Актуальные и перспективные решения Broadcom для хранения и передачи данных

20.11.2018, Intel Innovation Day

Ярослав Надпорожский, инженер Broadcom

2018: Дата-центры – рост мощностей в среднем на 14%

Драйверы роста:

1. Закон Яровой
2. Закон о хранении персональных данных в пределах РФ
3. Облачные сервисы

Ожидаем в 2019:

1. Нехватка мощностей ЦОДов
2. Рост цен на услуги хранения данных – до 15%

BROADCOM

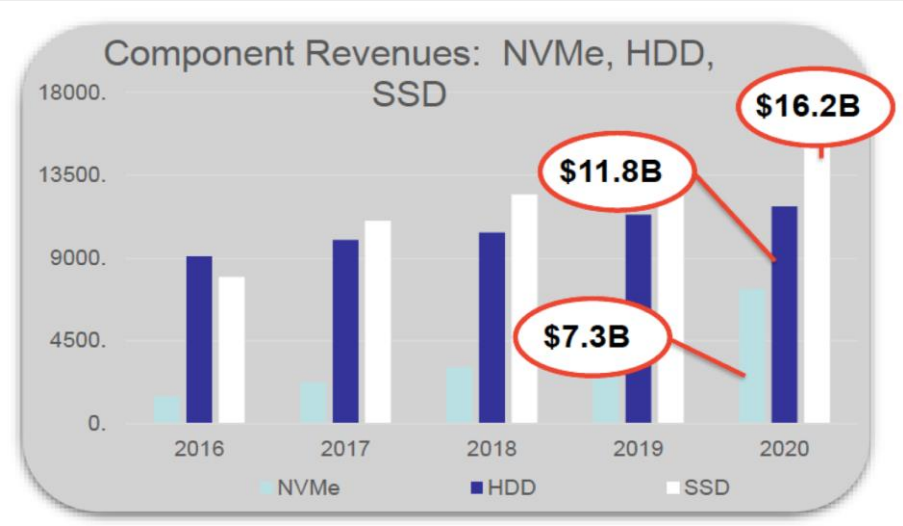
Основные требования к СХД – емкость и производительность

Самый быстрый Flash – на NVMe

- NVM Express (NVMe) или Non-Volatile Memory Host Controller Interface Specification (NVMHCIS) это открытый интерфейс для подключения логических устройств и доступа к системам хранения на энергонезависимых (Non-Volatile) устройствах через шину PCI Express (PCIe).
- Используемых форм-факторы
  - PCIe Slots (x4, x8)
  - U.2 (SFF-8639) (JBOFs) (x1,x2,x4)
  - M.2 (ноутбуки, настольные PC, загрузчики для серверов) (x1,x4)
- Высокая производительность
  - Пропускная способность. 4GB/s (x4)
  - IOPS – до 500kiops

NVMe против SAS

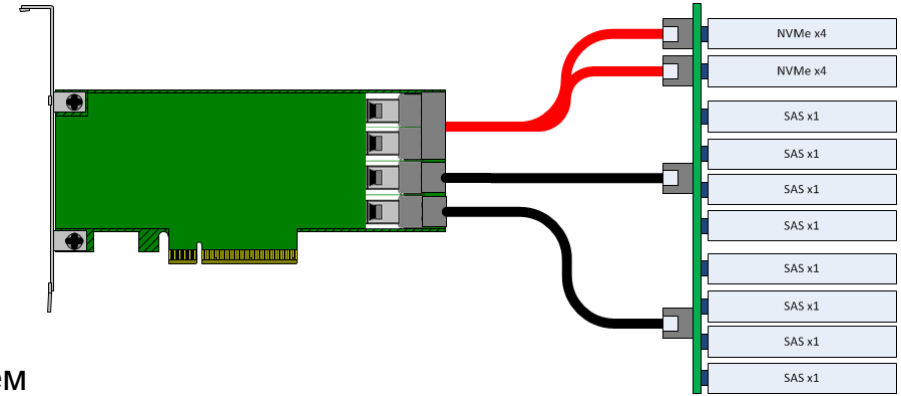
1. Высокая скорость и пропускная способность
2. Стандартный метод доступа к flash – низкая латентность (20 мкс и ниже) и параллелизм PCIe
3. Поддержка 64K очередей с глубиной до 64K команд (SAS поддерживает 1 очередь)
4. Для SAS устройств требуется преобразование между PCIe и стеком SAS/SCSI. NVMe устройства подключаются к PCIe напрямую.
5. Относительно легкий стек NVMe (13 команд, нет планировщика ввода-вывода, сложных уровней протокола SCSI для HDD). SAS – сотни команд.



BROADCOM

NVMe в вашем сервере (СХД) – альтернатива SATA и SAS SSD

- Один 16-портовый MegaRAID / HBA
  - 2 x4 NVMe SSD
  - 8 x SAS/SATA накопителей



- ✓ Убиваем 2x зайцев – производительность & объем
- ✓ Удобное управление и мониторинг: новые GUI (LSA = LSI Storage Authority) + CLI (StorCLI)
- ✓ NVMe устройства не занимают PCIe слоты
- ✓ Защита данных NVMe возможна за счет HW MegaRAID (те же доступные виды RAID, что и ранее)

- × PCIe x8 Gen3 могут стать узким местом (bottleneck) по пропускной способности. Альтернатива – HBA 9405 (PCIe Gen3 x16)

BROADCOM

Новое поколение трехрежимных 12G MegaRAID серии 9400

9460-16i, 9460-8i, 9440-8i, 9480-8i8e

	MegaRAID Adapters			
	9460-16i	9460-8i	9480-8i8e	9440-8i (iMR)
ROC	3516 (Ventura)	3508 (Harpoon)	3516 (Ventura)	3408 (Tomcat)
Connectors	HH 16i side SAS, SATA, PCIe	HH 8i side SAS, SATA, PCIe	HH 8i side 8e external SAS, SATA, PCIe	HH 8i side SAS, SATA, PCIe
Form Factor	STD LP-MD2	STD LP-MD2	STD LP-MD2	STD LP-MD2
Cache Protection	CacheVault (ONFI) on board	CacheVault (ONFI) on board	CacheVault (ONFI) on board	N/A
SuperCap	Remote	Remote	Remote	N/A
DDR Density*	4/8GB DDR4	4/8GB DDR4	2/4/8GB DDR4	N/A
Nominal Power (estimate)	<25W	<25W	<25W	<25W
Thermals	55C @ 300 LFM	55C @ 200 LFM	55C @ 300 LFM	55C @ 200 LFM
Warranty	3 Years	3 Years	3 Years	3 Years

BROADCOM

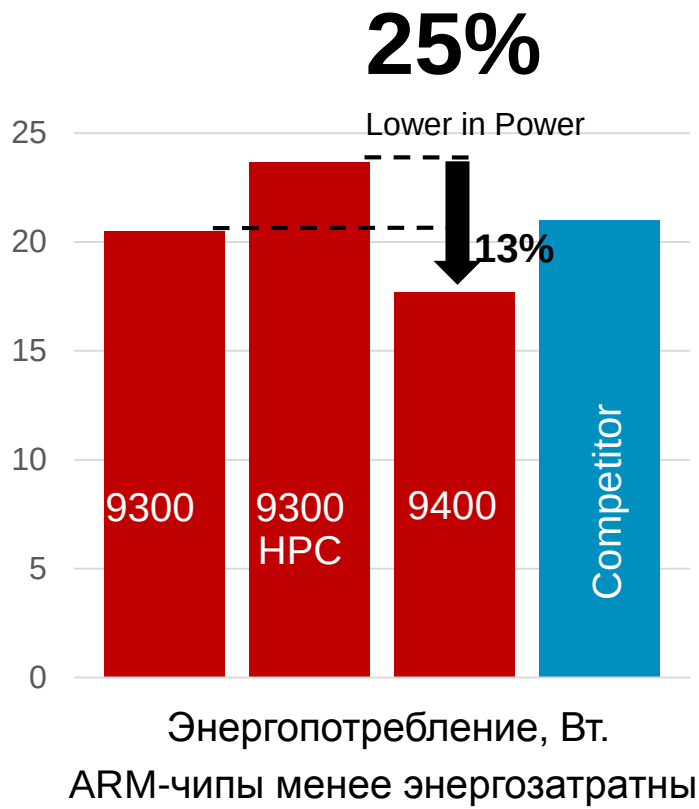
9460-16i, 9460-8i, 9440-8i, 9480-8i8e: новые возможности

Достоинства по сравнению с предыдущими сериями

- Многопортовые SAS/SATA/NVMe контроллеры для прямого (безэкспандерного) подключения
- Трехрежимное подключение обеспечивает дополнительные возможности при использовании в ЦОД
- Возможность смешанного подключения накопителей на 16-портовых контроллерах
- В 3+ раза улучшена производительность на RAID5/6 (4K RW с 50K IOPS до 185K IOPS)
- Поддержка режима JBOD, поддержка TRIM

Ключевые особенности

- Порты поддерживают 3 интерфейса подключения
  - SFF-8680 Bay
    - SAS/SATA
  - SFF-8639 (U.2) Bay
    - x2, x4 NVMe
- Поддержка 12, 6, и 3Gb/s SAS а также 6, 3Gb/s SATA уровней передачи данных
- 8x PCIe Gen3
- Поддержка x4 или x2 PCIe Gen3 подключаемых устройств
- CacheCade не поддерживается

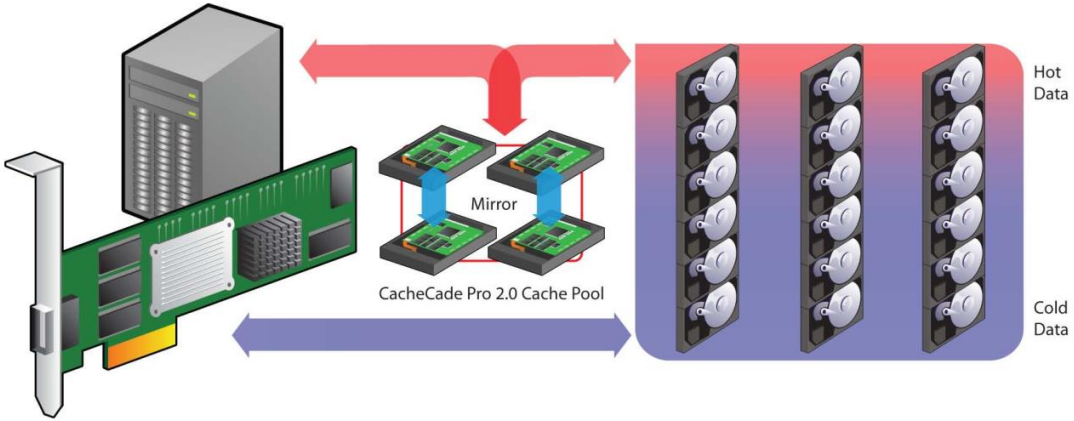


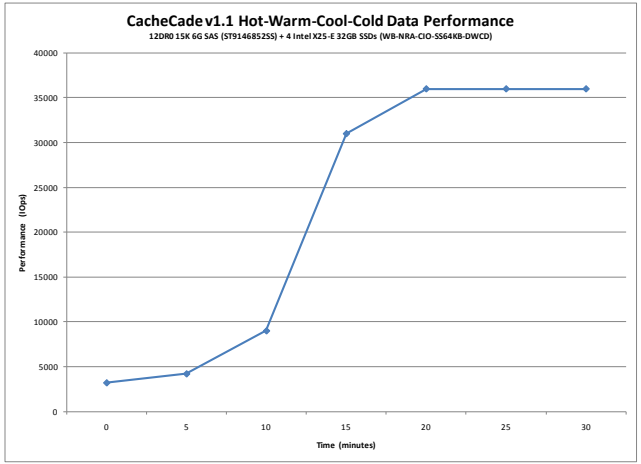
Энергопотребление, Вт.

ARM-чипы менее энергозатратны.

BROADCOM

CacheCade: с 20.11.2018 доступен только ФИЗИЧЕСКИЙ ключ





BROADCOM

Трехрежимные 12G HBA серии 9400

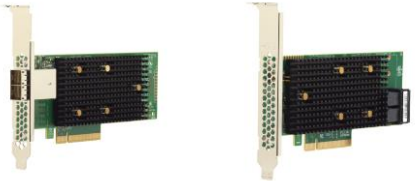
9400-16i, 9400-16e, 9400-8i, 9400-8e, 9405W-16i, 9405W-16e

16-port Tri-Mode HBAs

9405W-16e (05-50044-00) 9405W-16i (05-50047-00)

16-port Tri-Mode HBAs

9400-16e (05-50013-00) 9400-16i (05-50008-00)



8-port Tri-Mode HBAs

9400-8e (05-50013-01) 9400-8i (05-50008-01)

Области применения

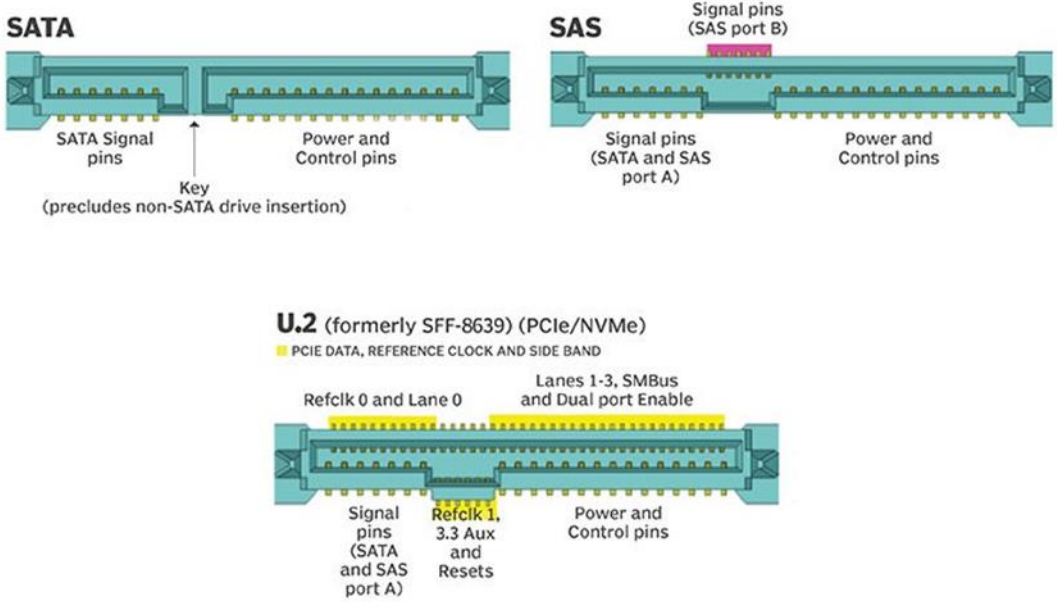
- Многопортовые SAS/SATA/NVMe адаптеры дают возможность прямого подключения
- Программно-определяемые системы хранения данных (SDS)
- HBA на x16 PCIe Gen3 являются удобным решением для потоковых задач (передача видео высокой четкости , big data аналитика, резервирование данных)
- Напрямую возможно подключение до 4 x NVMe (x4), через PCIe Switch – до 24 NVMe устройств
- Возможность смешанного подключения накопителей на 16-портовых контроллерах

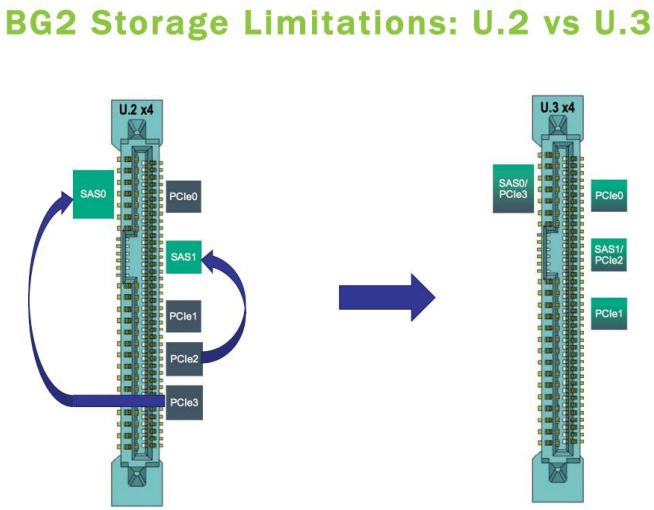
Key Features

- Порты поддерживают 3 интерфейса подключения
  - SFF-8680 Bay
    - x1 SAS/x1 SATA
  - SFF-8639 (U.2) Bay
    - x2, x4 NVMe
- Поддержка 12, 6, и 3Gb/s SAS а также 6, 3Gb/s SATA уровней передачи данных
- 8x PCIe Gen3
- Поддержка x4 или x2 PCIe Gen3 подключаемых устройств

BROADCOM

От U.2 (SFF-8639) к U.3 = SFF-TA-1001





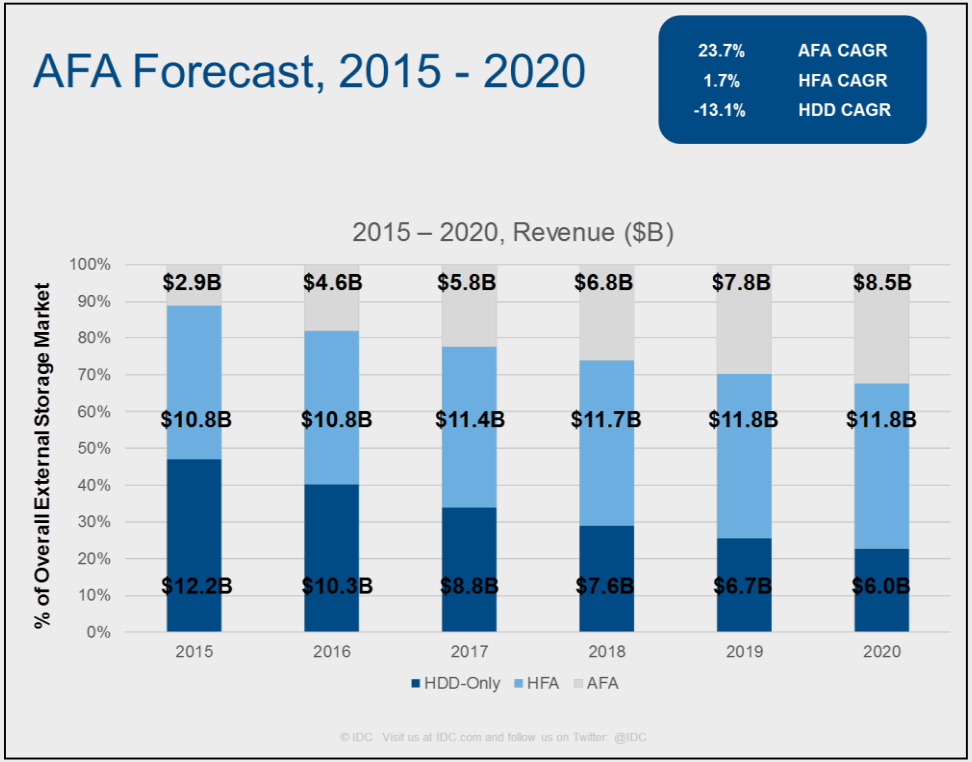
Основной плюс – возможность подключения как «обычных» SAS/SATA накопителей, так и NVMe(PCIe) в один слот на бэкплейне

Источник: https://ta.snia.org/higherlogic/ws/public/download/1272/SFF-TA-1001%20r1.0.pdf

BROADCOM

Распределенные СХД (SAN) на Flash - основной драйвер роста решений на Fibre Channel

- СХД, реализованные на SSD, принесли **$15B** на глобальный рынок СХД за период 2015-20 (от всего объема ~**24%**)
- **>85%** от AFA (All-flash Arrays) были подключены через Fibre Channel
- **>90%** от AFA были подключены через 16 либо 32G Fibre Channel



Source: The Evolving All Flash Array (AFA) Market, IDC Jan 2017

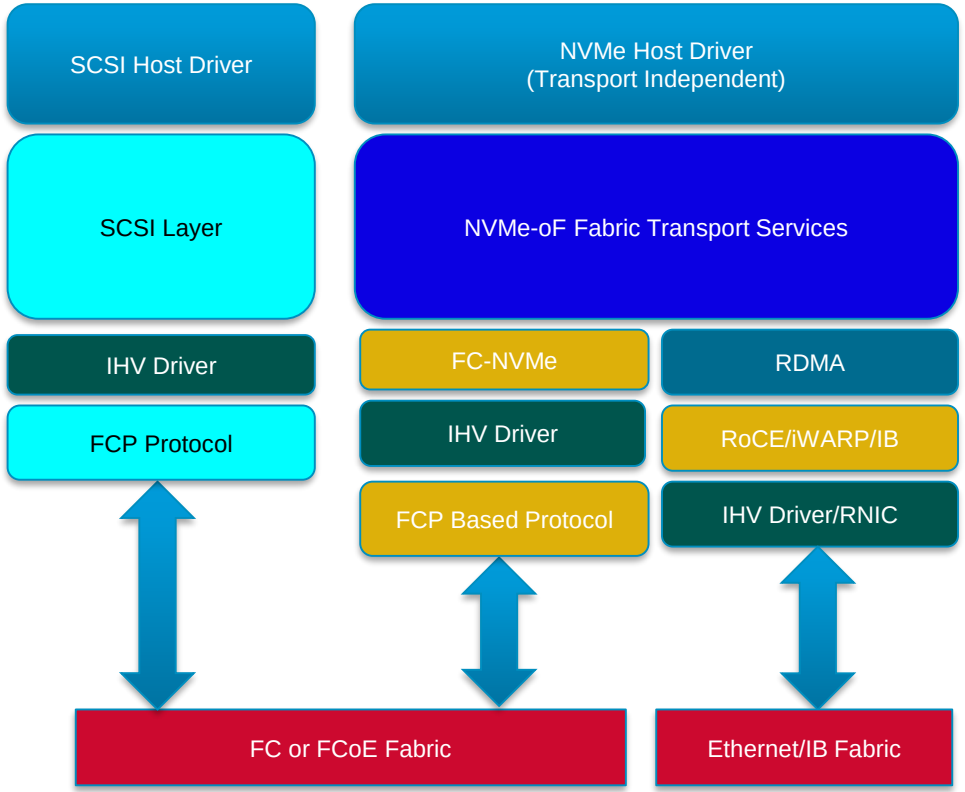
BROADCOM

NVMe over Fibre Channel: еще раз о стандарте

- NVM Express – Июнь 2016 опубликован стандарт NVMe over Fabrics
- INCITS T11 (Технический комитет T11 Международного комитета стандартов в ИТ) – развитие спецификации FC-NVMe
  - FC-NVMe спецификация утверждена “Technically Stable” – Август 2017
  - ведется работа над рекомендациями
  - Окончательное утверждение спецификации INCITS (2017)
- FCIA NVMe over Fibre Channel Plugfests:

как минимум 9 проведено

  - ноябрь 2018 – запланирован очередной (Университет Нью Хэмпшир, США, InterOperability Laboratory

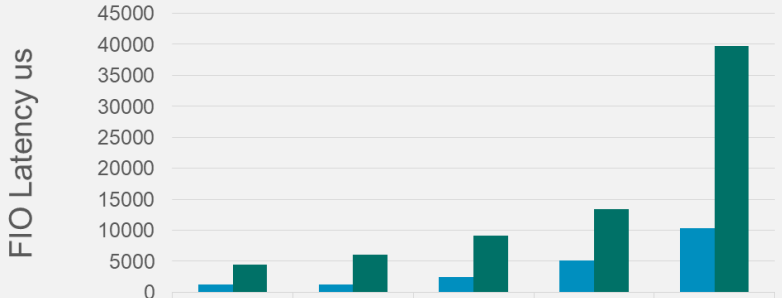


Ключевые особенности NVMe over Fibre Channel

- Поддержка NVMe и SCSI трафика в одной инфраструктуре
- NVMe поддерживает более низкие задержки и повышенную очередизацию запросов
- FC HBA Gen6 от Broadcom поддерживают оба протокола
  - FCP и FC-NVMe могут работать одновременно через одни и те же порты
  - Сохраняются вложения в существующую FC инфраструктуру
  - Достаточно обновить прошивку FC switch'ей (Brocade G610, G620 полностью поддерживают)
  - Данные могут легко мигрировать на новый уровень NVMe
- FC – закаленный в боях транспорт для SAN
  - Отработанная инфраструктура
  - Хорошо известны политики и процедуры
- Безопасность
  - К FC SAN трудно получить доступ через Интернет
- Традиционная модель тестирования и поддержки

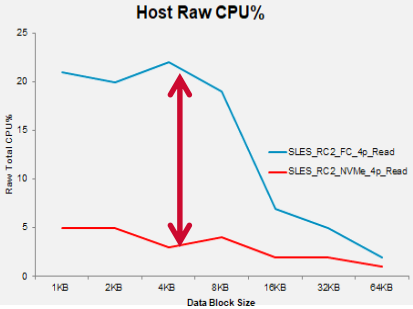
NVMe over FC: примеры выигрыша в производительности

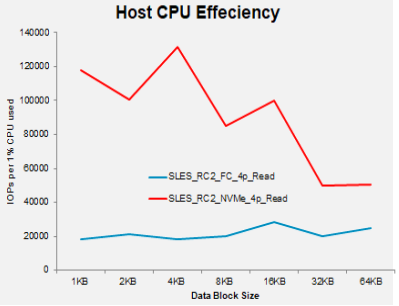
3x faster latency than SCSI over FC



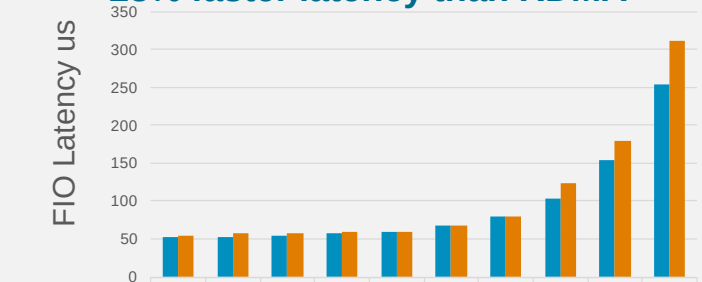
	4k	8k	16k	32k	64k
NVMe over FC	1225	1238	2517	5072	10319
SCSI over FC	4485	6012	9097	13442	39767

4x better CPU efficiency than SCSI over FC





23% faster latency than RDMA



	512B	1KB	2KB	4KB	8KB	16KB	32KB	64KB	128KB	256KB
ELX_Toshiba_Read	53	53	54	57	60	67	80	104	154	253
MLX_Toshiba_Read	54	57	58	59	60	67	80	123	180	312

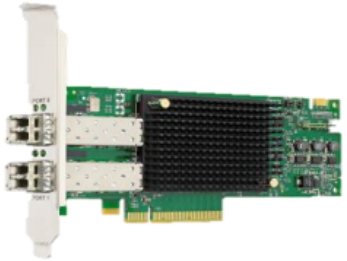
16x the work @only 16% latency cost

Queues x Queue Depth	Avg Response Time: NVMe	Avg Response Time: SCST
1 x 1	31 us	60 us
8 x 1	33 us	585 us
16 x 1	36 us	1294 us

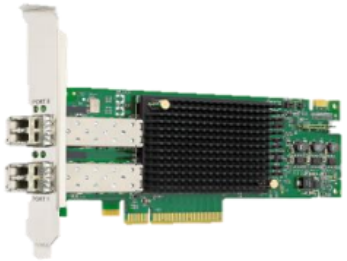
BROADCOM

Emulex Gen 6 Fibre Channel HBAs

LPe31000-series (16GFC, Upgradable); LPe32000-series (32GFC)



Emulex Gen 6 32GFC HBAs
Broadcom Ltd. MPN LPe32000-M2 (1p)
Broadcom Ltd. MPN LPe32002-M2 (2p)

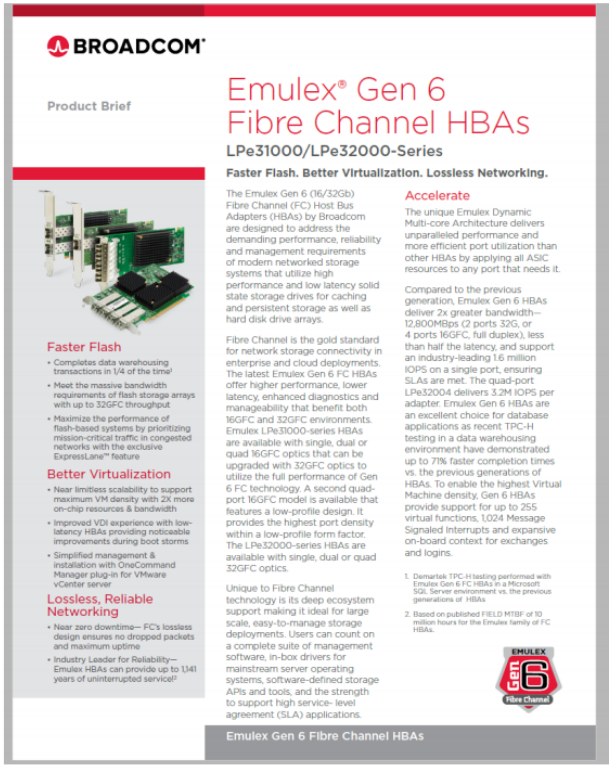


Emulex Gen 6 16GFC HBAs
Broadcom Ltd. MPN LPe31000-M6 (1p)
Broadcom Ltd. MPN LPe31002-M6 (2p)

- 1-port 16GFC; 2-port 16GFC; 1-port 32GFC; 2-port 32GFC SFP+ low profile, PCIe 3.0 adapters
- LPe31000-series supports 16GFC, 8GFC and 4GFC link speeds
  - Upgradeable to 32GFC with Emulex-branded optical transceiver kits only
- LPe32000-series supports 32GFC, 16GFC and 8GFC link speeds
- 71% faster completion times using TPC-H benchmark versus Gen 5 FC HBAs
- Maximize the performance of flash-based systems by prioritizing mission-critical traffic with Emulex ExpressLane™
- High performance hardware offloads for T10 Protection Information (T10-PI) for data integrity
- Designed to support emerging NVM Express and ANSI T11 FC-NVMe standards

Emulex Gen6 Quad-Port FC-HBA

- LPe31004-M6-SIO
  - Gen 6 16GFC, Quad-port
  - Full-height w/ standard bracket
  - Removable, standard SFP+ optical transceivers
    - Existing LPe31004-M6, 4-port Gen 6 16GFC HBA has embedded optics
  - Upgradeable to 32GFC
- LPe32004-M2-SIO
  - Gen 6 32GFC, Quad-port
  - Full-height w/ standard bracket
  - Removable Optics
  - 3.2M IOPS
- Availability
  - Now



BROADCOM

Product Brief

Emulex® Gen 6 Fibre Channel HBAs

LPe31000/LPe32000-Series

Faster Flash. Better Virtualization. Lossless Networking.

The Emulex Gen 6 (16/32Gb) Fibre Channel (FC) Host Bus Adapters (HBAs) by Broadcom are designed to address the demanding performance, reliability and management requirements of modern networked storage systems that utilize high performance and low latency solid state storage drives for caching and persistent storage as well as hard disk drive arrays.

Fibre Channel is the gold standard for network storage connectivity in enterprise and cloud deployments. The latest Emulex Gen 6 FC HBAs offer higher performance, lower latency, enhanced diagnostics and manageability that benefit both 16GFC and 32GFC environments. Emulex LPe31000-series HBAs are available with single, dual or quad 16GFC optics that can be upgraded with 32GFC optics to utilize the full performance of Gen 6 FC technology. A second quad-port 16GFC model is available that features a low-profile design. It provides the highest port density within a low-profile form factor. The LPe32000-series HBAs are available with single, dual or quad 32GFC optics.

Unique to Fibre Channel technology is its deep ecosystem support making it ideal for large scale, easy-to-manage storage deployments. Users can count on a complete suite of management software, in-box drivers for mainstream server operating systems, software-defined storage APIs and tools, and the strength to support high service- level agreement (SLA) applications.

Faster Flash

- Completes data warehousing transactions in 1/4 of the time[1]
- Meet the massive bandwidth requirements of flash storage arrays with up to 32GFC throughput
- Maximize the performance of flash-based systems by prioritizing mission-critical traffic in congested networks with the exclusive ExpressLane™ feature

Better Virtualization

- Near limitless scalability to support maximum VM density with 2X more on-chip resources & bandwidth
- Improved VDI experience with low-latency HBAs providing noticeable improvements during boot storms
- Simplified management & installation with OneCommand Manager plug-in for VMware vCenter server

Lossless, Reliable Networking

- Near zero downtime— FC's lossless design ensures no dropped packets and maximum uptime
- Industry Leader for Reliability— Emulex HBAs can provide up to 1,141 years of uninterrupted service[2]

Accelerate

The unique Emulex Dynamic Multi-core Architecture delivers unparalleled performance and more efficient port utilization than other HBAs by applying all ASIC resources to any port that needs it.

Compared to the previous generation, Emulex Gen 6 HBAs deliver 2x greater bandwidth— 12,800MBps (2 ports 32G, or 4 ports 16GFC, full duplex), less than half the latency, and support an industry-leading 1.6 million IOPS on a single port, ensuring SLAs are met. The quad-port LPe32004 delivers 3.2M IOPS per adapter. Emulex Gen 6 HBAs are an excellent choice for database applications as recent TPC-H testing in a data warehousing environment have demonstrated up to 71% faster completion times vs. the previous generations of HBAs. To enable the highest Virtual Machine density, Gen 6 HBAs provide support for up to 255 virtual functions, 1,024 Message Signaled Interrupts and expansive on-board context for exchanges and logins.

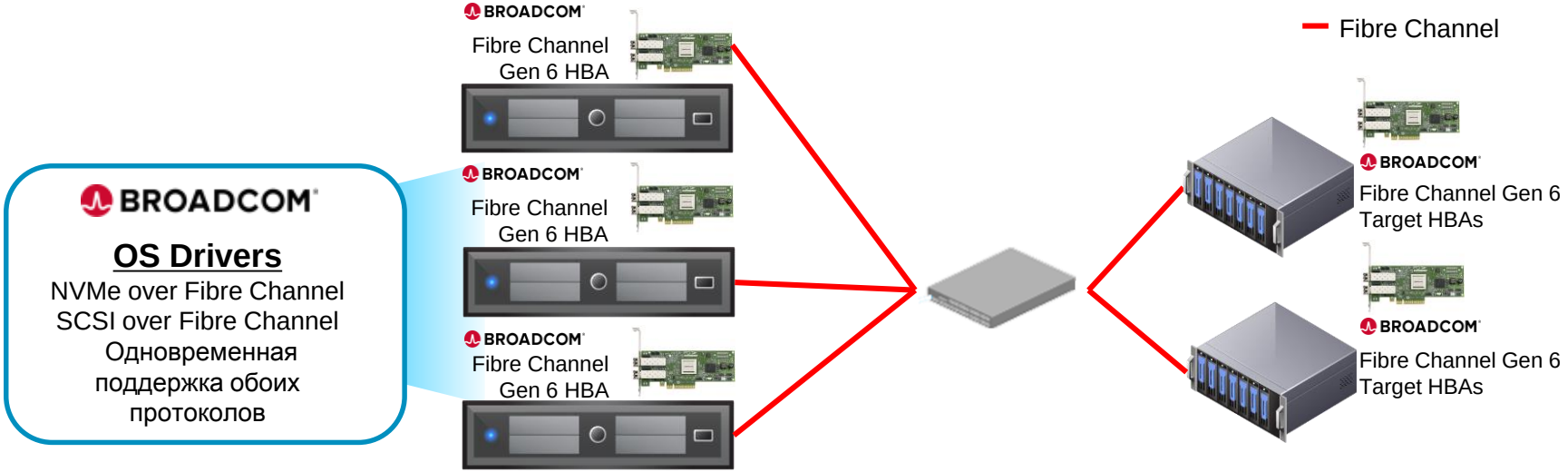
1. Demartek TPC-H testing performed with Emulex Gen 6 FC HBAs in a Microsoft SQL Server environment vs. the previous generations of HBAs
2. Based on published FIELD MTBF of 10 million hours for the Emulex family of FC HBAs.

EMULEX Gen 6 Fibre Channel

Emulex Gen 6 Fibre Channel HBAs

Сложившаяся инфраструктура для передачи NVMe over FC

HBA Emulex прекрасно справляются с транспортировкой NVMe over Fibre Channel и на стороне инициатора, и таргета одновременно



Operating System Partners

Windows Server · SUSE · Linux · redhat · vmware

Server OEMs

DELL Technologies · CISCO · Hewlett Packard Enterprise · IBM · HUAWEI · lenovo · ORACLE

Fibre Channel Switch Partners

CISCO · BROCADE

Disk and Flash Storage Arrays OEMs

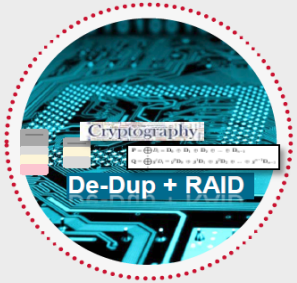
PURESTORAGE · DELL Technologies · Hewlett Packard Enterprise · IBM · HUAWEI · NetApp · nimble storage

BROADCOM

Stingray™ - Groundbreaking SmartNIC and Storage NVMeoF SoC

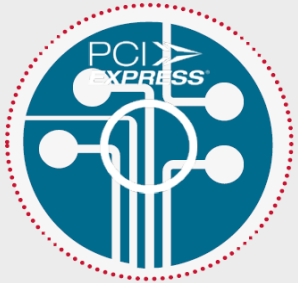
100G RDMA NIC
Low Latency RoCE
4x25G/2x50G/1x100G

HW Acceleration
RAID5/6, Crypto, Dedup

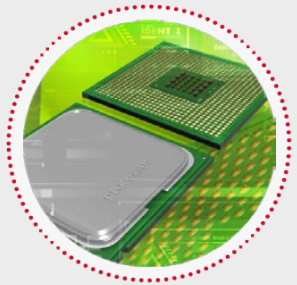


8-ARM Cores
A72@3GHz

100G PCIe BW
16 Gen 3 Lanes



460Gbps DDR BW
3 channels DDR4

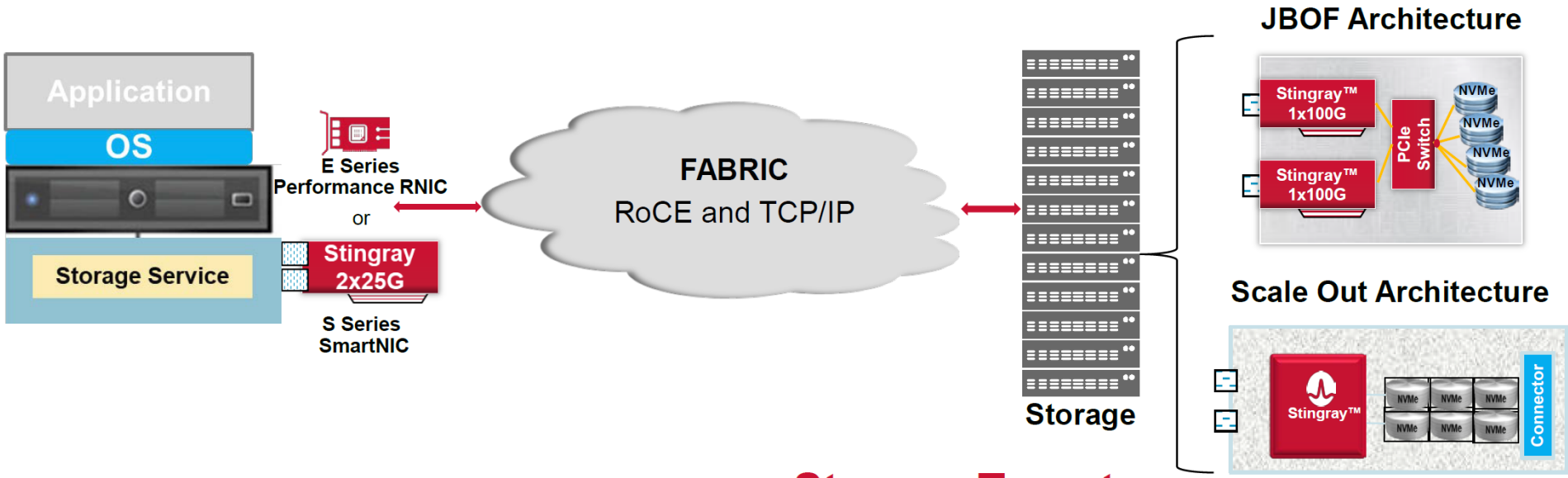


Key Applications

- SmartNIC adapter
- Storage services offload adapter
- Target storage controller

BROADCOM

End to End NVMe-oF Solutions based on NetXtreme E and S Series



Storage Initiator

- Offload services from x86
  - Security, Erasure/RAID, Compression etc.
- Manage storage separate from host

Storage Target

- ODM Partners

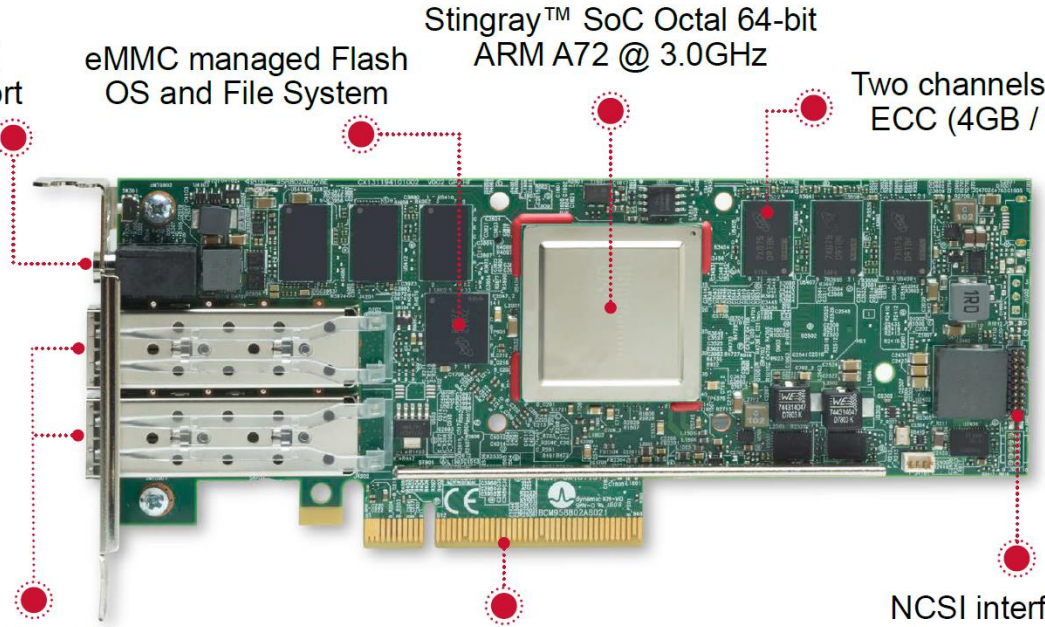
Celestica. wistron Inventec AIC quanta wiwynn

- Eco-system solutions

EXTEN Excelero H3 Platform CNEXLABS EIDETICOM BROADCOM

Stingray SmartNIC 2x 25G Adapter Card (PS225)

- **Breakthrough Perf/Watt**
- **Low Power:** 35W typical
- **Fully integrated NIC**
  - SR-IOV, RoCE, TruFlow™
- **Open Programming Model**

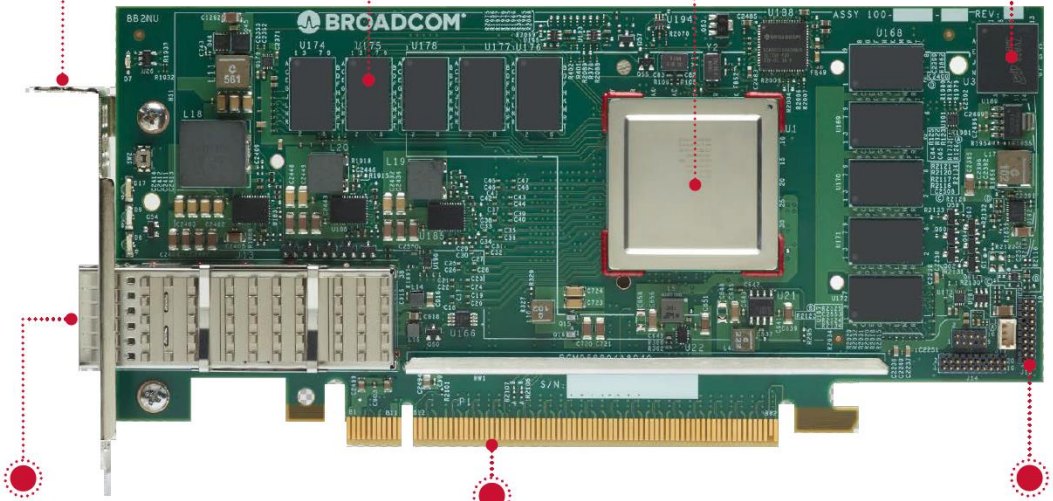


Breakthrough SmartNIC: OVS Offload, Bare Metal, Storage, NFV in Compact Form Factor

BROADCOM

100G Storage Target Adapter Card (PS1100R)

- **Storage services**
- **Low Power:**
  - 45W w/8c
  - 30W w/4c
- **Fully integrated rNIC**
  - NVMe oF
- **Turnkey solution**



BROADCOM

Broadcom NetXtreme E-Series PCIe NIC

Семейство сетевых карт 10, 25, 40, 50, 100GbE, 10GBase-T

Broadcom NetXtreme GbE
P1100p (1p, 40/50/1000GbE QSFP)
P150p (1P, 40/50GbE QSFP)
P225p (2P, 25GbE SFP28)
P210p (2p, 10GbE SFP)

Broadcom NetXtreme 10GBase-T
P210tp (2p, 10GBase-T RJ45)

Broadcom NetXtreme OCP

Key Features

- High Performance Network Controller
  - Full line-rate throughput on all ports
  - Low device latency for optimal application performance
  - Comprehensive stateless offloads to minimize host CPU workload
- Robust Virtualization Features
  - Single Root I/O Virtualization (SRIOV)
  - 128 VFs with fully flexible port assignment
  - VMQ support in hardware
- Optimized silicon design
  - Low device power using Adaptive Voltage Scaling (AVS)
  - Simplified device driver interface for ease of management
  - State machine device architecture for core functions
  - Integrated Ethernet PHY based on market-proven IP
- Reliability
  - Hardware architecture built on 10+ generations of Broadcom expertise in Ethernet
  - Robust driver for all major enterprise operating systems
- End-to-End 25GbE connectivity
  - Compliant with 25GbE standards (IEEE 802.3by and 25G Ethernet Consortium)
  - Assured performance and compatibility when paired with industry leading Broadcom StrataXGS® and StrataDNX™ based switches

NetXtreme 10/25/40/50/100G. Доступны OCP Mezzanine.

Form Factor	Name	RoCE	Part Number	Port Speed/Type	# Ports	Production
PCIe NIC 	P150c	No	BCM957304A3040C	50GbE	Single	now
	P225c		BCM957304A3041CC	25GbE	Dual	now
	P210c		BCM957302A3021AC	10GbE	Dual	now
	P150p	Yes	BCM957414A4140C	50GbE	Single	now
	P225p		BCM957414A4142CC	25GbE	Dual	now
	P210p		BCM957412A4120AC	10GbE	Dual	now
	P210tp		BCM957416A4160C	10GBase-T	Dual	now
	P1100p		BCM957454A4540C	100GbE	Single	now
OCP Mezzanine 	M150c	No	BCM957304M3041C	50GbE	Single	now
	M225c		BCM957304M3040C	25GbE	Dual	now
	M125c		BCM957302M3020CBK	25GbE	Single	now
	M210c		BCM957302M3022AC	10GbE	Dual	now
	M150p	Yes	BCM957414M4143C	50GbE	Single	now
	M225p		BCM957414M4142C	25GbE	Dual	now
	M125p		BCM957412M4122C	25GbE	Single	now
	M210p		BCM957412M4123C	10GbE	Dual	now
	M210tp		BCM957416M4163C	10GBase-T	Dual	now
	M1100p			100GbE	Single	now


BROADCOM®
connecting everything®