelle PCI-basierte Solid-State-Platten (SSD) auf den Leim gekrochen ist. Weil diese bis zu 1 GByte Druchsatz versprachen, bestellte er sich zwei Exemplare unter den letztjährigen Tannenbaum.

In der Folge stellte sich heraus, dass die PCI-Platten Fake-Raid-Festplattenverbünde waren, und nicht unter Linux unterstützt wurden (selbst ein Debian oder Ubuntu) liess sich um keinen Preis installieren). Immer wurden vier einzelne Platten angezeigt. Der Autor hat folglich Silvester ohne das neue SSD-Wunderfeuerwerk gefeiert. Zwar konnte er die Treiber einkompilieren, aber stabil liefen sie (Stand Dezember 2011) dennoch nicht. Ein RAID10 z.B. liess sich nicht einrichten, ein RAID0 schien (nicht ganz unberechtigt) zu riskant, RAID5 lief überhaupt nicht und bei RAID1 waren danach immer noch zwei Platten vorhanden.

Erst mit dem Rückgriff zu Software-RAID konnten die Platten zu einem Verbund mit RAID10 zusammengefügt werden. Der Autor fragte sich dabei, wie es wohl wäre, eine gespiegelte Platte bei einem Defekt auszutauschen. Bei teuren PCI-SSD-Karten könnten die RAM-Riegel einzeln ersetzt werden; allerdings nicht im laufenden Betrieb. Folglich wurden 10 SSD-2.5-Zoll-Platten zum Test bestellt.

Der Autor freute sich bereits daran, mit extrem hohem Speed ins SSD-Zeitalter zu fliegen, ehe er hart vom Januar-Kater getroffen wurden. Innerhalb zweier Tage waren vier von 10 SSD-Platten hinüber. Und wenn an dieser Stelle die Aussage erfolgt, derzeit auf SSD-Platten zu verzichten, so sei dies mit dem Hinweis verbunden, dass dies ja nicht für alle Zeit so bleiben muss; zu hoffen wäre es.

4 Software und mehr

Im Titel ist die Rede von Hochverfügbarkeit, von KVM, von DRBD und von ArchivistaVM. Folglich seien diese Technologien hier einzeln vorgestellt.

4.1 Begriff Hochverfügbarkeit

Virtualisierung und Hochverfügbarkeit sind in aller Munde. Der Autor hat viele viele hochverfügbare Systeme in der Praxis angetroffen (leider viele nicht mit ArchivistaVM). Alle wurden in irgendeiner Weise hochverfügbar bzw. ausfallsicher aufgebaut.

Wenn jewels nachgefragt wurde, ob ein Ausfall je getestet wude (z.B. durch Ziehen des Steckers), so wurde von den Verantwortlichen ausgeführt, dass sie dies lieber nicht simulieren wollten. Einzig ein ArchivistaVM-Kunde simulierte den Ausfall, indem er zwei Festplatten herauszog. Dummerweise waren es die falschen zwei Platten bei seinem RAID10-System.

Weil damals noch überhaupt keine Skripte für das Aufsetzten des ArchivistaVM-Clusters bestanden, durfte/musste der Autor die gesamte Arbeit (beinahe den gesamten Tag) von Hand nochmals durchführen. In diesem Sinne bedeutet Hochverfügbarkeit nicht einfach nur Ausfallsicherheit, sondern auch Denken in Varianten. D.h. für den Fall des Totalausfalles ist es immer gut bzw. besser, eine Lösung automatisiert aufsetzen zu können. Wer einfach eine hochverfügbare Lösung im KMU-Umfeld in Betrieb nimmt, ohne Ausfallszenarien durchgespielt zu haben, der darf sich nicht wundern, wenn er (schlimmer noch der Dienstleister) am Tag-X mit der Situation überfordert ist.

Der Autor hat beim Schreiben des Vortrages einen Selbstversuch unternommen und am Vorabend zum Vortrag den Stecker bei der für den Vortrag vorgesehenen Maschine gezogen. ArchivistaVM konnte problemlos wieder hochgefahren werden, doch mit einer VM-Datei gab es Probleme, der klassische Fall eines Problemes im Ernstfall. Die Frage stellte sich, Lösung suchen oder den Vortrag zu Ende schreiben? Der Autor hat sich für den Vortrag entschieden; dies wäre schwieriger gewesen, wenn der Vortrag in der 'beschädigten' Image-Datei gelegen hätte.

Was hätte der Autor im Ernstfall gemacht? Datei reparieren, Datensicherung zurückspielen? Wie lange dauert dies? Der Autor hätte wohl parallel versucht, die Datei zu reparieren und gleichzeitig die Datensicherung zurückzuspielen. Bei 2*500 GByte bzw. 1 TByte) und bei 200 MByte-Druchsatz wäre folglich nach ca. 40 Minuten (bei weniger Daten natürlich entsprechend früher) klar gewesen, ob die Variante Reparatur oder das Rückspielen der Datensicherung effizienter gewesen wäre.

Es darf die Frage gestellt werden, wo bleibt hier der Cluster? Natürlich sollte es bei einem hochverfügbaren Cluster nicht notwendig sein, die Datensicherung zurückspielen zu müssen. weil alle Daten (Gastsysteme) ja bereits in doppelter Ausführung auf laufenden Servern vorhanden sind. Aber, auch bei einem noch so hochverfügbaren Cluster kann es zu einem Totalausfall kommen.

Der Autor stellt immer wieder fest, dass ein Totalausfall der gesamten Server-Landschaft bei KMU-Unternehmen heute kaum mehr in Betracht gezogen wird. Der Lieferant habe ja die Hochverfügbarkeit (inkl. Wartungsvertrag) zugesichert. Es fragt sich einfach welche? Was bedeutet eine Verfügbarkeit von 99,99 Prozent? Gerne verweist der Autor auf die Erklärungen bei wikipedia.de bzw. dort bei Hochverfügbarkeit:

- **Verfügbarkeitsklasse 2:** 99 Prozent = 438 Minuten/Monat bzw. 7:18:18 Stunden/Monat = 87,7 Stunden/Jahr, d.h. 3 Tage und 15:39:36 h.
- **Verfügbarkeitsklasse 3:** 99,9 Prozent = 43:48 min/Monat oder 8:45:58 Stunden/Jahr.
- **Verfügbarkeitsklasse 4:** 99,99 Prozent = 4:23 Minuten/Monat oder 52:36 Minuten/Jahr
- **Verfügbarkeitsklasse 5:** 99,999 Prozent = 26,3 Sekunden/Monat oder 5:16 Minuten/Jahr
- **Verfügbarkeitsklasse 6:** 99,9999 Prozent = 2,63 Sekunden/Monat oder 31,6 Sekunden/Jahr

Die Nummer der Stufe entspricht der Anzahl der 9-er (z.B. Stufe 3 bei 99,9 Prozent). Ob bei 3*9 oder 4*9 von Hochverfügbarkeit gesprochen werden kann, darüber gehen die Meinungen auseinander. Für unsere Cluster-Lösung ArchivistaVM heisst dies, dass selbst bei einem Totalausfall aller Knoten die Stufe 4 erreichbar ist. Dies deshalb, weil der gesamte Cluster in weniger als 2 Minuten neu aufgesetzt werden kann und das Backup (auch bei 1 TByte) in weniger als 40 Minuten zurückgespielt werden kann. Beim Ausfall eines Knotens ist selbst Stufe 5 durchaus erreichbar.

Der Autor vertritt den Standpunkt, dass viele (wohl fast alle) Cluster-Lösungen (insbesondere bei KMU-Unternehmen) bereits von Beginn weg nicht hochverfügbar sind, weil der Ausfall aller Rechner nicht abgefangen werden kann. So gesehen bietet ein Setup von 1 bis 2 Minuten bei ArchivistaVM oder ein vollautomtisiertes Aufsetzen eines Clusters in der gleichen Zeit viel Sicherheit und Komfort für den Fall der Fälle.

4.2 Virtualisierung mit KVM

ArchivistaVM arbeitet mit KVM, genauer derzeit mit KVM 0.15.1. Was heisst das? An sich würde erwartet werden, dass KVM ja KVM ist und bleibt. Leider nicht ganz. Bei einer neuen

KVM-Version gibt es gut und gerne viele tollen Features, aber leider auch Dinge, die weniger toll sind.

Positiv zu vermerken bei KVM 0.15.1 ist: Verschachtelte Virtualisierung mit AMD-Prozessoren (bald auch mit Intel-CPUs), Lauffähigkeit alter 32-Bit-Betriebssysteme wie Windows9x, Emulierung wichtiger CPU-Merkmale wie Sockets, Cores und Threads. Weniger erfreulich war, dass die Default-Einstellungen mittlerweile für den Einstieg unbrauchbar sind. Die Disk-Formate qcow2 und raw bewirken neuerdings mi den Default-Einstellungen (cache=none) einen allesamt grottenschlechten Speed. Erst mit 'writeback' beim raw-Format und 'unsafe' beim qcow2-Format geht die Post im vergleichbaren Speed zu frühreren Versionen ab. Ob solche Änderungen bei den Default-Werten Sinn ergeben, diese Frage wäre durch die Entwickler bzw. das Marketing beim Hersteller von KVM zu beantworten.

4.3 Plattenspiegelung mit DRBD

Herzstück eines jeden ArchivistaVM-Clusters bildet DRBD. Diese Technologie stellt sicher, dass der Inhalt einer Festplatte immer auch auf eine Festplatte eines zweiten Rechners gesichert werden kann. Die Tecnnologie stammt von der österreichischen Firma LINBIT in Wien, die entsprechenden Module sind seit dem Kernel 2.6.33 fix im Kernel enthalten, d.h. es müssen keine Patches mehr eingespielt werden.

Dennoch gilt es zu beachten, dass DRBD aus zwei Teilen besteht. Einmal aus einen Kernel-Modul und weiter aus Tools, die für das Management eingesetzt werden. Nicht bewährt hat sich, verschiedene Versionen beim Kernel-Modul und bei den Management-Programmen zu verwenden. Es könnte dabei zu Performanz-Problemen kommen. Auf der ArchivistaBox findet sich dazu das Programm elevatordrbd.pl (Verzeichnis /home/cvs/archivista/jobs).

4.4 Management mit ArchivistaVM

Optisch kommt das Management-Tool von ArchivistaVM nunmehr seit mehr als zwei Jahren praktisch unverändert daher. Neben Kleinigkeiten (Übergabe beliebiger optionaler Parameter bei den Optionen) wurden vorallem die Parameter 'writeback' bzw. 'unsafe' hinzugefügt. Ebenfalls können bei den Netzwerkgeräten neu MAC-Adressen zugewiesen werden (unter Hardware, neue Netzwerkkarte hinzufügen).

Die Fortschritte an ArchivistaVM liegen uner der Haube. Nachfolgend seien die wichtigsten Neuerungen im letzten Jahr aufgezählt:

- Unterstützung aller Netzwekkarten (inkl. Karten mit Closed-Source-Treibern)

- Software-Raids mit bis zu 26 Festplatten. Im Regelfall werden dabei RAID10-Verbünde aufgebaut. Über die Kommando-Zeile können aber auch andere RAID-Level (z.B. RAID0) eingerichtet werden.

- Anschluss von USV-Geräten des Herstellers APC. ArchivistaVM erhält dabei Signale, sobald ein Stromausfall droht. Dabei können mit einem Hilfsprogramm die Instanzen oder die Server geordnet heruntergefahren werden.

- Verfügbarkeit von DRBD für VM-Instanzen. DRBD kann ab November 2011 automatisiert aufgesetzt werden. Die Skripte dazu finden sich unter /home/cvs/archivista/jobs.

- Datensicherung mit Cluster-Modus: Bei DRBD-basierten Clustern kann eine Maschine kurz gestoppt werden, die zweite DRBD-Instanz wird dabei ausgeklinkt, die erste Instanz arbeitet weiter, während auf der zweiten Instanz die Datensicherung in aller Ruhe abgearbeitet werden kann. Nach der Sicherung gleicht sich die zweite Instanz mit der ersten automatisch wieder ab.

- Switch-Over-Skripte, um Instanzen bequem auf dem zweiten Rechner hochzufahren und gleichzeitig dafür zu sorgen, dass die Instanzen nicht mehr auf dem ursprünglichen Rechner gestartet werden.

- Fail-Over-Skripte. Wie gesagt, für diesen Teil hat die Zeit bis zum Vortrag leider nicht mehr ganz gereicht, die Konzepte dazu stehen aber, sodass es einzig eine Frage der Zeit sein wird, bis Fail-Over vorhanden sein wird.

- Festplatten können in beliebiger Reihenfolge zwischen zwei Systemen ausgetauscht bzw. im gleichen System wieder eingeschoben werden.

Die obenstehende Liste erhebt keinen Anspruch auf Vollständigkeit, wohl aber auf die wichtigsten realisierten Punkte.

5 Details zur Technik

5.1 So funktioniert die Demo

Die Demo-Skripte zum Aufsetzen eines Clusters finden sich in der Datei scripts.tgz unter /home/cvs/archivista/jobs.

Diese Datei sollte zunächst nach /var/lib/vz/template/iso kopiert werden. Ebenso sollte die aktuelle ISO-Datei ab unserer www.archivista.ch heruntergeladen werden. Sie ist dort unter avtest1.iso abzulegen. Daran anschliessend können wir die Dateien `create_cluster.pl` und `create_clusternew.pl` bearbeiten. Bei `create_clusternw.pl` gilt es zu beachten, dass dabei bei den Instanzen 101, 102 und 103 die Image-Dateien beim Ausführen des Programmes gelöscht werden.

Allenfalls müssen auch noch isolinux.141, isolinux.142 und isolinux.143 bearbeitet werden; dies insbesondere dann, wenn nicht mit den IP-Adressen 192.168.2.141-143 gearbeitet werden soll. Ebenfalls ein Blick sollte in die conf-Dateien für die Instanzen geworfen werden.

Wenn die Datei `create_cluster.pl` ausgeführt wird, werden drei ISO-Dateien erstelt. Diese korrespondieren zu den Instanzen 101, 102 und 103. Mit `qm start 101`, `qm start 102` und `qm start 103` kann der Cluster angeworfen werden. Nach ca. 2 Minuten wid der fixfertige Cluster zur Vefügung stehen.

5.2 Netzwerkarten und MAC-Adressen

Der Autor wollte das Aufsetzen des Clusters virtualisiert zeigen, weil es ansonsten über einen Beamer nicht ganz so einfach gewesen wäre, gleichzeitig drei Bildschirme im richtigen Moment einzublenden.

Dieses Vorhaben sollte sich als weit schwieriger als angenommen präsentieren. Hauptfrage: Wie kann in virtualisierten Umgebungen eine Crossover-Kabel simuliert werden? Nebenfrage: Müssen es denn wirklich mehrere Karten sein bzw. wo liegt das Problem mit mehreren Netzwerkkarten?

Ja, es sind mehrere Karten einzurichten, nur so kann der Cluster überhaupt ohne redundante Switches betrieben werden. Das Problem mehrerer Netzwerkkarten liegt darin, dass diese nicht immer unter der gleichen PCI-Adresse gegenüber dem Linux-Kernel erscheinen. Kurz und schlecht, es kann passieren, dass die Karte, die für eth0 bestimmt ist, sich plätzlich unter eth1 oder eth2 meldet. Dies hat zur Folge, dass die gesamte Maschine nicht mehr von aussen erreicht werden kann.

Die Problematik besteht sowohl bei physikalischen als auch virtualisierten Umgebungen. Ursprünglich hatte ArchivistaVM eine sogenannte udev-Regel, damit die Netzwerkarten beim ersten Hochfahren bzw. bei weiteren Starts immer die gleiche Adresse zugewiesen bekamen. Diese Regel hatte aber das Problem, dass bei einem Austausch der Platten zu einem anderen Rechner sämtliche Netzwerkkarten aufgrund der udev-Regel gar nicht mehr erreichbar waren. Selbst bei einer einzigen Netzwerkkarte verhinderte die in der udev-Regel hinterlegte Mac-Adresse, dass der Server von aussen erreichbar war.

5.3 Crash-Kurs drbdadm

Die bei ArchivistaVM verwendeten DRBD-Verbünde sind auf allen Maschinen gleich aufgebaut. Die erste Maschine enthält die Kennung 'r1', die zweite 'r2' und die dritte Maschine 'r3' als primären Knoten (/dev/drbd0). Der primäre Knoten ist immer zum Arbeiten zu verwenden.

Gelichzeitig hält die zweite Maschine eine Kopie der ersten Maschine vor, die dritte Maschine jene der zweiten und die letzte Instanz enthält eine Kopie der ersten Instanz. Diese werden als secondary-Instanzen geführt (/dev/drbd1).

Eine Einführung in DRBD würde den Rahmen dieses Vortrages bei weitem sprengen, dennoch sollten hier in einer Art Crash-Kurs die wichtigsten Kommandos aufgeführt werden:

```
cat /proc/drbd
drbdadm down r1
drbdadm up r1
drbdadm primary r1
mount /dev/drbd0 /var/lib/vz
umount /var/lib/vz
drbdadm down r1
```

Wichtig zu wissen ist, dass erst nachdem auf einem Knoten eine DRBD-Instanz mit 'drbdadm primary rx' (x steht für die Nummer des Knoten wie z.B. 1,2 oder 3) die Platte formatiert und anschliessend zum Einsatz kommen kann.

Grundsätzlich kann jederzeit ein Wechsel beim primären Zustand auf beiden Disks erfolgen. Nicht vorgesehen (zumindest bei ArchivistaVM) ist einzig, dass beide Instanzen gleichzeitg den Status primary erhalten können.

5.4 Switch-Over

Bei Switch-Over geht es darum, dass ein Knoten mit möglichst wenig Aufwand ausgetauscht werden kann. Der Vorgang läuft dabei nicht vollautomatisiert ab, er kann aber über die

Fernwartung, d.h. ohne einen Einsatz vor Ort, durchgeführt werden. Dies ist aktuell (Stand November 2011) bei ArchivistaVM implementiert.

Unter /home/cvs/archivsita/jobs finden sich die beiden Programme `clusternodedown.pl` sowie `clsutersecondary.pl`. Das erste Programm schaltet den aktuellen Knoten aus. Mit dem zweiten Programm kann auf einem Knoten von primär auf secondary gewechselt werden. Dabei verschwinden die primären Instanzen (/dev/drbd0) und es werden die sekundären Instanzen (/dev/drbd1) eingebunden und gestartet, sofern der entsprechende Flag für das automatische Booten gesetzt ist.

5.5 Fail-Over

Bei Fail-Over entscheidet der Rechnerverbund ohne Interaktion eines Administratoren, was zu tun ist, wenn ein Knoten ausfällt. Die Schwierigkeit bei Fail-Over besteht darin, dass nur jene Szenarion funktionieren werden, die vor einem Ausfall eines Knotens auch angedacht bzw. implementiert wurden.

Wie bereits eingangs zum Vortrag genannt, konnte dieser Punkt nicht mehr fertiggestellt werden. Zwar sind die Arbeiten dazu realtiv weit fortgeschritten. Konzeptionell werden sich die (minimal drei) Rechner selber überwachen. Fällt ein Knoten aus, springt die Ersatzmaschine für die ausgefallene Maschine ein. Ebenfalls geplant ist, dass der ausgefallene Knoten im laufenden Betrieb wieder in den Cluster eingebunden werden kann.

6 Abschliessende Bemerkungen

6.1 Ausblick für ArchivistaVM

Das letzte Jahr hat viele Neuerungen bei ArchivistaVM gebracht. Kunden der DMS-Produkte sind zuweilen der Ansicht, die DMS-Lösung der ArchivistaBox käme zu kurz. Dies ist in zweifacher Hinsicht unrichtig. Einmal enthält eine jede DMS-Lösung die Virtualisierung als kostenfreie Zugabe und weiter können sämtliche Features von ArchivistaVM direkt auch von ArchivistaDMS verwendet werden.

Was das nächste Jahr bei der Entwicklung von ArchivistaVM bringen wird, das kann heute (mit der Ausnahme Fail-Over) noch nicht gesagt werden. Mit Bestimmtheit gesagt werden kann einzig, dass ein radikaler Umbau beim Interface nicht geplant ist. Der Autor, aber auch die ArchivistaVM-Kunden erachten das Interface als sehr intiutiv und einfach. Bei all den tollen Features für den Aufbau und Betrieb von Clustern darf nicht vergessen werden, dass ein einfaches Interface (mit oder ohne Cluster) fundamental zentral ist, wenn für KMU-Unternehmen (aber auch Private) Virtualisierung machbar sein soll.

6.2 Über den Autor und die ArchivistaBox

Urs Pfister kaufte sich mit 16 und dem ersten Lohn einen CPC464, gründete mit 30 die eigene Firma, und ist kurz darauf zu Linux hinübergeschwenkt. Seit dieser Zeit gilt seine Leidenschaft Open Source und allem, was dazugehört. Zur Zeit beschäftigt ihn die Weiterentwicklung der ArchivistaBox.

Die ArchivistaBox ist 2005 als Embedded-Box-Lösung für das seit 1998 bestehende Archiv-System Archivista entstanden. Ursprünglich als 32-Bit-Lösung konzipiert, besteht seit anfangs 2011 die 64-Bit-Version, welche neben einem Dokumenten-Management (DMS) und einem einfachen ERP-Modul immer auch ArchivistaVM als Virtualisierungslösung enthält.

Die ArchivistaBox hat im September 2011 den Swiss Open Source Award in der Kategorie Spezial gewonnen. Dazu Jury-Mitglied Matthias Günther, Dr. phil. nat., Mitglied der erweiterten Geschäftsleitung und CIO des Eidgenössischen Bundesamtes für geistiges Eigentum der Schweiz: 'Archivista war in der Schweiz ein Pionier im Bereich der Kombination Open Source und Businesslösung. Mit der zusätzlichen Bündelung mit Hardware ist der Firma Archivista ein Produkt gelungen, dessen lange Lebensdauer und Erfolg exemplarisch zeigt, wie Open Source Software in geschäftskritischen Bereichen verwendet werden kann.'

6.3 Copyright-Hinweise

Zu beachten gilt es die in diesem Skript verwendeten Produktenamen bzw. Warenzeichen. KVM ist ein RedHat Emerging Technology Projekt. DRBD ist ein eingetragenes Warenzeichen der Firma Linbit in Wien, Archivista ist eine geschützte Wort-/Bild-Marke der Firma Archivista GmbH.

Die Sourcen unter svn.archivista.ch/websvn unterliegen der GPLv2-Lizenz, die zur Verfügung gestellten ISO-Dateien unterliegen den folgenden Einschränkungen (siehe dazu auch die Hinweise unter www.archivista.ch):

Im Unterschied zu den Sourcen der ArchivistaBox, die der GPL-Lizenz unterstehen, ist dies beim Handbuch sowie den Logos nicht der Fall. Das Handbuch oder unsere Logos dürfen weder kopiert, verändert noch weiterverteilt werden. Archivista ist eine registrierte Wort-Bild-Marke. Es ist daher nicht gestattet, diese in anderer Form als auf den unmodifizierten ISO-Datei(en) zu verwenden.

Kein Problem stellt das unmodifizierte nicht kommerzielle Verteilen der ArchivistaBox-CD (samt Handbuch) dar. Nicht erlaubt dagegen sind (nicht abschliessend) das Anbieten der ISO-Dateien gegen Entgelt (insb. auch Unkostenbeitrag), das Einbinden der ArchivistaBox-CD in eine andere Distribution, das Anbieten von kommerzieller Schulung und Support sowie das Verwenden der Logos in irgendeiner Form.

Diese Punkte gelten auch für das Vortragsskript. Vielen Dank für die Aufmerksamkeit.