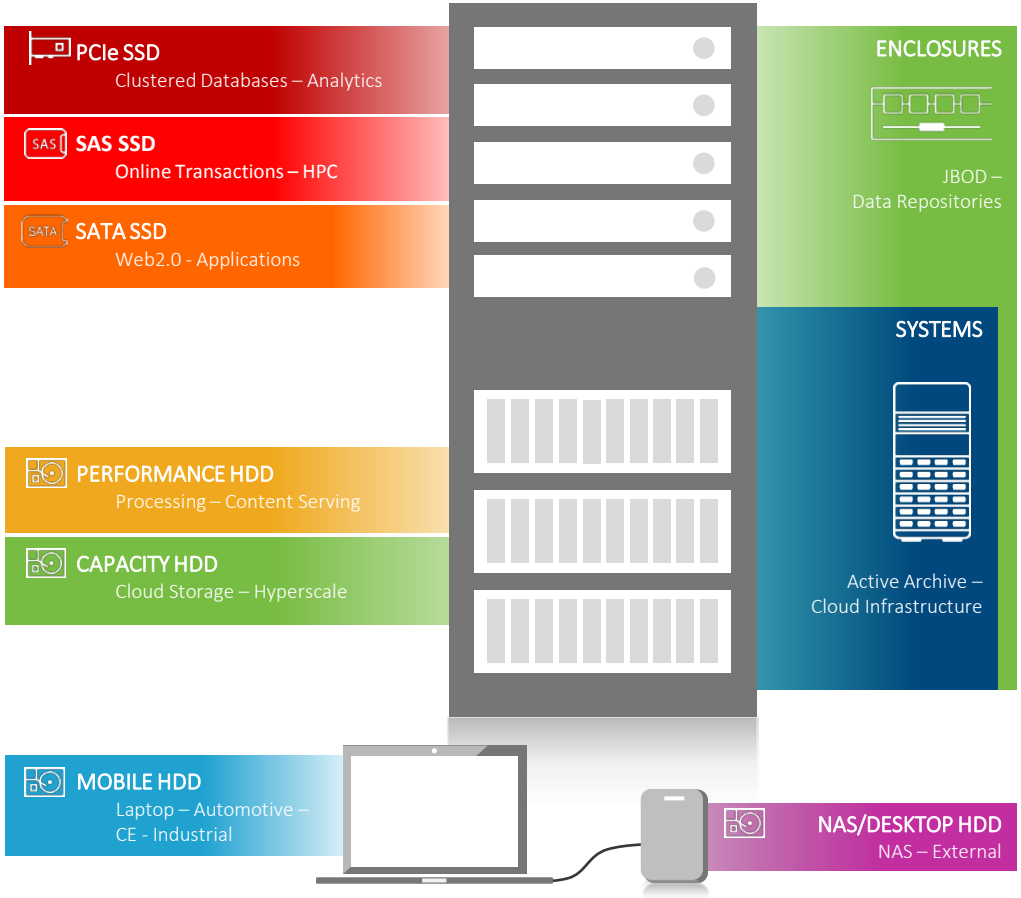
Технологии HGST для ЦОД NVMe SSD

Григорий Никонов, системный инженер

Сентябрь 2016



В 1956 году мы придумали первый в мире жесткий диск

На этом мы не остановились

Сегодня HGST предлагает инновации во всех сегментах продуктов для хранения данных – от супер-быстрых твердотельных накопителей до архивных систем с максимальной плотностью размещения в мире.

История SATA и SAS

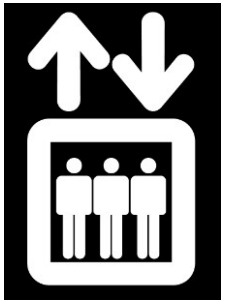
- 1984/1994 IDE -> ATA-1 (Master/Slave)
- 1996 ATA-2 (WD EIDE + STX FastATA, LBA, Block transfer)
- 1997 ATA-3 (SMART)
- 1998 ATAPI (SCSI over ATA - CD, DVD, tape)
- 2003 SATA (AHCI, NCQ)

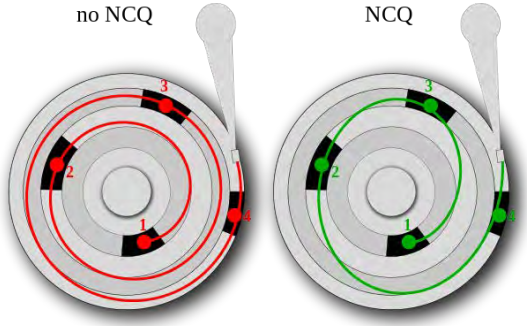
- 1978/1981 SCSI-1
- 1994 SCSI-2 (Fast)
- 2003 SCSI-3 (Ultra)
- 2005 SAS-1 (3 Gbps)
- 2009 SAS-2 (6 Gbps)
- 2013 SAS-3 (12 Gbps)
- 2017 SAS-4 (24 Gbps) ?

Hardware	SATA	SAS	PCIe
Software	AHCI	SCSI	NVMe

Очередь команд

- Принцип лифта
- 32 команды в очереди - AHCI vs NCQ
- 1 прерывание на команду / группу команд (interrupt aggregation)
- Sync I/O vs Async I/O
- Queue depth vs latency (few CPU cores)
- SCSI TCQ - 2^{64} - зависит от протокола, адаптера и т.д. SAS SSD/HDD max queue depth = 128
- ⬆ не подходит для SSD. SSD по природе обладают высоким уровнем параллелизма, а следующее поколение NVM - сверхнизким уровнем задержек (vs NAND)





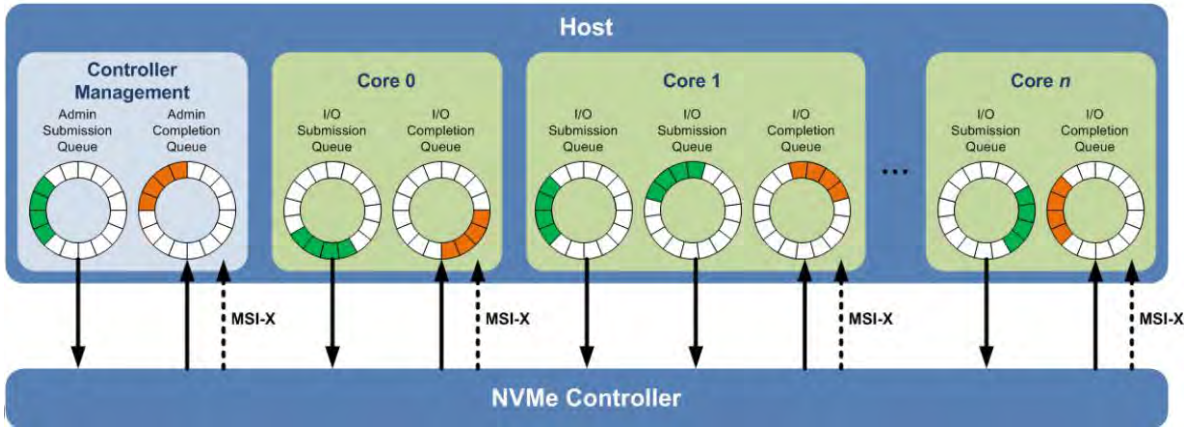
Появление NVMe

- 2011 - Появление Promoter Group, NVMe v 1.0
- 2012 - NVMe 1.1
- 2014 - NVMe 1.2
- 2014 - Образование компании NVM Express Organization

- Стандартизация регистров, функционала и набора команд
- Изначально создан для NAND и NVM следующего поколения
- Спроектирован для корпоративного и клиентского использования

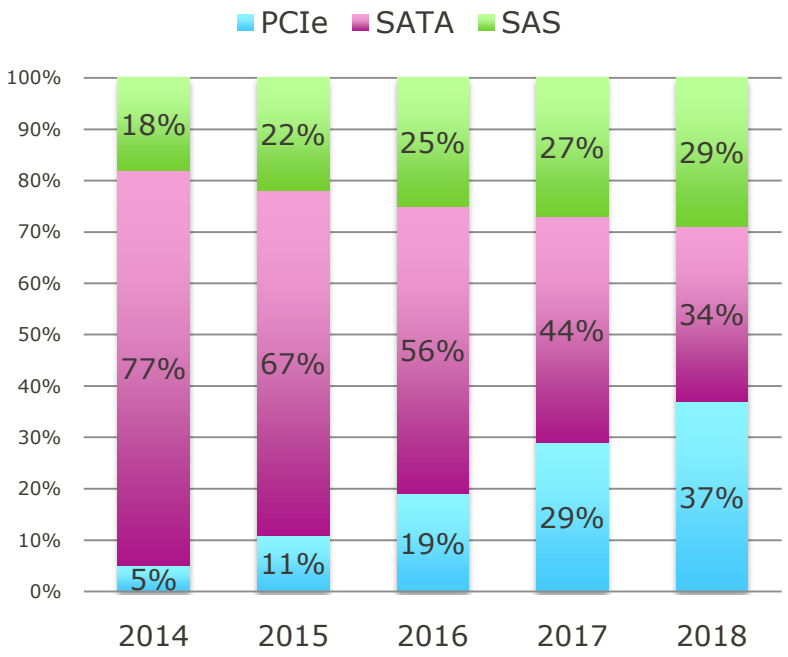
Технические основы NVMe

- Глубокие очереди 64K x 64K
  - Рекомендовано 32-128 Ent / 2-8 Client
- Компактный набор команд - 13 обязательных
  - ATA: 100+ legacy commands / SCSI: 250+
- Поддержка MSI-X
  - Снижение нагрузки на CPU, устранение узких мест
- Работа с любым видом NVM
  - Совместимость с NAND и будущими типами памяти
- Эффективная работа на 4К
  - Все параметры команды в одном запросе 64B
- Опциональный функционал для клиентских и корпоративных дисков



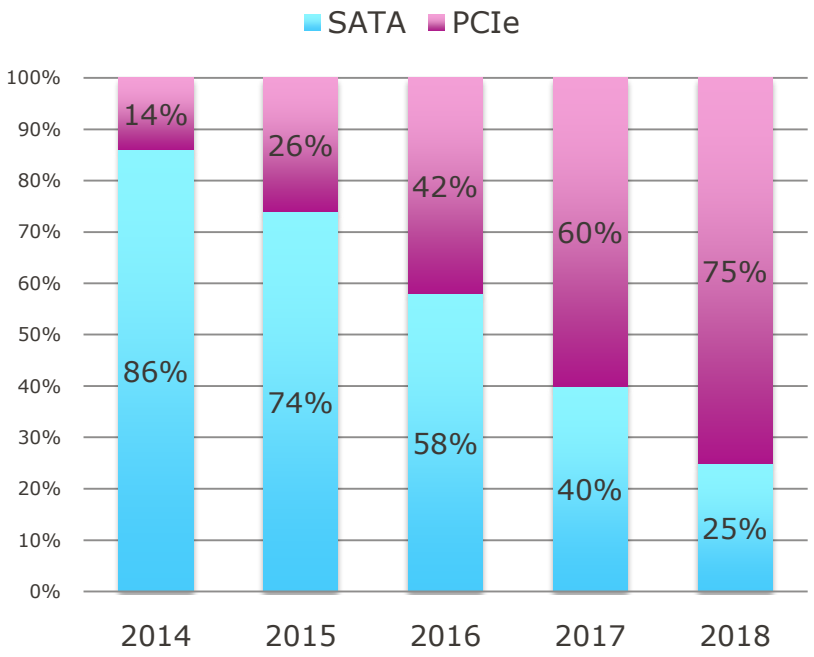
Наступление NVMe

Enterprise



IDC Worldwide Solid State Drive 2014-2018 June 2014

Client



Forward Insights

Знаете ли вы?.. iPhone 6s использует NVMe

Root > N66mAP > AppleARMPE > arm-io > AppleS8000IO > apcie > AppleS8003PCIe > pci-bridge0 > IOPP > s3e > AppleEmbeddedNVMeController

Properties

Physical Interconnect = PCI-Express

Controller Characteristics = {cell-type = 3,controller-unique-id = 0147086C65DE11030 ,pages-per-block-mlc = 258,capacity = 64000000000,pages-in-read-verify = 88,caus = 4,firmware-version = 12.22.01,sec-per-full-band-slc = 2752,ce-per-bus = 2,bytes-per-sec-meta = 16,num-dip = 8,dies-per-channel = 2,package_blocks_at_EOL = 16304,nand-marketing-name = 1Y128G-TLC-2p ,sec-per-full-band = 8256,cau-per-die = 2,page-size = 16384,pages-per-block-slc = 86,sec-per-page = 4,num-bus = 2,block-pairing-scheme = 0,chip-id = S3E,blocks-per-cau = 2108,Encryption Type = AES-XTS,vendor-name = Hynix ,pages-per-block0 = 0,default-bits-per-cell = 3,manufacturer-id = 0147086C65DE11030 }

http://forums.macrumors.com/threads/iphone-6s-64gb-tlc-nand-and-nvme-pcie-controller.1922221/

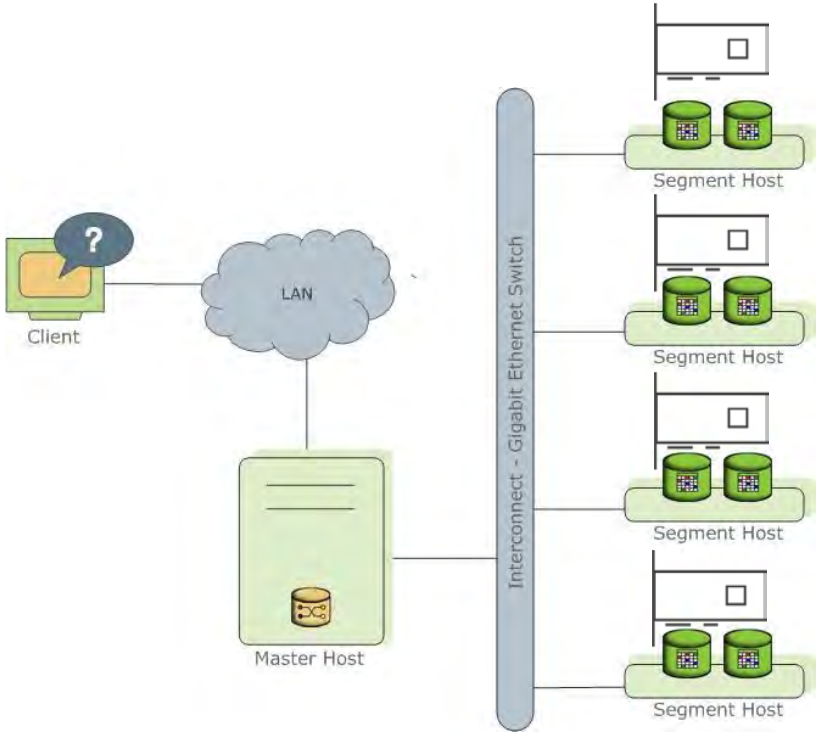
Ultrastar® SN100 Series

NVM Express™ Compatible PCIe SSD

- Высокая производительность: до 740K IOPS на чтение, 310K IOPS при 70/30 r/w, 3.0 GB/s
- Большой объем – до 3.2Тб
- Два форм-фактора – PCIe HH-HL (SN150) и 2.5” U.2 с поддержкой hot-plug (SN100)
- Возможность переформатирования на больший объем или производительность

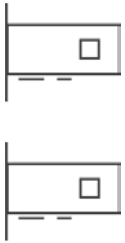
Пример использования: ретейл-банк, big data

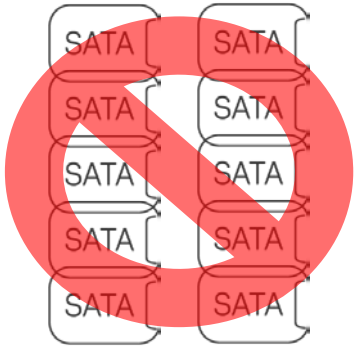
- Задача: ускорить работу кластера big data на базе Greenplum/PostgreSQL
- Конфигурация: 2 сайта, 24 узла на сайт, HP DL380 G8
- Решение: добавление одной карточки SN150 в каждый узел
- Результат: в тестовой среде - снижение продолжительности исполнения тяжелых запросов в 3-7 раз, снижение загруженности SAS дисков в RAID массиве в 2 раза (100% –> 50%)



Пример использования: онлайн кинотеатр

- Задача: загрузить 4x10GbE в сервере 1U / 1 socket
- Решение:
  - сервер DEPO Storm 1480P1 – 3 x PCIe 3.0 slots
  - 1 x Intel 710-based 4x10GbE NIC
  - 2 x SN150





- Результат:
  - 38Gbps при 100% загрузке NVMe карточек, полное насыщение 4 x 10GbE интерфейсов
  - Более экономичная и производительная альтернатива 8-10 SATA SSD
  - Уникальная сверхпроизводительная платформа в компактном исполнении
  - Потенциал для дальнейшего совершенствования – платформа на базе 1U со слотами U.2

Сравнение твердотельных накопителей: CERN

http://lvalsan.web.cern.ch/lvalsan/lfTnke7TrHowZmQxu1hsqI9Iuhg5LVr/

Filters

SSD manufacturer:

☐ HGST ☐ Intel ☐ OCZ ☐ Samsung

SSD family:

☐ 845DC Evo ☐ 845DC Pro ☐ DC P3600 ☐ DC P3700 ☐ DC S3500 ☐ DC S3700 ☐ PM853T ☐ SM843T ☐ SN100 fw 100 ☐ SN100 fw 110 ☐ SN100 fw 120 ☐ Vertex 3 ☐ X25-E

SSD capacity (GB):

☐ 64 ☐ 200 ☐ 240 ☐ 400 ☐ 480 ☐ 800 ☐ 960 ☐ 1600

SSD manufacturer:

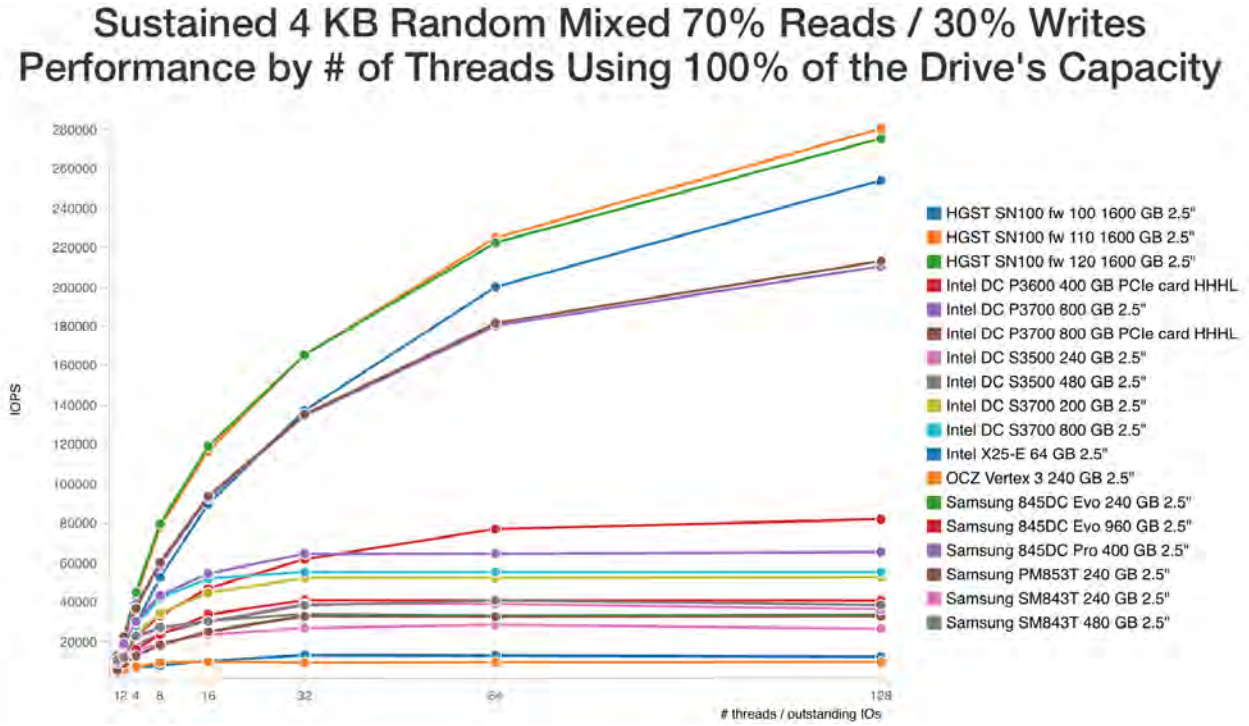
☐ 2.5" ☐ PCIe card HHHL

SSD inteface:

☐ NVMe ☐ SATA Revision 2.x ☐ SATA Revision 3.x

SSD flash type:

☐ MLC ☐ MLC V-NAND ☐ SLC ☐ TLC



Спасибо!

backup

Сравнение интерфейсов

	SAS SSD	SATA SSD	PCIe SSD	NVMe SSD	M.2
IOPs (R/W)	130K/100K	65K/15K	340K/60K	740K/110K	270K/90K
Sequential R/W (MB/s)	1100/750	500/450	2,700MB/1, 400MB	3000/1600	2100/ 1500
Latency µs	50µs	60µs	20µs	20µs	
Size (GB)	100-1920	80-1600	550-4800	400-3200	64-512
Power (Wt)	6-11	6-9	<25	12-25	
Price	$$	$	$$$	$$	$
Price/Read IOP	$$	$$	$$	$	$
Market	Enterprise	Client, Enterprise	Enterprise	Client, Enterprise	Client