100-Gbit/s-Ethernet-Verarbeitung: Neue Herausforderungen und ihre Lösungen

Ralph Schlenk

Der weltweite Netzverkehr explodiert. Um der Datenflut Herr zu werden, muss auch im Kernnetz massiv investiert werden: Die 10-Gbit/s-Ethernet-Technik, die in Transportnetzen weit verbreitet ist, reicht auf manchen Strecken schon heute nicht mehr für den exponentiell ansteigenden Datenverkehr aus. Der neue 100-Gbit/s-Ethernet-Standard verspricht Abhilfe, stellt jedoch die Technik im Bereich Paketverarbeitung vor große Herausforderungen.

Am 17. Juni 2010 wurde der IEEE-Standard 802.3ba für 100-Gbit/s-Ethernet (100 GbE) verabschiedet. Eine Kooperation zwischen der IEEE 100 GbE Task Force und der ITU-T-Studiengruppe 15 stellte dabei sicher, dass die neue Ethernet-Datenrate auch über optische Transportnetze (OTN) übertragen werden kann. Somit können erste Hersteller bereits jetzt komplette 100-GbE-Lösungen – Switching und Transport – anbieten [1].

Während die Anforderungen an das Transportnetz und die optischen Schnittstellen unbestritten sind [2], gibt es beim 100-GbE-Paketswitching ebenfalls große Herausforderungen: Datenraten von 100 Gbit/s bringen die heutige Chiptechnik insbesondere im Bereich Daten- und Speicherschnittstellen an die Grenzen ihrer Möglichkeiten.

Architektur eines Paketverarbeitungssystems

Die beiden Kernkomponenten im Datenpfad eines 100-GbE-Paketswitches sind der Network Processor (NP) und der Traffic Manager (TM), in **Bild 1** dunkel gezeichnet. Aufgaben eines Network Processors sind die Klassifizierung von Ethernet- bzw. IP-Paketen, die Manipulation von Paket-Headern sowie natürlich die Wegewahl (Switching/Routing). Die dafür benötigten Filterregeln und Adresstabellen sind aufgrund ihres Umfangs in einem externen Speicher untergebracht. Abhängig von Entscheidungen des Network Processors in Bezug auf Dienstgüte und Wegewahl setzt der Traffic Manager das Bandbreitenmanagement und die Steuerung des Datenflusses durch die Switch Fabric um. In Paketswitchen mit großer Kapazität ist diese Matrix auf einer zentralen Baugruppe untergebracht. Die Verbindung zu den einzelnen 100-GbE-I/O-Baugruppen laufen über Fabric-Adapter-Bausteine.

Der Standard IEEE 802.3ba deckt selber „nur“ die 100-GbE-PHY- und -MAC-Schichten ab, links in Bild 1; ein im OTN vorhandener ODU4-Framer ist in ITU-T G.709 standardisiert. Ethernet Switching (IEEE 802.1) bei 100 Gbit/s unterscheidet sich in seiner Funktion nicht vom Switching bei anderen Ethernet-Datenraten. Somit müssen weiterhin alle Paketoperationen im Datenpfad für eine minimale Paketlänge von 64 Byte ausgelegt sein. Dies bedeutet, dass im Datenpfad in jeder Richtung bis zu 150 Mio. Pakete pro Sekunde verarbeitet werden.

Auf einen Blick

Eine kosteneffiziente Realisierung von 100 GbE ist für eine schnelle Verbreitung dieser Technik unumgänglich. In herkömmlichen Paketswitchen machen Speicherchips bis zu einem Viertel der Gesamtkosten aus, daher spielen Verbesserungen am Speicher-Subsystem eine bedeutende Rolle.

Die Herausforderungen liegen in den Speicherschnittstellen

Das begrenzende Element bei der Verarbeitung dieser Datenmengen stellt dabei nicht notwendigerweise die Chiplogikdichte dar: Auf Kosten von Platz- und Stromverbrauch kann die Funktion

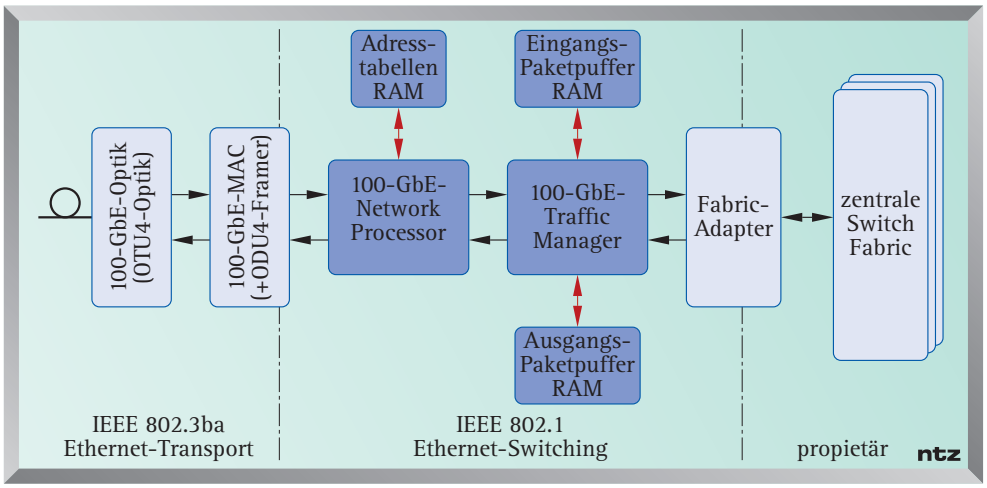


***Bild 1**. Architektur eines 100-Gigabit/s-Ethernet-Paketswitches*

Parameter	QDR-II+ SRAM	RLDRAM-II	RLDRAM-III	DDR3-SDRAM
Verfügbarkeit	verfügbar	verfügbar	angekündigt 2011	verfügbar
relative Kosten/Bit	50	20	—	1
Kapazität/Chip	72 Mbit	576 Mbit	bis 1 Gbit	4 Gbit
I/O-Bus	SIO	SIO	SIO	CIO
Datenrate/Pin	1,0 Gbit/s	1,1 Gbit/s	2,1 Gbit/s	1,6 Gbit/s
initiale Latenzzeit	5,0 ns	15 ns	< 10 ns	45 ns
t_{RC} *(Row Cycle Time)*	(2 ns)	15 ns	< 10 ns	45 ns
Speicherbänke	1	8	16	8
t_{RRD} *(Row-to-Row Delay)*	—	(1,9 ns)	(0,9 ns)	6 ns

***Tabelle 1**. Vergleich von Speichertechniken*

häufig auf mehrere Bausteine aufgeteilt werden, wie in Bild 1 angedeutet ist. Das Problem ist die Bandbreite der externen Daten- und insbesondere Speicherschnittstellen (rot gezeichnet), die in den letzten Jahren nicht mit dem Wachstum der Chiplogikdichte („Moore's Law") mitgehalten hat.

Zusätzlich zum Speicher für den Network Processor werden in einem 100-GbE-Paketswitch mehrere Gigabyte an RAM für die Ein- und Ausgangs-Paketpuffer des Traffic Managers sowie für die Verwaltung der Warteschlangen benötigt. Für Adresstabellen und Statistikfunktionen wird normalerweise schnelles SRAM (Static RAM) benutzt, während die großen Paketpuffer auch in langsamerem DRAM (Dynamic RAM) mit hoher Speicherkapazität realisiert werden. In der Praxis wird für letzteres in Anwendungen mit hohen Datenraten Reduced Latency DRAM (RLDRAM) verwendet, denn die Zugriffszeit von Standard-DRAM (Synchronous DRAM, SDRAM) war bislang nicht ausreichend.

Aktuelle Fortschritte in der DRAM-Technik ermöglichen seit neuestem nun auch die Realisierung von 100-GbE-Systemen auf der Grundlage von Standard-SDRAM. Die **Tabelle 1** zeigt einen Vergleich von DDR3-SDRAM mit den bereits erwähnten Speichertechniken SRAM und RLDRAM.

Neben den Kennzahlen für Kosten und Kapazität ist für eine 100-GbE-Lösung besonders der Parameter t_{RC} (Row Cycle Time) von Interesse, da er maßgeblich die Zugriffszeit bestimmt. Im 100-Gbit/s-Ethernet müssen Pakete alle 6,7 ns verarbeitet werden können (entsprechend 150 Mio. Paketen/s). Aus Tabelle 1 ist ersichtlich, dass dafür zunächst nur (teures) SRAM geeignet ist.

Die DRAM Architektur versucht, das t_{RC}-Problem durch Verwendung mehrerer voneinander unabhängiger Speicherbänke zu lösen: Da t_{RC} für jede Bank getrennt auftritt, kann mit einer Verteilung der Speicherzugriffe auf aufeinanderfolgende Bänke die für 100 GbE erforderliche Zugriffszeit erreicht werden (45 ns/8 < 6,7 ns). Der Bankwechsel ist jedoch durch den Parameter t_{RRD} (Row-

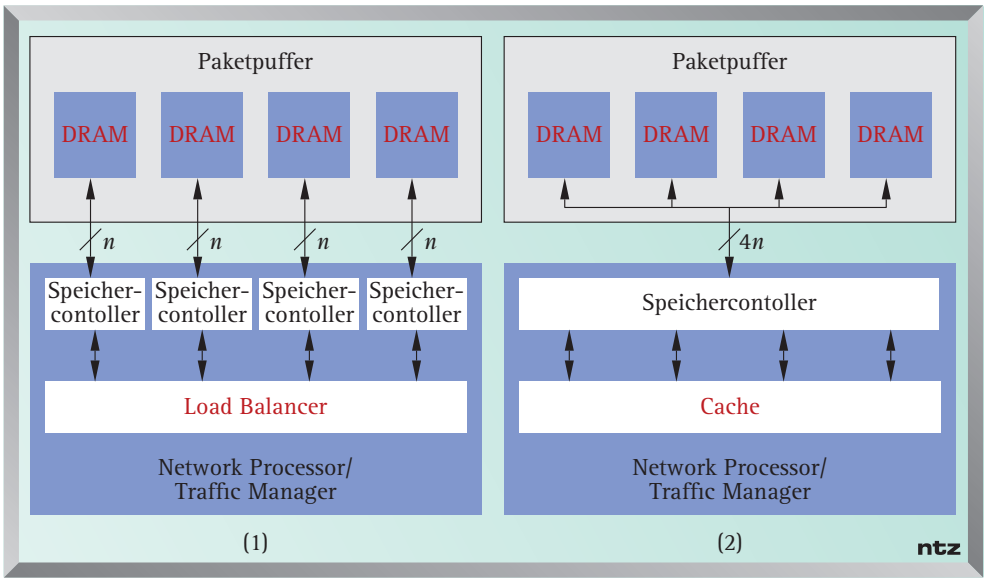


Bild 2. Erhöhung der Speicherbandbreite durch: mehrere Speicherschnittstellen (1) oder durch einen breiten Speicherbus (2)

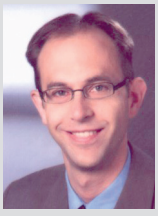
to-Row Delay) beschränkt. Zu beachten ist ferner, dass weiterhin eine initiale Latenz vorhanden ist. An diesem Punkt setzt die RLDRAM-Technik an: Die Latenzzeiten sind wesentlich geringer; zudem kann mit jedem Takt zwischen Speicherbänken gewechselt werden.

Die Kostenattraktivität von DDR3-SDRAM gegenüber den anderen Techniken sticht sofort ins Auge; eine Nutzung in 100-GbE-Paketswitchen ist daher wünschenswert. Es ist nun Aufgabe der Speichermanager in NP und TM, durch eine intelligente Verteilung der Speicherzugriffe die Beschränkungen von SDRAM zu umgehen. In den folgenden Abschnitten werden die zwei wichtigsten Realisierungsmöglichkeiten für 100-GbE-Speichermanager vorgestellt sowie ihre technischen Grenzen diskutiert.

Lösungsansätze für schnelle Pufferspeicher

DDR3-SDRAM-Chips werden mit Datenbusbreiten von 4 bit bis 16 bit angeboten. Tabelle 1 zeigt, dass der für 100 GbE benötigte Datendurchsatz ohnehin nur mit mehreren, parallel geschalteten SDRAM-Bausteine erreicht werden kann (4 … 16 × 1,6 Gbit/s < 100 Gbit/s). Weitere DRAM-Chips sind nötig, um die oben erläuterten Beschränkungen bei der Zugriffszeit auszugleichen.

Ralph Schlenk leitet das vom BMBF geförderte Celtic-Projekt „100GET-AL Layer2-Transportkonzept“ am Alcatel-Lucent-Standort in Nürnberg.
E-Mail: ralph.schlenk@alcatel-lucent.com



In **Bild 2** sind die zwei prinzipiellen Methoden zur Erhöhung der Speicherbandbreite dargestellt: Der erste Ansatz (links) zielt darauf ab, die Anzahl der Speicherschnittstellen zu erhöhen. Ein Speichermanager verteilt die Last auf die parallelen Speichercontroller, so dass die garantierten Zugriffsraten der einzelnen DRAM-Chips nicht überschritten werden. In der zweiten Lösung (rechts) werden die Speicherchips an einen einzigen Speichercontroller angeschlossen und bilden so einen extrem breiten Datenbus. Die Daten werden zunächst im Speichermanager gepuffert und aggregiert, um dann mit einer geringeren Zugriffsrate in das externe DRAM geschrieben zu werden.

Für einen Load Balancer, Bild 2 links, dessen Aufgabe es ist, die garantierten Zugriffsraten einzuhalten, besteht die Schwierigkeit darin, Schreib- und Leseoperationen geschickt zu sortieren. Zudem gilt es, Bus-Turnaround-Zeiten zu beschränken, denn DDR3-SDRAM ist nur als Common-I/O-(CIO-)Variante erhältlich. (Die Separate-I/O-(SIO-)Technik bei SRAM und RLDRAM erlaubt hingegen gleichzeitiges Lesen und Schreiben über getrennte Busse.)

Die beliebige Verteilung von Zugriffen auf Speichercontroller ist allerdings nur bei Schreiboperationen ohne Weiteres möglich. Eine Planung von Leseoperationen ist aus Sicht des Speichermanagers nicht möglich: Wenn sich zu viele Lesezugriffe auf einen Speichercontroller konzentrieren, kann die garantierte Zugriffsrate überschritten werden. Die Lösung besteht darin, die Daten in mehreren DRAM-Chips zu replizieren: Es wird dann immer nur von einem Speichercontroller gelesen, der aktuell keine Zugriffsbeschränkungen aufweist.

In der Praxis lässt sich diese Lösung jedoch nicht beliebig skalieren, denn die Anzahl der Speichercontroller in einem Chip ist beschränkt: Ein ideal implementiertes DDR3-SDRAM-Paketpuffer-Subsystem benötigt bereits mehr als 500 Pins pro Datenrichtung.

Die andere Möglichkeit zur Lösung des 100-GbE-Skalierungsproblems ist die Nutzung eines schnellen Cache-Speichers im Speichermanager, Bild 2 rechts. Dieser Speicher puffert im Beispiel des Traffic Managers den Anfang und das Ende der einzelnen Paketwarteschlangen. Die aggregierten Daten können daher mit einer geringeren Rate in den externen Speicher geschrieben (Ende einer Paketwarteschlange) bzw. aus ihm gelesen werden (Anfang einer Paketwarteschlange). Der Speichermanager stellt dabei sicher, dass der Cache-Speicher immer gut gefüllt ist, so dass er bei Zugriffen niemals leerläuft.

Die Realisierbarkeit Cache-gestützter Lösungen ist abhängig von der Anzahl der Warteschlangen, für die schneller Cache-Speicher bereitgestellt werden muss: Die mögliche Größe von Logikchip-internem Speicher ist durch die Chiptechnik begrenzt.

Es besteht weiterer Optimierungsbedarf

Durch die heute noch notwendige Aufteilung von Funktionen auf mehrere Bausteine (MAC, NP, TM, parallelisierter Speicher) ergibt sich in Zukunft Optimierungsbedarf hinsichtlich Größe, Energieverbrauch und Kühlung. Weitere Fortschritte in Bezug auf Logikchiptechnik (28-nm-FPGA, 40-nm-ASIC), Datenschnittstellen (25-Gbit/s-Transceiver) und Speicherschnittstellen (serielle statt parallele Schnittstellen) sind nötig, damit zukünftige Netzelemente die technischen Herausforderungen von 100 GbE kosteneffizient sowie platz- und energiesparend adressieren. Alcatel-Lucent hat sich als Marktführer bei 100-Gbit/s-Ethernet mit der „High Leverage Network (HLN)“-Strategie [3] der Lösung dieser Herausforderungen angenommen.

Literatur

[1] www.alcatel-lucent.com/100g
[2] *Winterling, P.:* OTN als Transportmedium der Zukunft und Anforderungen an die Messtechnik. **ntz** Fachzeitschrift für Informations- und Kommunikationstechnik 62 (2009) H. 6, S. 28 – 32
[3] www2.alcatel-lucent.com/hln/bandwidth.php