Hochverfügbare Virtualisierung mit KVM, DRBD und ArchivistaVM

Vortrag LinuxTag 2012, 23.5.2012, Berlin

Contents

1 Einleitung

Wenn heute von hochverfügbarer Server-Virtualisierung die Rede ist, dann denken zwar viele an Linux, aber gleichzeitig auch an Komplexität und hochkarätige Hardware. Wer denkt denkt schon daran, dass es für hochverfügbare Systeme im Prinzip nicht mehr als zwei oder drei einfache Rechner bedarf.

Der Vortrag stellt die Cluster-Lösung ArchivistaVM vor, mit der kleinere bis mittlere Cluster (zwischen 12 und 256 CPUs) in extrem kurzer Zeit aufgebaut werden können. Dabei wird der gesamte Aufbau des Clusters live gezeigt und es werden die Hintergründe dazu vermittelt. Eine Demo, die den Aufbau eines 3er-Clusters in ca. 100 Sekunden zeigt, ist verfügbar unter

http://www.archivista.ch/avcluster.gif

Aufgrund dieser Demo erlebt der Vortragende oft das folgende Szenario. Wow, toll, dass das so einfach geht. In einer durchaus positiven Euphorie werden Projekte gestartet, die bisher aus Kostenüberlegungen nicht realisierbar waren. Aber diese Einfachheit bringt auch gewisse Tücken mit sich, denn, wenn es so einfach ist, warum nicht gleich eine neue gute Hardware ordern, und ab geht die Post... Und bald drehen sich die Fragen nur noch um CPU, RAM, RAID-Controller und SAS, wo hingegen der Vortragende dazu einladen möchte, sich über Konzepte Gedanken zu machen - und zwar vor dem Ordern der neuen Hardware.

Ziel von ArchivistaVM ist es gerade, dass auf der Hardware-Seite keine komplexe Server-Infrastruktur notwendig ist, damit Zeit für das saubere Aufgleisen von Projekten bleibt. Das heißt nicht, dass keine High-End-Hardware eingesetzt werden kann oder darf. Vielmehr möchte der Vortragende dazu einladen, Projekte in der Virtualisierung (mit oder ohne Hochverfügbarkeit) mit einer realistischen Bedarfsabklärung zu starten.

Einige Fragen, die hier gestellt werden sollten: Welche Server gibt es derzeit? Welchen Zeithorizont gibt es? Sollen die alten Systeme migriert werden? Mit welchem Aufwand geht dies? Welchen Bedarf an Leistung (Hard- wie Software) besteht über die nächsten Jahre? Wie werden die Daten und die Systeme gesichert? Wie lange kann der Ausfall eines oder mehrerer Systeme 'verkraftet' werden. Was, wenn alle Server gleichzeitig ausfallen? Reicht es, wenn ich einen 7/24-Stunden-Wartungsvertrag für die Software abschließe?

Der Vortrag zeigt daher anhand von Praxisbeispielen auf, was sich bewährt hat, und was eher nicht - vor allem aber, warum und womit bessere Erfolge zu erreichen sind. Dabei geht es um Grundregeln und Ausnahmen, um Vernunft und Unvernunft, sowohl bei Menschen wie Maschinen. Und es wird selbstverständlich gespielt, live mit dem anlässlich des Vortrages aufgebauten Cluster. Ziel des Vortrages ist, dass a) Cluster-Virtualisierung mit einfachsten Mitteln aufge-

baut werden kann und dass b) die eigenen Bedürfnisse so zusammengetragen werden können, dass c) sich dabei die Frage nach der richtigen Hardware von alleine beantwortet.

Soweit der eingereichte Text für den diesjährigen Vortrag. Die schelchte Nachricht gleich vorweg, der Vortrag und die neue Version ArchivistaVM sind 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 Projekte zu.

Die Demo zeigt auf, wie ein Cluster (3 Instanzen) komplett automatisiert in ca. 100 Sekunden aufgebaut werden kann.

2 Einichten des Clusters

Dass für die Demo eine virtualisierte Umgebung verwendet wird, 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 wird ein handelsübliches Notebook (¡500 Euro) mit einer QuadCore CPU und 16 GByte RAM. Dadurch wird die Demo verlangsamt ablaufen. Dafür sei um Verzeihung gebeten.

Im übrigen ist dieser Vortrag die Weiterentwicklung des Vortrages vom letzten November anlässlich des linuxday.at; auch dies sei hier nicht verschwiegen.

2.1 Was ist neu?

Neu bei der Demo anlässlich des LinuxTages 2012 ist, dass der Cluster komplett im RAM aufgebaut wird. Ebenfalls neu ist, dass ArchivistaVM bereits mit einer ISO-Datei mit etwas über 200 MByte machbar ist. Bisher war der Cluster im RAM nicht machbar und die ISO-Datei benötigte etwa 350 MByte. Auch ist es nicht mehr möglich, im RAM-Modus den Hauptspeicher so in Beschlag zu nehmen, dass sich das System mit einem Speicherüberlauf verabschiedet.

2.2 Welche Vorteile ergeben sich daraus?

Zunächst ist der Speed auf Seiten von ArchivistaVM zu vermerken. Dank dem Aufsetzen im RAM ist aber vorallem keine Installation mehr notwendig, ebenso können sämtliche Updates vorgelagert werden; der Cluster kann schneller aufgebaut werden. Im Endeffekt können ArchivistaVM-Cluster deutlich vereinfacht gewartet werden.

2.3 Cluster-Aufbau von Hand (Stand 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
```

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

Drei Rack-Rechner mit 10-GBit-Karten sind nicht im unteren vierstelligen Beriech machbar. Daher wurde das Budget-Konzept entwickelt.

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.

4 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 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.

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 avtest0.iso 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 1` und `create_clusternew.pl 1` bearbeiten. Bei `create_clusternw.pl 1` gilt es zu beachten, dass dabei bei den Instanzen 141, 142 und 143 die Image-Dateien beim Ausführen des Programmes gelöscht werden. Wer nicht mit den IP-Adressen 192.168.0.141..143 arbeiten möchte, möge bitte das Skript `create_cluster.pl` anpassen und anschliessend die ISO-Dateien mit den alternativen IP-Adressen mit `perl create_cluster.pl 1 1` erstellen.

Hinweis: Wird nur `perl create_cluster.pl` (ohne die 1 am Schluss) aufgerufen, so wird der Cluster nicht im RAM, sondern auf den virtuellen Festplatten eingerichtet. Ebenfalls führt der Aufruf von `perl create_clusternew.pl` (ohne die 1 am Schluss) dazu, dass die Festplatten so eingerichtet werden, dass der Cluster auf den virtuellen Festplatten aktiviert wird.

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 1` ausgeführt wird, werden drei ISO-Dateien erstelt. Diese korrespondieren zu den Instanzen 141, 142 und 143. Mit `qm start 141`, `qm start 142` und `qm start 143` 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.

6 Virtualisierung im KMU-Umfeld

Der Vortragende erlebt immer wieder die gleichen Szenarien. Ein Projekt wird aus dem Boden gestampft, irgendwo treten Probleme auf, es gibt Zeitverzögerungen und Kostenüberschreitungen. Steht die Lösung, ist es 5 vor oder nach 12; zu kurz kommen Fragen bei der Schulung und letztlich das Bewusstsein für die Wartung. Daher seien an dieser Stelle diese Fragen in ganz allgemeiner Form erörtert.

6.1 Welche Server?

Wer an dieser Stelle eine Übersicht über Server-Landschaften erwartet, wird wohl enttäuscht sein.

Etwas salopp gesagt vertritt der Vortragende die Ansicht, dass Server-Landschaften in KMU-Unternehmungen im Grundsatz fast immer überdimensioniert sind. Dazu ein Beispiel: Wenn in einer KMU-Unternehmung pro Tag 200 Rechnungen erfasst werden, dann sollten dazu nicht zwingend 4 CPUs und 16 GByte RAM erforderlich sein. Und selbst wenn dem so ist, so müssen es nicht zwingend Rack-Rechner sein.

Natürlich bieten 19-Zoll-Racks gewisse Features, die bei Desktop-Rechnern fehlen. Dazu zählen beispielsweise redundante Netzteile sowie eine eigene Netzwerkkarte, um Remote das BIOS anzusprechen.

Nur, was nützen redundante Netzteile im KMU-Umfeld, wenn diese beinahe so hoch wie ein kompletter Desktop-Rechner zu liegen kommen? Und auch der Remote-Zugriff ins BIOS ist an sich eine nette Sache, doch ist dieses Feature nur mit erhöhter Komplexität zu haben, neben den BIOS-Einstellungen ist ein zusätzliches Interface zu erlernen.

Ein weiteres Beispiel: Einer unserer Kunden benötigte einen halben Tag, um innerhalb eines Fake-Raid-Kontrollers (ohne Software nicht lauffähig) vier einzelne Platten festzulegen, sodass danach ArchivistaVM mit Software-RAID installiert werden konnte.

Nochmals: Nach Ansicht des Vortragenden eignen sich bereits Desktop-Rechner mit AMD-Prozessoren (ECC-Memory mit Prüfbits jederzeit einbaubar) oder preiswerte zwei HE-Racks (zwei Höheneinheiten), wenn es denn Racks sein sollen. Ganz allgemein wird die Hardware bei der Virtualisierung im Verhältnis zur Software überschätzt. Der Autor hat bereits derart viele VM-Lösungen gesehen, bei denen die edelste und teuerste Hardware verwendet wurde, ohne dass die Software auch nur annähernd optimal konfiguriert gewesen wäre.

Gerne gibt der Vortragende auch dazu ein (mehrfach in der Praxis angetroffenes) Beispiel: Bei Umgebungen im Bereich jenseits von 50000 Euro/Franken scheint es dem Vortragenden etwas

peinlich, wenn keine 10 GBit-Karten für die interne Kommunikation für die VM-Server eingebaut sind; der maximale Speed für alle VM-Instanzen kann so nie über ca. 100 MByte/Sekunde (eine 1-GBit-Karte gibt nun mal nicht mehr her) liegen - und dabei waren Plattenverbünde eingebaut, die problemlos plus/minus mehrere hundert MByte/Sekunde hergeben würden.

6.2 Zeithorizont

Vor nicht allzu langer Zeit hat uns ein Community-User per Mail mitgeteilt, aufgrund unserer Nicht-Reaktion im Community-Forum habe er sich zwischenzeitlich für eine andere Lösung entschieden. Dabei ging es um eine Zeitspanne von einigen Tagen. Eine solche Zeitspanne scheint etwas gar kurz gegriffen.

Wer keine Tests vor dem Produktivbetrieb fährt, wird vielleicht danach unter 'scharfen' Bedingungen feststellen müssen, dass die Datensicherung zu lange dauert oder dass die Performanz auf Anhieb noch nicht überzeugt. In einem Testbetrieb ist dies nicht weiter tragisch. Im produktiven Betrieb allerdings müssen plötzlich Efforts geleistet werden, die vermeidbar wären. Deshalb die Empfehlung, testen, testen und nochmals testen...

6.3 Alte Systeme migrieren?

Auch nach unzähligen von Migrationen (insbesondere in der Windows-Welt) kann der Vortragende nicht pauschal sagen, es lohnt sich oder es lohnt sich nicht.

Sehr gute Dienste leistet bei Migrationen Clonezilla, dennoch kann es jederzeit und völlig unerwartet zu Problemen kommen. Als Beispiel sei hier jener Terminal-Server erwähnt, welcher lebensnotwendige Informationen jenseits der Partierungstablle ablegte. Clonezilla kopierte dabei zwar sämtliche Partitionen, aber ohne den unpartionierten Bereich zu Beginn der Festplatte liess sich der virtualisierte Server anschliessend nicht hochfahren. Letztlich leistete im konkreten Fall `dd if=/dev/sdx of=/dev/sdy/eins.img` die besseren Dienste.

In jedem Falle sollten aber von den zu migrierenden Platten zunächst Sicherungskopien erstellt werden. Anschliessend können die Platteninhalte zunächst offline kopiert werden. Lassen sich diese Images danach nicht hochfahren, sollten (wiederum bzw. insbesondere bei Windows-Transformationen) im alten laufenden System sämtliche hardwarespezifischen Treiber entfernt werden (Stichwort Zurückfahren auf IDE-Treiber).

6.4 Aufwand für ArchivistaVM

Die Virtualisierung mit ArchivistaVM ist heute extrem schnell implementiert, die Konzepte bei einer hochverfügbaren Virtualisierungslösung bleiben aber gleichwohl in einem gewissen Masse

anspruchsvoll. Immerhin sind bei bei einem minimalen Cluster von 3-Rechnern 6 Festplatten, 9 Netzwerkkarten, je drei externe und interne Kabel anzuschliessen. Bewährt hat sich, sowohl Rechner, Netzwerkkarten als auch Kabel entsprechend zu beschriften.

Offen gestanden, und vielleicht durchaus im Widerspruch zum Vortragstitel 'Hochverfügbare Virtualisierung mit ArchivistaVM, KVM und Debian' erscheint dem Vortragenden eine verfügbare einfache Virtualisierung sinnvoller als eine den Administratoren überfordernde hochverfügbare Lösung.

Bei der Planung sollte berücksichtigt werden, dass eine Maschine nicht produktive Instanzen enthalten sollte, da ansonsten beim Ausfall eines Rechners nicht einfach geswitcht werden kann. Über den Daumen gepeilt sollten auch nicht mehr Instanzen produktiv und unter Last laufen als CPU-Kerne zur Verfügung stehen.

Eine weitere Grundregel: Bei zwei bis drei einzelnen Rechnern (bis ca. 24 Cores) müssten mit ArchivistaVM ein paar Stunden reichen, um eine Testumgebung aufzubauen bzw. die Konzepte dahinter zu verstehen (Lesen des Handbuches wird sehr empfohlen). Danach sollten für einige Wochen Erfahrungen gesammelt werden, ehe an einen produktiven Betrieb zu denken ist.

Wer die gesamte Umstellung selber bewerkstelligen möchte, wird zu Beginn die Frage zu beantworten haben, welches Produkt zum Einsatz kommen soll. Selbstverständlich wird der Autor hier nicht andere Lösungen propagieren wollen. Grundsätzlich gilt aber, dass Virtualisierungsprodukte, von den Daten her betrachtet, praktisch beliebig austaushbar sind.

Die verwendeten Festplatten-Images der verschiedenen Produkte lassen sich mittlerweile fast alle mit mehr oder weniger Aufwand zwischen den Produkten hin- und herschieben (qemu-img convert). Sollten Sie mit der Zeit mit einem Produkt nicht zufrieden sein, so können die Instanzen mit relativ wenig Aufwand migriert werden.

6.5 Bedarf an Leistung

Im Prinzip sollte es keinen erhöhten zusätzlichen Bedarf an Hard- und Software geben. Dank der Virtualisierung sollten um den Faktor 1:2 oder 1:3 weniger Rechner zur Anwendung kommen.

In der Praxis dürfte es aber eher so sein, dass mit der Virtualisierung nicht wesentlich weniger Rechner im Betrieb stehen werden. Bei der hochverfügbaren Virtualisierung sollten minimal drei Rechner zur Anwendung kommen, bei separaten Rechnern für Daten und VM-Instanzen müssen es sogar minimal vier oder mehr Server sein.

Als Grund, weshalb letztlich nicht weniger Rechner zum Einstz kommen, wo es doch möglich wäre, kann der Vortragende aus dem eigenen 'Nähkästchen' plaudern. Dank der Virtualisierung

können in unserer Firma die Dinge derart viel eleganter gelöst werden (z.B. automatisierte Test-Suites für alle erstellen ISOs), dass diese Features mit Wonne in Anspruch genommen werden. Folglich stehen nicht weniger sondern eher mehr Rechner im Einsatz, denn plötzlich besteht die Möglichkeit, Dinge zu realisieren, die zuvor undenkbar waren.

6.6 Datensicherung

Extrem bewährt haben sich Voll-Datensicherungen, die täglich erfolgen. Beim ArchivistaVM-Cluster kann sich dabei die zweite Instanz (sämtliche Instanzen werden ja gleichzeitig auf zwei Rechnern vorgehalten) ausklinken, die Datensicherung in aller Ruhe druchführen, um sich danach wieder einzuhängen. Dadurch entstehen extrem kurze Latzenz-Zeiten bei der Datensicherung.

Das Anlegen von Snapshots ist an sich eine elegante Möglichkeit, um extrem schnell einen x-beliebigen Zustand einer Maschine zu sichern. Es darf dabei aber nicht vergessen werden, dass der aktuelle Zustand des RAMs mit dem aktuellen Zustand der Festplatte gesichert wird, und dass bei einem Hervorholen dieses Zustandes gerade bei komplexteren Umgebungen (z.B. Datenbanken mit externen Requests) das Recovery scheitern kann. Bei einer Gesamtsicherung besteht dieses Problem nicht; die Latenzzeiten bei einer Cluster-Lösung und Vollbackups sind nicht länger, es stehen dafür aber jederzeit komplette Sicherungen zur Verfügung, die jederzeit ohne Probleme hochgefahren werden können.

6.7 Ausfallzeiten

Der Autor ist gestern mit dem Zug von Zürich nach Berlin gefahren. Ab Basel stand ein moderner ICE zur Verfügung; während der gesamten Zeit stand das Reservationssystem nicht zur Verfügung. Es sei an dieser Stelle nichts gegen die deutsche Bahn gesagt, Ausfälle können immer und jederzeit auftreten, passieren sollte es trotzdem nicht. Immerhin führte der Ausfall des Reservationssystems im Zug zu keinem Chaos.

Nachtrag: Zwei Tage später, bei der Rückfahrt, bestand das Problem noch immer. Irgendwann wurde per Lautsprecher verkündet, die Diskette sei nun korrekt eingelesen, das Problem sei behoben. Kurz danach zeigten alle Anzeigen praktisch allesamt 'Berlin - Interlaken' an. Ein Chaos brauch auch diesmal nicht aus, wohl aber waren viele Reisende nunmehr sichtlich genervt...

Wäre das gleiche 'Malheur' beim Online-Ticket-Verkauf der Bahn aufgetreten, wäre es deutlich schmerzlicher gewesen; der Autor buchte das Ticket kurzfristig, und ein Ausfall des Verkaufssystems über 8 Stunden (bzw. 40 Stunden) hätte saftige Umsatzeinbussen zur Folge gehabt.

In Erinnerung ist dem Autor auch der Ausfall bei einem der grössten Detailhändler geblieben. An einem Morgen konnten sämtliche Filialen (immerhin einige Hundert) keine Produkte mehr verkaufen, weil die Server für die Kassensysteme nicht mehr hochgefahren werden konnten. Später teilte die Pressesprecherin mit, der Ausfall habe nicht vermieden werden können, auf der Test-Umgebung sei es zu keinem Ausfall gekommen.

Etwas mehr Übung des Ernstfalles hätte in beiden Fällen wohl geholfen. Aber, das Üben am Objekt hilft letztlich nicht wirklich, wenn die Übungsobjekte (Test- und Produktiv-Rechner) nicht identisch sind. Bereits das Verwenden einer anderen CPU (AMD zu Intel) kann dazu führen, dass die Windows-Aktivierung erneut losgetreten wird.

6.8 Totalausfall

Klassische Szenarien bei der Hochverfügbarkeit gehen immer davon aus, dass nur einzelne Rechner (Nodes) ausfallen. Was, wenn alle Server den Geist aufgeben? Der/die gestandene Administrator/in wird einwenden, dass dies sehr unwahrscheinlich sei, und dass dafür ja ein Wartungsvertrag bestünde. Mag sein, aber unter Verwnedung von Standard-Hardware wird es dennoch viel leichter sein, den gesamten Maschinenpark (sofern Datensicherungen existieren) auf neue Rechner zu migrieren.

Bei OS-Produkten kann dieser Ernstfall auch problemlos (abgeschottet vom Netzwerk) durchgespielt werden. Dies ist bei kommerziellen Umgebungen nicht ohne weiteres möglich; da schlicht und einfach die Lizenzen nicht doppelt zur Verfügung stehen.

6.9 Wartung und mehr

Die meisten Firmen werden Informatik-Lösungen nicht ohne Wartungsverträge in Betrieb nehmen. Es kann aber nie schaden, Ausfallszenarien unabhängig von den Wartungsverträgen in eigener Regie durchzuspielen.

Der Wartungsvertrag alleine denkt nicht mit. Auch hier ein aktuelles Beispiel aus der Praxis. Der Kunde meldet den Ausfall eines Rechners und sendet das Gerät umgehend per Eilpacket (Express) zu. Es stellt sich heraus, dass das Problem nicht am Rechner lag. Vielmehr wurden (dummerweise) die Netzwerkkabel zwischen zwei Switches vertauscht.

7 Empfehlungen für den Alltag

7.1 Standardhardware

Verwenden Sie Standardhardware. CPU und RAM, redundante Festplatten. Für Cluster mit ArchivistaVM sind minimal Rechner mit jeweils drei Netzwerkkarten und zwei Festplatten erforderlich (3er-Cluster).

7.2 Standardsoftware

Verwenden Sie keine Systeme, welche Sie von Hand installieren und konfigurieren, es sei denn, Sie betreiben Virtualisierung als Freizeitbeschäftigung.

7.3 Kleine Lösungen für Ansprüche

Der Anspruch von KMU-Unternehmungen ist nicht unwesentlich kleiner als jener von grossen Unternehmungen. Erliegen Sie trotzdem nicht der Versuchung, überdimensionierte Lösungen aufzubauen.

7.4 Spielen Sie den Ernstfall durch

Übung macht bekanntlich den Meister. Nur weil ArchivistaVM in einer Minute aufgesetzt ist, heisst dies noch lange nicht, dass Sie es nur einmal tun sollten, ehe der Ernst- bzw. Ausfall zuschlägt.

7.5 Denken Sie in Ausbauplänen

ArchivistaVM bietet derzeit im Cluster-Modus kein Tool an, um den Ausbau des Clusters vollkommen automatisiert durchzuführen. ArchivistaVM-Cluster können aber jederzeit automatisiert in geänderter Formation neu aufgesetzt werden. Unter Verwendung von Standard-Hardware lassen sich problemlos auch mehrere kleinere Cluster betreiben; die Kosten für Hard-, Software und Wartung sind minim.

7.6 Live-Migrationen gut überdenken

Wenn Sie Instanzen zügeln möchten, so müssen Sie derzeit das QCOW2-Format verwenden. Diese Dateien wachsen mit jedem Byte an. Beim RAW-Format wird ein maximal fixer Bereich zugewiesen. Sie haben dann zwar keine Live-Migrationen mehr, dafür aber werden die Instanzen über Jahre den vollen Speed bringen.

7.7 Datensicherung

Führen Sie eine tägiche Datensicherung durch. Verwenden Sie dabei USB3-Geräte; Sie erhalten dabei deutlich mehr Speed, als wenn Sie über das Netzwerk die Datensicherung durchführen.

7.8 Hochverfügbarkeit heisst Annäherung

Alle sprechen heute von Hochverfügbarkeit, und doch gibt es diese nur in der Annäherung. ArchivistaVM bietet hier dank Automatisierung und Einfachheit eine Annährung, nicht mehr und nicht weniger.

8 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.