Kernel-Log

Linux 4.10: Rucklervermeidung und RAID-Cache

Der dieser Tage erwartete Linux-Kernel 4.10 enthält eine Technik, mit der Linux- und Windows-VMs die 3D-Beschleunigung des Grafikprozessors im Host nutzen können. Die Entwickler haben sich zudem der Linux-Inkompatibilität einiger Lenovo-Notebooks gewidmet und die Unterstützung für einen Schlafzustand aktueller Notebooks verbessert.

Von Thorsten Leemhuis

Kurz nach Verkaufsstart dieser c't soll der Linux-Kernel 4.10 erscheinen. Er bietet Virtual Machines (VMs) neue Möglichkeiten, ausgewählte Hardware des Wirtsystems zu nutzen. Linux- und Windows-VMs sollen auf diesem Weg etwa Beschleunigungsfunktionen des Grafikprozessors im Host ohne größere Geschwindigkeitseinbußen verwenden und ohne dass es die Sicherheit des Wirts gefährdet. Die Technik vermeidet dabei Inkompatibilitäten, derentwegen frühere Ansätze zum VGA-Passthrough auf so manchem System mehr schlecht als recht funktionieren.

Die Grundlage für den neuen Ansatz legt das neue „Mediated Device Interface". Damit können Treiber im Wirt ein „Mediated Device" (Mdev) erzeugen, das limitierten oder nahezu kompletten Zugriff auf kontrollierte Hardware gewährt; der Treiber muss dabei verantworten, dass die Sicherheit des Hosts gewahrt bleibt. Die Kontrolle über das Mdev kann Qemu dann an eine Virtual Machine (VM) übergeben, die unter KVM oder Xen läuft. Das dort laufende Betriebssystem sieht Hardware genauso wie andere Komponenten, die Qemu für die VM emuliert.

AMD und Nvidia wollen diesen Weg schon bald nutzen, um Teile von Radeon- oder GeForce-GPUs vom Wirt an VMs durchzureichen. Der Intel-Grafiktreiber des Kernels kann das bereits bei 4.10, sofern KVM als Hypervisor dient. Intel nennt das Konzept „Graphics Virtualization Technology" (GVT). Bei Linux 4.10 erfordert das Ganze einen Prozessor aus einer der drei neuesten Generationen (Broadwell/Core-i-5000 und neuer). Außerdem funktioniert bislang lediglich ein Headless-Mode, denn die Unterstützung zur Bildausgabe ist noch in Arbeit. Selbst mit der Technik vertraute Entwickler kämpfen beim Einsatz derzeit mit Kinderkrankheiten; es dürfte daher noch etwas dauern, bis das Ganze alltagstauglich wird und ohne Basteleien bei Host- und Gast-Betriebssystem funktioniert.

Ruckler-Vermeidung

Bislang geraten Linux-Desktops und Anwendungen häufiger mal ins Stocken, wenn der Kernel im Hintergrund große Datenmengen schreibt. Besonders häufig passiert das, wenn man Firefox verwenden will, während man beispielsweise ein ISO-Image auf einen USB-Stick transferiert.

Das neue „Writeback Throttling" soll solche Aussetzer vermeiden. Dazu drosselt Linux das Rausschreiben der im Cache zwischengespeicherten Daten, wenn andere Anwendungen einige Lese- oder Schreiboperationen absetzen. Letztere kommen dadurch zeitnah an die Reihe. Bislang wurden sie hinten angestellt, was zum Stocken der Anwendungen führte. Die Drosselfunktion ist auch für Server wichtig, damit beispielsweise Datenbankabfragen nicht plötzlich deutlich länger dauern, bloß weil der Kernel gerade viele Daten schreibt.

Cache für RAID 4, 5 & 6

Seit Linux 4.4 kann der Kernel einen Datenträger als „Log-Device" an Software-RAIDs der Level 4, 5 und 6 koppeln, um bei Abstürzen die Datenintegrität besser zu gewährleisten. Ab 4.10 kann der Kernel einen solchen Datenträger auch als Write-back-Cache nutzen.

Das steigert die Schreib-Performance eines Festplatten-RAID deutlich, wenn eine SSD als Log-Device dient: Anwendungen sehen die Daten bereits als geschrieben an, sobald die SSD sie gespeichert hat. Der Kernel überführt die Daten erst später auf das typischerweise langsamer arbeitenden Festplatten-Array. Dadurch kann er größere Datenmengen zum Schreiben ansammeln und so die Stripes des RAID häufiger in einem Rutsch füllen, was zeitaufwendige Read-Modify-Write-Zyklen vermeidet. Die neue Funktion gilt aber noch als experimentell und lässt sich nur mit Entwicklerversionen von Mdadm nutzen.

NVMe-SSD-Problemhinweis

Linux warnt jetzt in den Kernel-Meldungen, wenn es NVMe-Datenträger eines Systems nicht ansprechen kann, weil dessen BIOS die Storage-Adapter auf eine untypische Weise konfiguriert. Das ist eine Reaktion auf die Linux-Inkompatibilität des Lenovo Yoga 900-13ISK2, über die zahlreiche Medien berichtet haben. Zwischenzeitlich hat Lenovo für dieses Modell ein spezielles „Linux-BIOS" freigegeben: Mit ihm gelingt die Linux-Installation, sofern man im BIOS-Setup vom RAID- auf AHCI-Modus umschaltet.

Auf dieselbe Weise muss man auch einige Notebooks anderer Hersteller umkonfigurieren; der Kernel weist daher jetzt in den per `dmesg` abrufbaren Meldungen darauf hin. Ursprünglich hatten die Kernel-Entwickler das Problem aus der Welt schaffen wollen, indem der Kernel die Hardware beim Start selbstständig umkonfiguriert. Zur zuverlässigen Umsetzung dieser Methode fehlen den Entwicklern aber Informationen, die Intel geheim hält.

Neuer Schlafzustand

Vorwiegend für neue Notebooks relevant ist die neue Sysfs-Datei /sys/power/mem_sleep. Über sie kann man festlegen, in welchen Zustand das System wechseln soll, wenn man es schlafen schickt – etwa über die Desktop-Umgebung oder durch Zuklappen des Notebook-Displays. Bislang wechseln Systeme dann typischerweise per ACPI S3 in einen Suspend-to-RAM.

Einige neuere Notebooks bieten mit Suspend-to-Idle (auch S2I oder s2idle genannt) einen moderneren Schlafmodus. Dessen Leistungsaufnahme ist zwar etwas höher, dafür wachen Systeme deutlich schneller auf und behalten auch ihre Netzwerkverbindung; so sind Notebooks voll einsatzbereit, sobald man das Display aufgeklappt. In der Windows-Welt wird dieser aus „Connected Standby“ hervorgegangene Schlafzustand „Modern Standby“ genannt. Neuere Ultrabooks verwenden ihn teilweise standardmäßig. Bei einem Kurztest mit der neuesten XPS-13-Generation haperte S2I unter Linux noch, weil der Kernel keine Aufwachereignisse konfigurierte; dieses Problem wollen die Entwickler in einer der nächsten Linux-Versionen korrigieren.

WLAN: Schneller & robuster

Seit Jahren kämpft eine Reihe von Entwicklern gegen exzessives Puffern bei Netzwerkübertragungen, denn das verlängert die Latenzen, was Verbindungsabbrüche und andere als „Bufferbloat“ geführte Probleme nach sich ziehen kann. Entwickler der Initiative „Make-wifi-fast“ gehen diese Problematik jetzt auch im WLAN-Stack an und haben Umbauten am WLAN-Treiber Ath9k vorgenommen, um zu puffern, ohne dass lange Latenzen entstehen.

Das steigert die Reaktionsgeschwindigkeit von Webseiten und macht die Kommunikation mit Webservern robuster, was Verbindungsabbrüche bei schlechten WLAN-Verbindungen vermeidet. Bislang profitieren davon allerdings nur die vom Ath9k-Treiber unterstützten WLAN-Chips von Atheros/Qualcomm. Die Entwickler arbeiten aber an Änderungen, um Bufferbloat auch beim Ath10k-Treiber zu vermeiden, der modernere Chips des Unternehmens unterstützt.

Prozessor-Effizienz

Der Kernel unterstützt die „Turbo Boost Max Technology 3.0“ neuerer Intel-Prozessoren jetzt besser. Hierbei kann ein CPU-Kern etwas höhere Frequenzen erreichen als die anderen. Das macht sich der Prozess-Scheduler jetzt zunutze und delegiert Tasks an den schnelleren Kern, um die verfügbare Rechenkraft besser auszunutzen.

Auch Intels Cache Allocation Technology (CAT) wird nun unterstützt. Mit CAT können Admins den Cache einiger Xeon-Prozessoren partitionieren. Dadurch kann eine vorübergehend laufende Anwendung nicht mehr die Daten eines wichtigeren Programms aus dem Cache drücken, was diese verlangsamen würde.

Die Performance-Analyse-Infrastruktur `perf` bringt jetzt das Kommando `perf c2c` mit, das Probleme bei der Nutzung der Prozessor-Caches finden kann, die die Verarbeitungsgeschwindigkeit reduzieren. Ebenfalls neu ist `perf sched timehist`, das eine bedarfsgerechte Analyse der Scheduler-Aktionen ermöglicht.

Der Kernel soll Thunderbolt-Geräte auf Macbooks jetzt „voll unterstützen“. Ein mit der Standard-Konfigurationsdatei für ARM64-Systeme kompilierter Kernel bootet jetzt auch auf dem Raspberry Pi 3.

Dachschindel-Platten

Der Block-Layer des Kernels weiß jetzt die Zoned Blocks in Festplatten mit Shingled Magnetic Recording (SMR) handzuhaben. SMR unterteilt Platten in Zonen, bei denen eine neu geschriebene Spur die zuvor geschriebene ein wenig überlappt, ähnlich wie übereinanderliegende Dachschindeln (englisch „Shingle“). Dadurch lassen sich auf der Plattenfläche mehr Daten unterbringen; allerdings muss durch diesen Ansatz die komplette Zone neu geschrieben werden, um ein zuvor gespeichertes Bit zu ändern. Für optimale Performance sollte das Dateisystem auf diese Eigenart Rücksicht nehmen; bislang beherrscht das aber noch keines der bei Linux gängigen Desktop- und Server-Dateisysteme.

Ext4 soll jetzt deutlich robuster gegen Angriffe mit gezielt korrumpierten Dateisystemen sein. Das XFS-Dateisystem hat eine neue Direct-IO-Implementation auf Basis der jüngst eingeführten Iomap-Codes bekommen, die einfacher und schneller arbeiten soll. OverlayFS, das mehrere Dateisysteme übereinander schichtet, unterstützt nun das Umbenennen von Dateien und Verzeichnissen.

Netzwerkrouten lassen sich nicht mehr nur systemweit festlegen, sondern für jeden Nutzer individuell. Der Kernel von Android bietet eine solche Funktion schon länger; Google-Entwickler haben den dafür zuständigen Code jetzt überarbeitet und in den offiziellen Kernel eingebracht. Unabhängig davon haben die Kernel-Entwickler einige Funktionen beim Paketfilter Nftables (NFT) nachgerüstet, die bislang Iptables vorbehalten waren, das NFT zu beerben versucht.

Grafiktreiber

Neben den Grafikprozessoren Polaris 10 und 11, die bei Radeon RX 460 bis 480 zum Einsatz kommen, unterstützt der Amdgpu-Treiber jetzt auch die Polaris-12-GPUs; wo AMD diese einsetzen will, ist noch unklar. Der Kernel kann nun ohne weitere Treiber den TV-Ausgang verschiedener Raspberry-Pi-Modelle aktivieren. Der Support für Fragment Shader Threading in VC4 verspricht, die Grafik-Performance von Raspis zu steigern.

Der Nouveau-Treiber unterstützt jetzt auch den Grafikchip GP106, der auf der GeForce 1060 sitzt. Jetzt bietet der Treiber auch Basis-Support für Multi-Stream Transport (MST) und kann so 4K/HiDPI-Displays per DisplayPort ansteuern, die sich beim Betriebssystem als zwei unabhängig arbeitende Bildschirme melden.

Über ein neues Userspace-LED-Interface lässt sich bei Titan-Grafikkarten jetzt regeln, wie hell das Nvidia-Logo der Karte das Gehäuseinnere illuminiert. Das ist sicherlich ein weniger wichtiges Kernel-Feature – aber auch Entwickler von Linux-Treibern müssen mal entspannen und arbeiten daher manchmal an Details, nach denen ihnen gerade der Sinn steht. *(thl@ct.de)* c't