# 120 TByte und 120 CPUs in 10 Rack-Höheneinheiten

***Pfaffhausen, 12. Mai 2011:*** *Bis vor zwei Jahren lieferten wir keine Rechner aus, die mehr als 320 GByte an Daten aufnehmen konnten. Im vorletzten Jahr haben wir mit den ArchivistaBoxen Summit und Matterhorn erstmalig die 8 TByte Marke erreicht. Wir freuen uns sehr, Ihnen die neue ArchivistaVM-Cluster-Generation SwissRocket vorstellen zu dürfen.*

## Herkömmliche Cluster-Konzepte bieten wenig Ausfallsicherheit

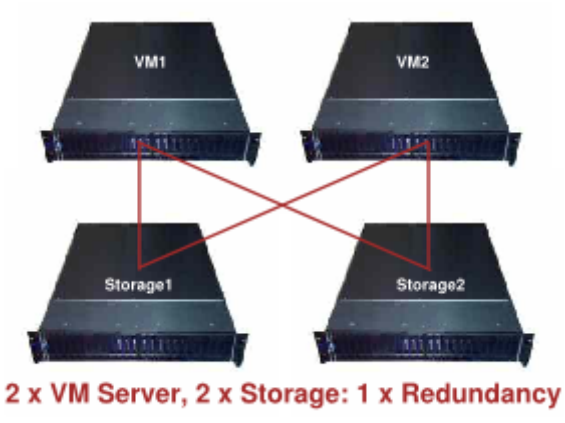
Einleitend sei hier festgestellt, ein Cluster ist ein Verbund von Servern. Bei Ausfallsicherheit (Hochverfügbarkeit) geht es darum, dass wenn eine oder mehrere Komponenten ausfallen, der Rechnerverbund (Cluster) ohne Unterbrechungen weiterlaufen kann. Wir sprechen hier auch von Redundanz (jede Teilkomponente ist doppelt verfügbar).

Bevor wir das neue SwissRocket-Cluster-Konzept vorstellen, möchte ich einige bestehende Cluster-Konzepte vorstellen, die mir in den letzten beiden Jahren bei KMU-Firmen zu Augen und Ohren gekommen sind. Gerade bei kleineren Firmen besteht die Problematik, dass aus Kostengründen oft Kompromisse gemacht werden, die später im Betrieb bzw. für den weiteren Ausbau Folgen, mitunter fataler Art, haben.

Bei der Virtualisierung beherbergt ein physikalischer Server mehrere Instanzen (virtuelle Rechner). Dabei werden die CPU-Kerne und die Festplatte(n) geteilt. Die einzelnen Instanzen liegen dabei als Abbild auf dem physikali


schen Server vor. Die Instanzen können gesichert und auch wieder zurückgespielt werden. Nun sind bei einem Ausfall sämtliche virtualisierten Instanzen betroffen. Aus diesem Grunde werden oft Konzepte gewählt, die Ausfallsicherheit enthalten. D.h, es werden mehere Server für die Virtualisierung (VM-Server) und die Speicherung der Daten (Storage-Server) aufgesetzt. "Stribt" ein Server, so läuft der Betrieb uneingeschränkt weiter. Oft ist dabei von "Failover" die Rede. Die verbleibenden Rechner erkennen den Ausfall eines Mitgliedes und verteilen die Last auf die übrigen Rechner.

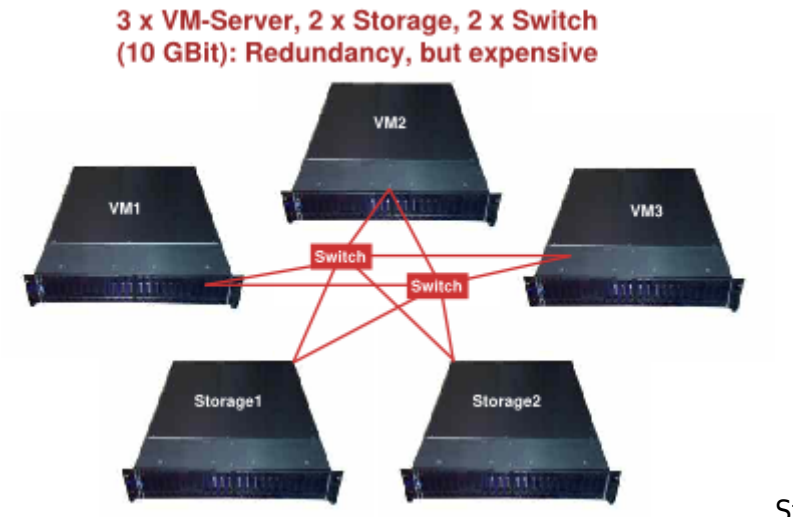
Da bei der Vritualisierung die laufenden Maschinen die Kapazität der Festplatten teilen müssen, werden meist separate Server für die Virtualisierung und das Speichern der Daten aufgesetzt. Die Server für die Virtualisierung enthalten möglichst viele CPU-Kerne und RAM, die Speicher-Server (Storage, NAS, SAN etc) bieten schnelle Festplattenverbünde an. Was auf den ersten Blick einleuchtend erscheint, ist mitunter nicht zu Ende gedacht. Betrachten wir eine Implementierung mit 3 Servern. Zwei Server arbeiten für die Virtualisierung, der dritte Rechner speichert mit zwei separaten Kontrollern in zwei Festplattenverbünden die Daten. Wir haben bei dieser Konstellation das Problem, dass eben gerade keine Redundanz besteht. Sollte es je ein Problem beim Server für das Speichern der Daten geben, steht der gesamte Cluster still.




Im Minimun sollten daher zwei Server für die Virtualisierung und zwei Server für das Speichern der Daten implementiert werden. Beide VM-Server speichern die Daten redundant auf den Storage-Servern. Fällt ein VM-Server aus, übernimmt der zweite VM-Server die Arbeit, fällt ein Storage-Server aus, übernimmt der zweite "Speicher-Knecht" die Arbeit. Ausfallsicherheit haben wir damit erreicht. Allerdings muss bei diesem Konzept beachtet werden, dass gängige

Netzwerkkarten (1 GBit) nur ca. 100 MByte pro Sekunde an Daten übertragen können. D.h. die schnellsten Storage-Server bringen nichts, wenn die Daten auf der Leitung steckenbleiben. Es gilt also, bei diesem Konzept mit schnellen 10 Gbit Netzwerkkarten für genügend Bandbreite (ca. 1 GByte pro Sekunde) zu sorgen.

Nun sind 10 GBit Netzwerkkarten mittlerweile nicht mehr teuer, wohl aber enstprechende Switches. Bei 2 VM- und 2 Storage-Server können Dual-Port-Netzwerkkarten mit Crossover-Kabeln zum Einsatz kommen. Damit lässt sich der Einsatz von Switches vermeiden. Allerdings, ein solcher Cluster lässt sich nicht ohne weiteres erweitern. Streng genommen haben wir nun vier Server, um den Ausfall eines Rechners verschmerzen zu können.




Standen bisher vier nicht virtulisierte Server im Hause, so stehen nun noch immer vier Server im Server-Raum. Immerhin, fällt ein Server aus, so können wir den Betrieb ohne Unterbruch auf den verbleibenden drei Servern fortführen. Nur, ohne die Vritualisierung konnten wir früher bei Bedarf jederzeit weitere neue Server aufsetzen, mit einem Cluster für die Virtualisierung geht dies nicht mehr so einfach. Daher werden oft noch Cluster mit 3 oder mehr VM-Servern implementiert.

Nachfolgend gehen wir von einem Szenario mit 3 VM-Servern und 2 Storage-Rechnern aus. Dabei können wir zwei VM-Server auslasten, den dritten VM-Server können wir zu Testzwecken verwenden. Um diese VM-Server hochverfügbar zu halten, sind redundante 10-GBit-Switches erforderlich, denn auch ein Switch kann ausfallen und in diesem Falle würde die gesamte Lösung stillstehen. 10-Gbit Switches sind nicht günstig, ein 6er-Switch kostet nach wie vor irgendwo zwischen 6000 und 8000 Franken, bei zwei Switches dürfte mit Kosten zwischen 12000 und 15000 Franken zu rechnen sein. Wohl verstanden, wir können bei einem 6er-Switch maximal 6 Server anschliessen, 12er oder 24er Switch mit 10 GBit kosten schnell einmal 20'000 Franken pro Switch.

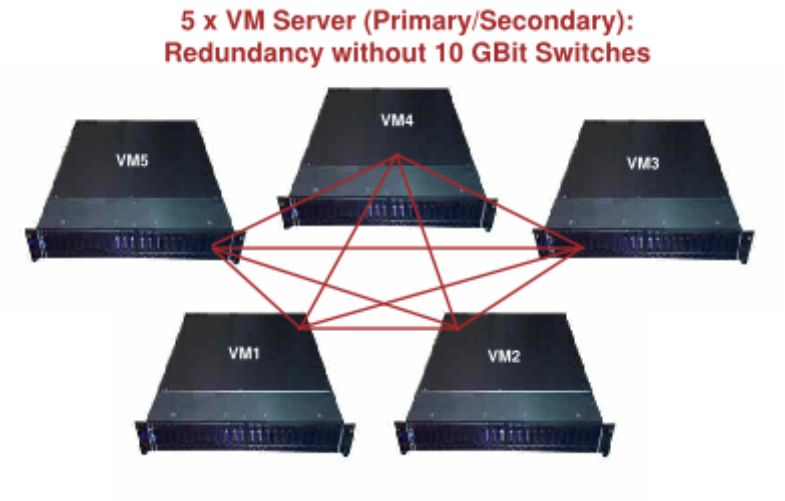
## Archivista SwissRocket: Ausfallsichere Cluster mit

## ArchivistaVM


Hier setzt das Konzept unserer Archivista SwissRocket Cluster ein. Bereits der 3er-Cluster ist ausfallsicher (2 Rechner im Betrieb, 1 Rechner StandBy). Bei einem Ausfall eines der drei Server kommt der StandBy-Rechner zum Zuge. Doch wie können wir mit 3-Servern die Daten der zwei Produktiv-Server verfügbar halten? Ganz einfach, all unsere SwissRocket-Server sind so ausgelegt, dass ein jeder Server in einem 2ten-Festplattenverbund die Daten eines anderen Rechners automatisch mitspeichert. Jeder SwissRocket-Server ist einmal Primary (produktiv) und einmal Secondary (standby). Bei einem Festplattenverbund mit Raid10 werden dafür ca. 2 bis 3 Prozent einer CPU an Rechenleistung benötigt, bei einem Raid5-Verbund sind es ca. 40 bis 50 Prozent einer CPU.

Hochverfügbarkeit ist immer relativ. Die Frage, ob ein Server auch während der Datensicherung zur Verfügung steht, ist von der Verfügbarkeit auf Stufe Virtualisierung zu unterscheiden. Bei den Archivista SwissRocket-Clustern klinkt sich der Secondary-Server aus, erledigt die Datensicherung und klinkt sich danach wieder ein. Um die Datenintegrität nicht zu gefährden, wird dabei auf dem primären Server eine jede Instanz entweder neu gestartet oder kurz standby gesetzt. Danach stehen die Server uneingeschränkt für die Weiterarbeit zur Verfügung. Bei einem Neustart beträgt die Ausfallsicherheit weniger als 1 Minute, bei standby sind es einige wenige Sekunden.

Wir bieten bei der Modellserie SwissRocket pro Server maximal 24 CPUs und 24 TByte an Daten an. Minimal sind es 6 CPUs sowie 3 TByte Festplattenkapazität, letztlich benötigt ja nicht jede KMU-Unternehmung im Grundumfang 24 CPUs (entspricht ca. 12 bis 24 Server). Bei einem 5er-Cluster können bis zu 120 CPUs und 120 TByte erreicht werden. Selbstverständlich können die Cluster später erweitert werden. Weiter ist Redundanz bereits mit 2 Servern möglich (1 x produktiv, 1 x standby). Einem solchen 2er-Cluster können Sie jederzeit zu einem 3er, 4er und 5er-Cluster umrüsten, genauso wie Sie einen 5er-Cluster bis auf maximal 24 Cluster (24×48=1152 CPUs) ausbauen können.




Die Software der Archivista SwissRocket-Cluster untersteht der GPL-Lizenz. Bei der Hardware setzen wir auf Markenkomponenten, bauen die Server aber bewusst bei uns zusammen. Die Erfahrung mit eingekauften Servern hat uns gelehrt, dass wir weit bessere Konzepte hinkriegen, wenn wir Standardkomponenten verwenden. Ich möchte hier aber auch erwähnt haben, dass sich das Konzept ebenso mit Marken-Servern realisieren liesse, sofern die Festplatten einzeln angepsrochen werden können und 10 GBit-Netzwerkkarten verfügbar sind. Die SwissRocket Cluster können bei uns live in Aktion beschnuppert werden. Sie werden von der Leistung begeistert sein, als Beispiel sei hier die Zeit genannt, die auf einem SwissRocket-Server benötigt wird, um 1 TByte an Daten zu schreiben:

```
dd if=/dev/zero of=/var/lib/vz/1TB.img bs=256M count=3760
oflag=direct
3760+0 records in
3760+0 records out
1009317314560 bytes (1.0 TB) copied, 861.507 s, 1.2 GB/s
```

Für all diejenigen, die lieber die Rechner rechnen lassen, 861.5 s entsprechen 14 Minuten und 21 Sekunden. Der Autor erinnert sich an Zeiten, wo für 300 MByte eine Minute benötigt wurden. Zugegeben es ist länger her (richtig erinnert müsste es um die Jahrtausendergrenze herum gewesen sein). Dennoch der Vergleich, umgerechnet auf 14 Minuten 21 Sekunden ergeben sich ca. 4.3 GByte. Mit anderen Worten, der SwissRocket-Server ist um den Faktor 232 schneller.

## Eckdaten und Preise der neuen Archivista SwissRocket-Server

Die ArchivistaSwiss-Rocket-Cluster decken einen Leistungsumfang zwischen 12 CPUs und 120 CPUs ab, wobei einem weiteren Ausbau (bis 24 Cluster mit 1152 CPUs) nichts im Wege steht. Die Cluster können automatisiert aufgesetzt werden. Je nach Ausbaustufe und Leistungsklasse kommen Rack- und/oder Desktop-Gehäuse zum

Einsatz. Eine 24 CPU-Maschine mit 24 Festplatten benötigt im Leerlauf ca. 150 Watt, unter Vollast ca. 300 Watt, d.h. wir benötigen pro CPU plus/minus zwischen 6 und 12 Watt. Der 2er-Cluster startet bei unter 7000 Franken, der 3er-Cluster kostet keine 10000 Franken, ein 5er-Cluster mit 120 CPUs und 120 TByte Speicherkapazität (nutzbar maximal 96 CPUs und 110 TByte) liegt irgendwo zwischen 50'000 und 60'000 Franken. Wohl verstanden, die **Preise umfassen die gesamte benötigte Hard- und Software.** Gerne **beraten wir Sie** bei der Wahl des für Sie geeigneten Clusters.