Kernel-Log

Linux 4.9: Internet-Beschleuniger und Performance-Analysen wie mit Dtrace

Das XFS-Dateisystem von Linux 4.9 kann doppelt gespeicherte Daten zusammenführen und große Dateien in Sekundenbruchteilen kopieren. Außerdem wurden die Sicherheit verbessert und die Möglichkeiten zur Performance-Analyse erweitert, um System- oder Programmoptimierungen zu erleichtern.

Von Thorsten Leemhuis

Der zum Erscheinen dieser c't erwartete Linux-Kernel 4.9 enthält das von Google entwickelte BBR (Bottleneck Bandwidth and RTT). Dabei handelt es sich um einen Congestion-Control-Algorithmus zum Regeln der Übertragungsgeschwindigkeit von Netzwerkverbindungen. Davon enthält der Kernel schon einige – BBR soll allerdings Latenzen und Datenstaus zuverlässiger vermeiden und zugleich die maximale Datenübertragungsrate effizienter ausschöpfen.

Dazu orientiert sich BBR beim Ausloten der optimalen Sendegeschwindigkeit nicht mehr maßgeblich an Paketverlusten, wie es traditionelle Congestion-Control-Algorithmen tun. Als Hauptmaßstab dienen vielmehr die von der Gegenseite zurückgesendeten Bestätigungspakete (ACKs). Laut seinen Entwicklern arbeitet BBR dadurch deutlich robuster auf Verbindungen, bei denen zufällige Paketverluste auftreten. Bei Simulationen einer typischen Letzte-Meile-Anbindung zeigte sich zudem, dass die Latenzen deutlich kürzer ausfallen. Letztlich will Google das Internet so reaktionsschneller machen. Abzuwarten bleibt, wie sich BBR im Feldtest schlägt. Außerdem ist unklar, ob Distributionen den neuen Ansatz in ihren Kerneln standardmäßig aktivieren.

Dtrace-ähnliche Analysen

Über den Berkeley Packet Filter (BPF) ausgeführte Performance-Analyse-Programme können nun autark regelmäßige Messungen durchführen (Timed Sampling) und die erfassten Daten auch gleich zusammenfassen. In Kombination mit vielen anderen Performance-Monitoring-Funktionen, die in den letzten ein bis zwei Jahren entstanden sind, soll Linux damit jetzt endlich grundlegenden Analysefähigkeiten bieten, die jenen der Performance-Analyse-Lösung Dtrace ähneln. So stellt es zumindest der Performance-Optimierungs-Experte Brendan Gregg dar – er hat an einigen der Funktionen mitgearbeitet und pflegt einen ganzen Satz von Werkzeugen zum dynamischen Performance-Monitoring mithilfe von BPF.

Gregg, der auch Bücher zum bei Solaris-Admins geschätzten Dtrace verfasst hat, lieferte jüngst in einem Blog-Eintrag einen guten Überblick zum Stand dynamischer Performance-Analyse mit Linux (siehe c't-Link). Ihm zufolge gibt es trotz der großen Fortschritte nach wie vor viel zu verbessern. So müssten die Funktionen leichter nutzbar werden. Ein Baustein dazu könnte eine von Systemtap abgeleitete Hochsprache zum Programmieren individueller Analysen sein. Ferner erläutert Gregg in dem Beitrag auch die Programme seiner Werkzeugsammlung Bcc-Tools, mit denen sich einige gängige Analyseaufgaben leicht ausführen lassen.

XFS lernt Copy-on-Write

Ähnlich wie das Btrfs-Dateisystem beherrscht auch XFS jetzt den Funktionsaufruf `reflink`. Mit ihm kann `cp` bei Angabe von `--reflink` selbst große Dateien in Sekundenbruchteilen kopieren, sofern die Kopie innerhalb eines Dateisystems erfolgt. Ferner ermöglicht XFS jetzt Deduplizierung, sodass externe Programme doppelt gespeicherte Daten zusammenlegen können, um Speicherplatz zu sparen.

Diese vorerst experimentellen und daher mit den regulären Dateisystemwerkzeugen von XFS noch nicht nutzbaren Funktionen gelingen dank Copy-on-write (CoW). Durch diesen Mechanismus können mehrere Dateien jetzt denselben Speicherbereich referenzieren (Shared Extents). Trotzdem lassen sich die Dateien vollkommen unabhängig voneinander verändern: Sobald über einen Dateieintrag auf einen von mehreren Dateien verwendeten Speicherbereich geschrieben wird, legt XFS die neuen Daten dank CoW an einem anderen Platz ab; anschließend passt es die Metadaten der veränderten Datei an, um den neuen Bereich zu referenzieren. Die anderen Dateien bleiben dabei vollkommen unberührt. Das ist der große Unterschied zu Hardlinks, denn die machen ein und dieselbe Datei lediglich über weitere Dateinamen verfügbar.

Sicherheit

Durch die „Virtually Mapped Kernel Stacks" alloziert Linux 4.9 auf x86-Systemen jetzt den Speicher des Call Stack anders. Dadurch kann der Kernel nun Stack-Overflows besser erkennen und abfangen. Das erhöht die Sicherheit, denn Angreifer lösen diese manchmal gezielt aus, um den Code-Pfad zu verändern und so höhere Rechte zu erlangen. Die Änderung fängt aber auch durch nachlässige Programmierung entstehende Stack-Overflows ab und liefert zugleich mehr Informationen, um die Ursache zu identifizieren.

Das neue „latent_entropy plugin" hilft früh im Startprozess mehr Entropie zu sammeln, um die Erzeugung von Zufallsdaten zu verbessern. Das soll fehlende oder schlechte Zufallsdaten vermeiden, die immer mal wieder zu größeren Sicherheitslücken bei Verschlüsselung oder Schlüsselerzeugung führen. Das Plug-in wurde ursprünglich beim Grsecurity-Projekt entwickelt. Es konnte in 4.9 einfließen, nachdem ein unabhängiger Entwickler es überarbeitet hat, damit es den Qualitätsansprüchen der Kernel-Entwickler genügt. Der Code klinkt sich bei der in Linux 4.8 geschaffenen GCC-Plug-in-Infrastruktur ein.

Bei Intels Server-Prozessoren der Skylake-Generation können Anwendungen in Zukunft „Memory Protection Keys for

Userspace“ nutzen. Mit dieser als MPK oder Pkeys abgekürzten Funktion können Prozesse die im Arbeitsspeicher hinterlegten Daten effizienter vor Zugriffen schützen. Darüber lässt sich etwa ein Chiffrierungsschlüssel so ablegen, dass nur ein Thread eines Verschlüsselungsprogramms ihn lesen darf.

Über neue Dateien im Procfs kann man jetzt begrenzen, wie viele Namespaces ein Anwender innerhalb der ihm zugeteilten Namespaces anlegen darf. Das soll das System vor Ausfällen durch Überlastung schützen – etwa falls eine Software innerhalb eines Containers versehentlich oder bewusst so viel Dateisysteme mountet, dass es das System in die Knie zwingen würde.

Dateisysteme

Das Overlay-Dateisystem beherrscht jetzt erweiterte Attribute (EAs/xattrs). Das ist zum Einsatz von SELinux bei Docker & Co. wichtig, die Overlayfs häufig einsetzen.

Das Virtual File System (VFS), auf dem der Code der Kernel-eigenen Dateisysteme aufbaut, bietet eine neue Schnittstelle zur Datumsabfrage. Darauf portierte Dateisysteme liefern auch in 22 Jahren noch korrekte Werte. Das ist ein weiterer Baustein, um das Jahr-2038-Problem von 32-Bit-Kerneln zu beseitigen.

Der im Kernel enthaltene NFS-Server (NFSd) unterstützt nun die bei NFS 4.2 definierte COPY-Funktion. Mit dieser Fähigkeit ausgestattete NFS-Clients können Dateien Server-seitig kopieren, um ein zeitaufwendiges Übertragen der Daten zum Client und wieder zurück zum Server zu vermeiden. Zusammen mit der Reflink-Unterstützung von Btrfs und XFS lässt sich das Duplizieren der Dateiinhalte sogar komplett vermeiden.

Netzwerk

Admins oder Netzwerk-Management-Programme können außer TCP-Sockets nun auch UDP-Sockets per SOCK_DESTROY-Operation beenden. Anwendungen bekommen dann deutlich schneller mit, wenn Netzwerkanbindungen nicht mehr bestehen – das ist etwa wichtig, damit Video-Streams beim Wechsel von WLAN- auf Mobilfunkdatenanbindung nicht ins Stocken geraten.

Auf schwachbrüstigen Prozessoren soll der Kernel in bestimmten Konstellationen jetzt deutlich mehr Netzwerkdaten annehmen und zugleich den Prozessor weniger belasten. Das ist einigen Änderungen an der Handhabung von Soft-IRQs zu verdanken. Bei Tests auf einem System, auf dem sich zuvor Probleme gezeigt hatten, steigerten diese Änderungen die Zahl der pro Sekunde angenommenen UDP-Pakete von 2000 auf 900.000.

Das Tracing-Subsystem bringt jetzt einen Hardware Latency Tracer mit. Über ihn lässt sich erkennen, wie oft und lange das System mit der Abarbeitung von Non-maskable Interrupts (NMIs) und den damit verwandten System Management Interrupts (SMIs) beschäftigt war. Diese stören vor allem auf Realtime-Systemen, weil sie das Betriebssystem und die darunter laufenden Anwendungen für eine kurze Zeit komplett unterbrechen, während das BIOS oder die Management Engines von AMD und Intel sie abarbeiten.

Neue Treiber

Linux 4.9 unterstützt knapp 500 Geräte oder Geräteklassen mehr als sein direkter Vorgänger. Das geht aus den Datenbanken der Linux Kernel Driver DataBase (LKDDb) hervor, laut derer 4.9 Support für 180 weitere PCI/PCIe- und USB-Geräte bringt. Darunter sind beispielsweise einige neue WLAN-Module von Intel, denn neben Unterstützung für weitere Varianten der Modelle 8265 und 9460 enthält der Kernel nun auch Treiber für die Modelle 8275, 9170, 9270, 9460 und 9560.

Neu dabei ist auch ein bei der Firmware Test Suite (FWTS) eingesetzter Treiber, über den sich die standardkonforme Interaktion zwischen UEFI-Firmware und Betriebssystemen testen lässt. Ferner unterstützt Linux nun auch Amazons Elastic Network Adapters (ENA) – das sind Netzwerkkarten, die Amazon seit dem Sommer in seiner Elastic Compute Cloud (EC2) zur Kommunikation zwischen EC2-Instanzen einsetzt.

Bus für modulare Phones

In das Staging-Subsystem, wo unausgereifter und verbesserungswürdiger Code einfließen darf, wurden Infrastruktur und Treiber von Greybus integriert. Sie implementieren das Interconnect-Protokoll UniPro in der Form, wie es Google beim Project ARA nutzen wollte – dem modularen Smartphone, dessen Entwicklung im Sommer eingestellt wurde. Der Code floss jetzt dennoch in Linux ein, weil Motorola ihn bei einem Gerät einsetzt; auch Toshiba und einige andere Firmen erwägen, Greybus einzusetzen.

Greybus ist der Hauptgrund, warum Linux 4.9 den Ruf hat, mehr Änderungen als alle Versionen davor zu enthalten. Greybus wurde nämlich mit seiner gesamten Entwicklungshistorie integriert, die über 2300 Commits umfasst. Zusammen mit anderen für 4.9 vorgenommenen Änderungen summierte sich das auf über 17.000 Commits; das sind 3000 bis 4000 Änderungen mehr als zuletzt üblich und rund 2500 mehr als beim bisherigen Rekordhalter Linux 3.15.

Am Ende ist das aber alles unwichtige Statistik: Andere Entwickler hätten Greybus vielleicht in einem oder einem Dutzend Commits angeliefert, woraufhin 4.9 dann nur ungefähr so viele Änderungen gehabt hätte, wie der bisherige Rekordhalter. Außerdem sagt die Zahl der Änderungen nichts über Qualität oder Umfang. Ein Blick auf Letzteren zeigt auch: Bei der Zahl der neuen oder entfernten Codezeilen bleibt 4.9 weit hinter den bisherigen Rekordhaltern zurück und ist ein ganz typisches Release. *(thl@ct.de)* c't

Performance-Analyse mit BPF:
ct.de/y57d