Vorhersagbare Performance für Enterprise IT - At-Scale

Intelligente Infrastrukturen sind dort erfolgreich, wo herkömmliche und HCI-Architekturen scheitern

Hyperkonvergente Infrastrukturen (HCI) vereinen Rechen- und Speicherfunktionen in einem einzigen Gehäuse. Die Idee hinter HCI-Lösungen von Anbietern wie Nutanix und Simplivity
besteht darin, die IT-Abteilung von der Integration von Einzellösungen zu befreien und skalierbare, garantierte Leistung bei geringerem Risiko zu bieten.

Die Vorstellung ist, dass die Kunden aufgrund der integrierten Technologieebenen und Funktionen der konvergenten Plattform in der Lage sind, Ressourcen schnell bereitzustellen, einfach zu skalieren und die Kosten erheblich zu senken. Die Realität ist komplizierter - es ist schwierig
Es ist schwierig für HCI, Skalierbarkeit, Einfachheit und Kostenvorteile ohne Leistungseinbußen zu bieten. Und während HCI einfacher erscheinen mag, macht es die eng gekoppelte Architektur auch sehr schwierig, Leistungsprobleme zu beheben, da alles auf jedem Knoten überlagert ist.

Obwohl HCI in der Fachpresse viel Aufmerksamkeit erhalten hat, bleiben HCI-Architekturen mit konventionellem Speicher in einer Reihe wichtiger Leistungskennzahlen hinter den besten externen Speichersystemen - sowohl Hybrid-Flash- als auch All-Flash-Angeboten - zurück:

- **Latenzen**
- **IOPS (im besonderen im Vergleich zu All-Flash Systemen)**
- **Vorhersagbarkeit**

Infolgedessen haben sich die Rechenzentren von Unternehmen schnell für All-Flash-Storage entschieden, um die IO-Leistung bereitzustellen, die für Anwendungen aller Art erforderlich ist, insbesondere für Analysen und neue mobile und kundenorientierte Anwendungen.

Die Hauptgründe für die Entscheidung für All-Flash-Speicher sind:

- **Signifikante Reduzierung der Latenzen für sämtliche IO Operation**
- **Massive Erhöhung der Gesamt-IOPS**
- **Mehr vorhersagbare Performance für alle IO's**

Herausforderungen

- **Die meisten Infrastrukturanbieter können keine vorhersagbare Leistung für die reibungslose Skalierung von Unternehmens-Workloads liefern**
- **IOPS Potenzial Flash SSDs Können vergeudet werden**
- **Latenzen sind meist unvorhersehbar unter Last**
- **Neue Projekte können die Leistung noch weiter beeintrachtigen**

FEDERAL TECH BRIEF

Tintri VMstore liefert mühelos alle drei dieser Ergebnisse. Die intelligente Infrastrukturarchitektur von Tintri geht sogar noch weiter: Sie weist jeder virtuellen Maschine, Datenbank und jedem Container automatisch eine eigene IO-Lane zu, um Ressourcenkonflikte zu vermeiden.
Das macht es einfach, minimale und maximale Quality of Service (QoS)-Stufen für einzelne virtuelle Maschinen festzulegen und die Anwendungsperformance für jede einzelne Maschine autonom zu garantieren. Wenn die Anwendungs- oder Datenbankleistung eine wichtige Rolle spielt, sollten Sie alle Optionen sorgfältig prüfen und sich die Entscheidung für eine HCI-Lösung gut überlegen.

Vorhersagbarkeit

Für viele Anwendungen ist eine vorhersehbare Leistung genauso wichtig wie niedrige Latenzzeiten oder reine IOPS, aber sie ist oft viel schwieriger zu erreichen. Vorhersagbare Leistung ist noch schwieriger mit HCI-Lösungen zu erreichen, die Speicher und Rechenleistung kombinieren und beide Aktivitäten auf denselben Knoten ausführen. Und warum? Betrachten wir virtualisierte Umgebungen, in denen die Speichersoftware in der Regel innerhalb einer VM als virtuelle Appliance auf der Gast-Ebene läuft. In diesen Fällen durchläuft jeder IO-Vorgang den CPU-Scheduler des Hypervisors viermal: zweimal für den IO, zweimal für die IO-Bestätigung.

Bei geringer Systemauslastung mag dies kein Problem darstellen, aber wenn CPU-Ressourcen gemeinsam genutzt werden - wie bei HCI - führt dies bei mittlerer oder hoher Auslastung häufig zu einem erheblichen Engpass. Eine Lösung ist die Anwendung von CPU-Reservierungen, aber Reservierungen pro VM führen zu einer Reihe neuer
Die Reservierungen pro VM führen jedoch zu einer Reihe neuer Herausforderungen, die sich bis zu dem Punkt auswirken, an dem clusterweite HA-Failover-Richtlinien betroffen sein können. Außerdem garantieren CPU-Reservierungen nicht, dass die virtuelle Appliance sofortigen Zugriff auf die CPU hat. Wenn eine andere vCPU geplant ist und läuft, muss ihr erlaubt werden, ihren Betrieb zu beenden, was zu IO-Verzögerungen innerhalb der virtuellen Appliance führen kann.

Ob mit oder ohne CPU-Reservierung, das Ergebnis ist eine weniger vorhersehbare Latenz mit unerwarteten Spitzen, wenn ein Knoten oder Cluster ausgelastet ist. Bei Anwendungen können die Latenzzeiten von einem IO zum nächsten stark schwanken, was für besonders latenzempfindliche Anwendungen katastrophal sein kann. Dieses Problem wird in einer Enterprise-Cloud-Umgebung, in der ein Unternehmen möglicherweise Tausende von virtuellen Maschinen und/oder Containern verwaltet, noch verschärft. Trotz der Behauptung, „Web-Scale"-Kapazität zu liefern, sind HCI-Lösungen nur selten in der Lage, die Leistungserwartungen in großem Umfang zu erfüllen - sei es vor Ort oder in der Cloud.

Latenzen

Die Latenz von IO-Operationen auf herkömmlichen HCI-Speicherimplementierungen leidet im Vergleich zu externen Speichersystemen. Wenn beispielsweise Daten gespiegelt oder zur Datensicherung auf andere Knoten kopiert werden, werden mehrere Kopien jedes Datenblocks über das Netzwerk gespeichert, was sich direkt auf die Latenz auswirkt. Einige HCI-Anbieter unterstützen Erasure Coding als Alternative zur Spiegelung. Diese Technologie bietet zwar mehr „Neunen" an Verfügbarkeit, ist aber auch mit einen hohen Nachteil bei Leistung und Latenz. Es gibt Umgehungslösungen, die Post-Process Erasure Coding unterstützen, allerdings nur für kalte Daten, was für dynamische Unternehmens-Workloads nicht sinnvoll ist.

Sowohl Mirroring als auch Erasure Coding wirken sich auf die Schreiblatenz aus und können auch die Leselatenz beeinflussen. Die Enterprise Strategy Group (ESG) verglich kürzlich die Leistung verschiedener Speicherplattformen unter verschiedenen Bedingungen verglichen. Die beste Latenz, die von einer HCI-Lösung erreicht wurde, lag bei etwa 5 ms, was deutlich langsamer ist als bei den besten All-Flash-Systemen.

Abgesehen von Mirroring und Erasure Coding können Aktivitäten wie VMware vMotion, HA-Ereignisse, Wartungsarbeiten an Knoten und Knotenausfälle zu einer erhöhten Latenz für Workloads führen, da aufgrund der verrauschten, ressourcenintensiven Natur von vMotion und HA-Ereignissen und der Verringerung der insgesamt verfügbaren Ressourcen.

IOPS Performance

Die IOPS-Leistung, die ein Speicher liefern kann, insbesondere All-Flash-Speicher, korreliert direkt mit der verfügbaren CPU-Leistung. Die meisten Standalone-All-Flash-Arrays verwenden 28-40 Kerne pro Controller für 13-24 SSDs. (Obwohl einige Arrays noch höher skalieren, kann sich diese Progression negativ auf die IO-Dichte von All-Flash-Medien und damit auf die Leistungsvorhersagbarkeit auswirken).

Bei einigen HCI-Anbietern ist die Menge der für die Speicherung verfügbaren CPU begrenzt. Bis zu 8 vCPUs oder 20 % der verfügbaren CPU sind die typischen Grenzen. Das ist nicht genug Leistung, um die volle Leistung der Flash-Laufwerke auf jedem Knoten (6-24) zu erreichen, was zu einer Menge ungenutzter, verschwendeter Flash-IOPS führt.

Bei anderen HCI-Anbietern können Sie die Anzahl der für die Speicherung vorgesehenen CPUs erhöhen, was sich jedoch stark auf die Lizenzkosten auswirken kann. Sie möchten nicht auf teuren Hypervisor-, SQL Server- und/oder Oracle-Lizenzen für CPUs sitzen bleiben, auf denen letztendlich Speicherfunktionen ausgeführt werden.

Vorsicht ist geboten

Viele Anbieter von Speicherlösungen versprechen eine Reihe fortschrittlicher Funktionen und damit verbundener Vorteile, aber wie bei jeder größeren Infrastrukturentscheidung sollte der Käufer aufpassen. IT-Teams in Unternehmen wollen – und müssen in vielen Fällen - alle neuesten Funktionen nutzen, die die Infrastruktur bieten kann.

Die Realität bei vielen Lösungen ist, dass die Aktivierung neuer Speicherfunktionen die Ressourcenauslastung über ein akzeptables Maß hinaus erhöhen kann. Wenn Funktionen wie Snapshots, Replikation, Deduplizierung, Komprimierung usw. aktiviert werden, stehen Sie vor der Wahl: mehr Hardware hinzufügen oder die Vorhersagbarkeit und Leistung Ihrer Infrastruktur opfern. Diese Abwägungen sind für viele Administratoren inzwischen fast alltäglich.

Die Aktivierung von Datenreduzierungsfunktionen beispielsweise führt dazu, dass HCI-Plattformen noch mehr CPU-Ressourcen verbrauchen. Aus diesem Grund ist die Datenreduzierung bei vielen HCI-Implementierungen optional.

Tintri Takeaway

Intelligente Workload-Mobilität. Bei der Bereitstellung und Verwaltung der Leistung stoßen HCI-Lösungen an eine Reihe von Grenzen. Die zusätzliche Konvergenz und gemeinsame Nutzung von Speicher- und Rechenressourcen führt zu einzigartigen Engpässen, die sich auf Latenz, IOPS und Leistungsvorhersage auswirken. HCI kann im Vergleich zu konventionellen Einzelspeichersystemen eine bessere Workload-Mobilität bieten. Ja, Sie können Dinge auf verschiedene Knoten in Ihrem Cluster verschieben, aber Sie haben nicht die Intelligenz, Dinge zu verschieben, ohne dass dies Auswirkungen auf die Leistung anderer Workloads hat. Das ist etwas, das Tintri VMstore in einzigartiger Weise bietet.

Das Beste aus beiden Welten. Tintri Intelligent Infrastructure hebt Flash-basierten Speicher in eine andere Dimension, mit unvergleichlichen Leistungsvorteilen gegenüber HCI in Bezug auf Transparenz, Analyse und Servicequalität. Tintri verfolgt einen anderen Ansatz als HCI und andere Array-Anbieter, indem es das Beste aus beiden Designs bietet: die Einfachheit von HCI und die Skalierbarkeit und Leistung von All-Flash. Im Gegensatz zu HCI konzentrieren wir uns auf die Optimierung von Storage und unterstützen gleichzeitig eine enge, einfache Integration über mehrere erstklassige Rechen- und Netzwerkplattformen hinweg. Unsere VMstore-Architektur besteht aus Modulen, die eine nahtlose, skalierbare Leistung mit einer erstklassigen externen Infrastruktur ermöglichen.

Auto-QoS für vorhersagbare Leistung. Unsere autonomen QoS-Funktionen stellen sicher, dass jeder Workload - sei es eine Anwendung, eine Datenbank oder sogar ein einzelner virtueller Desktop - immer die benötigten Ressourcen zum richtigen Zeitpunkt erhält, sodass Leistung und Latenz nie ein Problem darstellen. Dies gilt unabhängig davon, ob Sie zehn oder zehntausende von VMs haben, ob Sie erweiterte Funktionen nutzen oder nicht. Sie erhalten immer eine vorhersehbare Leistung mit einer Latenz, die **bei jeder Arbeitslast konstant unter 1 ms** liegt.

Unterbrechungsfreies, verbessertes Benutzererlebnis. Und da wir sowohl umfassende als auch granulare Sichtbarkeit und Kontrolle von einem einzigen Bildschirm aus bieten, erhalten Sie Informationen, die mit HCI - oder jeder anderen Lösung - einfach nicht verfügbar sind. Sie können sofort herausfinden, wo genau die Probleme für jede VM oder Anwendung in den Rechen-, Netzwerk- und Speicherebenen liegen, um die geringste Unterbrechung zu vermeiden und ein hervorragendes Benutzererlebnis zu gewährleisten.

Erleben Sie den Unterschied. Erleben Sie Tintri Intelligent Infrastructure.

© 2024 Tintri by DDN, Inc., All Rights Reserved