

CLUSTER-ADMINISTRATION MIT CLUSTWARE

Kern-Kompetenz

ROBERT HOMMEL

Energiesparen ist auch bei den großen Cluster-Architekturen angesagt. Die effiziente Nutzung der Ressourcen ist einer der wichtigsten Ansatzpunkte zum Senken des Stromverbrauchs. Dazu sind geeignete Betriebssysteme und die richtigen Administrationswerkzeuge für die HPC-Cluster erforderlich. Auf der ISC9 in Hamburg stellt das Chemnitzer Unternehmen Megware unter anderem die dafür ausgelegte Lösung Clustware vor.

Die Administration von HPC- (High-Performance-Computing-) Clustern erfordert mehr als nur die Überwachung von laufenden Jobs. Gerade durch die aufgekommenen Debatten über grüne Technologien spielen effizientes Energiemanagement und der gezielte Einsatz von Ressourcen eine nicht mehr wegzudenkende Rolle. Die effiziente Nutzung der Ressourcen spielt auch im Engineering-Umfeld eine Rolle, da dort HPC-Systeme immer häufiger von verschiedenen Abteilungen mit ihren jeweiligen Anwendungsprogrammen genutzt werden.

Cluster, das heißt Hochleistungscomputer, die Anordnungen einer Vielzahl herkömmlicher Rechner sind, haben in den letzten Jahren an Bedeutung gewonnen und die traditionellen Vektorrechner aus der Top-500-Liste fast schon verdrängt. In den Clustern arbeiten in der Regel mehrere CPUs der neuesten Generationen mit derzeit bis zu 128 GByte Arbeitsspeicher. Schon bei der Auswahl der Komponenten entscheiden sich die Folgekosten, die durch den Betrieb und die Wartung des Systems entstehen werden.

Mit der drohenden Klimakatastrophe und dem Trend zu „grünen“ Technologien ist der Energieverbrauch auch bei HPC-Systemen zu einem wichtigen Thema geworden, und mit den richtigen Komponenten lässt sich bereits viel Energie sparen. Viel wichtiger ist aber, die Gegebenheiten im laufenden Betrieb zu überwachen. Dazu zählt neben der Überwachung von Temperaturen in den Rechenknoten, Temperaturen der CPUs, des Mainboards und der Gehäu-

se auch die Temperatur der Kaltluftzufuhr im Schrank und der Luftfeuchtigkeit. Das gibt einerseits eine Möglichkeit, die Raumklimatisierung optimal einzustellen, andererseits schützt es die teure Rechentechnik vor Beschädigungen bei Ausfällen der Klimatisierung. Dazu werden vielfach PDUs (Power Distribution Units), das sind schaltbare Steckdosen,



Intelligenter Stromschalter ClustSafe.

verwendet, die in der Regel per Ethernet fernsteuerbar sind. Ein solches Gerät ist Clustsafe von Megware.

Neben der autonomen Überwachung, die auch dann noch funktioniert, wenn alle anderen Systeme versagen und auch dann noch die Rechenknoten schützen kann, ist die Berechnung der Leistungsaufnahme aller angeschlossenen Geräte unverzichtbares Hilfsmittel, wenn man sich im Rechenzentrum über den Energieverbrauch vergewissern möchte, denn einsparen kann man nur mit dem Wissen, wo zu viel Energie verbraucht wird.

Steckkarten für die Knoten des Clusters erweitern die Überwachung. Sie bieten die Möglichkeit neben dem Auslesen von Sensordaten wie Temperaturen, CPU-Spannungen und Lüfterdrehzahlen auch

den Rechner aus der Ferne ein- und auszuschalten, einen Reset auszulösen und sich per Serial-Over-LAN oder Video-Over-LAN direkt den Bildschirm ins Büro zu holen und von dort Einfluss zu nehmen.

Die Erweiterungskarten sind softwareunabhängig und nicht auf ein funktionierendes Betriebssystem angewiesen. Idealerweise stellen sie auch bei ausgeschaltetem Knoten ihre Dienste zur Verfügung. Für alle anderen zu sammelnden Daten, die das Betriebssystem betreffen, kommt ein Software-Dämon zum Einsatz. Er ermittelt CPU-Auslastungen, Speicherbelegung und überwacht Prozesse und deren Benutzer. Die Möglichkeit für das Ausführen einfacher Befehle erleichtert

viele Aufgaben. Wichtig ist ein ressourceneffizientes Konzept, denn der Dämon muss auch unter hoher Belastung des Rechenknotens einwandfrei funktionieren, um eine lückenlose und sichere Überwachung zu gewährleisten.

Archivieren

Nachvollziehbarkeit ist ein wichtiger Aspekt der Messung von Maschineninformationen. Dazu ist es notwendig, alle gesammelten Sensordaten von allen Rechenknoten zentral aufzuzeichnen und zu archivieren. Meist wird ein dedizierter Head-Node oder eine Appliance eingesetzt. Eine spezielle Datenbank, für Sensordaten konzipiert, ist dafür das ideale Werkzeug. Das sind meist so genannte Round-Robin-Datenbanken, die

eine festgelegte Anzahl von Datensätzen speichern können und automatisch ältere Daten akkumulieren. Zur Darstellung der gespeicherten Werte bietet sich einerseits eine eigenständige Software an, andererseits bieten heute systemunabhängige Weboberflächen fast unbegrenzte Möglichkeiten.

Durch grafische Darstellungen lassen sich schnell Fehler und Ursachen finden und Ereignisse aus den Sensoraufzeichnungen zusammenführen. Erhöht sich etwa die Temperatur in einem Rechenknoten durch einen ausgefallenen Lüfter und treten dadurch Probleme mit dem Hauptspeicher oder der CPU auf, so lässt sich das leicht nachvollziehen und der Zeitpunkt des Auftretens genau erkennen.

Die Systeme sind zudem in der Lage, Grenzwertüberschreitungen selbst zu erkennen, dem Administrator Warnmeldungen per E-Mail oder SMS zuzustellen oder in der Management-Software darzustellen und gegebenenfalls voreingestellte Gegenmaßnahmen zu ergreifen. Zu den gesammelten Daten des Betriebssystems und den Sensordaten kommen noch die Daten von anderen Geräten wie den PDUs dazu, zum Beispiel die Leistungsaufnahmen der Rechenknoten. Hier bietet sich die Möglichkeit, Vergleiche zwischen Energieverbrauch, Auslastung und Klimatisierungsstrategien zu ziehen.

Überwachung aller laufenden und anstehenden Jobs

Um die Administration von Clustern noch einfacher und übersichtlicher zu gestalten und mögliche Fehler zu vermeiden, ist es sinnvoll, das Queuing-System, das bei den meisten HPC-Clustern Standard ist, in die Management-Oberfläche einzubinden. Es

geht aber nicht um die Nutzung der vollen Funktionalität, sondern nur um die Überwachung aller laufenden und anstehenden Jobs und Prozesse. Will man Rechenknoten ausschalten, so sollten dort nicht gerade Jobs laufen, denn ein Abbruch des Jobs dürfte den Nutzer nicht gerade erfreuen. Werden aber die Jobs übersichtlich zusammen mit den wichtigsten Zustandsinformationen der Rechenknoten dargestellt, sind Fehleingriffe in den Betrieb des Clusters praktisch ausgeschlossen.

Die Anbindung bietet noch weitere Möglichkeiten. Zum einen kann anhand anstehender und laufender Jobs ermittelt werden, ob in nächster Zeit weniger Rechenknoten benötigt werden. Nach definierten Vorgaben wird dann eine entsprechende Anzahl von Rechenknoten automatisch ausgeschaltet. Besteht wieder Bedarf, werden Rechenknoten wieder zugeschaltet.

Da ein HPC-Cluster nur selten wirklich zu 100 Prozent ausgelastet ist, sind Energieeinsparungen in Größenordnungen möglich, die weder durch geschickte Klimatisierung noch durch Auswahl energieeffizienterer Komponenten erreicht werden können.

Administration von überall her

Neben der Bedienmöglichkeit vom Arbeitsplatz oder von jedem beliebigen Ort aus über die Weboberfläche ist auch die Bedienung am Cluster wichtig. Nicht immer steht ein Rechner bereit, der Zugriff auf die Managementoberfläche per Webbrowser bietet. Abhilfe können klei-



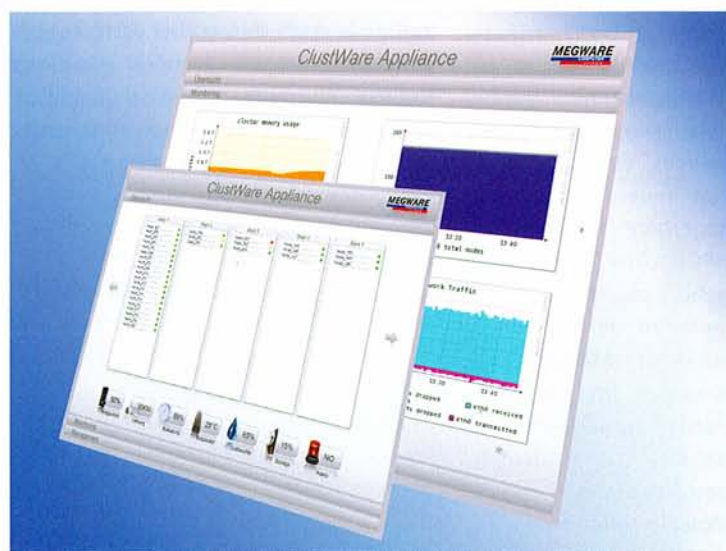
Cluster mit Touchdisplay.

ne Touchdisplays schaffen, die an den Clusterschränken angebracht sind. Hier hilft eine übersichtliche Darstellung mit zusammengefassten Informationen und Warnhinweisen mehr, statt Informationen bis ins kleinste Detail anzubieten. Intuitive Bedienoberflächen und Menüführungen sind notwendig, um die Benutzung zu erleichtern und zu verhindern, dass eine Scheu vor etwaigen Fehlbedienungen gänzlich von der Benutzung abhalten. Zusätzlich können administrative Eingriffe bequemer erfolgen, indem unter anderem die PDUs über die Touchdisplays mit gesteuert werden können. Das erleichtert die Arbeit, da auch nicht alle PDUs über eine Steuermöglichkeit direkt am Gerät verfügen, sondern meist nur per Ethernet Zugang bieten.

Installieren

Backup, Einspielen von Patches, Neuinstallationen und Konfigurationsverteilung sind weitere Aspekte, die ein Managementsystem beherrschen muss. All diese Dinge auf jedem Rechenknoten einzeln vorzunehmen ist schon bei kleinen Systemen keine Option. Schnittstellen zu vielen der marktüblichen Installationssysteme bieten dabei ein hohes Maß an Kompatibilität und können so den individuellen Ansprüchen jedes einzelnen Kunden gerecht werden.

Bedienfreundlichkeit und eine gute Übersicht, weniger die Masse an Detailinformationen oder ein unübersichtlicher Funktionsumfang sind letztlich entscheidende Kriterien. Das höchste Ziel für jeden Entwickler ist auch, dass die Benutzung dem Anwender Spaß macht. ■



Bedienoberfläche für das Clustermanagement. Bilder: Megaware

KENNZIFFER: DEM17532