Was ist eigentlich NVMe und NVMe-oF?

Protokolle für die Datenübertragung in Massenspeichermedien gibt es viele.
Eines ist **N**on-**V**olatile **M**emory **E**xpress oder **NVMe**, das im Jahr 2011 eingeführt wurde, also noch relativ neu ist. In der vollständigen Benennung heißt es **N**on-**V**olatile **M**emory **H**ost **C**ontroller **I**nterface **S**pecification (**NVMHCIS**). Bei der Entwicklung von NVMe lag der Fokus auf der Optimierung für Flash-basierte, nichtflüchtige Massenspeichermedien (z. B. **S**olid-**S**tate **D**rives, **SSD**). NVMe verbindet als logische Schnittstelle Flash-Speicher mit einem Rechner, in der Mehrheit Server und Workstations, aber zunehmend auch Gaming PCs.

Vorteile von NVMe

Eigenschaften von NVMe sind neben der Flash-Optimierung die Multithreading-Fähigkeit, hohe IOPS-Leistung und geringe Latenz sowie die bessere Skalierbarkeit gegenüber älteren Protokollen wie SATA. Ein großer Vorteil ist, dass es ohne herstellerspezifische Treiber auskommt.
Das NVMe-Protokoll kann für eine Vielzahl von Betriebssystemen verwendet werden, da es Daten parallel verarbeitet und den bestehenden PCIe-Standard nutzt. Es unterstützt verschiedene Ausführungen von nichtflüchtigen Flash-Massenspeichern, wie PCIe, **SFF** (**S**mall **F**orm **F**actor) mit U.2.-Anschluss oder den Standard **NGSFF** (**N**ext **G**eneration **S**mall **F**orm **F**actor), der besser unter dem Kürzel M.2. bekannt ist. Die Hot-Swap-Fähigkeit und der kleine Befehlssatz mit geringem Overhead sind weitere Vorzüge, durch die NVMe sich auch bestens für Workstations, Hosting und High Performance-Tasks eignet.

Die NVMe-Schnittstelle verringert erheblich Eingabe- und Ausgabebefehle im Speicher und unterstützt im Interrupt- oder Polling-Modus laufende betriebssystemseitige Gerätetreiber. So können eine bessere Leistung und niedrigere Latenzen sichergestellt werden.

Die NVMe-Architektur verfügt über einen High-Performance-Warteschlangenmechanismus, der die extrem hohe Menge von bis zu 65.535 E/A-Warteschlangen mit jeweils bis zu 65.535 Befehlen unterstützt! Dies ist gegenüber SATA mit 1 Warteschlange und 23 Einträgen bzw. SAS mit 1 Warteschlange und 256 Einträgen eine gigantische Steigerung – und da die Warteschlangen den CPU-Kernen zugeordnet sind, kann die Performance skaliert werden.

	SATA	SAS	NVMe
DURCHSATZ	6 Gbit/s	12 Gbit/s	16 Gbit/s (Gen3 x16) 32 Gbit/s (Gen4 x16)
IOPS	60.000 – 100.000	200.000 – 400.000	200.000 – 10.000.000
LATENZ	≤1 ms bis >100 ms	<100 µs bis >100 ms	<10 µs bis 225 µs
QUEUE	1	1	max. 65.535
EINTRÄGE PRO QUEUE	32	256	max. 65.535

Leistungsvergleich SATA vs. SAS vs. NVMe (eigene Darstellung)

Ausgehend von der NVMe-Spezifikation wurden weitere Spezifikationen für bestimmte Zwecke abgeleitet, wie zum Beispiel **NVMe-MI** und **NVMe-oF**.

- **NVMe M**anagement **I**nterface (**NVMe-MI**):
  Architektur für Monitoring/Konfiguration von Storage-Hardware (Temperatur, Leistung, Health, etc.) per In-Band- und Out-of-Band-Management

- **NVMe-o**ver-**F**abrics (**NVMe-oF**, mit z. B. NVMe-over-RDMA oder NVMe-over-TCP):
  Schnellere Datenübertragung bei Massenspeichermedien in Netzwerken via InfiniBand, FC (Fibre Channel) oder Ethernet

Wir stellen an dieser Stelle das NVMe-oF-Prinzip vor.

NVMe over Fabrics (NVME-oF)

Um NVMe über direkt angebundene flashbasierte Speicher hinausgehend auch für die Datenübertragung in Netzwerken per Ethernet, Fibre Channel oder InfiniBand nutzbar zu machen, wurde es als **NVME o**ver **F**abrics (**NVME-oF**) weiterentwickelt.

Die traditionellen Standard für Speichersysteme sind das veraltete SCSI, gefolgt von Fibre Channel (FC) und SAS. Später kam die Internetvariante von SCSI: iSCSI. Damit konnten Daten auch an geografisch entfernten Speicher übertragen werden. Der Nachteil von iSCSI ist jedoch der große Overhead der Datenpakete, der die Übertragungsleistung drosselt. Die Alternative Fibre Channel (FC) wiederum ist von der Anschaffung teurer FC-Hardware abhängig.

Mit NVMe-oF wurden diese Probleme gelöst:
Es ermöglicht bei SSD-Direktanschluss eine 10 bis 15 Prozent bessere Transportleistung einerseits und aufgrund der NVMe-Controller und -Fabric im Netzwerk die Verbindung von Tausenden NVMe-Hosts andererseits. Dank Multi-Queue mit getrennten Warteschlangen, die ca. 65.000 Aufträge bewältigen können, werden die Ressourcen optimal gemeinsam genutzt: Bottlenecks und lange Wartezeiten gehören der Vergangenheit an!

Vor NVMe gab es das Problem, dass SATA und PCIe die potenzielle Leistung von Flash-Speichermedien noch nicht vollständig ausschöpfen konnten. Erst das speziell für die Verbindung von SSDs mit PCIe entwickelte NVMe machte dies möglich und die Erweiterung um Ethernet- und FC-Eigenschaften bei NVMe-oF liefert schließlich die ersehnte hohe Übertragungsleistung in Netzwerken mit Flash-Speicher mit schnellerer und effizienterer Server-Storage-Konnektivität und geringerer Server-CPU-Auslastung.

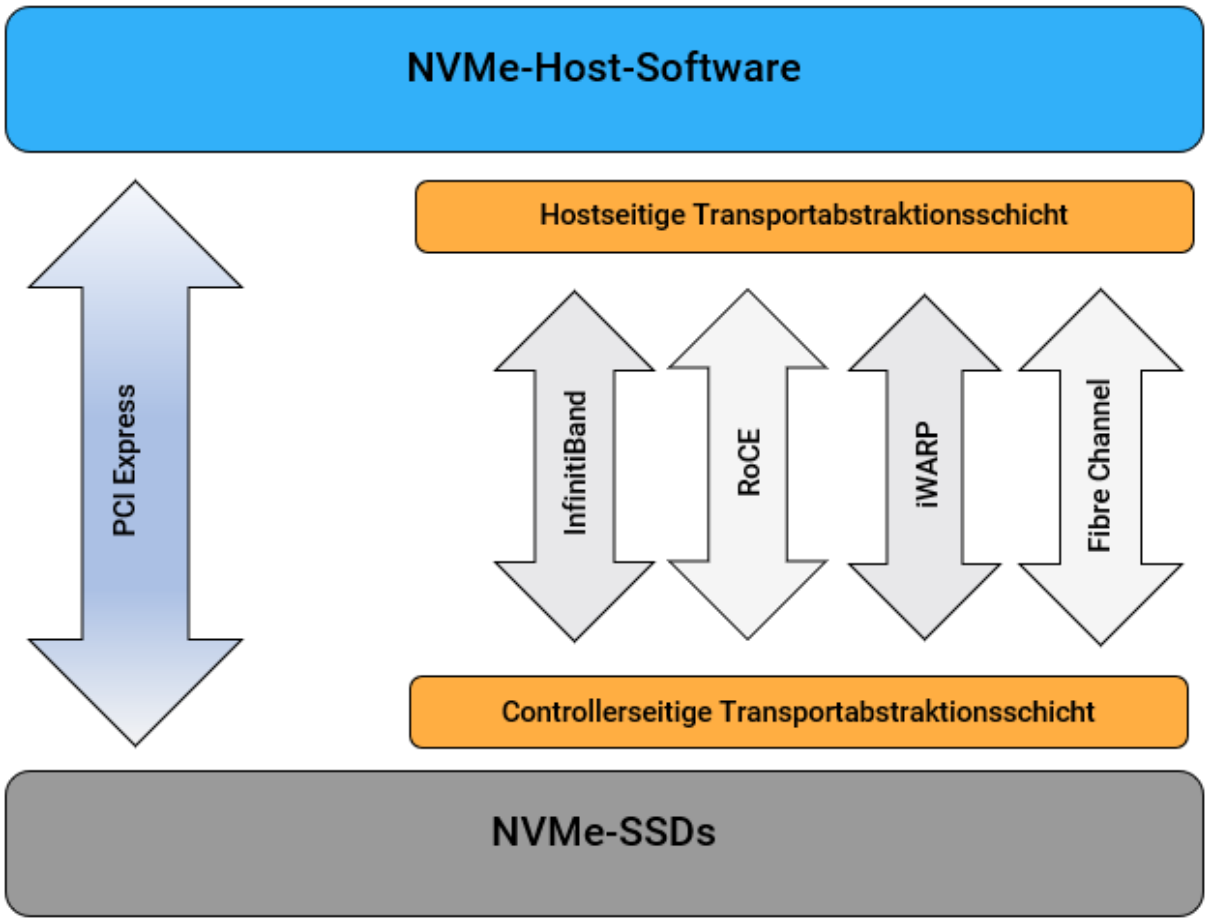
NVMe-oF stellt allgemein sicher, dass NVMe auch in bereits existierenden Fibre-Channel-Netzwerken funktioniert.

Weitere NVMe-oF-Varianten

Der NVMe-oF-Standard wurde für verschiedene Zwecke und Szenarien weiterentwickelt, so gibt es:

- NVMe over TCP:
  Erweiterung auf **TCP/IP**-basierte (**T**ransmission **C**ontrol **P**rotocol/**I**nternet **P**rotocol) Netzwerke

- NVMe over **RDMA** (**R**emote **D**irect **M**emory **A**ccess):
  Erweiterung auf verschiedene Netzwerke/Netzwerkprotokolle ohne Notwendigkeit eines Storage-Controller-Layers. NVMe over RDMA beinhaltet die Untervarianten **RoCE** (**R**DMA **o**ver **C**onverged **E**thernet), **iWAR**P (**I**nternet **W**ide **A**rea **R**DMA **P**rotocol) und InfiniBand.



NVMe-Schema (eigene Darstellung)

Auf dem Weg zu Software-Defined: All Flash und JBoF mit NVMe-oF

Indem das NVMe-oF-Konzept auf das gesamte Unternehmen oder Rechenzentrum angewendet wird, ergeben sich entsprechend erweiterte Möglichkeiten:
Durch die Zusammenfassung der vorhandenen flashbasierten Laufwerke zu einem „Pool“ (Virtualisierung) wird der physische Speicher vom Hardware-Controller logisch abgetrennt. Diese Abstrahierung und virtuelle Vereinigung mehrerer Laufwerke ergibt ein **JBoF** („**J**ust a **B**unch **o**f **F**lash-drives“), das eine bessere Performance als SATA/SAS erzielt. Die wesentliche Innovation dabei ist jedoch, dass dank höherer Geschwindigkeit und größerer Bandbreite im Netzwerk ein gemeinsam genutzter Netzwerkspeicher entsteht, d. h. Anwendungen können statt auf der Host-Hardware bzw. Host-Software nun auf diesem zentralen Network Storage laufen. Dies und der Umstand, dass Zugriffe auf NVMe-Controller, also auf netzwerkintegrierte Laufwerke, möglich sind, so als ob es lokale Laufwerke wären, bringen Netzwerk und Data Center näher zusammen. NVMe-oF entspricht also ganz dem „Software-Defined“-Trend, dessen Ziel es ist, nur noch mit virtualisierten, d. h.

ausschließlich als Code ausgeführten Lösungen, anstelle von hardwarebasierten Konzepten zu arbeiten.

Seit einiger Zeit gibt es Hersteller, die mit (All-)Flash-Arrays (AFA) reine SSD- bzw. NVMe-Systeme anbieten.

Willi Meutzner
Technischer Redakteur
Serverhero GmbH Köln

SERVERHERO