

Лучше с NVMe™ OpenFlex – сетевые хранилища NVMeoF

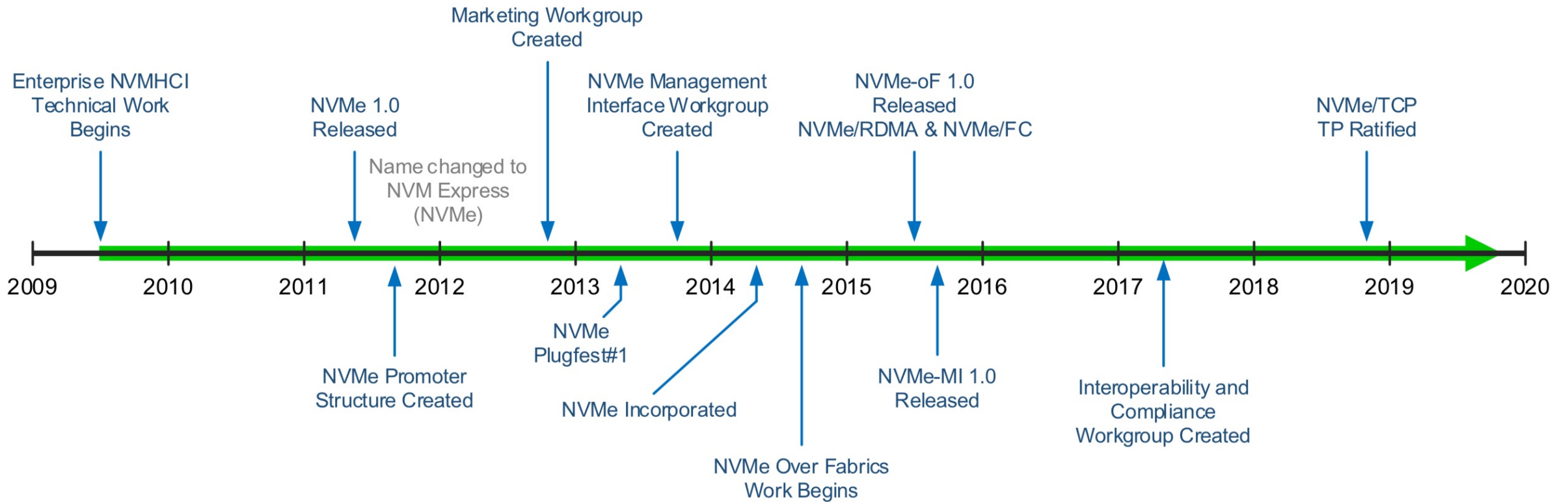
Григорий Никонов (grigory.nikonov@wdc.com)

Системный инженер

Сентябрь 2021

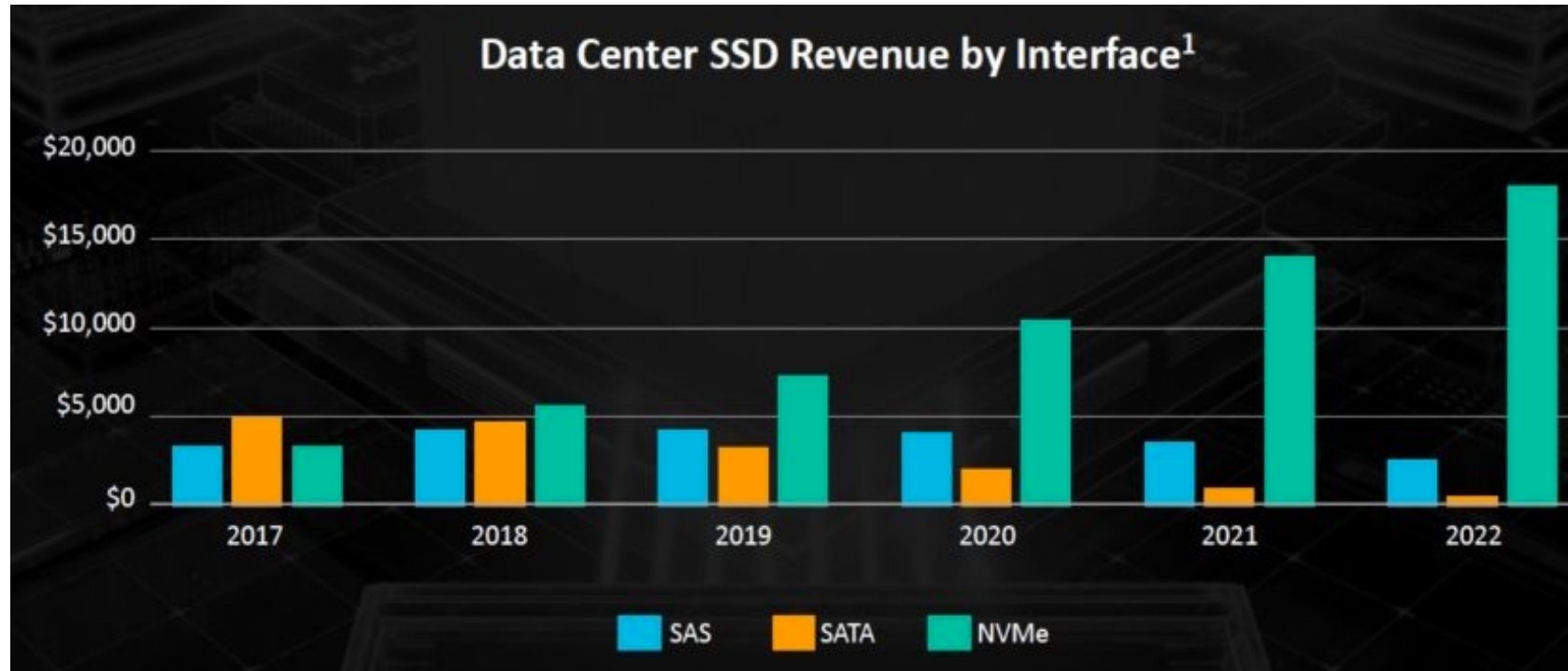
NVMe идет в 5 класс

Ten Years of NVMe



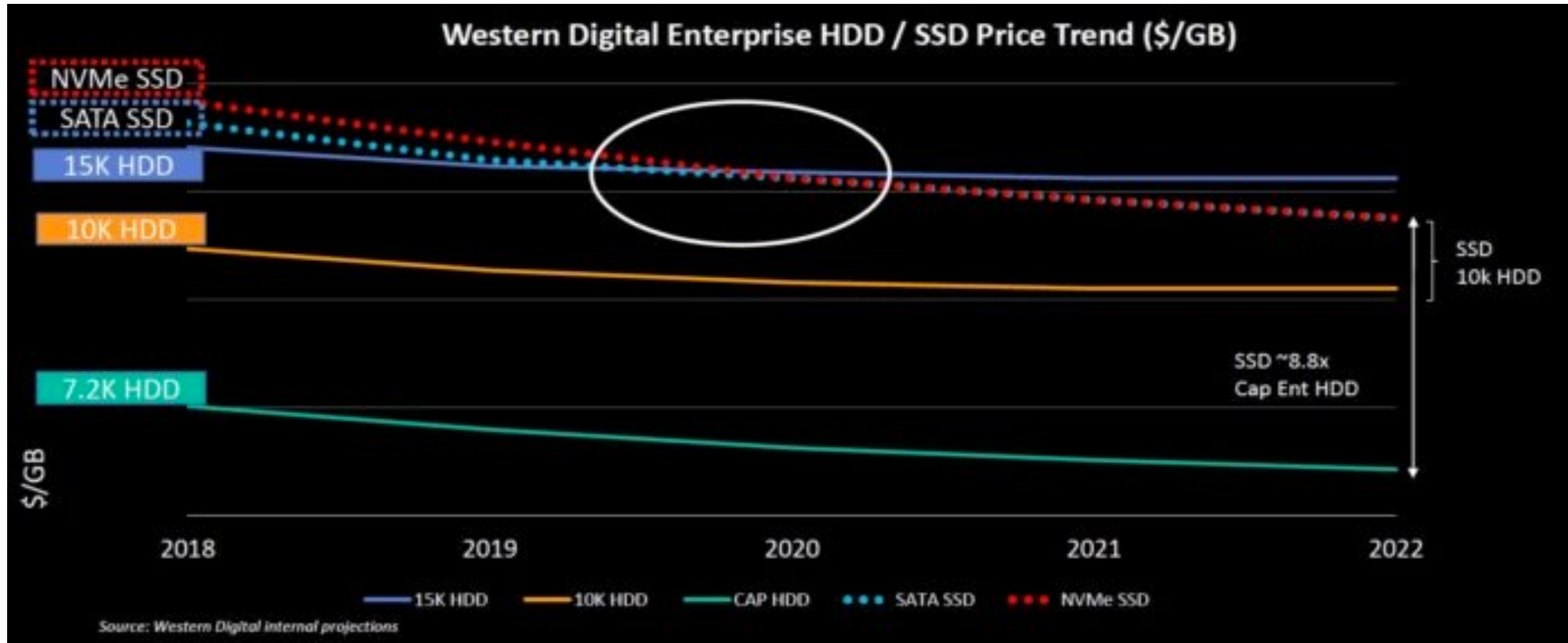
<https://www.snia.org/educational-library/nvme-state-union-2019>

... и становится мейнстримом



K Units	2016	2017	2018	2019*	2020*	2021*
Enterprise	364	749	1,048	2,774	5,740	11,192
Cloud	2,051	3,861	10,231	17,338	25,891	31,050
Client	33,128	50,385	82,613	111,888	187,689	243,889

Ценовой паритет с SATA eSSD

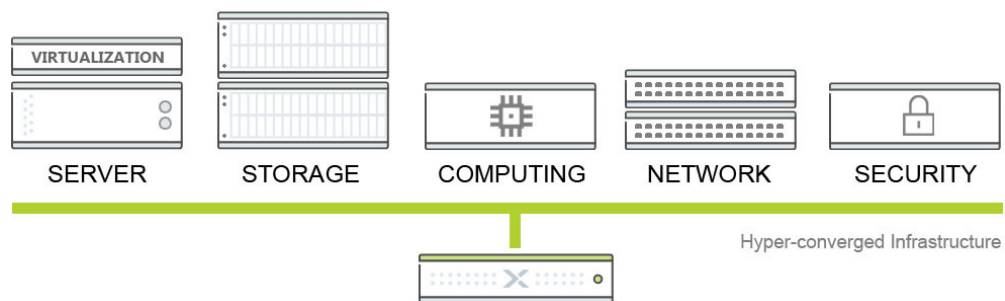


Гонка за эффективностью

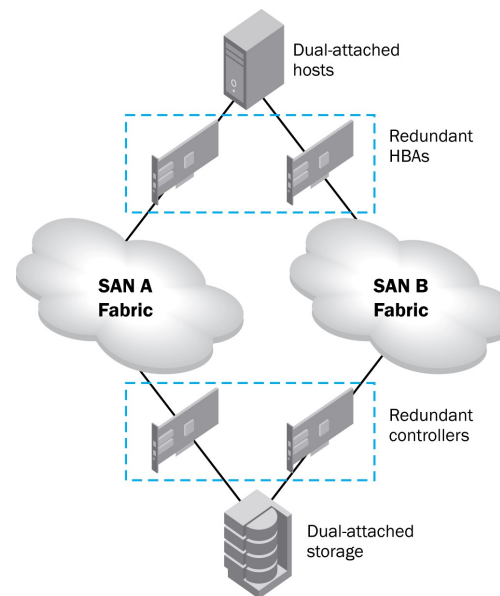
Мейнфрейм / DAS



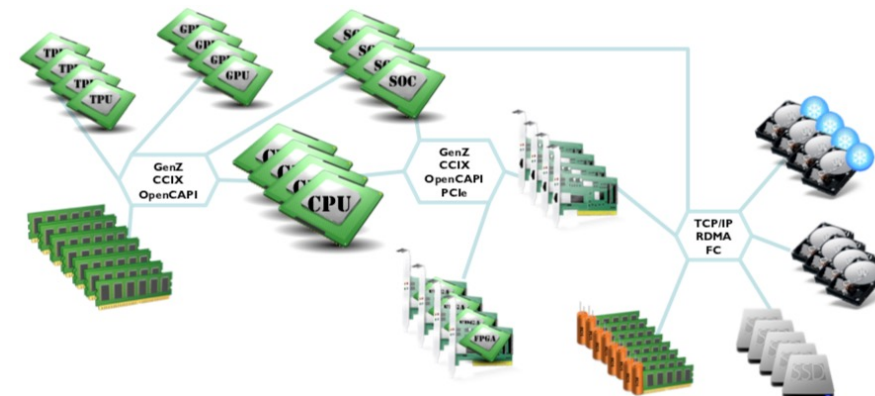
Конвергенция/гиперконвергенция



SAN



Компонуемая архитектура



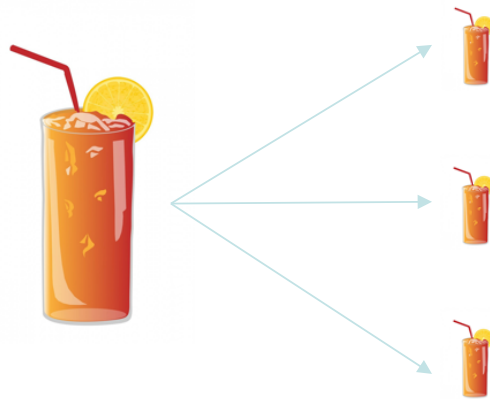
Задача: повысить утилизацию SSD до уровня облачных провайдеров
Решение: нужно вынести NVMe в сеть, избавиться от «темного флэша»

Виртуализация и компонуемость

Virtualization & Containerization vs. Composability

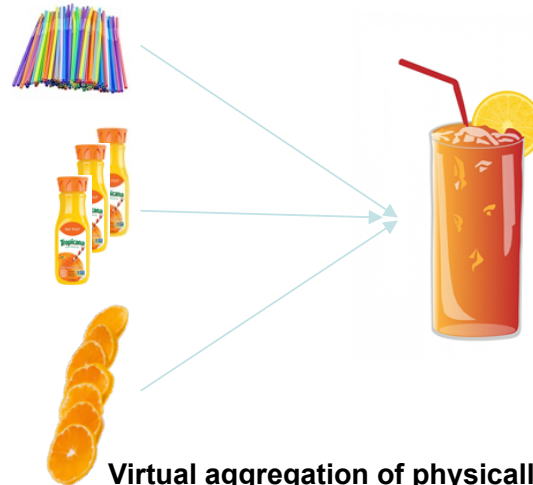
SNIATM | CLOUD STORAGE
CSTI | TECHNOLOGIES

Virtualization/Containerization



Virtual disaggregation of physically aggregated components

Composability



Virtual aggregation of physically disaggregated components

© 2019 Storage Networking Industry Association. All Rights Reserved.

<https://www.snia.org/sites/default/files/CSI/Why-Composable-Infrastructure-Final.pdf>

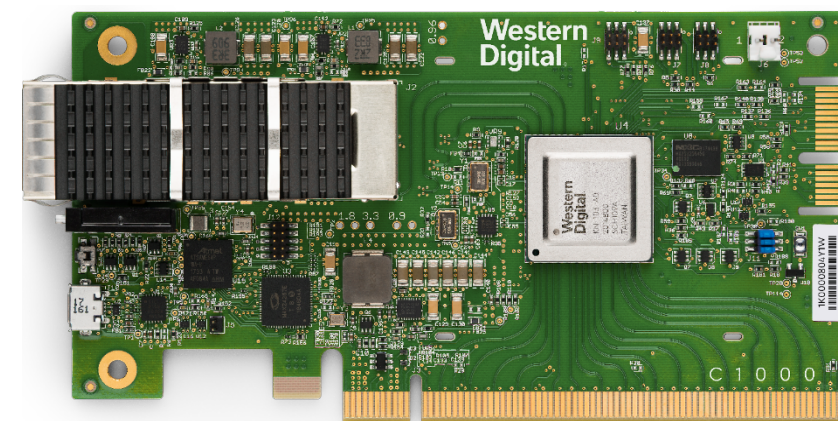
Выбор транспорта – сетевой холивар

- PCIe
 - FC
 - Ethernet – lossy/DCB? RDMA or TCP?
 - Infiniband
-
- Мы выбираем Ethernet – скорость, масштабирование, распространенность
 - Всегда остаются вопросы выделенной сети для трафика СХД, администрирования, безопасности и тд.

RapidFlex™ High-Performance NVMe-oF Controller

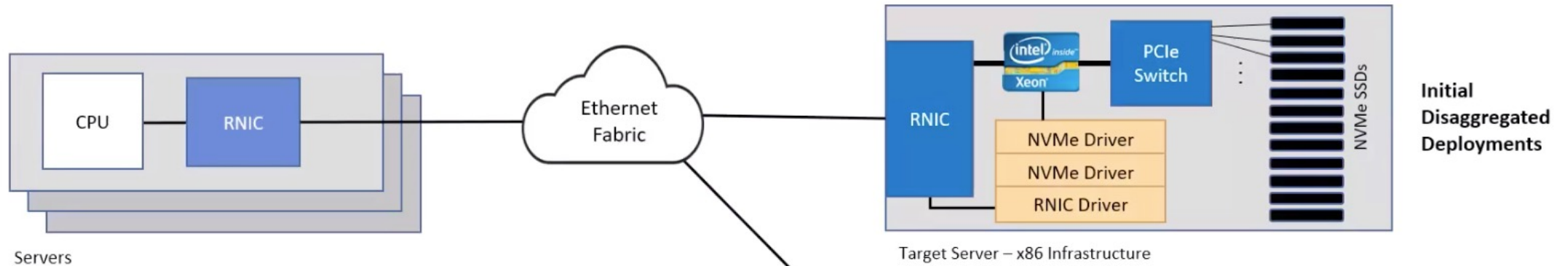
Kazan Networks acquisition accelerates key NVMe-oF building block

- An optimized NVMe-oF target JBOF solution
 - Extremely low power
 - Minimal latency penalty
- Advanced levels of hardware acceleration
- Enables efficient compute/storage disaggregation
 - DAS-like performance with all the benefits of composability
- Available as 21mm BGA packaged target ASIC
Or compact PCIe target adapter card
- 100Gb Ethernet connectivity
- 7W avg, no ARM cores



NVMe-oF JBOF

Enabled via WDC's Bridge



"Just a Bunch of Flash" (JBOF)

Attributes Enabled:

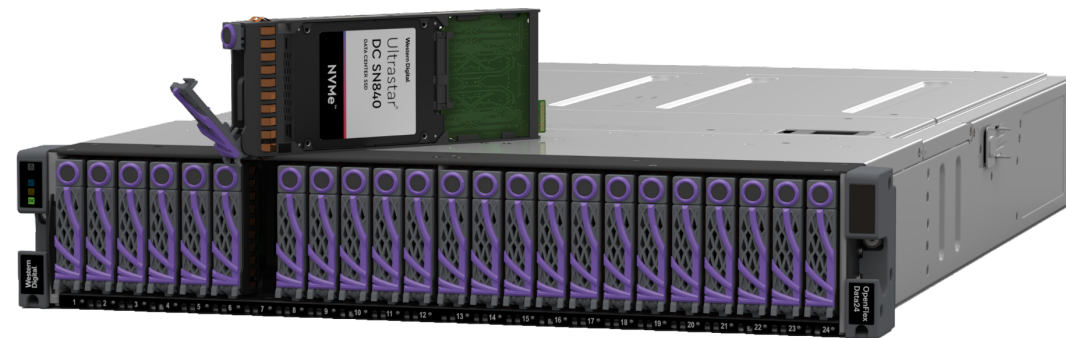
- Pooled, shared SSDs
- No loss in IOPS, BW, or latency performance (vs DAS)
- Can be scaled to virtually any level
- Low cost, low power method for SSD deployment



OpenFlex™ Data24 NVMe-oF JBOF

NVMe™ storage shared over Ethernet fabric

- Enables servers to share NVMe flash storage as if it were local
 - Up to 368TB in 2U 24-bay enclosure
 - Dual Port and Single Port* SSD options
 - Uses low latency fabric to fully utilize IOPs and capacity
 - Provides more efficient use of large capacity SSDs
 - Realizes the performance promise of NVMe
- Provides industry-leading connectivity options
 - Fan out to six connections without a switch
 - Balanced design to eliminate over-subscription and maintain NVMe advantage
 - Architected for extensibility with future enhancements
- Disaggregates NVMe storage for open composability
 - Allows multiple hosts concurrent access at low latency
 - Leverages mature standard NVM Express over Fabrics 1.0a
 - Provides open composability

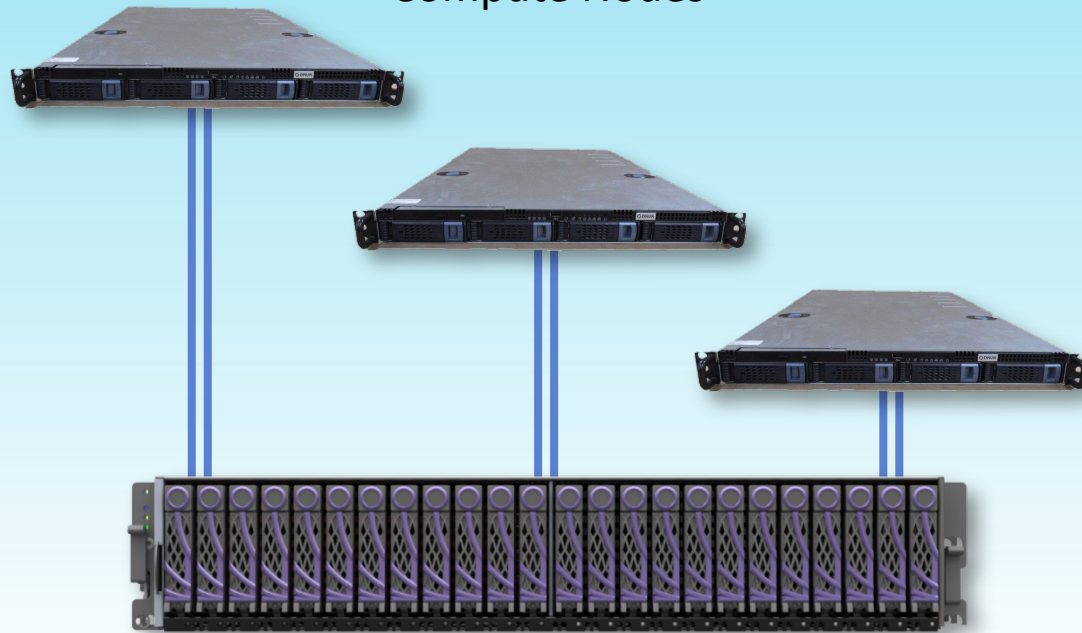


OpenFlex Data24 NVMe-oF Storage Platform

Deployment options

Direct Connect – No Switch Required

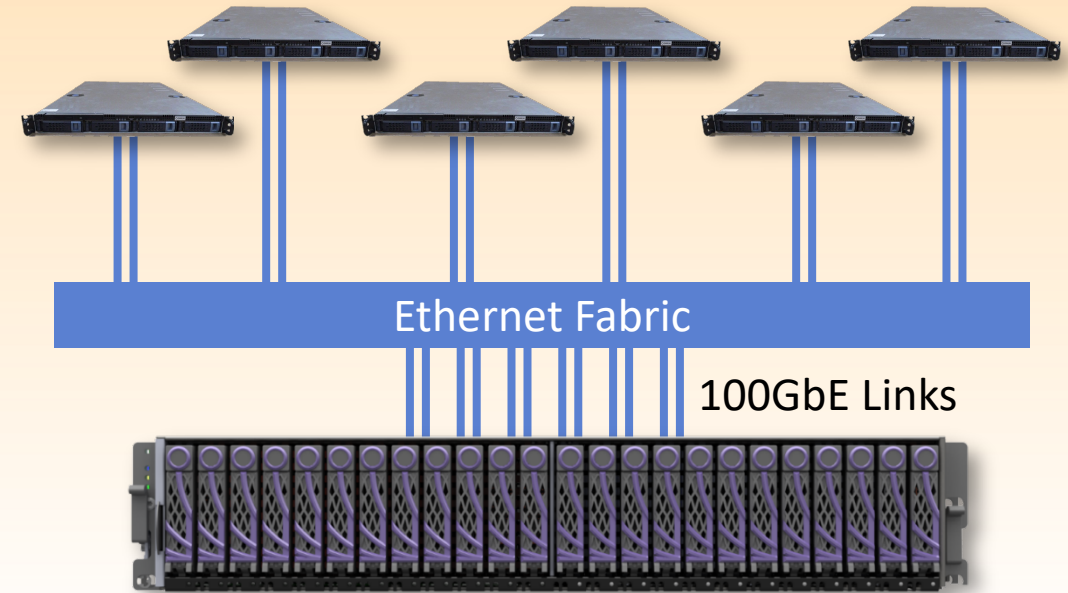
Compute Nodes



OpenFlex Data24 NVMe-oF JBOF

Fabric Connect

Compute Nodes



OpenFlex Data24 NVMe-oF JBOF

== 100GbE Links

Composability / SAS Replacement

Expanded Data24 Options

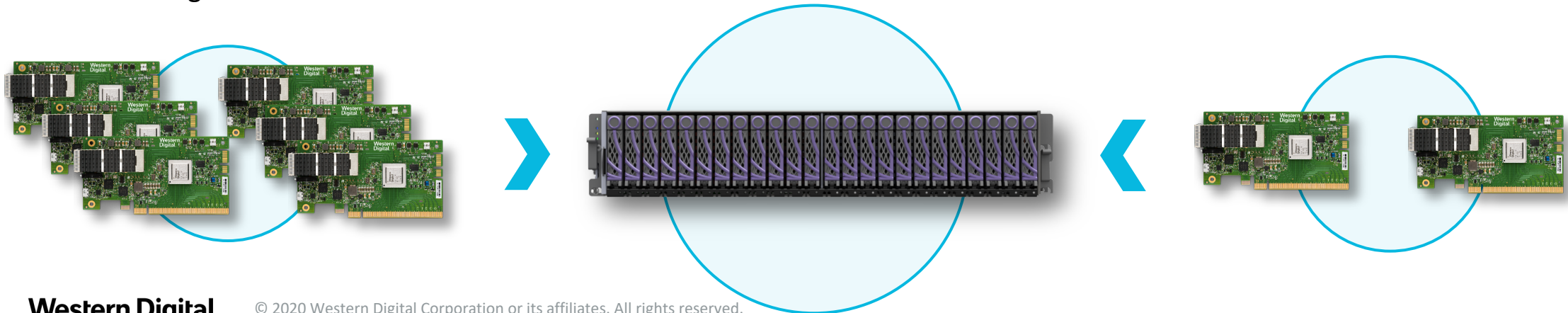
NEW

High Performance

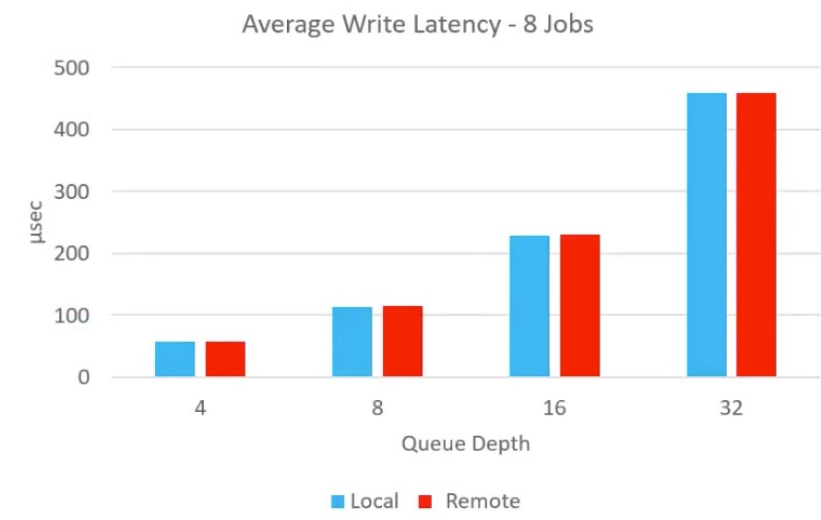
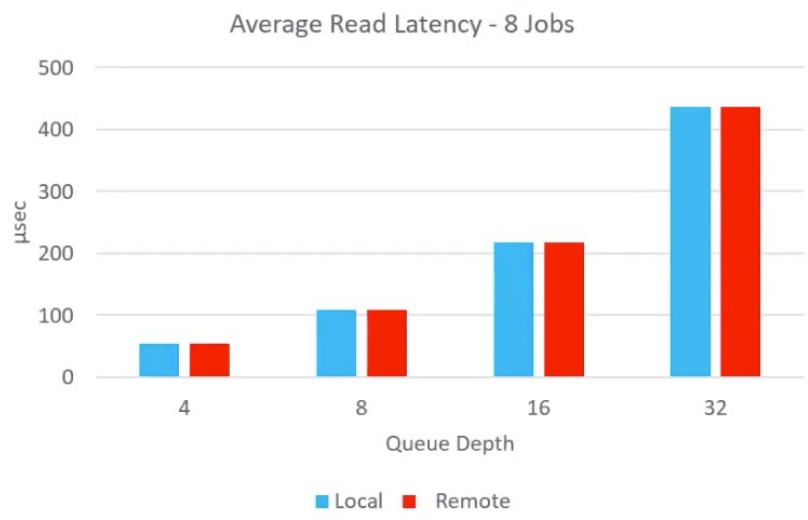
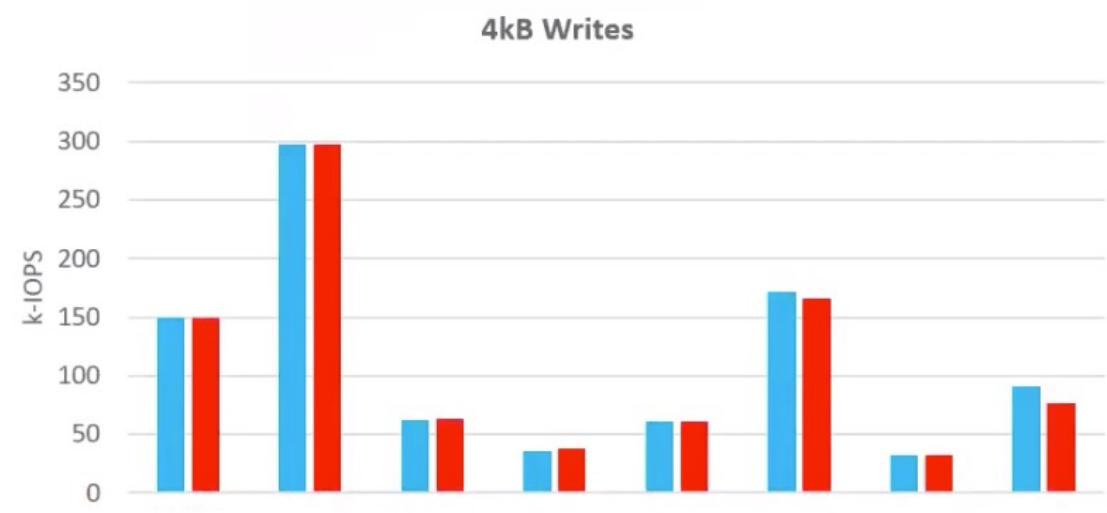
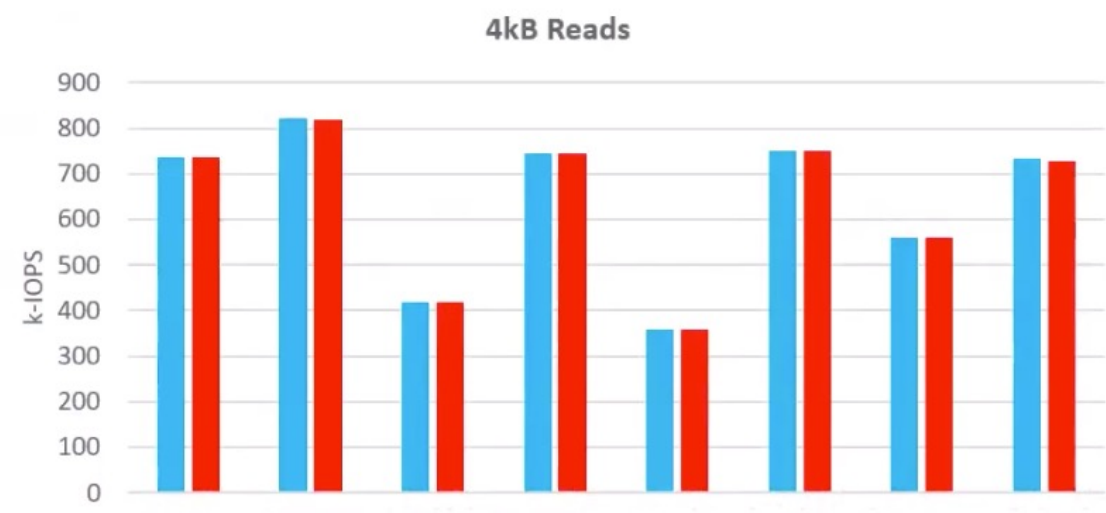
- Highest possible performance
- Matches PCIe lanes to the drives
- Each RapidFlex card sees only 8 drives
- Best fit for:
 - Maximum performance
 - Highest I/O rates
 - A larger environment with more servers

Performance

- Needs only a single PCIe slot per server (like SAS)
- Faster performance than SAS
- Each port can see all 24 drives
- Best fit for:
 - SAS JBOF replacement
 - Cost sensitive environments
 - Servers that cannot accommodate additional NICs



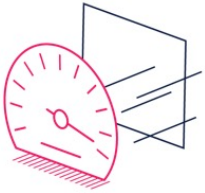
Производительность NVMeOF vs Local



RAIDIX ERA

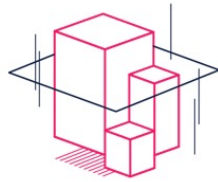
Software RAID for flash

Product Benefits



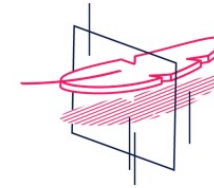
High Performance

- 55.8GB/s and 10 400 000 IOps
- Latency less 100 us
- No performance penalty in mixed I/O workloads
- Low performance penalty in degraded mode



Usability and compatibility

- Wide range of RAID levels (including RAID 7.3)
- Works with local and remote drives
- Clients get array as a block device
- POSIX API support: no needs to adjust apps and file systems



Lightweight

- Less 20% CPU usage under maximum array loading
- Minimum RAM utilization: less 4 GiB for full-featured work

Производительность ERA + SN640

```
root@localhost:~/raidix — ssh root@192.168.101.100 — 80x22
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]#
[root@localhost raidix]# fio fio.cfg
dimec: (g=0): rw=randread, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T) 4096B-4096B,
ioengine=io_uring, iodepth=64
...
fio-3.20-66-g797e
Starting 144 processes
Jobs: 144 (f=144): [r(144)][1.3%][r=38.2GiB/s][r=10.0M IOPS][eta 04m:56s]
```

- Great performance 4x competition
- Easy to configure and maintain
- Good feature set
- Free license for up to 8 WD drives
- Hardware agnostic
- No extra hardware
- Works with most Linux systems

WDC + RAIDIX promo coming soon



Western Digital[®]