Nonvolatile Memory Express

درس: مدارهای واسط

استاد مربوطه: مهندس امین فصحتی

ارائه دهنده: شمیم رحیمی

زمستان ۱۴۰۳

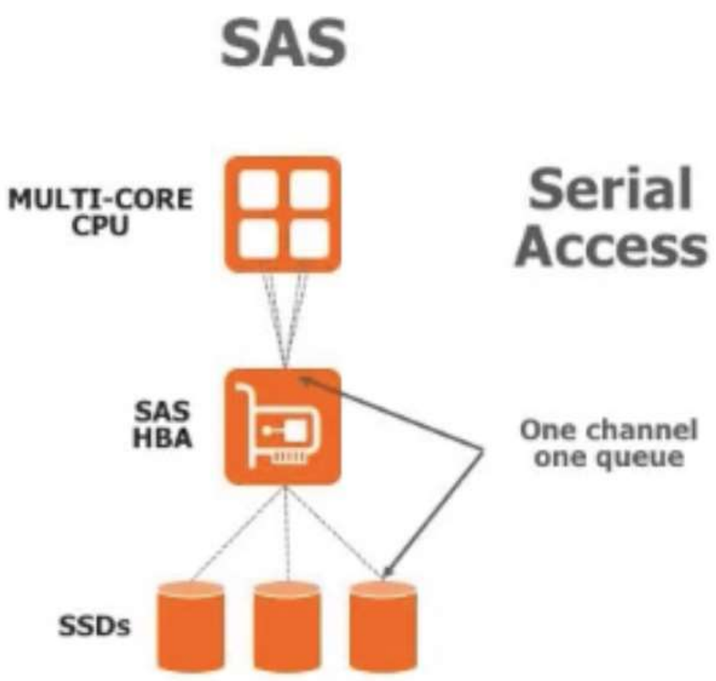
فهرست مطالب

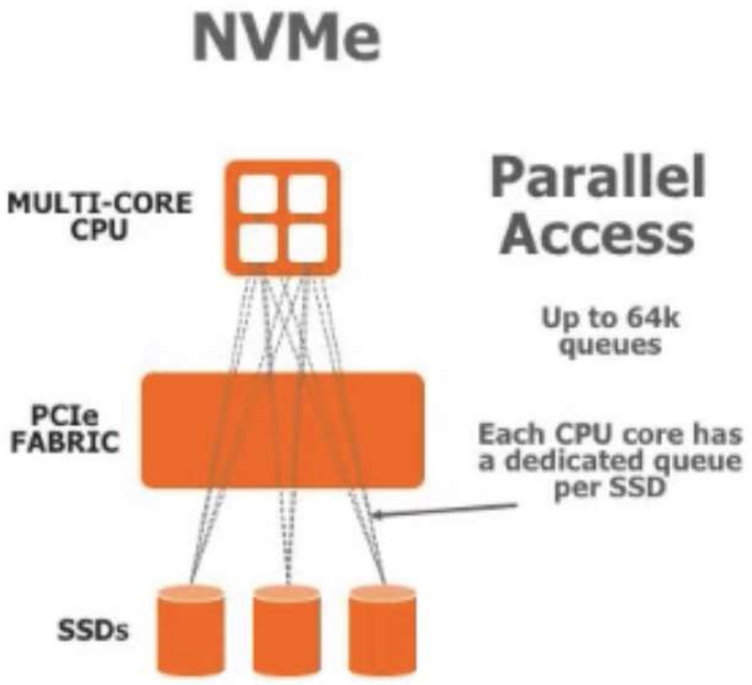
کاربرد و چرایی توسعه NVMe

- یک پروتکل ذخیره‌سازی پرسرعت است.
- به‌صورت خاص برای حافظه‌های SSD طراحی شده است.
- از رابط PCIe برای انتقال داده استفاده می‌کند.

- کاربردهای NVMe:
  - دیتاسنترها و سرورها
  - سیستم‌های گیمینگ و ویرایش ویدئو
  - هوش مصنوعی و یادگیری ماشین
- چرایی توسعه:
  - محدودیت‌های شدید پروتکل‌های قدیمی مانند SATA و AHCI که برای هارد دیسک‌های مکانیکی (HDD) طراحی شده بودند.
  - پهنای باند محدود SATA
  - تأخیر پایین‌تر و بهره‌وری انرژی بالاتر

Queue Depth and commands		
Protocol	Queue Depth	Commands
SAS	1	254
SATA (AHCI)	1	32
NVMe	65,000	65,000

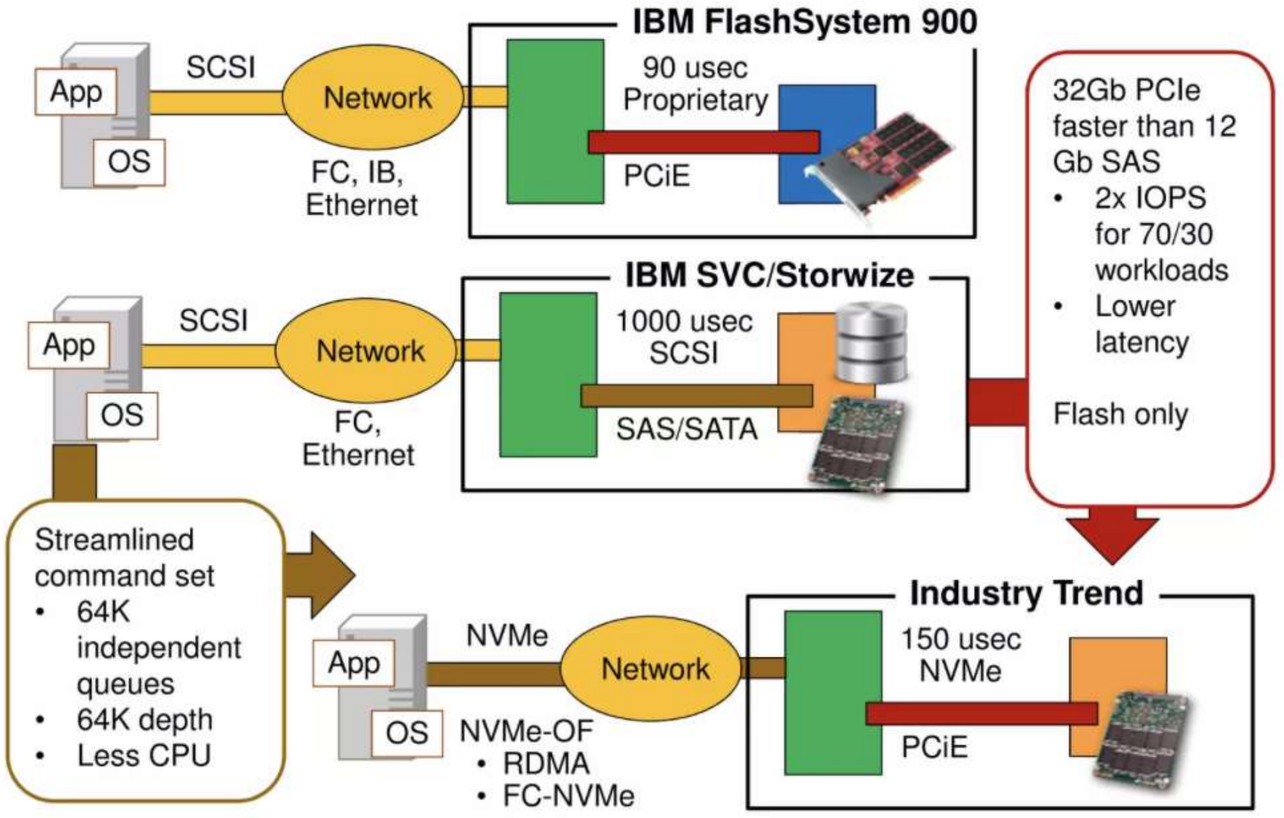

SAS
MULTI-CORE CPU
Serial Access
SAS HBA
One channel one queue
SSDs


NVMe
MULTI-CORE CPU
Parallel Access
Up to 64k queues
PCIe FABRIC
Each CPU core has a dedicated queue per SSD
SSDs

لایه‌ی فیزیکی، اتصالات و مدارات

- لایه فیزیکی و سیگنالینگ
  - NVMe از PCIe برای انتقال داده‌ها استفاده می‌کند.
  - لایه فیزیکی PCIe شامل دو سیم برای هر سیگنال است (یک سیم داده مثبت و یک سیم داده منفی).
  - این روش باعث:
    - کاهش نویز الکترومغناطیسی
    - افزایش سرعت انتقال داده
    - کاهش تداخل بین سیگنال‌ها

- اتصالات ضروری و اختیاری در NVMe
  - اتصالات ضروری:
    - PCIe Lane(s): برای انتقال داده
    - Clock Signal: برای همگام‌سازی داده‌ها
    - Power Supply: برای تأمین برق دستگاه
  - اتصالات اختیاری:
    - GPIOs



نوع ارتباط و انکودینگ

- ارتباط NVMe سریال است.

- NVMe از انکودینگ ۱۲۸b/۱۳۰b استفاده می‌کند که باعث کاهش Overhead و افزایش کارایی نسبت به روش‌های قدیمی مانند ۸b/۱۰b در PCIe ۲.۰ می‌شود.

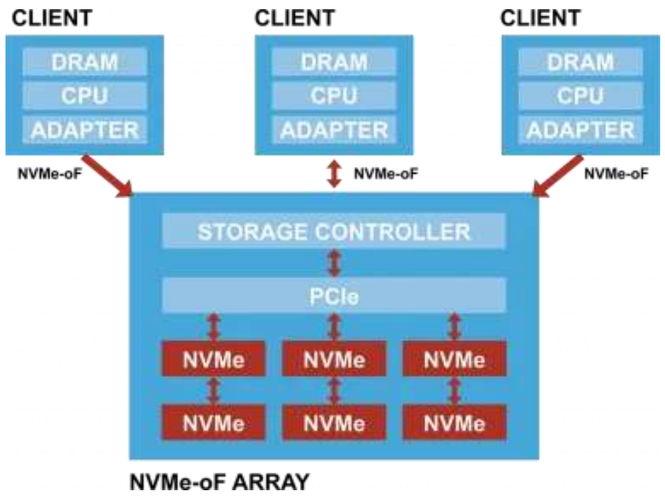
- روش انتقال: هم‌زمان یا ناهم‌زمان؟
- NVMe از انتقال ناهم‌زمان استفاده می‌کند.
- در روش هم‌زمان، پردازنده باید منتظر تکمیل یک درخواست بماند.
- در روش ناهم‌زمان، پردازنده می‌تواند چندین درخواست را به‌طور هم‌زمان پردازش کند.
- NVMe با صف‌های چندگانه و پردازش موازی باعث کاهش تأخیر و افزایش سرعت انتقال داده می‌شود.

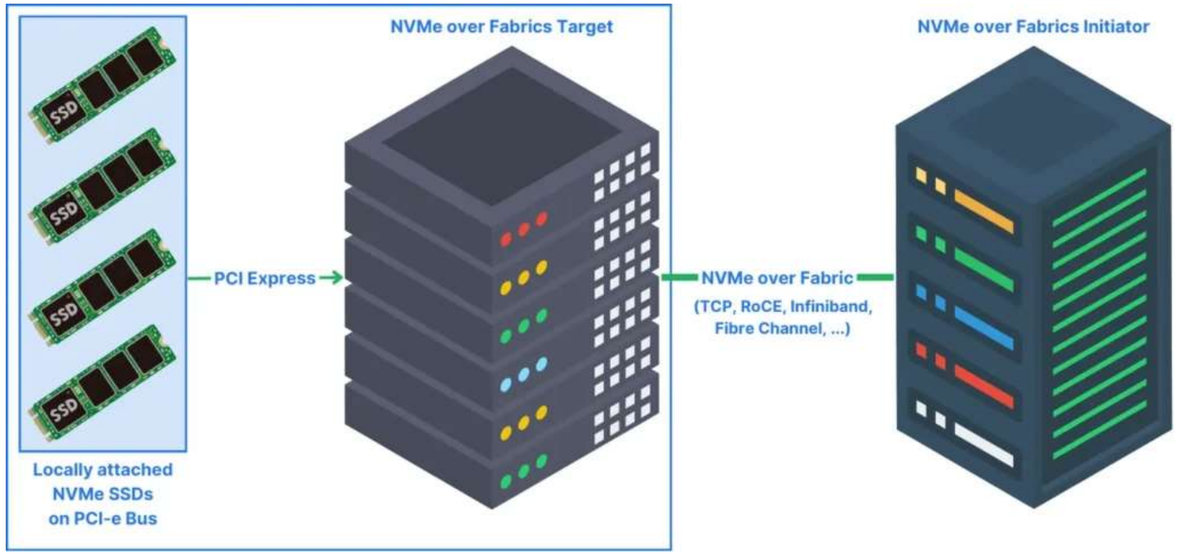
اتصالات بین چندین دستگاه

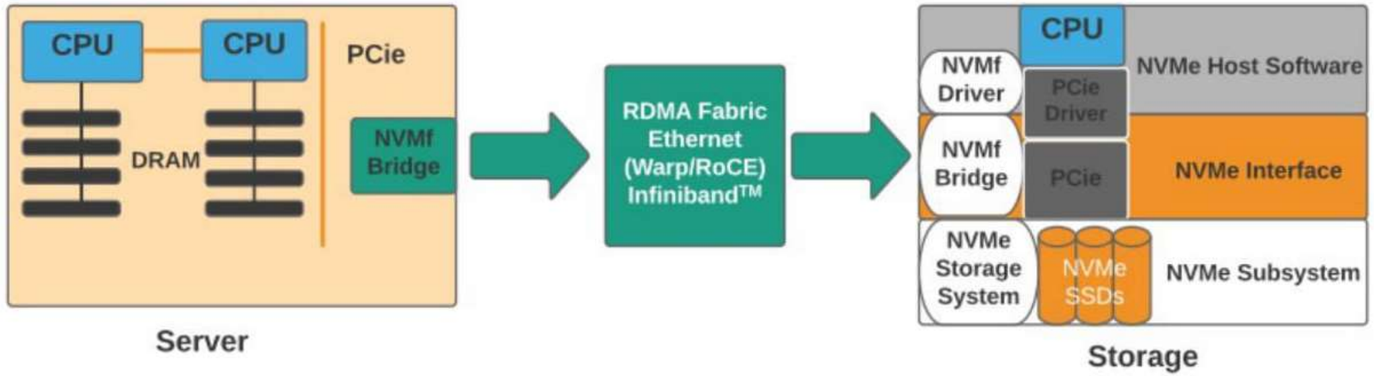
- NVMe از چندین دستگاه پشتیبانی می‌کند و می‌تواند از طریق PCIe Bus چندین دستگاه را به یکدیگر متصل کند.

- روش اتصال چندین دستگاه در NVMe
  - استفاده از چندین اسلات M.2 یا U.2 در مادربردهای مدرن.
  - NVMe over Fabrics (NVMe-oF): امکان اتصال SSDها به چندین سرور از طریق شبکه.
  - PCIe Switches: امکان استفاده از چندین SSD بر روی یک گذرگاه.

- چرا بعضی از SSDهای NVMe فقط یک دستگاه را پشتیبانی می‌کنند؟
  - PCIe به‌طور پیش‌فرض برای ارتباط نقطه‌به‌نقطه طراحی شده است.
  - برخی از SSDها از Multiplexing پشتیبانی نمی‌کنند.




NVMe over Fabrics Target
NVMe over Fabrics Initiator
SSD
SSD
SSD
SSD
PCI Express
NVMe over Fabric
(TCP, RoCE, Infiniband,
Fibre Channel, ...)
Locally attached
NVMe SSDs
on PCI-e Bus


CPU
CPU
PCie
DRAM
NVMf
Bridge
Server
RDMA Fabric
Ethernet
(Warp/RoCE)
Infiniband™
CPU
NVMf
Driver
PCie
Driver
NVMe Host Software
NVMf
Bridge
PCie
NVMe Interface
NVMe
Storage
System
NVMe
SSDs
NVMe Subsystem
Storage

آدرس‌دهی و مسیریابی

- آدرس‌دهی در NVMe روی PCIe
  - آدرس‌دهی از طریق Bus Number، Device Number و Function Number در PCIe انجام می‌شود.
  - هر SSD NVMe می‌تواند چندین Namespace (NSID) داشته باشد که فضای ذخیره‌سازی را جدا می‌کند.

- آدرس‌دهی در NVMe-oF (NVMe over Fabrics)
  - در NVMe-oF که روی شبکه‌هایی مثل Ethernet و TCP/IP اجرا می‌شود، آدرس‌دهی پیچیده‌تر است.
  - دستگاه‌ها از Qualified Name (QN) و Transport Address برای شناسایی استفاده می‌کنند.
  - مثال: در NVMe over TCP یک دستگاه NVMe می‌تواند آدرسی مانند ۱۹۲.۱۶۸.۱.۱۰:۴۴۲۰ داشته باشد.

- چرا NVMe روی PCIe به مسیریابی نیاز ندارد؟
  - چون PCIe به‌صورت نقطه‌به‌نقطه است و داده‌ها مستقیماً بین کنترلر و SSD NVMe منتقل می‌شوند، بدون نیاز به مسیرهای پیچیده.

مدیریت جریان داده

- NVMe از روش‌های مختلفی برای مدیریت جریان داده استفاده می‌کند تا عملکرد بهینه و تأخیر کم داشته باشد.

۱. صف‌های چندگانه (Multiple Queues)
- هر NVMe دستگاهی چندین صف ارسال (SQ) و دریافت (CQ) دارد که به پردازش چندین درخواست هم‌زمان کمک می‌کند.
- برخلاف SATA که فقط یک صف دارد، NVMe می‌تواند تا ۶۵۵۳۶ صف موازی ایجاد کند.

۲. مدیریت جریان داده با محدودیت‌های QoS
- برخی SSDهای NVMe قابلیت Rate Limiting دارند که محدودیت سرعت خواندن/نوشتن را اعمال می‌کند.
- Priority Classes برای اولویت‌بندی داده‌ها استفاده می‌شوند.

۳. جریان داده در NVMe-oF
- در NVMe over Fabrics، بسته‌های داده با TCP یا RDMA مدیریت می‌شوند.
- Credit-Based Flow Control از ترافیک بیش از حد جلوگیری می‌کند.

تشخیص و تصحیح خطا

- NVMe در لایه‌های مختلف از روش‌های تشخیص خطا استفاده می‌کند.

۱. لایه فیزیکی
- PCIe CRC برای تشخیص خطاهای انتقال داده استفاده می‌شود.
- در NVMe-oF، پروتکل‌های شبکه‌ای مانند TCP Checksum یا RDMA CRC خطاها را تشخیص می‌دهند.

۲. لایه داده
- End-to-End Data Protection (DIF/DIX)

۳. لایه فرمان (Command Layer)
- Status Code Fields در پاسخ‌ها، اطلاعات خطا را گزارش می‌دهند.

- NVMe نه‌تنها خطاها را تشخیص می‌دهد، بلکه از روش‌هایی برای تصحیح آن‌ها نیز استفاده می‌کند.

۱. تصحیح خطا در حافظه SSD
- از LDPC (Low-Density Parity-Check Code) برای تصحیح خطای سلول‌های حافظه استفاده می‌شود.

۲. تصحیح خطا در انتقال داده
- در PCIe، Replay Buffers در صورت خرابی داده‌ها، آن‌ها را مجدداً ارسال می‌کنند.
- در NVMe-oF، Retransmission Mechanisms در TCP برای ارسال مجدد داده‌های خراب‌شده به کار می‌رود.

۳. تصحیح خطا در فرمان‌ها
- Error Recovery Mechanisms در سطح نرم‌افزار، فرمان‌های نامعتبر را شناسایی و اصلاح می‌کنند.

انواع پیام

- پیام‌های NVMe به سه دسته اصلی تقسیم می‌شوند:

۱. دستورات (Commands): توسط میزبان Host به کنترلر ارسال می‌شوند و شامل موارد زیر هستند:
- دستورات خواندن: برای خواندن داده از SSD
- دستورات نوشتن: برای نوشتن داده روی SSD
  - شامل آدرس منطقی و داده ارسالی است.
- دستورات مدیریتی: برای مدیریت دیسک، بررسی سلامت، و تغییر تنظیمات
  - شامل Identify، Format NVM و Firmware Update است.

فیلد	توضیحات
Opcode	نوع دستور (Read, Write, Identify)
Namespace ID (NSID)	شماره namespace مورد استفاده
Logical Block Address (LBA)	آدرس بلوک مورد نظر برای خواندن/نوشتن
Data Length	مقدار داده انتقالی
Metadata	اطلاعات جانبی برای محافظت از داده‌ها
Command Identifier	شماره یکتای دستور

انواع پیام

- پیام‌های NVMe به سه دسته اصلی تقسیم می‌شوند:

۲. پاسخ‌ها (Responses)

- این پیام‌ها توسط کنترلر NVMe ارسال می‌شوند و شامل اطلاعات زیر هستند:
  - Status Code: مشخص می‌کند که دستور موفق بوده یا خطایی رخ داده است.
  - Data Buffer: شامل داده‌های خوانده‌شده در پاسخ به دستورات Read.

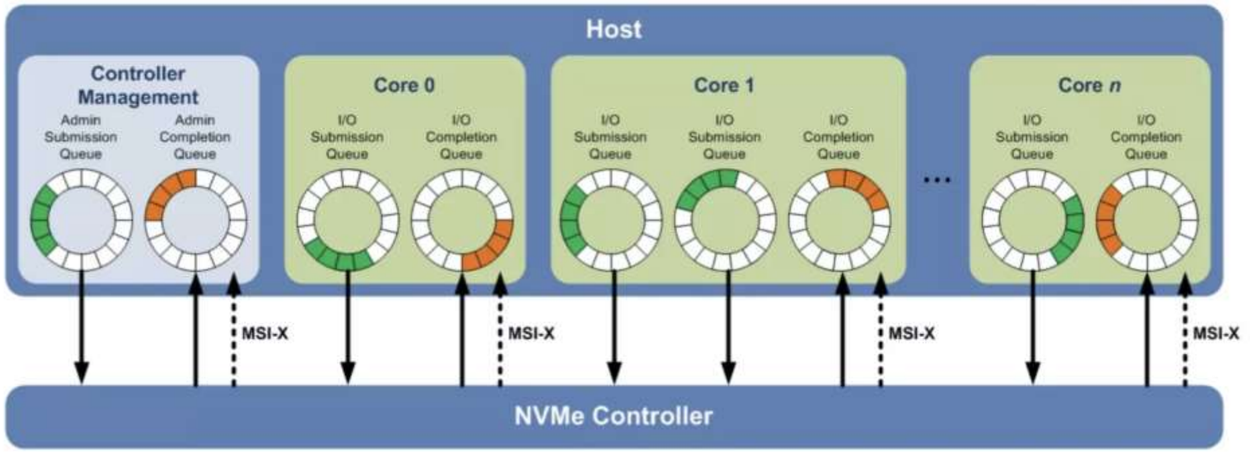
فیلد	توضیحات
Status Code	وضعیت اجرای دستور (موفق/ناموفق)
Data Buffer	در پاسخ به خواندن، شامل داده‌های خوانده‌شده
Command Identifier	شناسه دستور مربوطه
Completion Queue Entry (CQE)	نشانه پایان پردازش دستور

انواع پیام

- پیام‌های NVMe به سه دسته اصلی تقسیم می‌شوند:

۳. پیام‌های خاص (Special Messages):

- این پیام‌ها برای هماهنگی بین کنترلر و Host به کار می‌روند:
  - Completion Queue Entries (CQEs)
  - Asynchronous Event Notifications (AENs)



جمع‌بندی

- پروتکل NVMe یک تحول اساسی در دنیای ذخیره‌سازی است.
- هدف استفاده بهینه از حافظه‌های فلش و SSD است.
- این پروتکل، جایگزین رابط‌های سنتی مانند SATA و SAS شده و از گذرگاه PCIe برای ارتباط مستقیم با CPU استفاده می‌کند، که منجر به کاهش تأخیر و افزایش پهنای باند می‌شود.

- از نظر لایه فیزیکی، NVMe روی PCIe اجرا می‌شود.
- این پروتکل به دلیل ساختار سریال و ناهمزمان خود، عملکرد بهتری نسبت به گذرگاه‌های موازی قدیمی دارد.

- نسخه استاندارد NVMe برای اتصال مستقیم طراحی شده است.
- نسخه NVMe over Fabrics (NVMe-oF) امکان اتصال چندین دستگاه از طریق شبکه‌های ذخیره‌سازی را فراهم می‌کند.

- NVMe با استفاده از مکانیسم‌های پیشرفته صف‌بندی و اولویت‌بندی درخواست‌ها، پهنای باند را بهینه‌سازی می‌کند.
- این پروتکل دارای مکانیزم‌های تشخیص و تصحیح خطا در سطوح مختلف است.

- در مجموع، NVMe یک پیشرفت کلیدی در دنیای ذخیره‌سازی محسوب می‌شود که با افزایش سرعت، کاهش تأخیر و بهینه‌سازی پردازش موازی، نیازهای ذخیره‌سازی مدرن را برآورده می‌کند.

منابع

- SlideShare, "NVMe Overview," Available: https://www.slideshare.net/slideshow/nvme-overview/۸۸۲۴۹۳۸۷?from_search=۰
- SlideShare, "NVMe Revolution," Available: https://www.slideshare.net/slideshow/s۱۰۴۸۷۸-nvmerevolutionjburgv۱۸۰۹b/۱۱۵۶۶۲۸۶۵?from_search=۱
- StorageReview, "NVMe & NVMe-oF Background & Overview," Available: https://www.storagereview.com/review/nvme-nvme-of-background-overview
- NetApp, "What is NVMe?" Available: https://www.netapp.com/data-storage/nvme/what-is-nvme/#:~:text=NVMe٪۲۰(nonvolatile٪۲۰memory٪۲۰express)٪۲۰is,all٪۲۰types٪۲۰of٪۲۰enterprise٪۲۰workloads
- Wikipedia, "NVM Express," Available: https://en.wikipedia.org/wiki/NVM_Express