Kurz erklärt: Programmierbare Netzpaketverarbeitung mit Express Data Path

Schnellspur

Uwe Schulze

Da der betagte TCP-Software-Stack viele neue Anforderungen nicht erfüllen kann, erhält der Linux-Kernel moderne Erweiterungen. Eine davon ist der Netzpaketprozessor Express Data Path, der sich mit beliebigen Netzkarten benutzen lässt.

Bei der Freigabe des Linux-Kernels 4.12 ging neben den zahlreichen Neuerungen die Netzwerkerweiterung Express Data Path (eXpress Data Path, XDP) beinahe unter. Der Netzpaketprozessor wurde vom IO-Visor-Projekt unter Schirmherrschaft der Linux Foundation entwickelt und nutzt den Berkeley Packet Filter (BPF und extended BPF) fürs Abfangen der Datenpakete direkt an der Schnittstelle zur Netzkarte. Das Ausführen elementarer Funktionen an dieser Stelle soll Prozessor und Hauptspeicher des Systems entlasten. Etwas Ähnliches existiert bereits mit Intels Data Plane Development Kit, das jedoch hardwareabhängig ist und außerhalb des Linux-Kernels arbeitet.

XDP ist kein Ersatz für den TCP/IP-Stack und auch kein Bypass, denn der Großteil der Datenpakete wird weiterhin zur Verarbeitung an diesen weitergeleitet. Nur wenige Ausnahmebehandlungen nimmt XDP selbst vor – dafür jedoch bedeutend schneller. Dies ist möglich, weil es die Datenpakete (Raw Packets) nicht entpackt und alle Steuerungsinformationen auswertet, sondern nur wenige Felder im Header ausliest. Deshalb können an dieser Stelle lediglich einfache Entscheidungen getroffen werden, hauptsächlich das Filtern und bedarfsweise Löschen von Datenpaketen sowie deren schnelle Weiterleitung (Forward), wobei der Schreibzugriff auf empfangene Pakete zum Modifizieren des Headers genutzt wird. Außerdem können Monitoring- und Flow-Sampling-Prozesse XDP nutzen. Die Bearbeitung erfolgt statuslos pro Netzpaket.

Die Guten ins Töpfchen

Das Verwerfen von Datenpaketen (Drop) war die Hauptmotivation bei der Entstehung von XDP, insbesondere bei DDoS-Attacken, die Server mit Datenpaketen fluten und das Betriebssystem bei der Verarbeitung überfordern. Dem kann man nur begegnen, indem die Pakete zuvor abgefangen und verworfen werden, beispielsweise anhand von Sperrlisten für IP-Adressen. Deshalb war ein erklärtes Ziel bei der Entwicklung von XDP, die Komponente möglichst nah an der Netzhardware anzusiedeln, um direkt auf den Empfangspuffer der Netzkarte zuzugreifen.

Der zweite Anwendungsfall von XDP ist das direkte Weiterleiten von Datenpaketen zum Sendepuffer ohne Bearbeitung im Betriebssystem. Damit lassen sich Load-Balancing-Konzepte mit geringem Aufwand umsetzen. Facebook-Entwickler waren maßgeblich an Konzeption und Programmierung von XDP beteiligt: Der Konzern nutzt die Mechanismen zum Load Balancing seiner Serverfarmen. Hierbei gibt es keinen dedizierten Load Balancer für einen Cluster; stattdessen teilen Server die Datenpakete durch Weiterleitung selbstständig unter sich auf. In dieser ungewöhnlichen Konstellation fungiert jeder Server trotz nur einer Netzkarte als Load Balancer. Dieses Konzept skaliert sehr einfach, ist von Hause aus redundant und benötigt keine zusätzliche Hardware.

XDP ist im Userspace programmierbar, sodass sich die Funktionen einfach und ohne Modifizierungen im Kernel erweitern lassen. Ein Beispiel für neue Anwendungen liefert das Cilium-Projekt, das die Kommunikation zwischen Microservices in Containern mit XDP sicherer machen will. Es arbeitet mit dem Container Network Interface (CNI), Kubernetes und Docker zusammen und soll das Modifizieren der Filterprogramme ohne Unterbrechung der Netzverbindung erlauben.

Die XDP-Entwicklung begann 2015 auf ausgewählter Netzhardware; die ersten Treiber lieferte Mellanox für seine NICs mit 10 und 40 GBit/s. Doch erst der Linux-Kernel 4.12 ermöglichte die Zusammenarbeit mit beliebigen Netzkarten per generischer Schnittstelle (Generic XDP). Voraussetzung ist die Unterstützung durch die Treiber der Netzkarte, um direkten Zugriff auf die DMA-Puffer zu erhalten. Außer Mellanox wurden QLogic (Cavium), Netronome (Agilio SmartNIC), Broadcom und Intel getestet.

In Zukunft dürfte XDP zur Pflichtausstattung von Netzadaptern für Server gehören. Es ist zu erwarten, dass Chiphersteller einige Funktionen in ihrer Hardware implementieren (XDP Offload); von Netronome existieren bereits entsprechende Designs. Dies ist ein allgemeiner Trend bei der Zusammenarbeit von Netzhardware und Software-Stack und wird etwa in Form von TCP Offload praktiziert – auf diese Weise ist XDP möglichst nah an der Netzhardware implementiert. Aber auch beim Aufbau von Software-defined Networking ist XDP große Bedeutung zuzutrauen.

Unter Entwicklern wird XDP jedoch auch kontrovers diskutiert, insbesondere wegen der Möglichkeit, den gesamten Netz-Stack im Kernel zu umgehen. Hinzu kommt grundsätzliche Kritik am Konzept: So sehen einige Programmierer etwa Packet Drop direkt in der Hardware besser aufgehoben – und andere Funktionen wiederum auf der Applikationsebene oder im Netz-Stack des Kernels. Neben rein technischen Argumenten mögen bei der Diskussion aber auch Grundsätze eine Rolle spielen: reine Lehre des Softwaredesigns versus pragmatische Lösung. (tiw)

Uwe Schulze

ist Fachautor in Berlin.

Alle Links: www.ix.de/ix1709116