Speicherarchitektur für Massendaten einer Webapplikation

Bachlorthesis

HOCHSCHULE
FÜR TECHNIK
ZÜRICH

„Die im DNS-Molekul realisierte immense Speicherdichte an Information ist noch um Zehnerpotenzen den Chips in unseren modernsten Computern uberlegen."
- Werner Gitt, Äm Anfang war die Information", Yale, 2002

Abgabe:	22. Juni 2011
Präsentation:	6. Juli 2011

Student:	Lucien Stucker, lstucker@hsz-t.ch
Dozent:	Beat Seeliger, bseliger@hsz-t.ch
Studienbereich:	Informatik

Inhaltsverzeichnis

Abbildungsverzeichnis iv

Tabellenverzeichnis v

Source Code Verzeichnis vi

1 Vorwort 1

1.1 Was bedeutet diese starkte Wachstum für uns? . . . 1
1.2 Konequenzen für Service Provider . . . 1
1.3 Fazit . . . 2

2 Zusammenfassung 3

3 Ist-Analyse 4

3.1 Defintionen . . . 4
3.1.1 Speichereinheit . . . 4
3.1.2 Raid-5 . . . 4
3.1.3 MTTR . . . 5
3.1.4 MTBF . . . 6
3.1.5 Verfügbarkeit . . . 6
3.1.6 Datenverfügbarkeit . . . 6
3.1.7 Hochverfügbarkeit . . . 7
3.2 Bestandesaufnahme . . . 7
3.2.1 Applications-Architektur . . . 7
3.2.2 System-Architektur . . . 9
3.2.3 Speicher-Architktur . . . 9
3.2.4 Speicherkapazität . . . 9
3.2.5 Datenwachstum . . . 9
3.2.6 Zugriffswachstum . . . 10
3.2.7 Skalierbarkeit Datenvolumen . . . 10
3.2.8 Skalierbarkeit Datenzugriffe . . . 10
3.2.9 Daten Durchsatz I/O . . . 10
3.2.10 Daten Redundanz . . . 10
3.2.11 Datenverfügbarkeit . . . 11
3.2.12 Daten Sicherheit . . . 11

3.2.13 Daten Integrität 11
3.2.14 Backup 11
3.2.15 Wirtschaftlichkeit 11
3.2.16 Lokalität 11
3.3 Analyse-Ergebnisse 12
3.3.1 Datenwachstum 12
3.3.2 Skalierbarkeit Datenvolumen 12
3.3.3 Skalierbarkeit Datenzugriffe 13
3.3.4 Redundanz 14
3.3.5 Datenverfügbarkeit 14

4 Soll-Analyse **16**
4.1 Szenario 1 - Schwaches Wachstum gespeicherten Daten / schwaches Wachstum der Abfragen 16
4.2 Szenario 2 - Schwaches Wachstum Daten / starkes Wachstum der Abfragen 16
4.3 Szenario 3 - Starkes Wachstum Daten / schwaches Wachstum der Abfragen 16
4.4 Szenario 4 - Starkes Wachstum Daten / starkes Wachstum der Abfragen 16

5 Speicher-Markt **17**
5.1 Speicherarchitekturen 17
5.2 Block-Basierend 17
5.2.1 ISCSI 18
Datenverfügbarkeit / Redundanz 18
Skalierbarkeit Datenvolumen / Datenzugriffe 18
Integität 18
Durchsatz I/O 18
Lokalität 18
Backup 18
5.3 Datei-Basierend 18
5.3.1 Network File System 18
5.3.2 NAS Appliance 18
Datenverfügbarkeit / Redundanz 20
Skalierbarkeit Datenvolumen / Datenzugriffe 20
Integität 20
Durchsatz I/O 20
Lokalität 20
Backup 20
5.4 Objekt-Basierend 20
Datenverfügbarkeit / Redundanz 20
Skalierbarkeit Datenvolumen / Datenzugriffe 20
Integität 20
Durchsatz I/O 20

Lokalität 20
Backup 20

6 Machbarkeitsnachweis **21**

Glossar **i**

Anhang **iii**

Abbildungsverzeichnis

3.1 Raid 5 Architektur *Raid 5* [14] 5
3.2 Rails Architektur *Raid 5* [14] 8
3.3 Verhätnis von Speicheplatzverbrauch zur Speicherkapazität in Prozent in einer Zeitspanne von einen Jahr 10

Tabellenverzeichnis

Listings

3.1 Report Dateisystem Speicherplatz Belegung in Dezimal Prefix 9

1 Vorwort

"The Expanding Digital Universeßo bezeichnet das Internationale US-amerikanscihe Marktforschung und Beratungsunternehmen International Data Corporation (**IDC!** (**IDC!**)) das Starke Wachstum an generierten und gespeicherten Daten in einen Jahr. IDS hat seit beginn 2007 Ihrer Studie festgestellt, das sich der Speicherbedarf alle zwei Jahre verdoppelt hat. Bleibt die Zunahme der Daten die nächsten 10 Jahre konstant, haben wir so viele digitale Bits erstellt wie es Sterne im Universum gibt. Die grösste Teil an erstellten Daten sind unstrukturierte Daten, was das ganze gleich komplex macht wie das Universum.[**?**]

1.1 Was bedeutet diese starkte Wachstum für uns?

Neben der zunehmenden Digitalisierung, hat sich auch das Volumen des Speicherplatzes der eingesetzten Speichermedien weiter entwickelt, die Bedürfnisse konnte aber bis anhin nie gedeckt werden. Weshalb wir uns Seit des digitalen Zeitalters damit beschäftigen wie wir unsere generierte digitalen Daten am besten Speichern. Um die Daten effizient Speichern und austauschen zu können haben, wir zum Beispiel Algorithmen entwickelt welche unsere Daten ohne Verlust der wesentlichen Informationen komprimieren, oder wir haben Verfahren entwickelt um die Speichergrenze eines einzelnen Medium zu überwinden.

1.2 Konequenzen für Service Provider

Das Zürcher Startup Unternehmen "Reference Image AG"betreibt und Entwickelt eine Webapplikation zur Speicherung, Archivierung, Verwalten und Wiederverwenden von qualitativen hochauflösenden digitalen Bildern. Zu Ihrem Kundensegment gehören Galerien, Museen, Künstler und Fotographen die sehr hohe Ansprüche an die Qualität Ihrer Bilder haben. Die Reference Image AG erlaubt die Speicherung von Bildgrössen bis zu 2Gbyte. Das heutige Speichersystem ist bereits zu über 50

1.3 Fazit

Um unsere Zukünftigen unvorstellbaren bedarf an Speicherplatz decken zu können werden wir auch zukünftig ....

2 Zusammenfassung

3 Ist-Analyse

Ziel ist es mit der Situationsanalyse, den Aktuellen Stand der bestehende System bzw. Speicherinfrastruktur zu beschreiben und zu bewerten. Die Angaben für die Ist-Analyse wurde zusammen in Gesprächen und von gelieferten Schriftlichen Dokumenten mit dem Auftraggeber erstellt.

3.1 Defintionen

3.1.1 Speichereinheit

In der Arbeit werden die Speichereinheiten anhand des SI-Standards [12] zwischen Dezimal Präfixe wie sie bei Festplatten verwendet werden und Binär Präfixen verwendet werden unterschieden.

3.1.2 Raid-5

Ein Raid 5 Festplatten-Array besteht aus N Identische Festplatten auf welchen Daten verteilten gespeichert sind. Die Einheit an Daten welche auf der selben Festplatte platziert sind, bevor weitere Daten auf eine andere Festplatte platziert wird, bezeichnet man eine Stripe Einheit. In der **Abbildung 3.1** wird eine Stripe Einheit durch die Blöcke dargestellt, z.B. stellt A1 eine Stripe Einheit dar.

Stripe Einheiten welche auf allen Festplatten den selben Physikalischen Platz bzw. Adressen gespeichert sind bezeichnet man als Strip. In der **Abbildung 3.1** durch die gleich farbigen Blöcke dargestellt, hier bilden z.B. A1, A2, A3 und A_p einen Stripe dar. Jeder Strip enthält mehrere Daten Strip Einheiten und eine Parität Strip Einheit. Eine Parität Strip Einheit ist eine XOR-Verknüpfung aller Daten Strip Einheit eines Strips. Durch die Speicherung und Berechnung einer Parität Strip Einheit pro Strip, ermöglicht es in einen Raid-5 den Ausfall einer einzelnen Festplatte zu kompensieren. Die verlorenen Strip Einheiten der ausgefallene Festplatte können durch lesen alle vorhandenen Daten Strip Einheiten und der Parität Strip der noch intakten Festplatten wiederberechnet werden. Die Anzahl der Stripe Einheiten eines Stripe ist durch die Stripe Width (W_s) definiert, wo bei die Anzahl der Daten Stripe Einheiten in jeden Stripe durch $W_s - 1$ definiert ist.

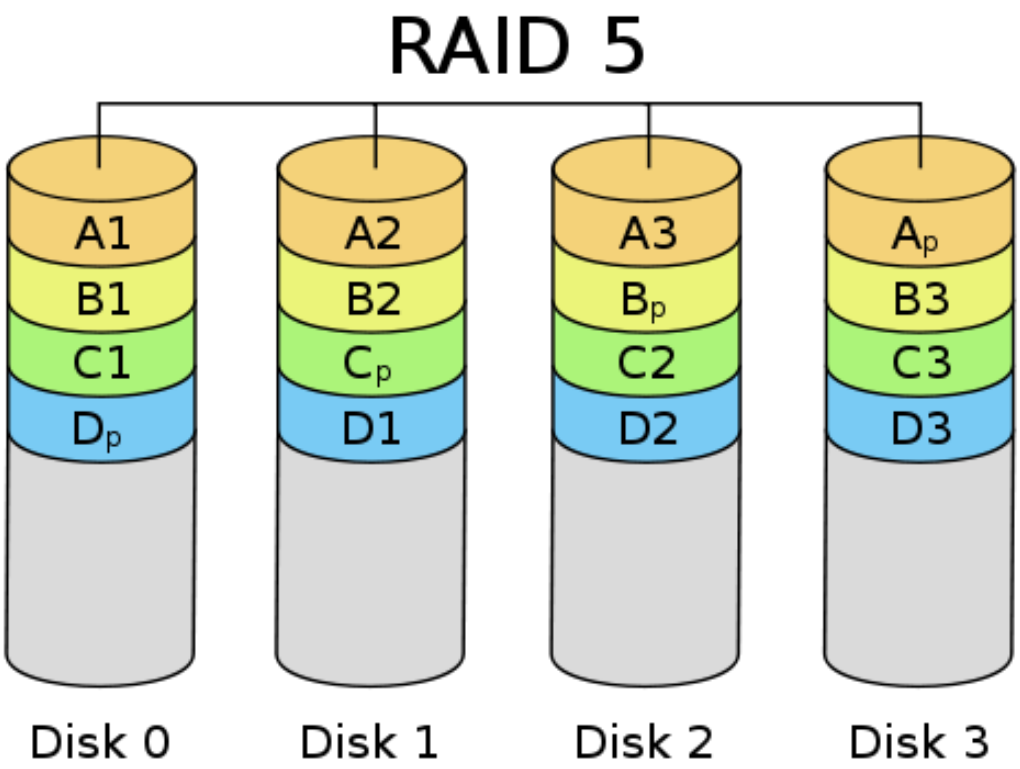

Abbildung 3.1: Raid 5 Architektur *Raid 5* [14]

Die einzelnen Parität Strip Einheiten werden rotierend über die Festplatten verteilt. Dieses Verfahren trägt dazu bei die I/O Last, welche durch Anfragen die eine Aktualisierung der Parität veranlassen auf die einzelnen Festplatten besser zu verteilen. Auf Grund, dass die Daten in einen Raid-5 auf mehre Festplatten verteilt werden, erhöht sich die Wahrscheinlichkeit das eine Festplatte bei einer I/O Operation beteiligt ist, was wiederum den Datendurchsatz und I/O Rate eines Array erhöht [7].å

Für die Speicherung der Parität, muss eine Teil der gesamt Kapazität aller Festplatten verwendet werden. Die für die Daten zur Verfügbaren Nettokapazität eines Raid 5 Array mit n Festplatten und s Speicherkapazität einer Festplatte lässt sich mit folgende Formel [7] berechnen:

$$\text{Speicherkapaziät Raid-5} = (N-1) * S \tag{3.1}$$

3.1.3 MTTR

MTTR ist die Abkürzung von "Mean Time To Recovery“ und bedeutet so viel wie die durchschnittliche Zeit die eine Komponente benötigt um sich nach einen Fehler wiederherzustellen. Die MTTR in einen Raid-5 berechnet sich aus der Zeit bis eine Ersatz Festplatte installiert wurde und der Disk Wiederherstellung Zeit.

$$MTTR_{Raid} = Replacement + Rebuild_{Raid} \tag{3.2}$$

Die Disk Wiederherstellungszeit in einen Raid hängt von Faktoren ab, wie die Anzahl der Festplatten, die Speicherkapazität der Festplatten, die Periodisierung der Rekonstruktion gegenüber dem normalen IO, die normale I/O Last während der Rekonstruktion und CPU Last. Für die Ist-Analyse stehen keine aussagekräftige Statistiken für die Disk Wiederherstellung Zeit zur Verfügung, die MTTR für die Ist-Analyse wird deshalb mit 25 Stunden für die Ist-Analyse definiert unabhängig von weiteren Faktoren.

3.1.4 MTBF

MTBF ist die Abkürzung von "Mean Time Between Failure“ und bedeutet so viel wie die durchschnittliche Betriebs Zeit einer Komponenten bis ein Fehler auftrifft. Hersteller von Festplatten geben diesen Wert an um die durchschnittliche Lebensdauer einer Festplatte anzugeben.

Die MTBF mit n Festplatten berechnet man in einen Raid-5 [4] wie folgt:

$$MTBF_{Raid5} = \frac{MTBF_{Disk}^2}{N * (N-1) * MTTR_{Disk}} \tag{3.3}$$

3.1.5 Verfügbarkeit

Eine Service bzw. eine System gilt als Verfügbar, wenn es seine Tätigkeit vollständig erfüllt für die es bestimmt wurde. Die Wahrscheinlichkeit in welchen eine Service in einer definierten Periode verfügbar ist bezeichnet man als Verfügbarkeit [5]. Im Idealfall darf eine Service nie ausfallen und ist immer Verfügbar. In der Realität ist eine Perfekte Verfügbarkeit nie gewährleistet, es wird jedoch bestrebt die erforderliche oder gewünschte Verfügbarkeit möglichst genau auszudrücken, damit mit Kennzahlen und Metriken die Verfügbarkeit definiert und gemessen werden kann.

Die Verfügbarkeit wird aus dem Verhältnis der Verfügbare Zeit (Updatime) und der nicht Verfügbare Zeit (Downtime) eines Services [5] bemessen:

$$\text{Verfügbarkeit} = \frac{\text{Uptime}}{\text{Downtime} + \text{Updatime}} \tag{3.4}$$

3.1.6 Datenverfügbarkeit

Die Datenverfügbarkeit ist eine Begriff der von Computer Speicherhersteller und Speicher Dienstleister Anbieter “Storage Service Provider (SSP)“ die von Ihren Produkte

oder Services gewährleisteten Verfügbarkeit der Daten und die zugesicherte Antwortzeit beim Zugriff auf die Daten im normalen Betrieb beschreibt [13].

In dieser Arbeit wird der Begriff dazu verwendet die Gewährleistete Verfügbarkeit für den Datenzugriff des Speichersystems zu beschreiben. Die Daten Verfügbarkeit für den Endbenutzer der Webapplikation wird nicht mit diesen Begriff definiert.

3.1.7 Hochverfügbarkeit

Die Autorin Andrea Held beschreibt in Ihren Buch Hochverfügbar wie folgt:

> *Ein System gilt als hochverfügbar, wenn ein Anwendung auch im Fehlerfall weiterhin verfügbar ist und ohne unmittelbaren menschlichen Eingriff weiter genutzt werden kann. In der Konsequenz heisst dies, dass der Anwender kein oder nur eine kurze Unterbrechung wahrnimmt [5].*

Hardvard Research Group hat die Verfügbarkeit in Verfügbarkeitsklassen eingeteilt:

- Conventional (AEC-0): Funktion kann unterbrochen werden, Datenintegrität ist nicht essentiell
- Highly Reliable (AEC-1): Funktion kann unterbrochen werden, Datenintegrität muss jedoch gewährleistet sein.
- High Availability (AEC-2): Funktion darf nur innerhalb festgelegter Zeit bzw. zur Hauptbetriebszeit minimal unterbrochen werden.
- Fault Resilient (AEC-3): Funktion muss innerhalb festgelegter Zeiten bwz. während der Hauptbetriebszeit ununterbrochen aufrechterhalten werden.
- Fault Tolerant (AEC-4): Funktion muss ununterbrochen aufrechterhalten werden, 24*7 Betrieb (24 Stunden, 7 Tage die Woche) muss gewährleistet sein.
- Disaster tolerant (AEC-5): Funktion muss unter allen Umständen verfügbar sein.

[5]

In der Arbeit werden zur Verfügbarkeitsbestimmung die Verfügbarkeitsklassen von Harvard Research Group verwendet.

3.2 Bestandesaufnahme

3.2.1 Applications-Architektur

Die genaue Applications-Architektur wird hier nicht beschrieben, da diese Geschäftsgeheimnisse des Auftraggeber verletzten würden. Es werden deshalb nur teile davon

welche das Geschäftsgeheimnisse nicht verletzten beschrieben.

Die Applikation ist eine Web-Applikation, welche auf dem Web-Framework Ruby On Rails aufsetzt. Wie dem Namen Ruby On Rails zu entnehmen ist basiert das Framework auf der Programmiersprache Ruby. Ruby ist eine dynamische, Objekt-Orientierte, interpretierbare Programmiersprache, hervorgebracht von Yukihiro Matsumoto 1995. Eine Ruby Programm benötigt keine Kompilation bevor es ausgeführt werden kann.

Ruby On Rails auch Rails oder RoR genannt, implementiert eine Model-View-Controller Architektur. Die drei unter Framework, spielen dabei eine signifikanten teil in der Separation: Active Record, Action View, und Action Controller. Die beiden unter Framework sind auch im refabbabb:Rails-architektur dargestellt (nach [3]).

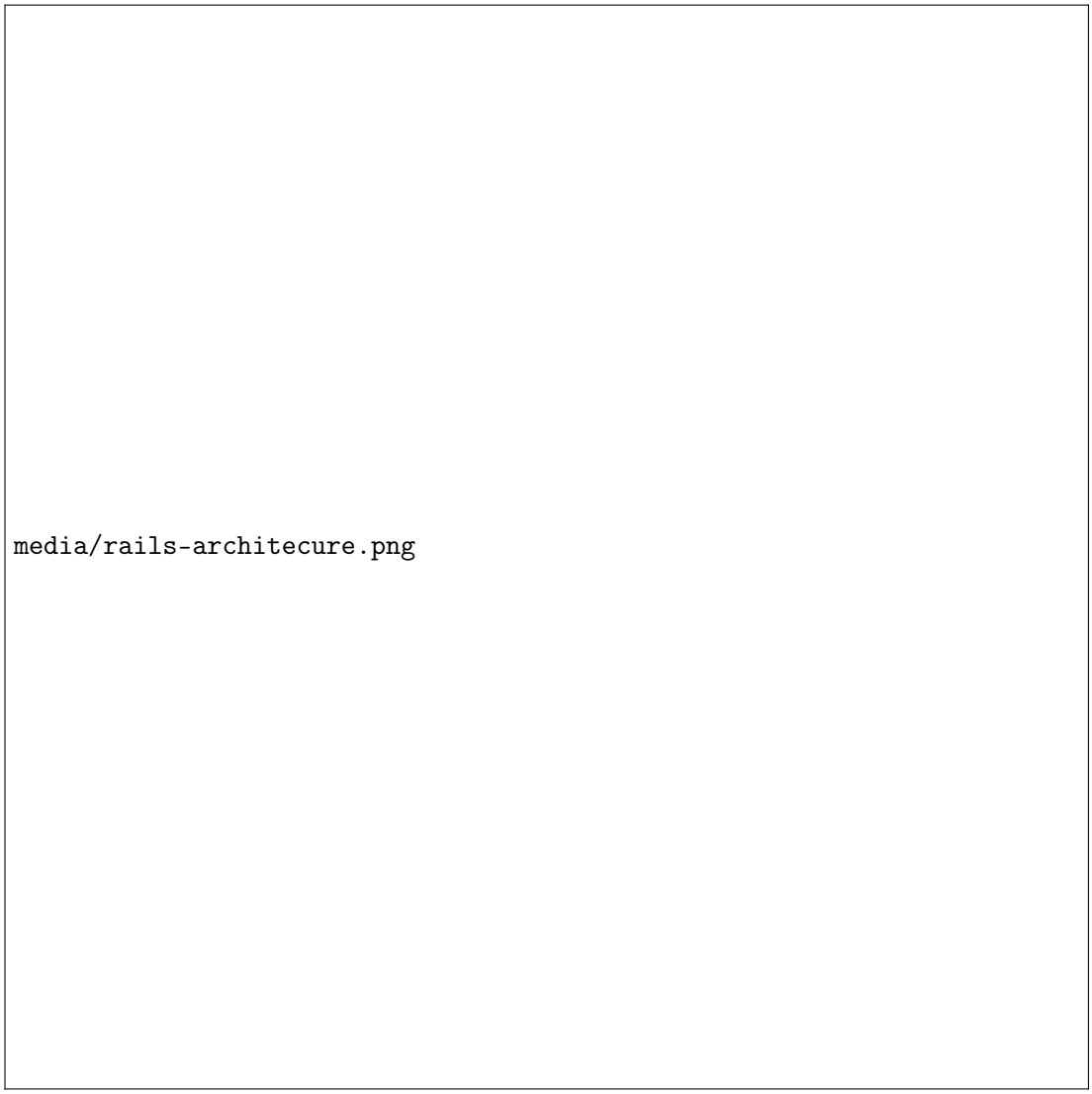

Abbildung 3.2: Rails Architektur *Raid 5* [14]

Beides Ruby On Rails, auch nur Rails genannt, und Ruby sind Open-Source das bedeutet quelloffen. Ruby wird unter der Lizenzen “Ruby License“ und GPL veröffentlicht während Rails unter der "MIT License"veröffentlicht ist.

3.2.2 System-Architektur

Die bestehende System-Architektur besteht aus einen einzelnen Server welche von einen bekannten Deutschen Webhosting-Dienstleister gemietet wird. Beim gemieteten System handelt sich um einen x64 Mikroarchitektur.

3.2.3 Speicher-Architktur

Die Speicher-Architektur besteht aus einen internen Block-Basierenden Raid-5 Speicher. Das bedeutet für das Raid-5 werden die im Server verbauten Festplatten mittels dem Linux Software Raid Kernel[1] und dem Verwaltungstool mdadm[2] zu einer Logischen Einheit zusammengefasst. Für das Raid-5 werden drei "SATA 2" Festplatten mit einer Festplatten Speicherkapazität von je 2 Terabyte bzw. 1.818 Tebibyte verwendet.

Im Raid Speicher ist eine root Dateisystem für Betriebsystem, Webapplikation, Datenbank, Logdateien und Bilddaten installiert.

3.2.4 Speicherkapazität

Die von Raid-5 zur Verfügung gestellte Speicherkapazität beträgt gemäss Berchnung

$$\text{Speicherkapazität Raid-5} = (3-1) * 1.818\,\text{TiB} = 3.636\,\text{TiB} \tag{3.5}$$

Listing 3.1: Report Dateisystem Speicherplatz Belegung in Dezimal Prefix

```
1 root@www1:~# df -h
2 Filesystem Size Used Avail Use% Mounted on
3 simfs 3.8T 2.5T 1.3T 67% /
```

3

3.2.5 Datenwachstum

Das **Abbildung 3.3** Zeigt den Speicherzuwachs von mitte Juni bis Ende November.

[1] https://raid.wiki.kernel.org/
[2] http://neil.brown.name/blog/mdadm
[3] http://www.debianadmin.com/manpages/dfmanpage.htm

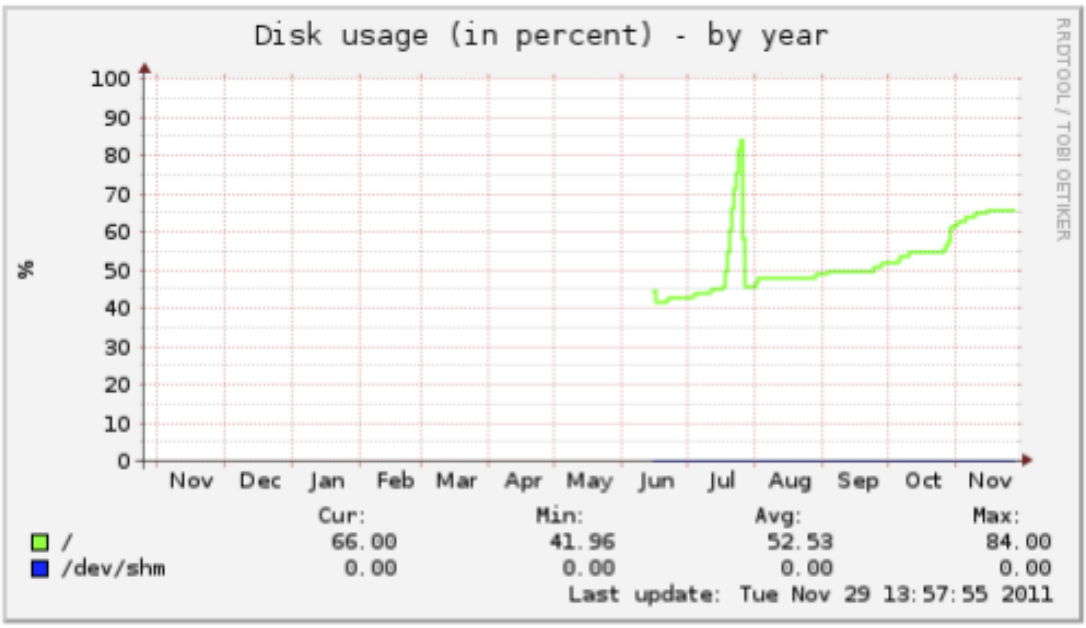

Abbildung 3.3: Verhätnis von Speicheplatzverbrauch zur Speicherkapazität in Prozent in einer Zeitspanne von einen Jahr

3.2.6 Zugriffswachstum

Für die Ist-Aufnahme konnten keine Historischen Messdaten zur Verfügung gestellt werden.

3.2.7 Skalierbarkeit Datenvolumen

Keine Skalierbarkeit während des Betriebs möglich, grössere Speicherkapazität ist nur durch eine Wechsel auf eine anderes Hosting Produkt möglich. Diese bedeutet jedoch eine Migration auf eine neue Server Plattform.

3.2.8 Skalierbarkeit Datenzugriffe

Keine Skalierbarkeit möglich.

3.2.9 Daten Durchsatz I/O

Für die Ist-Aufnahme konnten keine Messdaten zur Verfügung gestellt werden.

3.2.10 Daten Redundanz

Daten Redundanz wird durch Raid-5 gewährleistet.

3.2.11 Datenverfügbarkeit

Stromausfälle und Netzwerkstörungen im Rechenzentrum des Hosters, führten zu Störungen und Ausfällen in der Webapplikation und in der Datenauslieferung. Die Störungen wurden jedoch nicht dokumentiert weshalb keine Statistische messbare aussage über die erreichte Datenverfügbarkeit gemacht werden kann.

3.2.12 Daten Sicherheit

Die Daten sind durch Physischen Fremdzugriff durch eine Verschlüsselung geschützt. Für die Verschlüsselung wird das Device Mapper Module dm-crypt eingesetzt. dm-crypt verschlüsselt somit die Daten bereits auf Blockebene und ist somit für das Dateisystem transparent.

3.2.13 Daten Integrität

Zu jedem Bilder wird eine Hash Checksumme gespeichert, die bei jeden Backup geprüft wird.

3.2.14 Backup

Es wird Täglich mittels dem Tool ccollect[4] ein R-Sync Backup der Daten an einen anderen Standort erstellt.

3.2.15 Wirtschaftlichkeit

Die Kosten für die Speicherung der Daten inklusive Webinfrastruktur sind vergleichsweise tief.

3.2.16 Lokalität

Die Web-Applikation inklusive dessen online Daten werden an einen Standort betrieben. Die Backupdaten werden an einen seperaten Standort gespeichert.

[4] `http://www.nico.schottelius.org/software/ccollect/`

3.3 Analyse-Ergebnisse

3.3.1 Datenwachstum

Die Messdaten aus dem **Abbildung 3.3** zeigen, dass sich das Daten Volumen seit Mitte Juni 2011 bis Ende November 2011 mit Ausnahme einem grösseren Wachstumsschub Ende Oktober von 5% kontinuierlich zugenommen. Die Grosse Spitze lässt sich durch vorübergehende Technische Änderung am System erklären und hat somit für die Auswertung keine besondere Bedeutung. Betrachtet man die erwähnte Zeitspanne hat das Datenvolumen in Bezug zur Speicherkapazität von 40% (1556 Gibibyte bzw. 1.5 Tebibyte) auf beinahe 67% (2568 Gibibyte bzw. 2.5 Tebibyte) in 5 Monaten zugenommen. Das durchschnittliche Datenwachstum beträgt somit gemäss Messdaten pro Monat ca. 5% (195 Gibibyte bzw. 0.19 Tebibyte). Geht man davon aus, das sich das Datenwachstum auf dem gleichen Niveau vorsetzt, ist die Speicherkapazität innert 5 Monate ausgeschöpft.

Ein geplantes neues Feature der Web-Applikation würde den Speicherplatzbedarf verdoppeln. Aufgrund der Vorhanden Speicherkapazität kann das Entwickelte Feature zur Zeit nicht Produktive verwendet werden. Nach Inbetriebnahme des Feature würde sich das Datenwachstum um den Faktor 2 auf 390 Gibibyte steigern.

3.3.2 Skalierbarkeit Datenvolumen

Eine Skalierung kann nur durch hinzufügen von weiteren Festplatten oder durch Migration auf grösseren Festplatten erfolgen. Die maximale Anzahl an Festplatten wird durch die Anzahl vorhandenen SATA Anschlüsse begrenzt. Da es sich jedoch um eine Hosting Produkt handelt ist, eine Skalierung nur mit einen wechseln auf eine besser ausgebautes Hosting Produkt möglich. Der selbe Hosting Provider bietet eine Produkt mit 15 Festplatten SATA mit je 3 Terrabyte Speicherkapazität an. Die Maximale Raid-5 Speicherkapazität 3.7 ist aufgrund der Anzahl Disk und der Speicherkapazität nicht die Ideale Konfiguration, da die Gefahr eines Doppelten Disk Ausfall durch die geringere MTTF 3.9, welche durch die Zunahme der Anzahl Festplatten sinkt 3.8 und der höheren Rebuild Zeit steigt. Speichersystem Hersteller wie NetApp setze bei dieser Konfiguration auf Raid-6 welche doppelte Parität bietet und somit zwei Festplattenausfälle kompensieren können.

Festplattenkapazität in Tebibyte:

$$3\,\mathrm{TB} = 3 * \frac{1000^4}{1024^4} = 2.7285\,\mathrm{TiB} \tag{3.6}$$

Maximale Raid-5 Speicherkapazität:

$$(15 - 1) * 2.7285\,\text{TiB} = 38.199\,\text{TiB} \tag{3.7}$$

MTBF 1 Festplatten (ST33000650SS):

$$1'200'000\,\text{h} \tag{3.8}$$

MTBF 15 Festplatten (Raid-5) (ST33000650SS):

$$\frac{1'200'000\,\text{h}}{15} = 8'000\,\text{h} \tag{3.9}$$

3.3.3 Skalierbarkeit Datenzugriffe

Die Skalierung im Bereich Datendurchsatz können durch schnellere Festplatten, durch Verteilung der IO Operationen auf verschiedene Festplatten oder durch Verteilung der Daten auf Verschiedene Systeme erreicht werden.

Skalierung durch schneller Festplatten: Modernere Festplatten mit Ausnahme von den teuren Solid State Disk (SSD) Festplatten bieten zwar eine grössere Speicherdichte aber aufgrund der erreichten Physikalischen Grenzen keine bedeutend grössere IOPS. Eine Festplatte mit 7200 RPM erreicht durchschnittlich eine IOPS zwischen 75 und 100 IOPS bei SSD Disk wurden schon 1'190'000 IOPS in einen einzelnen PCI Device erreicht .[11] [6] Ein Wechsel auf die genannten SSD Festplatten, ist jedoch aktuelle noch mit höheren Kosten pro Gigabyte verbunden. Gartner geht davon aus das 2012 die Preise pro Gigabyte bei SSD auf durchschnittliche 1$ betragen wird, wobei heute die Preise bei Festplatten bei 30 Cents pro Gigabyte sind. Aus diesen Grund wird diese bei grossen Datenvolumen meist noch nicht eingesetzt, wenn nicht spezielle Anforderungen an die Datenzugriffs Performance gestellt werden.[2]

Skalierung duch Verteilung der IO Operationen: Ein Raid oder Volume Manager bietet die Möglichkeit, die IO Operationen auf mehre Festplatten zu verteilen. In einen Raid-5 verkleinert sich wie**Kapitel 3.3.2** beschrieben die MTTF mit jeder weiteren Festplatte. Zudem sind die verfügbaren Anschlüsse in einen Server ebenfall eine Limitierenden Faktor.

Skalierung durch Verteilung der Daten: Nicht alle Kategorien von Daten haben die gleiche Anforderungen an den Durchsatz. Datenbanken z.B. benötigen in der Regeln einen höheren IOPS als Statische Daten welche wenige Abgefragt werden. Für die Verteilung der Daten ist eine Änderung in der System Architektur und allenfalls in der Applications Architektur notwendig. Ein Beispiel für eine Anpassung der System Architektur wäre die Auslagerung der Datenbank auf eine separates System welches mit schnellen Solid State Disk ausgerüstet wäre.

In einer Webapplikation kann mittels Casching

3.3.4 Redundanz

Eine Raid-5 System bietet wie im **Absatz 3.1.2** beschrieben, keine 1:1 Redundanz. Die Daten sind im Raid-5 nicht doppelt gespeichert, sondern werden bei einen Datenverlust mittels XOR Operation aus den Parität Stripe und den vorhandenen Daten Stripe berechnet. Die Berechnung erfolgt auch bei einen Online Zugriff auf einen Verlorenen Daten Strip, durch die Berechnung ist

Bei einen Zugriff auf einen Verlorenen Daten Strip werden die Daten Online berechnet.

Treten bei einer Festplatten Fehler auf, werden die verlorenen Daten bei einen Zugriff aus dem Parität Stripe Einheit . Der Zugriff auf die Daten ist durch die Berechnung der Daten aus dem Parität Stripe Einheit Online möglich. Wird die ausgefallen Festplatte durch eine neu intakte Festplatte ersetzt können die Daten durch eine Rebuild während des Betriebs wieder hergestellt werden. Durch die Berechnung und Schreib/Lese Operationen im Raid währende des Wiederherstellung Prozess, verschlechtert sich der Datendurchsatz und I/O Rate für den Datenzugriff.

Einen Ausfall einer weiteren Festplatten kann das Raid-5 System nicht kompensieren und führt zu einen Datenverlust aller Online Daten. Die Daten müssten durch eine Backup wiederhergestellt werden, was nur im Offline Betrieb möglich ist.

Bei Festplatten welchen aus dem gleichen Produktionszyklus stammen, kann die Wahrscheinlichkeit eines weiteren Ausfall höher sein, im vergleich zu Festplatten die aus unterschiedlichen Produktionszyklen stammen.

3.3.5 Datenverfügbarkeit

Der Hosting Anbieter gewährleistet gemäss Allgemeiner Geschäfts Bedingungen nur eine Netzwerkverfügbarkeit. Die eine Verfügbarkeit von 99% im Jahresmittel gewährleistet. Eine Verfügbarkeit der Restlichen Infrastruktur und System Verfügbarkeit wird nicht erwähnt. [1] Der Auftraggeber muss in diesen fall selber für die Verfügbarkeit sorgen.

Die im **Absatz 3.2.2** Beschriebene System-Architektur gewährleistet keine Hochverfügbarkeit der Daten. Gründe dafür sind, das die Applikation nur auf einen einzelnen Server betrieben wird, dessen Hardware mit Ausnahme der Festplatten keine Redundanz und für Hochverfügbar ausgelegt sind.

Die Applikation wird auf einen einzelnen Server betrieben. Der Server und dessen Hardware mit Ausnahme der Festplatten stellen einen “Single Point of Failure“ (SPOF) dar.

Fällt der Server wegen eines Hardware Defekts oder Software Fehler aus, ist kein Zugriff auf dem Server möglich.

Zudem werden die Daten und Applikation nur an einen Standort Betrieben, treten unvorhersehbare Ereignisse, wie z.B. eine lokale Stromausfall, Brand, Naturkatastrophen, kann der Betrieb nicht ohne Wiederherstellung des ganzen Systems inklusive Daten von Backup weiter geführt werden. So fern nicht bereits ein gemietet Ersatz Server zur Verfügung muss ebenfalls die Bestelldauer und Auslieferung bei einen anderen Hoster zur Ausfallzeit mit einberechnet werden.Der Service wäre werden dieser Zeit nicht verfügbar.

— Skalierbarkeit Datenvolumen — Bei Hosting Angeboten bezieht man in Normalfall Fertige Produkte, die Sich im Server Typ, Grösse des Speichers, der Anzahl Festplatten, CPU und Memory unterscheiden. Benötigt man für seine Applikation mehr Leistung oder einen grösseren Speicher muss man das Produkt wechseln, was meist mit einer Migration auf einen anderen Server verbunden ist.

Die jetzige Speicher- und Applikations- Architektur ist auf dem Betrieb auf einen Server ausgelegt. Für die Skalierbarkeit des Datenvolumen bedeutet diese, dass man das Datenvolumen nur durch grössere Festplatten oder durch eine Grösse Anzahl der Festplatten ausgebaut werden kann.

4 Soll-Analyse

4.1 Szenario 1 - Schwaches Wachstum gespeicherten Daten / schwaches Wachstum der Abfragen

Im Szenario wird davon ausgegangen,

4.2 Szenario 2 - Schwaches Wachstum Daten / starkes Wachstum der Abfragen

4.3 Szenario 3 - Starkes Wachstum Daten / schwaches Wachstum der Abfragen

4.4 Szenario 4 - Starkes Wachstum Daten / starkes Wachstum der Abfragen

5 Speicher-Markt

5.1 Speicherarchitekturen

Die heutigen Speicherarchitekturen können in Block- (Block-Based), Datei- (File-Based) und Objekt-Basierende Adressierende unterteilt werden.

5.2 Block-Basierend

> *Since the first disk drive in 1956,1 disks have grown by over six orders of magnitude in density and over four orders in performance, yet the storage interface (i.e., blocks) has remained largely unchanged. Although the stability of the block-based interfaces of SCSI and ATA/IDE has benefited systems, it is now becoming a lim- iting factor for many storage architectures. As storage infrastructures increase in both size and complexity, the functions system designers want to perform are fundamentally limited by the block interface.*

[8]

Vergleicht man die erste Festplatte welche von IBM Produziert wurde mit einer Seagate von 2011, hat sich die Speicherdichte von 2000 bit per Quadratzoll auf 625 GByte und in der Geschwindigkeit von 8 kbytes auf 600 MB verbessert.[9][10]

Zu den Block-Basierenden Speicher Systemen zählen Direct Attached Storage (DAS) und Storage Area Network (SAN).

Bei DAS die Festplatten direkt an den Server Angeschlossen, als übliche Schnittstelle kommen meist SCSI zum Einsatz. Da DAS Storage nicht mehreren Server zur Verfügung gestellt werden können, kommt diese Form von Speicher für

Blocks offer fast, scalable access to shared data; but without a file server to authorize the I/O and maintain the metadata, this direct access comes at the cost of limited security and data sharing.

Direct Attached Storage Storage Area Network

5.2.1 ISCSI

Datenverfügbarkeit / Redundanz

Skalierbarkeit Datenvolumen / Datenzugriffe

Integität

Durchsatz I/O

Lokalität

Backup

5.3 Datei-Basierend

5.3.1 Network File System

Das Network File System Protocol wurde von der Firma SUN () 1984 vorgestellt und ermöglicht es über das Netzwerk auf Dateisysteme eines anderen Host (Server) zu zugreifen als würde der Zugriff Lokal stattfinden. Das Protokoll von wurde mit der Version 2 1989 zum ersten mal von Internet Standard Request for Comments (RFC) unter der Nummer 1094[1] standardisiert. Die Version 2 von NFS verwendet ausschliesslich das UDP Transportprotokoll. Mit Version 3 RFC 1813[2] die im Jahr 1995 veröffentlicht wurde NFS Maschinen, Betriebsystem und Netzwerk Architektur, und Transport-Protokoll unabhängig. Die Unabhängigkeit wird mit der Verwendung von Remote Procedure Call () welches wiederum ein eXternal Data Representation (XDR) verwendet erreicht. Das wurde mittels dem separaten Protokoll Network Lock Manager (NLM) erreicht.

5.3.2 NAS Appliance

Network Attached Storage sind Speichersystem mit angepassten Datei System für den gemeinsamer Dateizugriff in einen Hetrogenen Computer Netzwerk welche über ein LAN angeschlossen sind. Als Speicher verwenden NAS je nach Typ interne Festplatten, Direct Attached Storage oder über eine SAN angefügten Speicher. An Clients stellen NAS Ihren Speicher über NFS, CIFS, ISCSI zur Verfügung. High-End NAS können Ihren Speicher wiederum über Fibre-Channel zur Verfügung stellen.

Gemäss Gartner gehören die Anbieter IBM, EMC und NetAPP zu den führenden NAS Anbieter in Midrange und High-End bereich. Wobei gemäss Garnter Magic Quadrant

1
2

Netapp zusammen mit EMC zu den innovativsten Anbieter.

Strengths

- *NetApp remains one of the few truly unified storage providers among all top-tier vendors, with its software features continuing to be industry benchmarks. The company was able to regain some of the NAS revenue market share that it had lost in 2009. Its fast revenue growth in 2010 was driven by its successful campaign targeted at midsize enterprises with the value propositions of NFS supporting VMware and unified storage in consolidating Windows application storage.*
- *In 2010, NetApp increased its aggregate up to 100TB with Data ONTAP 8.0.1 and introduced compression to complement its popular deduplication capability. It added a RESTful object storage interface (based on its acquisition of Bycast) to its unified storage, targeting global content repositories. On the hardware side, it launched new systems with better performance and denser disk shelves.*
- *NetApp's new software bundles have simplified the procurement process and made software pricing more affordable. For customers seeking converged infrastructure, NetApp launched FlexPod for VMware with its partners Cisco and VMware, offering packages including servers, storage and switches.*

Cautions

- *The vast majority of the Data ONTAP 8.0 adoption was on the 7 mode (instead of the cluster mode) for larger aggregates, while the early adoption of the cluster mode focuses on high- performance NFS file services. The cluster mode is not ready for mainstream enterprise customers who require those 7-mode features that are still missing in the cluster mode. The ONTAP 8.1 scheduled for release later this year will likely continue to support the two modes: clustered and nonclustered modes of operation.*
- *While NetApp continues to enjoy its leading edge in unified storage, it's facing fiercer competition in the high-end NAS market, where file systems larger than 100TB are required and where high performance without the expensive Flash Cache is desired. NetApp is also challenged in the low-end NAS and unified storage market with new products from both major and emerging competitors.*

[?]

Datenverfügbarkeit / Redundanz

Skalierbarkeit Datenvolumen / Datenzugriffe

Integität

Durchsatz I/O

Lokalität

Backup

5.4 Objekt-Basierend

Festplatte in 1956 von IBM 1956 erschienen ist,

Seit die erste pro Sekunde gesteigert, das Speicher Schnittstelle (d.h. Blöcke) blieb weitgehend unverändert. Auch wenn bisher die Systeme von der Stabilität Block-Basierende Speicher Schnittstellen wie SCSI und ATA/IDE profitiert haben, sind Sie heute mehr den mehr der limitierende Faktor von vielen Speicherarchitekturen geworden.

Datenverfügbarkeit / Redundanz

Skalierbarkeit Datenvolumen / Datenzugriffe

Integität

Durchsatz I/O

Lokalität

Backup

6 Machbarkeitsnachweis

Glossar

. 18

. 18

RFC

Request for Comments (RFC) sind Dokumente, über Internet, inklusive der technischen Spezifikation und Richtlinien, welche von der Organisation Internet Engineering Task Force entwickelt wurde. “Das RFC wird erst nach erfolgter Diskussion unter der Aussicht des Internet Architecture Board (IAB) herausgegeben und fungiert als Quasistandard. Jedes RFC enthält eine eindeutige, vorlaufende Nummer, die kein zweites Mal zu gewiesen wird.“ [**?**] `http://www.rfc-editor.org/`. 18

UDP

adfajsdfjadslkfjaödjfölaksdjfajsklfj. 18

XDR

Die eXternal Data Representation (XDR) Spezifikation stellt ein Standardisierte Verfahren zur Präsentation von gebräuchlichsten Daten Typen über das Netzwerk zur Verfügung. Dies löst das Problem der verschiedenen Byte-Reihenfolge (Big Endian), Speicherausrichtung auf unterschiedlichen Kommunikations Partner.. 18

-@alias

Literaturverzeichnis

[1] Ag, Hetzner Online: *Allgemeine Geschäftsbedingungen*. 2009.

[2] Agam Shah, IDG News: *Consumer SSDs to Break out in 2012, Gartner Says*. 2011.

[3] Bächle, Michael Paul Kirchberg: *Ruby on Rails*. Ieee Software, 24(December):105–108, 2007.

[4] Chen, Peter M.: *Reliable Secondary Storage*. Computing, 26(2), 1994.

[5] Held, Andrea: *Oracle 10g Hochverfügbarkeit*. Addison-Wesley, München, 2004.

[6] io, Fusion: *Maximizing Performance for Large Datasets*.

[7] Kuratti, Anand, William H Sanders W Main St: *PERFORMANCE ANALYSIS OF THE RAID 5 DISK ARRAY*. Access, 1995.

[8] Mesnier, Mike, Carnegie Mellon Gregory R Ganger: *Object-Based Storage*. IEEE Communications Magazine, 41(August):84–90, 2003.

[9] Seagate: *Barracuda XT Data Sheet*, 2011.

[10] Seagate: *Seagate Breaks Areal Density Barrier: Unveils The World's First Hard Drive Featuring 1 Terabyte Per Platter*, 2011.

[11] Symantec, Ianatkin: *Getting the hang of IOPS*, 2011.

[12] Technology, National Institute of Standards and: *Prefixes for binary multiples*, 1998.

[13] TechTarget: *Defintion: data availability*, 2001.

[14] Wikipedia: *RAID*, 2006.

Anhang