Führende HPC-Anbieter setzen auf QLogics InfiniBand-Technologie

Roy Krischer

Am 16. November 2009 gab QLogic Inc. eine Reihe von Partnerschaften bezüglich ihrer InfiniBand-Produkte bekannt. HP wird QLogics gesamte QDR InfiniBand Produktpalette an Host Channel Adaptern (HCA), Directors und Switches im Rahmen des HP Unified Cluster Portfolio vertreiben. Auch SGI wählte QLogics Directors und Switches der 12000-Serie für seine CloudRack HPC (High Performance Computing) Angebote. Desweiteren wird Dell QLogics QDR-Switches und Directors anbieten, zusätzlich zu den QLogic HCAs, die Dell schon seit einiger Zeit offeriert. Schließlich wird IBM, nach QLogics QDR InfiniBand-Switches der 12000-Serie, im IBM System Cluster 1350 nun auch QLogics QDR InfiniBand CHAs der Serie 7300 in den System x Servern verwenden.

Diese Entwicklungen verstärken QLogics Einfluss im Geschäft mit QDR InfiniBand und den Zugang der Firma zum HPC-Markt. Zudem stellen sie eine ernsthafte Herausforderung für die InfiniBand-Anbieter Mellanox und Voltaire dar.

Die Technologie

High-Performance Computing

Unter High-Performance Computing (HPC) versteht man den Zweig der Informatik, der sich mit der Berechnung extremer numerischer Probleme beschäftigt. Viele wissenschaftliche, staatliche oder industrielle Forschungszweige haben es mit umfangreichen Berechnungen zu tun, wie sie z.B. bei Strömungssimulationen auftreten. Man findet sie in der Wettervorhersage genauso wie in Klimamodellen, der Simulation von nuklearen Detonationen, simulierten Crashtests im Automobilbereich, der pharmakologischen Forschung, Finanzmodellen oder der computerisierten Produktentwicklung. Die benötigte Rechenleistung übersteigt bei Weitem die CPU-Leistung und den Speicherdurchsatz einzelner Computer. Deshalb sind HPC-Systeme normalerweise als Cluster aufgebaut, in denen eine große Anzahl Computer (Knoten) miteinander durch ein Netzwerk (LAN) verbunden sind. Um in einem solchen System ein Problem effizient berechnen zu können, muss dieses in Teilprobleme unterteilt werden, die auf den jeweiligen Knoten verteilt unabhängig voneinander gelöst werden. Leider erlauben die meisten interessanten Probleme keine perfekte Unabhängigkeit der Knoten untereinander, so dass diese ständig miteinander kommunizieren, d.h. Daten austauschen, müssen. Da Netzwerkverkehr um Größenordnungen langsamer ist als CPUs Daten verarbeiten können, steht die mögliche Beschleunigung eines Berechnungsproblems durch Nutzen eines solchen Cluster im umgekehrten Verhältnis zur Kommunikationsmenge zwischen den einzelnen Knoten, die das Problem erfordert. Die Wahl der Netzwerktechnologie zur Anbindung der Knoten ist daher entscheidend, denn abhängig vom berechneten Problem können sich Durchsatz- und (oft wichtiger) Latenzbeschränkungen dramatisch auf die Berechnungszeit auswirken. Durch die schiere

Größe solcher Clusterinstallationen sind die Skalierbarkeit des Netzwerks und dessen Stromverbrauch ebenfalls wichtig.

Als Beispiel für solch ein HPC-System kann IBMs Roadrunner im amerikanischen Los Alamos herhalten. Dieses war lange Zeit der schnellste Cluster (inzwischen jedoch auf dem zweiten Platz) und besitzt 6480 Opteron-CPUs mit je zwei Kernen, 12960 PowerXCell 8i Prozessoren mit je neun Kernen und insgesamt 103,6 TB Arbeitsspeicher (nur zu Rechenzwecken eingesetzte Hardware wird hierbei gezählt). Diese sind auf 270 Racks zu je 12 TriBlade-Knoten verteilt, also insgesamt 3240 TriBlades. Um all diese Knoten zu vernetzen wird eine zweischichtige InfiniBand-Topologie eingesetzt.

InfiniBand

InfiniBand (IB) ist eine Netzwerktechnologie für höchste Leistungen in lokalen Netzwerken und basiert auf offenen Standards. Es verwendet eine „switched-fabric"-Topologie, in der Hosts (oder besser: ihre Host Channel Adapter, HCA) miteinander über Switches verbunden sind. Peripherie wie beispielsweise Massenspeicher kann ebenfalls über sog. Target Channel Adapter (TCA) angeschlossen werden.

Die Datenübertragung erfolgt über serielle bi-direktionale Verbindungen mit Durchsätzen von 2,5 Gb/s, 5 Gb/s (Double Data Rate, DDR) oder 10 Gb/s (Quad Data Rate, QDR). Diese Verbindungen kann man zu viert (sog. 4x.) bündeln, was mit QDR eine maximale Bruttodatenrate von 50 Gb/s ergibt; für extremen Durchsatzbedarf kann man sie zu zwölft (12x) zu einer Datenrate von 120 Gb/s bündeln. Datenpakete heißen Nachrichten (engl. *Messages*) und haben bis zu 4 kB an Länge. Die Latenz einer Anwendungsnachricht (z.B. für MPI) zwischen zwei Endpunkten liegt im Bereich von 1-2 µs.

InfiniBand bietet Quality-of-service, Ausfallsicherheit (engl. *Failover*) und andere spezielle Features wie z.B. Remote Direct Memory Access (RDMA), mit welchem ein Netzwerkknoten direkt auf den Speicher eines anderen Knotens zugreifen kann, ohne dass Betriebssystem oder CPU des Ziels involvieren zu müssen, was zusätzlichen Overhead vermeidet. Durch diese Eigenschaften und die hervorragende Leistungsfähigkeit bzgl. Durchsatz und Latenz ist InfiniBand derzeitig der *de-facto* Standard in Top-HPC.

Laut der TOP500[1]-Liste vom November 2009 wird InfiniBand von 36% (rauf von 30%) aller gelisteten Systeme als Netzwerktechnologie verwendet. Gigabit Ethernet hat zwar einen Anteil von 52% (runter von 56%), doch handelt es sich hierbei vornehmlich um kleinere Systeme. Bei Highend-Systemen führt InfiniBand unangefochten mit neun der zwanzig schnellsten Systeme (einschließlich des bereits erwähnten Roadrunner). Gigabit

[1] Die TOP500-Liste (http://www.top500.org) wird zweimal im Jahr erstellt und ordnet Computersysteme nach eingereichten Benchmarkwerten. Sie gilt gemeinhin als die weltweite Supercomputer-Rangliste. Nicht alle Organisationen reichen Resultate für ihre Systeme ein, d.h. die Liste ist möglicherweise unvollständig. Der verwendete Linpack-Benchmark löst eine große Menge linearer Gleichungen, was relativ gut skaliert. Ob dies einen guten Anhaltspunkt für die Leistungsfähigkeit eines eingereichten Systems darstellt, hängt davon ab, wofür dieses System eigentlich eingesetzt wird.

Ethernet taucht in den Top-20 nicht auf. Zusammen haben die 181 InfiniBand-Systeme 70% mehr Rechenleistung als die 259 zusammengenommenen Ethernet-Systeme. Noch vielsagender ist das Bild, wenn man die tatsächlich gemessenen mit den theoretischen aufgrund reiner CPU-Leistung erreichbaren Benchmarkwerte vergleicht. Dann sieht man nämlich, dass Ethernet-Systeme nur 50% ihrer theoretischen Leistung ausschöpfen können, wohingegen InfiniBand-Systeme 77% erreichen. Da Netzwerkleistung in großem Maße zu Diskrepanzen zwischen theoretischer und tatsächlicher Leistungsfähigkeit beiträgt, untermalen diese Werte InfiniBands Überlegenheit gegenüber Ethernet.

Produkte

QLogics InfiniBand-Produkte basieren auf der *TrueScale*-Architektur, die mittels ASIC-Technologie hochintegrierte, energiesparende InfiniBand-Lösungen mit Topleistung ermöglicht. QLogics Kompetenz in der ASIC-Entwicklung zeigt sich über die gesamte Produktpalette, wie z.B. auch in den in [1] beschriebenen konvergierten Netzwerkadapter für Fiber Channel over Ethernet (FCoE).

Die x8 PCI-Express 2.0 QDR-Adapter aus QLogics 7300-Serie bieten einen uni-direktionalen Durchsatz von 3400 MB/s und eine Nachrichtenrate von bis zu 30 Millionen pro Sekunde. Laut QLogic ist diese Nachrichtenrate fünfmal so schnell wie die ihrer Mitbewerber, was durch die bessere Anpassung an moderne Multicore-Architekturen erreicht wird. Die 7300-Serie bietet eine sehr schnelle Latenz von 1 µs für MPI-Nachrichten und behält diese durch ihren verbindungslosen Ansatz auch dann bei, wenn die Knotenanzahl steigt. Ihre elektrische Leistungsaufnahme von typischerweise rund 6W ist derzeit die sparsamste im aktuellen Markt.

QLogics 12000-Serie an QDR InfiniBand Director Switches sind in einer Vielzahl flexibler Konfigurationen erhältlich, von 18 bis zu 864 Ports. Dieses Maximum übersteigt das der Mitbewerber, die höchstens 648 anbieten und diese Zahl nur durch Verbinden kleinerer Switches erreichen können. Installation, Konfiguration und Überwachung des Netzwerks werden alle durch QLogics umfangreiche *InfiniBand Fabric Software Suite* erleichtert.

Die Akteure

QLogic

QLogic ist ein führender Anbieter von Hochleistungs-Netzwerklösungen im Bereich FCoE, Ethernet, FC, iSCSI und InfiniBand, mit einem Produktportfolio, dass sowohl Geräte wie Netzwerkadapter und Switches umfasst als auch Technologie, mit der verschiedene Tier-1 OEM-Anbieter beliefert werden. Nach Marktforschung der Dell'Oro Group ist QLogic mit einem Marktanteil von über 50 Prozent der führende Anbieter von FC-Technologie. Das Unternehmen besitzt umfangreiches Knowhow in ASIC-Technologie, um hochintegrierte Produkte herzustellen, die höchste Leistung bieten und dabei wenig Strom verbrauchen.

Die Partner

Laut der TOP500-Liste vom November 2009 wurden 42% aller gelisteten Systeme von HP geliefert, was die Spitzenposition hinsichtlich Anzahl der Systeme darstellt. Dem folgt IBM mit 37% der Systeme. Allerdings muss man beachten, dass IBMs Systeme zusammengenommen eine Leistung von 9,7 PFlops aufweisen, was HPs Resultat von

6,3 PFlops deutlich übersteigt. Es folgen SGI auf Platz drei mit 3.8% der gelisteten Cluster und Dell auf dem fünften Platz (allerdings kann man Dell noch weitere 0.8% zuschreiben, die in Kooperation mit anderen Anbietern oder von Tochterunternehmen geliefert wurden, was Dell auf den dritten Platz schieben würde). Insgesamt sind diese QLogic-Partner für überwältigende 86% aller TOP500-Systeme verantwortlich.

Fazit

Der Bereich HPC wächst ständig. Den traditionellen Kunden wie Regierungen und Forschungseinrichtungen gehen die komplexen Berechnungsprobleme nicht aus. Gleichzeitig erscheinen immer mehr kommerzielle Anwendungsfelder für HPC auf der Bühne. In wie zuletzt wirtschaftlich angespannten Zeiten stehen sich Unternehmen schwierigen strategischen Entscheidungen gegenüber. Kostensenkungen können kurzfristig helfen, doch langfristig ist Innovation der bessere Ansatz, um die wirtschaftliche Zukunft zu sichern. Ohne den Anreiz innovativer neuer Produkte könnten sich sonst nämlich Kunden, die ja schon selbst ans Sparen gewöhnt sind, mit dem, was sie schon haben, zufriedengeben. Für Innovation sind Investitionen in Forschung und Entwicklung nötig, welche heutzutage verstärkt extreme Berechnungsprobleme beinhalten. Ergo werden HPC-Installationen weiter zunehmen, und da InfiniBand *der* Netzwerkstandard für HPC ist, sieht sich InfiniBand einer rosigen Zukunft entgegen.

Es ist häufig nicht einfach, Leistungswerte verschiedener Anbieter zu vergleichen. Zu sehr hängen diese von der Netzwerktopologie, dem Benchmark mit Parametern und einer Unzahl anderer Details ab. Einige Zahlen lassen sich jedoch durchaus vergleichen: So bieten z.B. Qlogics Directors der 12000-Serie mehr Ports als jedes vergleichbare Produkt, was bei immer wachsenden Clustergrößen wichtig ist. Tests mit dem SPEC-MP2007 Standardbenchmark bescheinigen darüber hinaus, dass QLogics neue QDR-Adapter überlegen sind, wenn die Anzahl der Kerne im Cluster steigt; bei 256 Kernen sind die HCAs der 7300-Serie über 20% schneller als vergleichbare ConnectX-Adapter von Mellanox, und das bei geringem Stromverbrauch. Bei wachsenden Clustergrößen und Strompreisen, die eher steigen als fallen werden (und Umweltsorgen), sollte geringer Stromverbrauch bei gleichzeitiger Topleistung oberstes Ziel für HPC-Installation sein.

Letztendlich spricht die schiere Breite an Anbietern, die nun QLogics QDR Infiniband-Technologie vertrauen, fast schon für sich. Wenn vier der fünf wichtigsten HPC-Anbieter, die zusammen immerhin für 86% der TOP500-Systeme verantwortlich sind auf QLogics Technologie setzen, dann stellt dies eine klare Empfehlung von denjenigen dar, die sich in diesem Bereich am besten auskennen. Insbesondere IBM ist hier hervorzuheben, der nach Rechenleistung größte HPC-Anbieter. Nachdem IBM kürzlich erst beschloss, QLogics FCoE-Technologie in seine Produkte einzusetzen, kommt nun auch noch QLogics QDR InfiniBand-Technologie hinzu. QLogics Geschäft nimmt offensichtlich in allen Netzwerksparten, in denen es präsent ist, Fahrt auf.

Quellenangabe

1. **Krischer, Josh.** *QLogic mit frischer Innovation in Konnektivität und Netzwerkinfrastuktur.* 2009. http://joshkrischer.com/files/QLogic_CNA_de.pdf.

