DC Express: Latenzarmer NVMe-Nachfolger

Nicht klingeln

Susanne Nolte

Heutige High-End-Anwendungen verlangen von Speichermedien vor allem kurze Antwortzeiten. Doch reaktionsschnelle Medien helfen wenig, wenn I/O-Busse mit hohen Latenzen sie ausbremsen.

Immer mehr Anwendungen, darunter Hypervisor, Datenbanken und -analysen, benötigen schnelle Lesezugriffe mit geringen Verzögerungen. Lange Antwortzeiten lassen aber nicht nur die Anwendung warten, sondern reduzieren vor allem bei hoher I/O-Last den Durchsatz dramatisch. Solche Latenzen entstehen zum einen im Speichermedium selbst, da dessen Controller eine Weile braucht, bis er die ersten Daten liefern kann, zum anderen auf dem I/O-Bus – beim Initialisieren und Beenden des Datentransfers oder beim Warten auf einzelne Pakete.

Reaktionsschnelle Medien gibt es inzwischen einige, etwa NAND-Flash, FeRAM/FRAM (Ferroelectric Random Access Memory), MRAM und STT-MRAM (Spin-Transfer Torque Magnetoresistive RAM) sowie PCM (Phase Change Memory). Während Flash bei hohen I/O-Anforderungen aufgrund seiner Latenz von 25 bis 80 µs bereits an seine Grenzen stößt, finden reaktionsschnellere FeRAM und MRAM wegen des hohen Preises nur in speziellen Systemen als wenige MByte große Chips Verwendung. Hoffnungen auf Chips mit deutlich höherer Kapazität schürt momentan die Weiterentwicklung STT-MRAM.

Derweil findet PCM Verbreitung als Flash-Ersatz in Smartphones und Tablets. Für x86-Systeme hat HGST vor wenigen Wochen einen Prototyp mit PCIe-Schnittstelle gezeigt. Mit ersten PCM-Serien ist 2015 zu rechnen. Die einzelnen Chips fassen bereits 1 GBit, ihre Reaktionszeit liegt bei 110 ns bis zum ersten gelesenen Byte. Das nützt aber wenig, wenn der I/O-Bus eine Latenz von mehreren bis mehreren Dutzend µs besitzt.

Die schnellste Verbindung zum Massenspeicher in Servern liefert momentan PCIe. Dafür existiert mit NVMe (Non Volatile Memory Express) ein herstellerunabhängiges Speicheranbindungsprotokoll analog zu SAS oder SATA. Es zog erst 2013 in erste PCIe-SSDs ein, da die Hersteller lange auf eigene proprietäre Techniken setzten. PCM-SSDs würde es aber mit seinem Overhead massiv ausbremsen.

Ohne lange Vorrede

Deshalb ist bereits ein NVMe-konformer Nachfolger für künftige PCM-SSDs namens DC Express entstanden (siehe „Alle Links“). Mit ihm konnten die Entwickler die Gesamtlatenz beim Lesen auf gut 1 µs drücken, indem sie die Zahl der PCIe-Pakete, der sogenannten Transaction-Level Packets (TLPs), und der Kontext-Switches reduzierten.

Bei NVMe beginnt die Initialisierung damit, dass die Host-CPU ein Read-Kommando im RAM vorbereitet und ein Doorbell-Paket über PCIe an das Speichergerät sendet, um es darüber zu informieren, dass ein Paket in der Queue des Host-RAM liegt. Es sendet daraufhin einen DMA-Request (Direct Memory Access) an den Host, um das Read-Kommando aus dessen Queue abholen zu dürfen, was der Host seinerseits mit einem DMA-Response beantwortet. Jedes dieser Pakete benötigt auf der derzeit schnellsten Hardware über 0,6 µs – also ein Vorspiel von über 1 µs für jede einzelne Leseabfrage.

DC Express dagegen verlagert den Overhead in die Idle-Zeiten: Es verwendet an dieser Stelle ein kontinuierliches Polling, mit dem sich das Endgerät laufend nach neuen Anfragen in der Queue erkundigt. Auf diese DMA-Requests kann der Host bei Bedarf sofort mit einem positiven DMA-Response reagieren. In die Datenlieferung streut das Endgerät zudem neue DMA-Requests mit hoher Priorität ein und stapelt eintreffende Leseanfragen, um sie in dem Moment zu bedienen, in dem die vorherige Datenlieferung abgeschlossen ist.

Ein NVMe-Gerät schreibt am Ende des Lesevorgangs einen Eintrag in die Completion-Queue und sendet ein Interrupt-Signal, das den auf Daten wartenden CPU-Thread weckt. Da die Pakete aber über das mehrkanalige, serielle PCIe unsortiert ankommen können, muss das Endgerät über ein Bit im TLP-Header zu Beginn den Transport-Modus „Strict Packet Ordering“ verlangen. Dadurch müssen alle folgenden Pakete auf ein säumiges warten. Zudem verursacht die Kommunikation per Interrupts einen zusätzlichen Overhead durch Kontext- und Mode-Switching in der CPU, verbunden mit einer Wartezeit von mehreren µs.

DC Express schreibt stattdessen am Anfang – über den deutlich schnelleren CPU-RAM-Link – lauter Incomplete-Tags in den Empfangspuffer des CPU-Thread. Sind alle verschwunden, also mit Daten überschrieben, ist die Lieferung komplett, unabhängig davon, in welcher Reihenfolge die Pakete ankommen.

Den erhöhten Energiebedarf der Protokollneuerungen vor allem im Idle-Betrieb haben die Entwickler anhand von Testsystemen mit etwa 6 % berechnet. Da DC Express mit NVMe kompatibel ist, wäre ein Umschalten zwischen beiden Verfahren aber denkbar.

Da sich die Entwicklung von DC Express an den Anforderungen von PCM orientiert hat, liegt der Fokus auf dem Verringern der Lese-Latenz. Beim Schreiben liegt PCM – bedingt durch die Grenzen heutiger Lithografie – mit der fünfzigfachen Antwortzeit im Bereich von Flash. Aufgrund ausreichend großer Schreib-Buffer in x86-Systemen sind Flash und PCM in der Regel schnell genug für heutige Anwendungen. Bis sich hier etwa durch neue Herstellungstechniken oder STT-MRAM wieder etwas verändert, wird noch einige Zeit vergehen. (sun)

Alle Links: www.ix.de/ix1410122