Isilon Solutions + OneFS

Anne-Victoria Meyer

Betreuer: Dr. Julian Kunkel

Proseminar:
Ein-/Ausgabe - Stand der Wissenschaft

16. Oktober 2013

Contents

1 Einleitung

Die Menge an Daten, die heutzutage produziert wird, sowie gespeichert und verwaltet werden muss, nimmt überall zu. Ein Beispiel von vielen ist die Filmindustrie. Mit zunehmender Auflösung der Aufnahmen, steigt auch der Speicherbedarf. Zu diesen Mengen an Rohmaterial addiert sich dann ebenfalls noch alles an Daten, was bei der aufwändigen, digitalen Nachbearbeitung entsteht. Bei diesen immer weiter zunehmenden Mengen an Daten, darf auch die effiziente Speicherung und Verwaltung dieser nicht außer Acht gelassen werden. Viele traditionelle Speichersysteme lassen sich nur mühsam erweitern und anpassen, um dieser Anforderung gerecht zu werden.
In dieser Arbeit soll es nun um die Speicherlösungen gehen, die hierfür von EMC Isilon angeboten werden. Diese sind eben dafür optimiert, auch mit großen Datenmengen zu funktionieren, und dass ohne größeren Aufwand. Die Besonderheit dieser Speicherlösungen ist eben gerade, dass man mit wenig Speicher anfangen, und ohne eine Veränderung der Infrastruktur diesen Speicher erweitern kann.
Im Folgenden werde ich zunächst auf die grundlegende Technologie dieser Speichersysteme eingehen. Anschließend werde ich die Hardware und das Datei- und Betriebssystem, das auf dieser operiert, vorstellen. Abschließend werde ich dann einen Einblick in die Administration des Speichersystems geben, und an einem Beispiel erläutern, wie eine Datei in den Speicher geschrieben wird.

1.1 Scale-Out-NAS

Bevor ich anfange, die Speichersysteme von EMC Isilon vorzustellen, möchte ich an diesem Punkt ersteinmal aufzeigen, was genau Scale-Out-Network-Attached-Storage (Scale-Out-NAS) ist. NAS ist Speicher, der in einem Netzwerk von einem Server zur Verfügung gestellt wird. Dieser Server verwaltet den Speicher und regelt den Zugriff. So können mehrere Clients auf die gleichen Daten zugreifen, ohne die Gefahr, dass dadurch korrupte Daten entstehen. Wenn nun mehr Speicher im Netzwerk benötigt wird und einfach weitere Datenträger an den Server angeschlossen werden, so steigt nur

die Speicherkapazität, nicht aber die Performance des Servers. Wenn nun beispielsweise mehr Zugriffe auf die Daten geschehen, so müssen die Daten immernoch durch diesen Server mit derselben Rechenleistung und dem gleichen maximalen Durchsatz zur Verfügung gestellt werden.

Hier kommt nun Scale-Out-NAS ins Spiel. Beim Scale-Out-NAS wird nicht nur die Speicherkapazität erhöht, sondern auch Rechenleistung und Durchsatz. Dies wird durch ein Cluster von Knoten realisiert. Ein Knoten ist ein Server, der über das Backend-Netzwerk mit allen anderen Knoten verbunden ist und kommuniziert, und über das Frontend-Netzwerk den Speicher des Clusters zur Verfügung stellt.
Wenn nun mehr Speicherplatz hinzukommen soll, so müssen weitere Knoten in das Cluster eingefügt werden. Da diese jedoch eine eigene CPU, Cache und Netzwerkschnittstellen mitbringen, erhöht sich automatisch auch die Performance des Clusters mit steigender Speicherkapazität. Dies ist der Grund dafür, dass sich Scale-Out-NAS-Systeme so gut skalieren lassen, ohne dass viel am System geändert werden muss. So müssen Administratoren beispielsweise nicht mehr bereits beim Einführen eines Speichersystems schon den Speicherbedarf der nächsten Jahre abschätzen, sondern können mit so viel, wie aktuell benötigt wird, anfangen und mit steigendem Bedarf weitere Knoten hinzufügen.

1.2 EMC Isilon

Das Unternehmen Isilon Systems wurde 2001 mit dem Ziel gegründet, möglichst effiziente, leicht zu verwaltende und einfach skalierbare Cluster-Speicherlösungen auf den Markt zu bringen. Der Administrationsaufwand bleibt auch bei steigender Anzahl von Knoten im Cluster auf einem ähnlichen, niedrigen Level, da das Cluster selbstständig neue Knoten einbinden und die Speicherplatzbelegung unter den Knoten ausgleichen kann. Diese Cluster sollen Thema dieser Arbeit sein.

2 Hardware

An dieser Stelle möchte ich nun die Isilon-Hardware vorstellen. Wie bereits beschrieben, setzt sich ein Cluster aus mehreren Knoten zusammen. Ein Isilon-Cluster muss immer aus mindestens drei Knoten bestehen. Diese Knoten müssen über ein Back-End-Netzwerk verbunden werden. Hierfür wird ein InfiniBand-Netzwerk verwendet, da InfiniBand relativ latenzarm ist. Dies ist wichtig, da die Knoten permanent über das Back-End-Netzwerk kommunizieren und Daten austauschen. Diese Kommunikation innerhalb des Clusters geschieht über das Internet Protocol (IP). Als Knoten im Cluster bietet EMC verschiedene Modelle an. Insgesamt gibt es vier verschiedene Serien, die für verschiedene Anforderungen optimiert sind. Diese Serien werden in den folgeneden Unterkapiteln im Einzelnen behandelt.

2.1 X-Serie

Die X-Serie ist die flexibelste und vielseitigste Serie. Sie wurde vor allem für Anwendungen mit hohem Datendurchsatz und einer großen Zahl gleichzeitiger Anfragen entwickelt. Sie stellt einen Mittelweg zwischen der leistungsorientierten S-Serie und der NL-Serie, die für die Langzeitspeicherung möglichst großer Datenmengen ausgelegt ist, dar. Die X-Serie bietet sich beispielsweise für die Verwendung in den Bereichen Biowissenschaften, Digitale Medien, Web 2.0 und Entwurfsautomatisierung elektronischer Systeme an.
Modelle der X-Serie gibt es in zwei verschiedenen Größenkategorien. X200-Modelle nehmen 2 Höheneinheiten (HE) in einem Rack ein. Die größeren X400-Modelle sind doppelt so hoch und nehmen somit 4 HE ein. Ein X200-Modell kann je nach gewählter Festplatten-Konfiguration bis zu 36 TB Speicherkapazität zum Cluster beitragen. Ein X400-Modell kann jedoch bis zu 144 TB Speicherkapazität besitzen. Folglich lässt sich so mit einem X400-Modell gegenüber einem X200-Modell die vierfache Speicherkapazität in der doppelten Menge an Platzbedarf unterbringen. Das entspricht einer Verdopplung der Speicherkapazität pro HE im Rack. Diese Skalierung gilt jedoch nicht für die Performance. Wenn mehr Wert auf eine Steigerung der Performance gelegt wird, sollten lieber zwei X200-Modelle statt einem X400-

Modell gewählt werden. Diese Entscheidungsmöglichkeiten, sowie die Kombinierbarkeit der verschiedenen Modelle tragen dazu bei, dass die X-Serie die vielseitigste und am flexibelsten einsetzbare der Serien ist.

2.2 S-Serie

Die S-Serie ist die leistungsorientierteste der drei Serien. Sie bietet sich vor allem für Anwendungen an, bei denen es auf einen hohen IOPS-Wert ankommt. Dafür ist jedoch die maximale Speicherkapazität pro Knoten geringer. Im Gegensatz zu einem X200-Modell, das bis zu 36 TB Speicherkapazität besitzen kann, kann ein Knoten der S-Serie maximal 21,6 TB Speicherkapazität besitzen. Dies hängt natürlich wieder von der gewählten Festplatten-Konfiguration ab, so gibt es beispielsweise auch Knoten der S-Serie mit nur 6,6 TB Speicherkapazität.
Ein Cluster mit Knoten der S-Serie bietet sich beispielsweise in den Bereichen Design und Simulation, Digitale Medien und Web 2.0 an. Eine mögliche Anwendung wäre z.B. Streaming in Echtzeit.

2.3 NL-Serie

Im Gegensatz zu der X- und der S-Serie ist ein Cluster der NL-Serie nicht als Primärspeicher konzipiert worden. Ein Cluster der NL-Serie bietet Nearline- bzw. Sekundärspeicher und macht somit am meisten Sinn als Ergänzung zu einem Primärspeicher-Cluster. So können Daten, die zwar noch gebraucht werden, aber in einem bestimmten Zeitraum kaum angefordert oder verändert werden müssen, von dem Primärspeicher-Cluster in das Nearline-Cluster übertragen werden. Der Vorteil dabei ist, dass Nearlinespeicher billiger ist und eine höhere Speicherdichte erreicht werden kann. Dafür ist jedoch die Zugriffszeit auf dort gelagerte Daten länger, als bei Daten, die sich im Primärspeicher-Cluster befinden.

2.4 Performance- und Back-up-Modelle

Neben den Modellen der X-, S- und NL-Serie gibt es weitere Modelle, die jeweils zu einem bestimmten Zweck in ein bereits bestehendes Cluster eingefügt werden können. Zum Beispiel kann es vorkommen, dass in einem Cluster noch mehr als genug Speicherkapazität vorhanden ist, es jedoch an Performance magelt.
Wenn beispielsweise die Übertragungszeit von Daten vom Cluster oder ins Cluster zu lang ist, dann können zusätzlich zu den eigentlichen Knoten des Clusters, in denen auch die Daten gespeichert sind, spezielle Performance-Knoten zum Cluster hinzugefügt werden. Diese bringen keine weitere Speicherkapazität mit, dafür jedoch tragen sie unmittelbar zur Verbesserung der Bandbreite und der Steigerung der Rechenleistung des Clusters bei.
Wenn jedoch eigentlich genügend Bandbreite vorhanden ist, ein großer Teil dieser jedoch ständig durch Back-ups in Anspruch genommen wird, so bieten sich die speziellen Back-up-Modelle an. Diese dienen lediglich dem Erstellen von Back-ups und bringen demnach auch keine weitere Speicherkapazität in das Cluster mit. Solche Back-up-Modelle bieten sich vor allem in Clustern der NL-Serie an. So können, ohne den Onlinebetrieb zu beeinträchtigen, über die Back-up-Knoten wichtige Daten auf Offline-Speichermedien übertragen und somit gesichert werden.

3 OneFS

OneFS ist das System, das auf Isilon-Clustern läuft. Dabei ist OneFS Dateisystem und Betriebssytem in einem. OneFS läuft auf allen Knoten des Clusters gleichzeitig und sorgt so dafür, dass diese überhaupt erst zu einem Cluster werden und miteinander kommunizieren. OneFS verwaltet die gesamte Speicherkapazität und präsentiert diese den Nutzern als ein einziges, großes Volume. So befindet sich alles innerhalb des Clusters in demselben Namensraum.
Als Betriebssystem-Grundlage nutzt OneFS BSD. OneFS nutzt zsh als Shell, es gibt jedoch weitere, spezielle Befehle, die auf OneFS zugeschnitten sind

und entsprechend alle mit dem Schlüsselwort "isi" beginnen.
Unterstützte Protokolle sind NFS, SMB/CIFS, HTTP, FTP, HDFS, iSCSI und REST API.

3.1 Dateisystem

Das Dateisystem basiert auf dem Unix File System (UFS) und agiert vollständig symmetrisch und verteilt. So ist beispielsweise kein dedizierter Metadaten-Server nötig, da die Metadaten von allen Nodes gemeinsam verwaltet werden. Gespeichert werden die Metadaten des Dateisystems in Inodes.
Die Platzierung der Daten wird von OneFS bis in die Sektoren der Datenträger aller Nodes kontrolliert, wodurch OneFS sehr effizient und flexibel mit dem Speicher arbeiten kann. So wird es auch möglich, dass Quota-Management nicht auf der Volumeebene, sondern auf Verzeichnisebene stattfindet.
Das Dateisytem ist weiterhin ein Journaling-Dateisystem, denn alle Änderungen werden vorerst in einem Journal abgelegt, bevor der eigentliche Schreibvorgang stattfindet. Dieses Journal wird im NVRAM der Nodes gespeichert, sodass auch nach einem Stromausfall stets alle Änderungen rekonstruiert werden können und sich alle Daten zu jedem Zeitpunkt in einem konsistenten Zustand befinden.

Zum Schutz der Daten beim Ausfall von Datenträgern oder kompletten Nodes kann zwischen zwei Methoden gewählt werden. Die eine Möglichkeit ist die Spiegelung der Daten. Dies wird beispielsweise für Metadaten verwendet. Die andere Möglichkeit basiert auf einem Fehler-korrigierenden Code und bietet einem N+M-Schutz. N ist die Anzahl der Nodes, die zur Speicherung der Daten an sich verwendet werden. M ist die Anzahl der Nodes, die zur Speicherung der Fehler-korrigierenden Codes verwendet werden. Eine Datei, die mit dieser Methode geschützt wird, ist immernoch sicher, wenn bis zu M Nodes gleichzeitig ausfallen, unabhängig davon, um welche der Nodes es sich handelt.

3.2 Administration

Zur Administration eines Isilon-Clusters gibt es verschiedene Schnittstellen. Zum einen gibt es ein Web-Interface, das einen Überblick über den aktuellen Zustand des Clusters gibt, sowie auch die Verwaltung des Clusters ermöglicht. Dies ist jedoch auch über die Kommandozeile mittels der seriellen RS232-Schnitstelle oder SSH im Netzwerk möglich. Des Weiteren wird auch eine Programmierschnittstelle angeboten, die RESTful Platform API. Letztendlich besitzen die Nodes selbst auch ein LCD Panel, dieses bietet jedoch nur eine einfache Funktion, um die Node in ein Cluster einzubinden oder wieder zu trennen. Darüber hinaus wird zur Vereinfachung der Administration eine ganze Palette an Verwaltungssoftware zur Verfügung gestellt.

3.3 Beispiel: Das Schreiben einer Datei

Abschließend möchte ich nun beispielhaft den Ablauf eines Schreibvorgangs darstellen. Angenommen ein Client ist über das Frontend-Netzwerk mit Node A und über Node A mit dem gesamten Cluster verbunden. Wenn nun der Client etwas schreiben möchte, so wird dieser Schreibvorgang von Node A geleitet. Zuerst wird die Datei in mehrere, gleich große Datenblöcke aufgeteilt. Nun werden entsprechend der gesetzten Regeln Paritäten oder Spiegelungen gebildet. Dann verteilt Node A die einzelnen Blöcke über das Backend-Netzwerk auf das Cluster. Jede Node, die so einen Block oder mehrere Blöcke erhält, schreibt diesen oder diese auf ihre Datenträger. Auch Node A selbst schreibt mindestens einen der Blöcke in ihren Speicher. Wie genau diese Aufteilung geschieht und mit welcher Methode die Datei vor Verlust geschützt wird, wird je nach eingestellten Regeln umgesetzt. Darüber hinaus kann der Administrator auch das voraussichtliche Zugriffsmuster auf eine Datei oder ein Verzeichnis angeben. Wenn beispielsweise für eine Datei als Zugriffsmuster "zufällig" ausgewählt ist, so wird sie auch nach mehrmaligem Zugriff nicht im Cache behalten.

4 Fazit

Die vorgestellten Nodes von Isilon bieten eine effiziente Möglichkeit, ein einfach zu verwaltendes und leicht skalierbares Speicher-Cluster aufzubauen. Es hat sich gezeigt, dass das Betriebs- und Dateisystem OneFS das ist, was das Cluster verbindet und die einfache Verwaltung ermöglicht. Für Unternehmen, die ein Speichersystem mit niedrigem Administrationsaufwand und einfacher Skalierbarkeit suchen, könnte ein solches Isilon-Cluster gut geeignet sein.
Wer dieses System nutzen möchte, ist jedoch vollständig an EMC gebunden, da alle Komponenten nur exklusiv untereinander einsetzbar sind. Wer beispielsweise mit der, in der Nodes verbauten, Hardware nicht zufrieden ist, und lieber andere verwenden möchte, der kann dieses System nicht nutzen, da es nur so voll funktioniert, wie es von EMC angeboten wird. Somit sollte sich jedes Unternehmen, das ein Speicher-Cluster aufbauen möchte, im Vorfeld überlegen, ob dieses System, so wie es angeboten wird, seinen Anforderungen genügt.

5 Quellen

Alle zuletzt abgerufen am 16.10.2013

- http://www.mitwa.org/node/8321
- http://arstechnica.com/business/2011/05/isilon-overview/
- http://www.enterprisestorageforum.com/storage-hardware/emc-isilon-buyers-guide.html
- http://www.hamburgnet.de/products/isilon.html?gclid=CPTLo7_Ew7cCFajKtAodyxQAhQ
- http://www.scribd.com/doc/131726467/Onefs-Command-Ref-6-5
- http://www.ndm.net/emcstore/isilon/emc-isilon-s-series-platform-node
- http://www.ndm.net/emcstore/isilon/emc-isilon-x-series-platform-node
- http://www.ndm.net/emcstore/isilon/emc-isilon-nl-series-platform-node
- http://www.storagenewsletter.com/news/business/isilon-iq-lightstorm-entertainment
- http://www.emc.com/collateral/hardware/white-papers/h10719-isilon-onefs-technical-overview-wp.pdf
- http://www.emc.com/collateral/hardware/specification-sheet/h10690-ss-isilon-s-series.pdf
- http://www.emc.com/collateral/software/specification-sheet/h10639-isilon-x-series-ss.pdf
- http://www.emc.com/collateral/software/specification-sheet/h10640-isilon-nl-series-ss.pdf
- http://www.emc.com/collateral/data-sheet/h11231-ss-isilon-perf-acc.pdf
- http://www.emc.com/collateral/hardware/specification-sheet/h10791-isilon-backup-accelerator-ss.pdf
- http://www.emc.com/collateral/software/technical-documentation/h10689-cm-isd-3rdparty-sw.pdf