Kernel-Log

Linux 4.14: Speicherverschlüsselung, neuer Schlafmodus, mehr Performance

Der neue Linux-Kernel überwindet eine Grenze, die auf die ersten 64-Bit-x86-Prozessoren zurückgeht. Außerdem kann der Kernel im Speicherbereich von Programmen liegende Daten jetzt effizienter versenden und Daten mit einem flexiblen Kompressionsalgorithmus packen.

Von Thorsten Leemhuis

Kurz nach Erscheinen dieser c't dürfte Linus Torvalds die Linux-Version 4.14 freigeben. Dieser Kernel wartet zwar mit keiner herausragenden Neuerung auf, von der Anwender sofort profitieren. Aber er bringt eine ganze Reihe wichtiger Verbesserungen, die ihr Potenzial langfristig oder bei genauerem Hinsehen offenbaren.

Grenzüberwindung

Vierzehneinhalb Jahre nach der Einführung des AMD Opteron samt der 64-Bit-x86-Befehlssatzerweiterung durchbricht der Kernel eine Grenze, die damals noch in weiter Ferne lag. Genau wie der erste x86-64-Prozessor unterstützen nämlich auch die aktuellen maximal 64 Terabyte physischen Arbeitsspeicher und einen virtuellen Adressraum von 128 Terabyte. Die Hersteller einiger High-End-Systeme stoßen gerade an eben dieses Limit.

Intel will die Grenze offenbar mit neuen Xeon-Prozessoren aus der Welt schaffen, die mithilfe einer „LA57“ genannten Erweiterung bis zu 4096 Terabyte (4 Petabyte) Arbeitsspeicher adressieren können sollen; Programme sollen sogar einen virtuellen Adressraum von 128 Petabyte bekommen. Das gelingt mit einer fünften Page-Table-Ebene, die Linux dank zahlreicher von Intel selbst eingebrachter Änderungen jetzt unterstützt; passende Prozessoren hat das Unternehmen aber bislang nicht ankündigt.

Nach mehreren Entwicklungsjahren ist Infrastruktur für Heterogeneous Memory Management (HMM) eingeflossen. Sie soll Programmierern eine deutlich simplere und zugleich performantere Nutzung der verschiedenen Arbeitsspeicherbereiche in modernen Systemen ermöglichen. Das ist etwa für Berechnungen mit C++, OpenCL oder CUDA wichtig, denn durch HMM kann der Grafikprozessor leichter Datenstrukturen lesen und schreiben, die im Hauptspeicher liegen; zugleich kommt der Hauptprozessor auch einfacher an Daten, die im Grafikspeicher liegen. Das Ganze ist nicht nur für Stream-Prozessoren relevant, sondern auch zum effizienteren Einbinden von Host Bus Adaptern (HBAs), Krypto-Beschleunigern, FPGAs oder DSPs.

Daten effizienter versenden

Der Kernel kann von Programmen aufbereitete und im Arbeitsspeicher liegende Daten jetzt direkt über TCP-Verbindungen verschicken, ohne diese zuerst aus dem Speicherbereich der Anwendung (dem „Userspace“) in den des Kernels („Kernelspace“) kopieren zu müssen. Die „Socket Sendmsg MSG_ZEROCOPY“ genannte Funktion vermeidet dadurch Overhead und steigert so die Geschwindigkeit. Dieser Trick ist keineswegs neu, denn per `sendfile()` gelingt das schon länger – allerdings nur mit Dateien.

Laut dem zuständigen Entwickler erzielt die Technik bei Tests mit weitgehend aus dem Produktionsbetrieb übernommenen Workloads einen Performance-Vorteil von 5 bis 8 Prozent. Der Kernel kann das

Ganze aber nicht automatisch nutzen: Programme müssen die Zerocopy-Funktion explizit anfordern, schließlich dürfen sie den Speicherbereich nicht modifizieren, bevor der Kernel die darin enthaltenen Daten verschickt hat.

Flexibler komprimieren

Das Btrfs-Dateisystem und das bei vielen Live-Linuxen verwendete Kompressions- und Image-Dateisystem SquashFS können Daten nun mit Zstandard (kurz: Zstd) komprimieren. Dieser verlustfreie Kompressionsalgorithmus eignet sich für ein breiteres Einsatzspektrum, weil sich Packdichte und der zum Komprimieren benötigte Rechenaufwand flexibler gegeneinander abwägen lassen als bei anderen Algorithmen.

Laut dem zuständigen Entwickler kann Zstd beispielsweise in einem niedrigeren Kompressionslevel ähnlich schnell packen wie LZ4, in einem höheren Level aber eine an LZMA heranreichende Packdichte erreichen; beim Dekomprimieren sollen alle drei auf ähnlichem Niveau liegen und damit knapp doppelt so schnell entpacken wie das von Gzip verwendete Zlib.

Der zuständige Entwickler hat auch das Xxhash-Modul nachgerüstet, das die von Zstd verwendeten Hash-Algorithmen xxh32 und xxh64 implementiert. Diese sind nicht für kryptografische Zwecke geeignet, sondern zur Prüfsummenberechnung – dabei sollen die neuen aber deutlich schneller arbeiten als etwa CRC32, daher sollen sie womöglich auch an anderen Stellen innerhalb des Kernels zum Einsatz kommen.

Performance-Zuwachs

Mit dem neuen ORC Stack Unwinder bietet Linux jetzt eine weitere Methode, um bei Kernel-Fehlern einen Stacktrace zu erzeugen – also eine Aufstellung mit den Namen der Funktionen, über die der Kernel-Code zum Fehlerpunkt gelangt ist. Fedora, Ubuntu und einige andere Distributionen nutzen dazu bislang den Frame Pointer Unwinder, obwohl mit ihm großer Overhead einhergeht, der das System verlangsamt. Laut dem federführenden ORC-Entwickler kann ein Umstieg auf ORC daher die Gesamtperformance um 1 bis 3 Prozent steigern; bei besonders Kernel-lastigen Aufgaben können es sogar zwischen 5 und 10 Prozent sein.

Die Performance-Steigerung war aber nicht der Hauptgrund zur Entwicklung von ORC. Vielmehr soll ORC letztlich Kernel Live Patching (KLP) robuster

Die 64-TByte-Grenze geht auf den ersten 64-Bit-x86-Prozessor AMD Opteron zurück.

und flexibler machen; auch Profiler profitieren vom neuen Ansatz. Einige Distributoren wollen aber offenbar noch abwarten, wie sich ORC im Feldtest schlägt, bevor sie umsteigen.

Anders schlafen

Mit dem neuen Kernel wechseln manche Notebooks nicht mehr in den Suspend-to-RAM (aka STR oder ACPI S3), wenn man sie in den Bereitschaftsmodus schickt, sondern in Suspend-to-Idle (S2I oder S2Idle). In diesem systemweiten Zustand legt der Kernel die Chips des Systems mit den zur Laufzeit nutzbaren Stromsparmodi so tief wie möglich schlafen, ähnlich wie es auch Android macht. Der Stromverbrauch beim Schlafen ist dadurch höher als im STR, wo außer dem Arbeitsspeicher nur die zum Wecken benötigten Chips noch Strom erhalten; dafür wachen die Geräte viel schneller auf und sind schneller einsatzbereit, weil sie etwa WLAN-Verbindungen nicht neu aufbauen müssen.

Das Ganze betrifft vorwiegend mit besonders sparsamen Prozessoren ausgestattete Notebooks aus dem oberen Preissegment, etwa das mit Linux und Windows erhältliche Ultrabook Dell XPS 13 (9360) oder aktuelle Surface-Notebooks von Microsoft. Im Windows-Betrieb nutzen diese einen von Microsoft „Modern Standby“ genannten Schlafzustand, der weitgehend S2I entspricht. STR funktioniert auf solchen Notebooks teilweise unzuverlässig, weil die Hersteller diesen Schlafzustand nicht testen und dadurch schwerwiegende Fehler im für ACPI S3 zuständigen Firmware-Code übersehen. Unter Windows ist daher kein Wechsel in den STR möglich; unter Linux kann man den Einsatz dieses Systemschlafzustands vorgeben, indem man `deep` in die Datei /sys/power/mem_sleep schreibt und danach ganz normal in den Bereitschaftsmodus wechselt.

Kaltstartattacken-Schutz

Linux implementiert jetzt eine EFI-Technik zum Schutz gegen Angreifer, die durch Neustarts an sensible Arbeitsspeicherinhalte zu gelangen versuchen. Bei solch einer Kaltstartattacke (Cold Boot Attack) löst ein Angreifer mit physischem Zugriff einen sofortigen Reboot aus, um dann mit einem von ihm kontrollierten Betriebssystem die noch verbliebenen RAM-Inhalte auszulesen. Der Support für die „TCG Platform Reset Attack Mitigation“ erschwert diesen Angriffsweg: Ein BIOS, das diese Spezifikation implementiert, löscht beim Neustart alle Speicherinhalte, bevor es wieder ein Betriebssystem startet.

RAM verschlüsseln

Der neue Kernel unterstützt AMDs „Secure Memory Encryption“ (SME), mit der AMDs Epyc-Prozessoren und der für Business-PCs gedachte Ryzen Pro weite Teile des Arbeitsspeichers verschlüsseln können. Anders als der erwähnte Kaltstartattacken-Schutz im EFI-Code hilft das auch gegen Angreifer, die Speichermodule zum Auslesen in ein anderes System verpflanzen. Außerdem schützt SME auch gegen Lauschen (Sniffing) am Speicherbus oder beim Diebstahl nichtflüchtiger Speichermodule (NVDIMMs).

AMD zielt mit der Technik aber offenbar vornehmlich auf Cloud-Provider, damit aus einer VM ausgebrochene Angreifer die Speicherinhalte anderer VMs nicht einsehen können. Das ermöglicht das auf SME aufbauende Secure Encrypted Virtualization (SEV). Patches zur Unterstützung dieses Ansatzes in Linux und dessen Hypervisor KVM sind noch in der Begutachtungsphase und sollen bald folgen.

Firmware entfernt

Die Entwickler haben sämtliche Firmware-Dateien entfernt. Diese vorkompilierten und von manchen Treibern benötigten Dateien lagen bislang unterhalb von firmware/, passender Source-Code fehlte in der Regel. Anwender erhalten die Dateien jetzt über die unabhängig vom Kernel gepflegte Firmware-Sammlung „linux-firmware“, die alle Mainstream-Distributionen schon länger einsetzen.

Die Datei /proc/cpuinfo zeigt jetzt wieder eine grobe Schätzung der Taktfrequenz von x86-Prozessoren an: Eine für 4.13 vorgenommene Änderung, durch die sich dort nur noch der Basistakt fand, wurde revidiert.

Treiber

Um die Hardware-Unterstützung zu verbessern, haben die Entwickler wieder zahlreiche Treiber überarbeitet und weitere integriert. So gab es größere Umbauten am DVB-S2-Treiber, der Probleme beseitigt, den Funktionsumfang erweitert und Support für die von Digital Device gefertigte DVB-Hardware CineS2 V7(A), DuoFlex S2 V4 und Max-S8 nachrüsten. Linux bringt jetzt auch einen Treiber für den 802.11ac-WLAN-Chip RTL8822BE von Realtek mit, der allerdings im Bereich für Code mit größeren Qualitätsmängeln liegt. Die Alsa- und ASoC-Treiber unterstützten jetzt die Audio-Funktion von Cannon-Lake-Prozessoren, die Intel voraussichtlich in den nächsten Monaten einführen will. Auch der Grafiktreiber unterstützt diesen Chip jetzt besser, der Support gilt aber weiter als unfertig.

Ferner gab es bei Intels Grafiktreiber einige Detailänderungen, die die Performance steigern und den Support für die seit Linux 4.8 mögliche GPU-Virtualisierung mit Intels Graphics Virtualization Technology (GVT) verbessern. Der Nouveau-Treiber des Kernels kennt jetzt auch die GeForce GT 1030; 3D-Beschleunigung gelingt nicht, weil Nvidia bislang keine passende Firmware veröffentlicht hat. Der für die verschiedenen Raspberry Pi zuständige Treiber VC4 unterstützt jetzt HDMI CEC (Consumer Electronics Control), durch das sich mehrere per HDMI verbundene Geräte über eine Fernbedienung steuern lassen. Neu dabei ist ein Treiber für das LCD-Panel des Roboterbaukastens LEGO Mindstorms EV3.

Der neue Kernel bringt einen Treiber für das Display des Roboterbaukastens LEGO Mindstorms EV3 mit.

Amdgpu: Land in Sicht

Der Amdgpu-Treiber, der für moderne Radeon-GPUs zuständig ist, kann jetzt große Speicherseiten (Huge Pages) nutzen. Solche vermeiden Overhead beim Speicherzugriff und steigern so die Performance – insbesondere beim Rechnen mit Grafikprozessoren. Ferner gab es allerlei Verbesserungen und Feintuning rund um den Support von Vega-Grafikkarten. Mit ihnen kann der Treiber bislang nur Rendern und Rechnen, aber keine Bildschirme ansteuern; dasselbe gilt für die GPUs des jüngst angekündigten Ryzen Mobile (siehe S. 12).

Bei diesen beiden Grafikprozessoren schafft der Treiber das erst mit der anfangs DAL (Display Abstraction Layer) und mittlerweile DC (Display Core) genannten Patch-Sammlung. Der zuständige Kernel-Entwickler hat sie anfangs rundweg abgelehnt, weil sie eine unnütze Abstraktionsschicht einziehen. AMDs Mitarbeiter haben diesen und weitere Kritikpunkte in den letzten 20 Monaten ausgeräumt. Der Subsystem-Betreuer ist jetzt weitgehend zufrieden und will Torvalds bitten, die Patch-Sammlung in Linux 4.15 zu integrieren, das Mitte Januar erscheinen dürfte. Damit würden AMDs quelloffene Grafiktreiber alle wesentlichen Funktionen aktueller Grafikchips unterstützen und dabei auch ordentliche 3D-Performance liefern. Das ist das erste Mal, denn in einem dieser beiden Bereiche haperte es bislang immer. Damit scheint AMD sein Ziel erreicht zu haben, zum Linux-Support verstärkt auf quelloffene Grafiktreiber zu setzen – zehn Jahre, nachdem das Unternehmen sich genau das groß auf die Fahnen geschrieben hat. *(thl@ct.de)* c't