Mai 2012 Blog Beitrag Website - Storage Consortium
Autor: Norbert E. Deuschle, Business Consulting & Research, Inh.

Thema: Kurze Analyse zur „stillen" Datenkorruption im Storage Stack oder: Lassen sich auch große Mengen an Daten in der Cloud fehlerfrei speichern?

Zum Hintergrund: Betreiber von Enterprise-Disk-Storage (Cloud-Anbieter wie die interne IT) verlassen sich zu 100 Prozent und implizit auf die Datenintegrität Ihrer Systeme, wenn wichtige Anwendungsdaten gespeichert bzw. archiviert werden. Storage-Arrays z.B. verfügen hierzu über mehrere Ebenen von Integritäts-Prüfungen, sowohl Controller- als auch Diskseitig. Trotzdem konnten einige Untersuchungen (siehe Quellen am Textende) die Entstehung von korrupter Daten bei großen Speicherumgebungen beobachten. Die Fehlerraten sind zwar bei (silent-)data corruption auf dem Papier nicht hoch (2% latente Sektoren-Fehler-Rate bei Enterprise Disks innerhalb von 2 Jahren; 19% im selben Zeitraum bei Nearline-Storage-Systemen und Fehlerraten durch stille Datenkorruption von 0.06% bei Enterprise Disks (0.6% bei Nearline), ebenfalls im Zeitraum von zwei Jahren), weisen aber auf ein Phänomen hin, das im Zusammenhang mit der raschen Zunahme von unstrukturierten Daten und Cloud-Archivlösungen (auf Festplattenbasis) aus Anbietersicht beachtet werden sollte.
Nicht unproblematisch ist eben der Trend zu mehr Daten, der zwangsläufig zu einem sehr ausgeprägten Kostenbewusstsein auf der Einkaufsseite führt (dies gilt natürlich analog auch für Cloud Storage Anbieter); nicht nur auf Grund steigender Storage-Verwaltungskosten bzw. Aufwendungen für Sicherungsverfahren etc. Und eines ist sicher: Die Anzahl an ausgelieferten Festplatten wird in den nächsten Jahren nicht abnehmen. Zitat „…Storage demand growth right now is over 50% in the cloud, in other businesses its 25%. Overall call it 40%. Areal density growth is right now under 25%. So you're going to say, well, wait a second. If this is growing at 40, and this is growing at 25, how are you going to fill that gap. The only way we can do it is more heads and disks... " Interview, Zitatauszug: Seagate CEO Steve Luczo, March 2012, Quelle Forbes.com, US.

In diesem Blogbeitrag versuche ich deshalb, die Problematik der sog. stillen Datenkorruption kurz aufzuzeigen und verweise in diesem Zusammenhang ausdrücklich auf einige wissenschaftliche Untersuchungen, welche die Phänomene ausführlich analysiert haben (University of Wisconsin-Madison, CERN, N. Bairavasundaram etc. - siehe Quellenangabe).

1. Silent Data Corruption, Parity Pollution, um was geht es?

Das Verschwinden von Bits ist zwar ein seltenes Phänomen, kann aber über die Zeit dazu führen, dass Informationen unbemerkt verloren gehen. Im Prinzip werden beim Schreiben der Daten Bits geändert und können beim Lesen nicht korrekt wiedergegeben werden. Storage- und Filesysteme sollten eigentlich das fehlerhafte Schreiben erkennen, den Fehler melden und dann beheben. Bei der schleichenden Datenkorruption können jedoch die geänderten Bits nicht erkannt werden; indirekt wird damit das Schreiben von fehlerhaften Daten unfreiwillig unterstützt. Bei stark steigenden Datenmengen, sprich Anzahl Festplatten steigt statistisch gesehen damit auch die Gefahr, Daten zu verlieren. Ohne permanente Backups und der Möglichkeit, Silent Data Corruption beim fehlerhaften Schreiben zu erkennen werden diese Fehler erst erkannt, wenn Daten bereits verloren sind.
Silent Data Corruption ist Hardware-seitig meist auf die große Anzahl von unterschiedlichen Controllern, Speicher, Chipsätze - im Nearline-Bereich aus Kostengründen gerne kombiniert mit preiswerten Disks wie z.B. (S)ATA Desktop-Varianten. Hinzu kommt, dass alle Daten in der Regel eine lange Kette von Geräten durchlaufen, was die Fehleranfälligkeit erhöht. Data Corruption kann dabei entstehen durch:

- Controller Versagen, während die Daten gerade im Cache sind
- Ein längerer Stromausfall mit kritischen Daten im Cache
- Plötzlicher Controller-Fehler während des Schreibvorgangs
- Fehler beim Lesen von Daten von Disk

Weitere Ursachen für die genannten Probleme können sein: Bootstorms, veraltete oder fehlerbehaftete Softwarefragmente in der Registry oder das Überschreiben von Treibern, statt diese sauber zu löschen.

a) Welche Arten von Datenkorruption sind möglich, welche Auswirkungen können diese haben und wie können sie identifiziert bzw. korrigiert werden? *(Quelle: An Analysis of Data Corruption in the Storage Stack / siehe Textende)*

Fehlerquelle: Checksum mismatch
Auswirkung: Bit-level corruption; misdirected, torn write
Erkennung: RAID block checksum beim Disk-Read

Fehlerquelle: Identity discrepancy
Auswirkung: Lost or misdirected write
Erkennung: File system-level block identity bei File Reads

Fehlerquelle: Parity inconsistency
Auswirkung: Memory corruption; lost write, bad parity calculation
Erkennung: RAID parity mismatch.
Lösung: Data Scrubbing.

Architekturseitig ist der Weg der Daten deshalb durch die Umsetzung verschiedener ECCs (Error Correction Code) und CRCs (Cyclic Redundancy Check) begleitet:

1. Der Speicher korrigiert einzelne Bitfehler
2. Der Cache ist ECC-protected
3. PCIe und SATA-Anschlüsse haben CRC implementiert
4. Der Festplatten-Cache arbeitet mit ECC und das Schreiben auf die Festplatte ist durch ECC-Mechanismen abgesichert.
5. CRC wird implementiert, um bis zu 32 Byte-Fehler (pro 256 Bytes) zu korrigieren. Die Daten sind dann meist fünf Mal gecheckt, bevor sie die eigentliche Festplatte erreichen.

Enterprise-SCSI-Laufwerke (z.B. Fibre Channel- oder SAS Drives), wie sie bei Disk-Arrays führender Anbieter verwendet werden, arbeiten normalerweise mit 520-Byte-Sektoren-Größe:
Für jeden Block stehen 32KB zur Verfügung und es existieren 64 Sektoren. Jeder Sektor besitzt 512 Bytes an Daten und zusätzlich 8 Bytes für DIF (Data Integrity Field) Prüfsummen sowie für weitere herstellerspezifische Verwendungen.
Bei ATA-Laufwerke (einschließlich SATA) passen in der Regel 65 Sektoren in einen 32KB-Block. Diese Sektoren sind 512 Byte groß und werden ausschließlich für die Datenspeicherung verwendet.

Es gibt jedoch in der Regel hier keine vergleichbaren Prüfsummen (checksum). Siehe hierzu auch die Ergebnisse der CERN-Studie:

Desktop disk:	10^-14 (1 bit in ~11.3 TiB)
Enterprise disk:	10^-15 (1 bit in ~113 TiB)

b) Parity-Pollution

Ein weiterer Grund für schleichende Datenkorruption liegt in der Zunahme der Festplattengrößen (3TB+) der letzten Jahre in Kombination mit Parity-basierten RAID-Verfahren begründet: Wenn (silent) korrupte Daten im System verwendet werden, um neue Parity Informationen im RAID-Array zu erzeugen, kann das RAID-System die Parity-Daten nicht mehr für den Rebuild von nicht-korrumpierten Daten verwenden (Parity Pollution). Durch die Kombination von Code-Bugs in der Firmware und möglichen Hardware-Defekten können einzelne Daten - während des I/Os vom Controller auf die Festplatte und wieder zurück - unentdeckt von der Applikation beschädigt werden (siehe auch > *http://insidehpc.com/2008/07/08/a-silent-sata-drive-failure-problem/*)

Dies gilt übrigens auch für Dateisysteme, welche auf einen defekten Datenblock auf der Festplatte zugreifen, diesen Defekt jedoch nicht identifizieren können... dieser wird dann erst von der Applikation registriert, wenn sie den beschädigten Datenblock nicht einlesen kann. Eine wissenschaftliche Studie „An Analysis of Data Corruption in the Storage Stack" [1] über 1,53 Millionen Festplatten im 41 Monats-Zeitfenster hat ergeben, dass 8,0% der dort untersuchten SATA-Festplatten potentielle Kandidaten für stille Korruption darstellen. Diverse Festplatten-Array-Prüfungen laufen deshalb im Hintergrund (Datenebene, Parity-RAID usw.) und können wie oben beschrieben diese Art von Fehler abfangen. Die Studie ergab auch, dass 13% der Fehler durch die parallel ablaufende Überprüfung verpasst wurden. Im einzelnen wurde dies auf die bereits oben beschriebenen drei Fehlerklassen zurückgeführt (Checksum mismatches, Identity discrepancies und Parity inconsistencies).

Die Auswirkungen sind laut o.g. Analyse zwar nicht gravierend, sollten aber für die Entwickler von Storagesystemen Anlaß genug sein, die beschriebenen Phänomene ernst zu nehmen... *„We find that only a small fraction of disks develop checksum mismatches.*

...An even smaller fraction are due to identity discrepancies or parity inconsistencies. Even with the small number of errors observed it is still critical to detect and recover from these errors since data loss is rarely tolerated in enterprise-class and archival storage systems.

Für eine RAID-5 (4 + P)-Konfiguration bei 930 GB pro 1 TB SATA-Laufwerk (Nutzdaten) rechnet die Analyse vor, dass 15 nichtentdeckte Fehler pro Petabyte auftreten könnten. Wenn ein System ständig Daten mit 200 MB/s liest, würde in dieser Konfiguration statistisch alle 100 Stunden jeweils ein Fehler auftreten.

Eine **weitere Untersuchung [2]** zu den Ausfallszenarien für komplette Storage-Systeme kommt zu folgendem Ergebnis: In den 39.000 Speichersysteme, die analysiert wurden, zeigte das Firmware-Protokoll zwischen 5% und 10% Ausfallraten. Dies sind Fehler, welche auf einen fehlerhaften Code in der Disksystem-Firmware zurückgeführt werden konnte. In Folge kann dies wiederum zu stiller Korruption führen. Umso wichtiger ist deshalb ein kontinuierliches Update-Procedere sowohl durch den Storage Hersteller als auch die Umsetzung beim Service und-/oder die Betreiber der Systeme vor Ort.

Die **CERN-Studie [3]** kam hierzu bereits 2007 in einer hausinternen Analyse zu folgendem Schluss: „Wir haben festgestellt, dass ein niedriger Level von Data Corruption vorhanden ist und dass dieser auf verschiedene Ursachen zurückzuführen ist. Man kann nun versuchen, diese Fehler zu reduzieren, aber es ist sehr unwahrscheinlich, dass sie ganz verschwinden werden. Wir beobachteten vielmehr über die Zeit (durch neue Hardware, Software, Firmware etc.) eine Zunahme von Datenkorruption und werden deshalb eine konstante und sorgfältige Überwachung der Situation anstreben... Die Umsetzung dieser Anstrengungen wird jedoch zu einer Verdoppelung der ursprünglichen geplanten I/O-Performance auf den Disk-Servern führen und zieht eine Erhöhung der verfügbaren CPU-Kapazitäten dieser Server (50%?!) nach sich. Dies wird natürlich einen Einfluss auf die Kosten und Dimensionierung der CERN-Computing-Systeme nach sich ziehen. Zusätzliche steigt damit die operative Belastung durch den erhöhten Administrationsaufwand… *"Zitatübersetzung ND aus „Data integrity, Bernd Panzer-Steindel, CERN/IT Draft 1.3 vom 8. April 2007 (siehe Quellenangabe unten).*

3. Bisheriges Fazit:

Es gibt wenige veröffentlichte Reports zu Festplattenfehlern. Die meisten Studien untersuchen die Fehler entweder von Laufwerksausfällen oder der bereits angesprochenen latenten Sektorfehler. Eine Ausnahme ist die Analyse der Universitäten von Wisconsin-Madison und Toronto, übrigens mit Unterstützung der Firma NetApp. Interessant ist in diesem Zusammenhang auch: Bei der groß angelegten Analyse von Festplattenausfällen von Schröder und Gibson (Analyse der Daten von rund 100.000 Platten über einen Fünf-Jahres-Zeitraum) kam heraus, dass die Ausfallrate bzw. Fehlerquote nicht zwingend mit dem Festplattenalter korreliert.
Aufgrund der geringen Fehlerwahrscheinlichkeit erinnert mich deshalb das Thema auch etwas an die Möglichkeit von Hash-Kollisionen bei Daten-Deduplizierung. Auch dort können unter bestimmten Bedingungen Daten verloren gehen, wenngleich dies erst im Exabyte-Bereich statistisch relevant wird (weitere Randbedingung: herstellerspezifische Implementierungen).

Andererseits bleibt als Fazit vor dem Hintergrund des raschen Datenwachstums und der Zunahme von Cloud Services - Angeboten die Forderung an die Hersteller von Storage-Lösungen um so deutlicher bestehen, so weit wie möglich allen technologiebedingten potentiellen Datenverlusten vorzubeugen.

Im zweiten Teil dieses Blogs -folgt in Kürze- wird auf die Behandlung von korrupten Daten bei (Storage-)Filesystemen eingegangen. Das Erkennung und die Wiederherstellung von diesen Daten ist wie gesehen kein triviales Problem und Dateisysteme können nicht immer zweifelsfrei erkennen, ob ein Festplattenfehler aufgetreten ist. Viele Filesysteme verwenden eine proprietäre Form der Festplatten-Blocktyp-Überprüfung, allerdings arbeitet die Mehrzahl derzeit noch auf Basis von Disk ECCs und Checks, um interne Fehler zu melden. Zwei der neueren Filesysteme, die selbst eigene Prüfsummen verwenden, um beschädigte Daten zu erkennen, sind Sun Oracle ZFS und Google GFS. Mehr dazu aber im nächsten Blog...

Quellenangaben:

[1] Bairavasundaram et al. “An Analysis of Latent Sector Errors in Disk Drives.” Proceedings of the International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'07). June 2007.
B. Schroeder and G. A. Gibson
Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?
In Proceedings of the 5th USENIX Symposium on File and Storage Technologies (FAST '07), San Jose, California, Feb. 2007.
“An Analysis of Data Corruption in the Storage Stack.” Proceedings of the 6th USENIX conference on File and Storage Technologies (FAST'08). February 2008
http://static.usenix.org/events/fast08/tech/full_papers/bairavasundaram/bairavasundaram_html/main.html
[2] Jiang et al. “Are Disks the Dominant Contributor for Storage Failures?” Proceedings of the 6th USENIX conference on File and Storage Technologies (FAST'08). February 2008
Krioukov et al. “Parity Lost and Parity Regained.”
Proceedings of the 6th USENIX conference on File and Storage Technologies (FAST'08). February 2008
[3] Panzer-Steindel. “Data Integrity.” Internal CERN/IT study. 8. April 2007
http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797
W. Jiang, C. Hu, A. Kanevsky, and Y. Zhou.
Is Disk the Dominant Contributor for Storage Subsystem Failures? A Comprehensive Study of Failure Characteristics. In Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST '08), San Jose, California, Feb. 2008.

end of document 11/05/2012.