Фабрика для NVMe

Сергей Платонов

Заместитель генерального директора по стратегии, «Рэйдикс»

Технология

RDMA

RDMA (Remote DMA) позволяет обеспечить прямой доступ к оперативной памяти другого компьютера и к данным, хранящимся в удалённой системе, без привлечения средств ОС обоих компьютеров. RDMA - метод пересылки данных с высокой пропускной способностью и низкой задержкой сигнала.

- Нулевая копия (Zero-copy)
- Kernel bypass
- Без привлечения ЦПУ
- Транзакции на основе сообщений (message-based)
- Поддержка разбросанного ввода-вывода (Scatter/gather)

RAIDIX

Технология

NVMe

Независимый от вендора интерфейс хранения:

- Многопоточный режим без привязки
- Упрощенный набор команд
- 64 000 очередей и 64 000 команд в очереди.

NVMe over Fabrics

- Способ передачи NVMe-команд поверх сетевых протоколов (“Fabrics”)
- Основан на той же архитектуре и использует то же ПО хоста, что и локальные устройства
- Спецификация только дополняет спецификацию NVMe.

RAIDIX

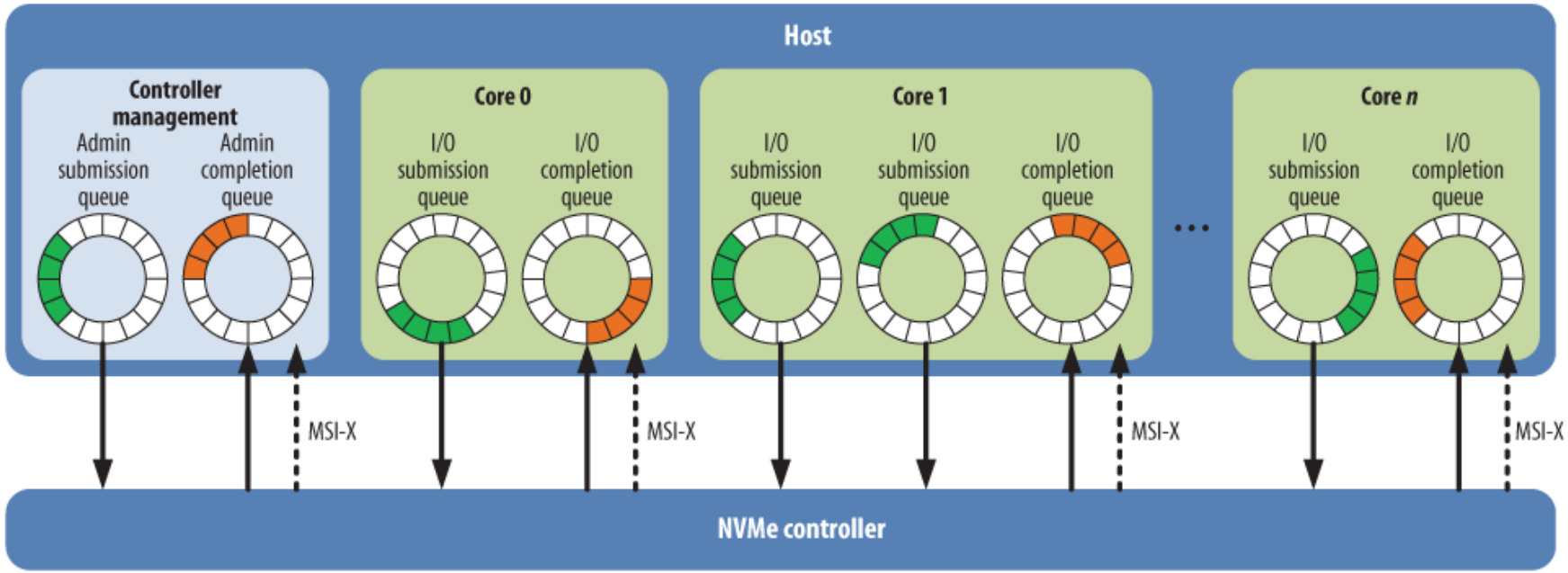
Технология

Преимущества NVMe и NVMe over Fabrics

- (64K – 1) I/O очередей и (64K – 1) команд в очереди
- Комплексная защита данных
- Приоритизация исполнения команд и описанный механизм арбитража
- Эффективный и расширяемый набор команд
- Поддержка множества адресных пространств
- Поддержка multi-path I/O и разделяемого доступа к адресным пространствам.

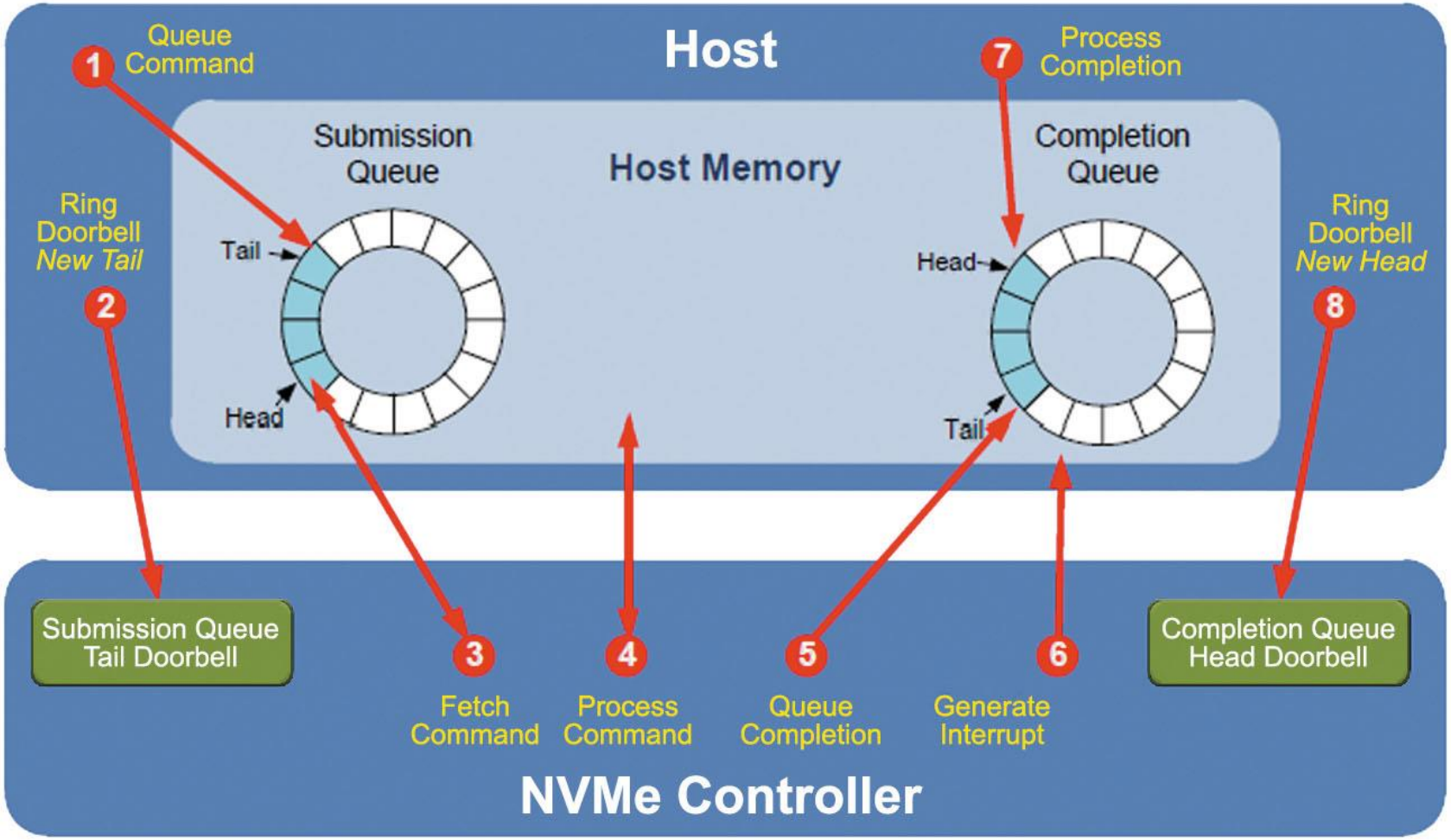
RAIDIX

Технология



RAIDIX

Технология



RAIDIX

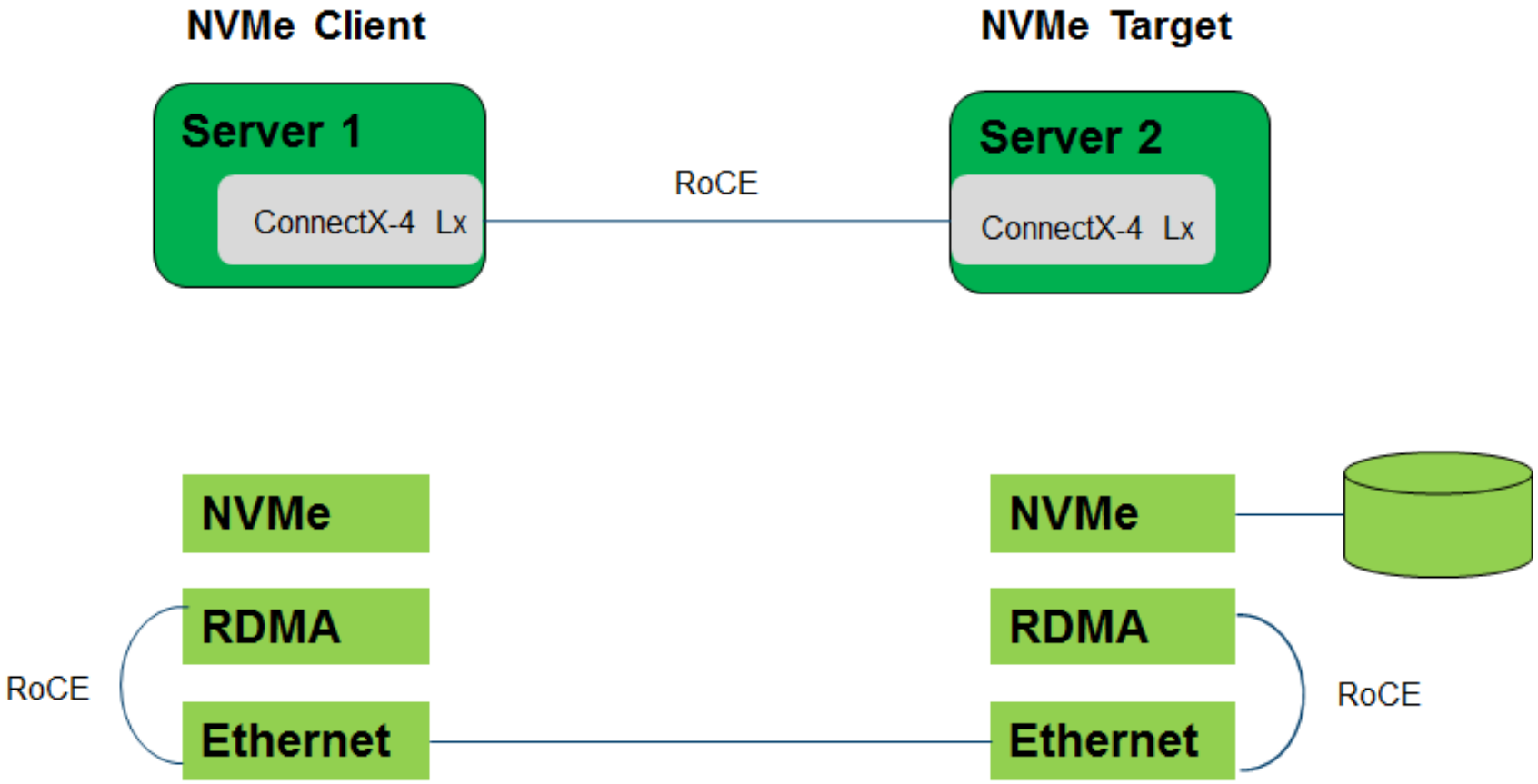
Фабрики

RDMA (в основном развивается компаниями HGST и Mellanox)

FC (в основном развивается компанией Broadcom)

RAIDIX

Технология

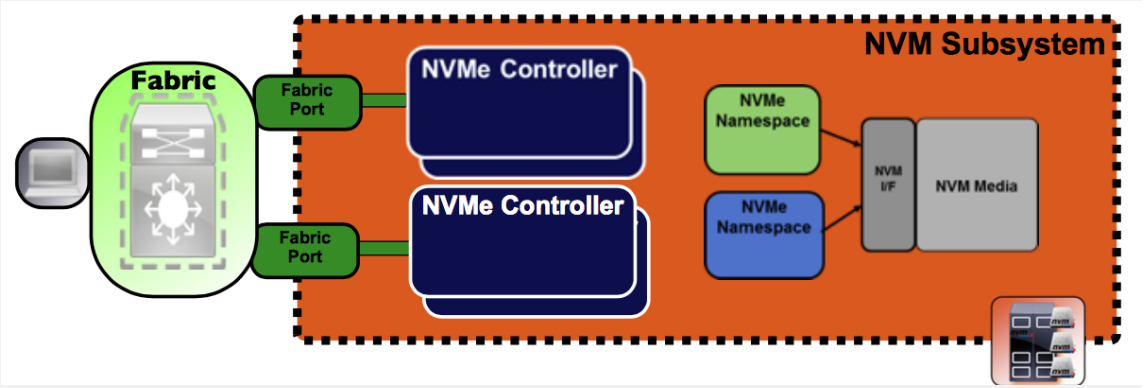


RAIDIX

Технология

Основные определения

- Subsystem (подсистема)
- Namespace (адресное пространство)
- Port
- NQN (NVMe Qualified Name)



Protocol	Type	Example
NVMe	NQN	nqn.2014-08.com.vendor:nvme:nvm-subsystem-sn-d78432
iSCSI	IQN	iqn.1991-05.com.microsoft:dmrtk-srvr-m

RAIDIX

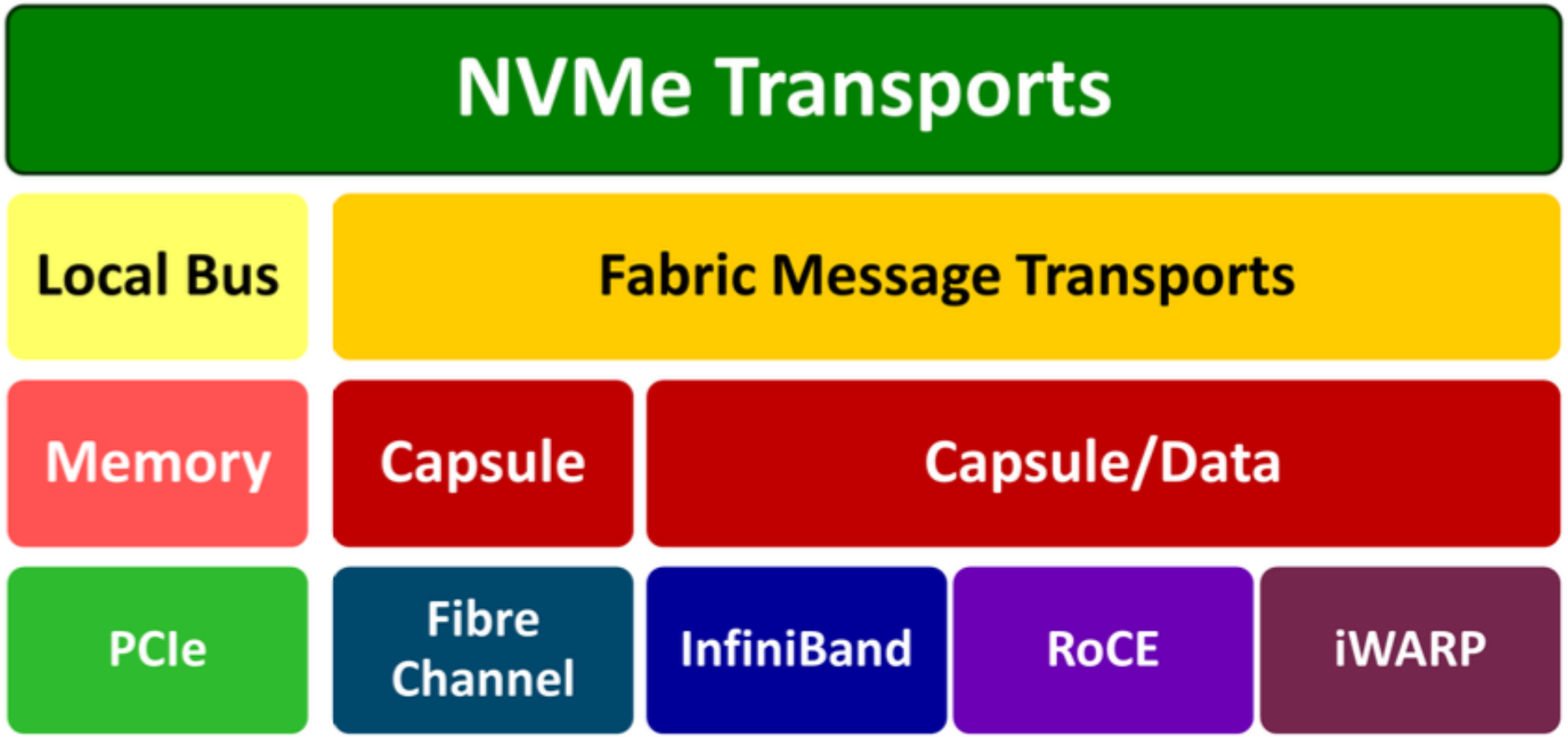
Технология

Differences	PCI Express (PCIe)	NVMe over Fabrics
Identifier	Bus/Device/Function	NVMe Qualified Name (NQN)
Discovery	Bus Enumeration	Discovery and Connect commands
Queueing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

PRP: Physical Region Page (physical memory page address, PCIe transport only)
SGL: Scatter-Gather List (list of locations and lengths for read or write requests)

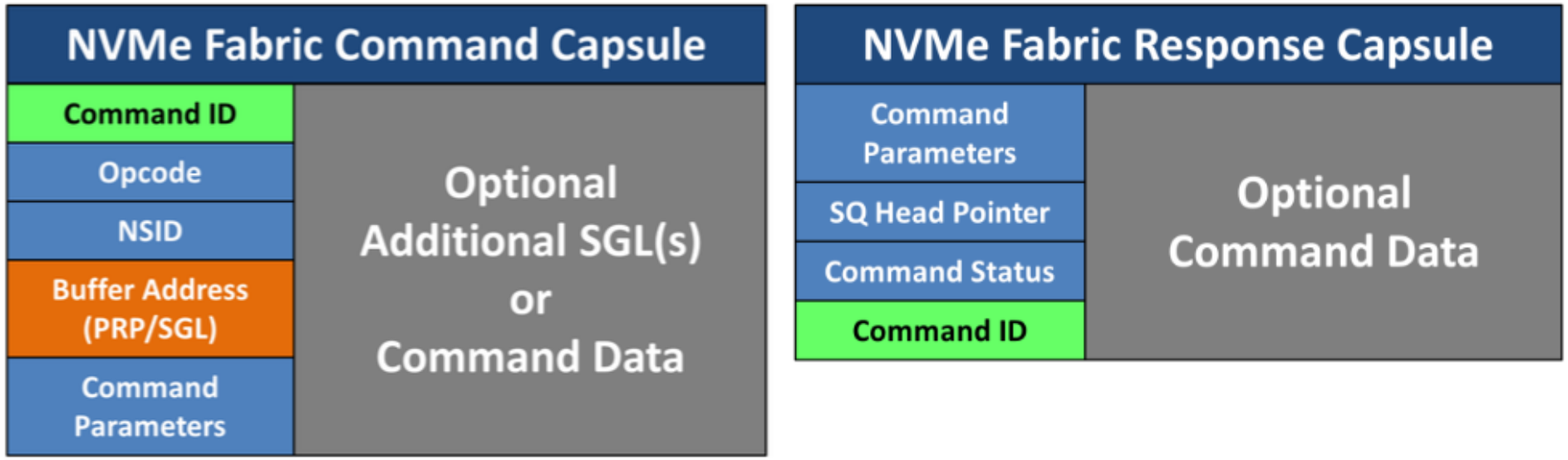
RAIDIX

Технология



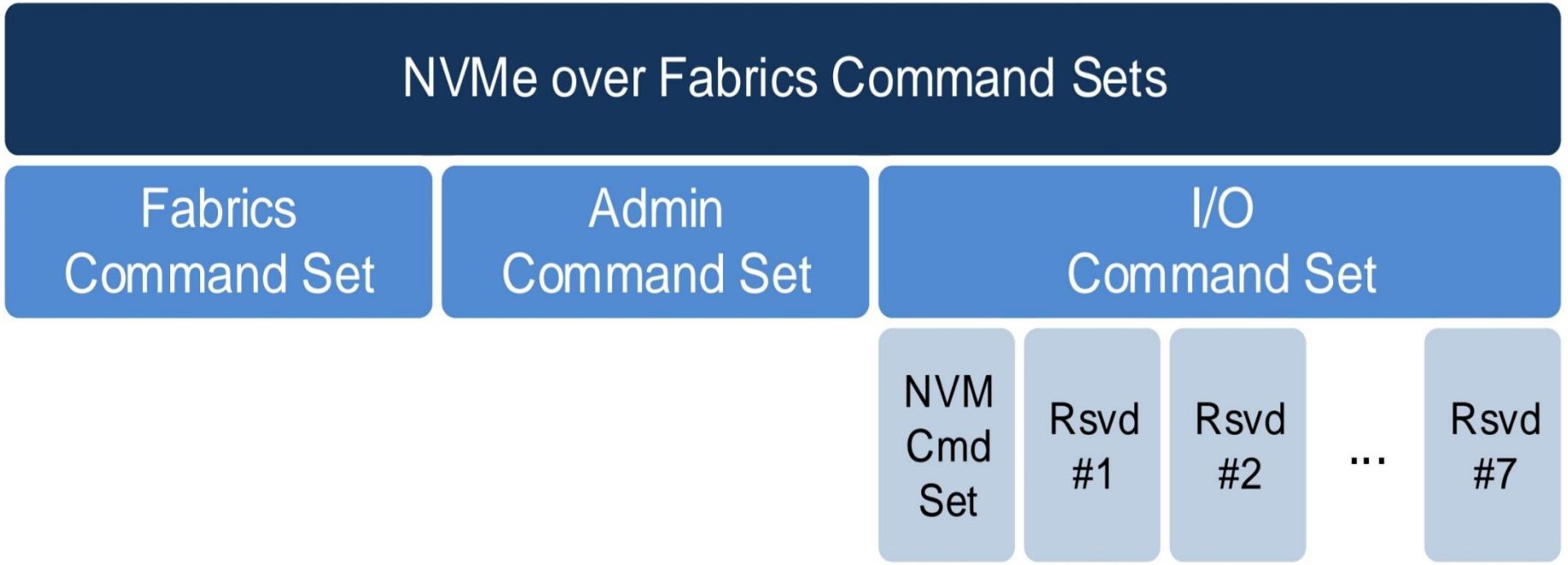
RAIDIX

Технология



RAIDIX

Технология



RAIDIX

Описание кластера

- 32*CPU Intel E5-2620v4
- 2048GB of RAM
- 64* IB 100Gb/ EN 100Gb (Mellanox ConnectX-4) ports
- 32* NVMe HGST SN150

Ограниченное время

RAIDIX

Конфигурация

Target

- 2xE5-2620v4
- 128GB DDR4 2133 of RAM
- 3xNVMe SN150 1600
- 2xMellanox ConnectX-4

Host

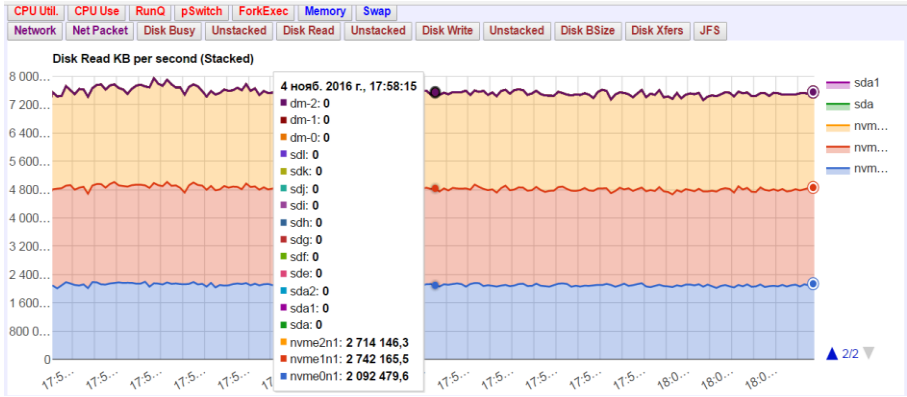
- 2xE5-2620v4
- 128GB DDR4 2133 of RAM
- 2xMellanox ConnectX-4

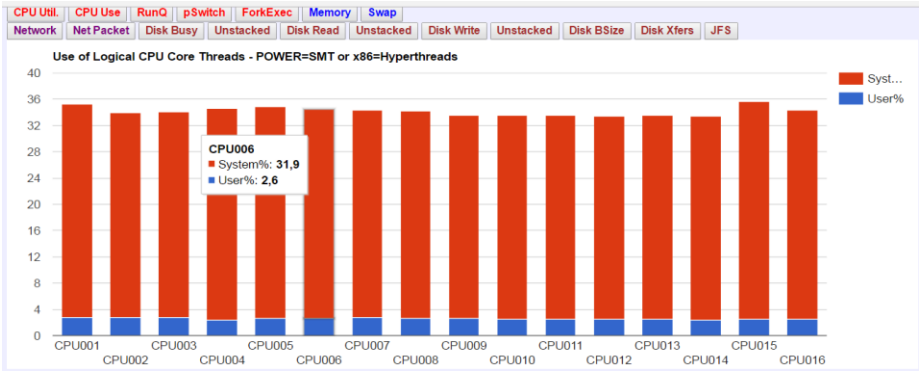
RAIDIX

Test bench description

ПО

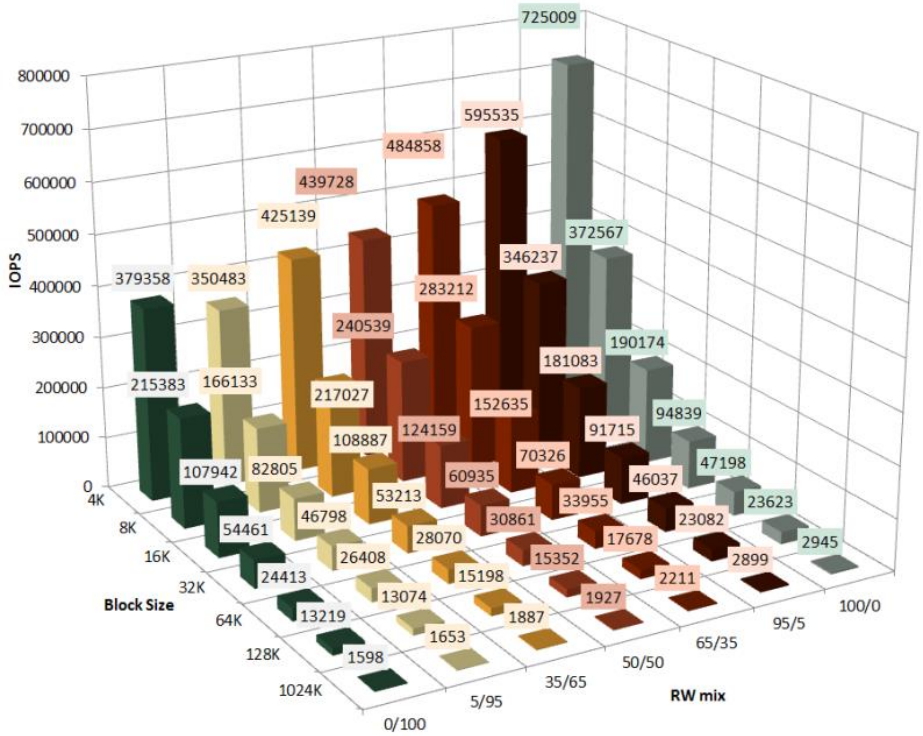
Centos 7.2
fio-2.15
nmon
nvme-cli
nvmet-cli
hdm-3





Процесс тестирования

SNIA SSS PTSe v.1.1
-IOps test
-Throughput test
-Latency test



http://www.truesystem.ru/review/362731/

RAIDIX

Процесс тестирования

Дополнительные параметры теста
Количество потоков – 8
Глубина очереди – 16

Не ограничимся SSS PTSe !
IOps с ограничениями по задержкам (latency)
Latency target = 300usec
Latency percentile=99.999

RAIDIX

Подготовка теста

Шаг 1 – Ядро
Шаг 2– Настройка устройств
Шаг 3- Настройка BIOS, ОС и сети

•HGST PCIe Solid State Drives Performance Measurement Guide
•RHEL 7 Power management Guide
•Performance Tuning Guide for Mellanox Network Adapters

RAIDIX

Подготовка теста

Самые значимые параметры

BIOS

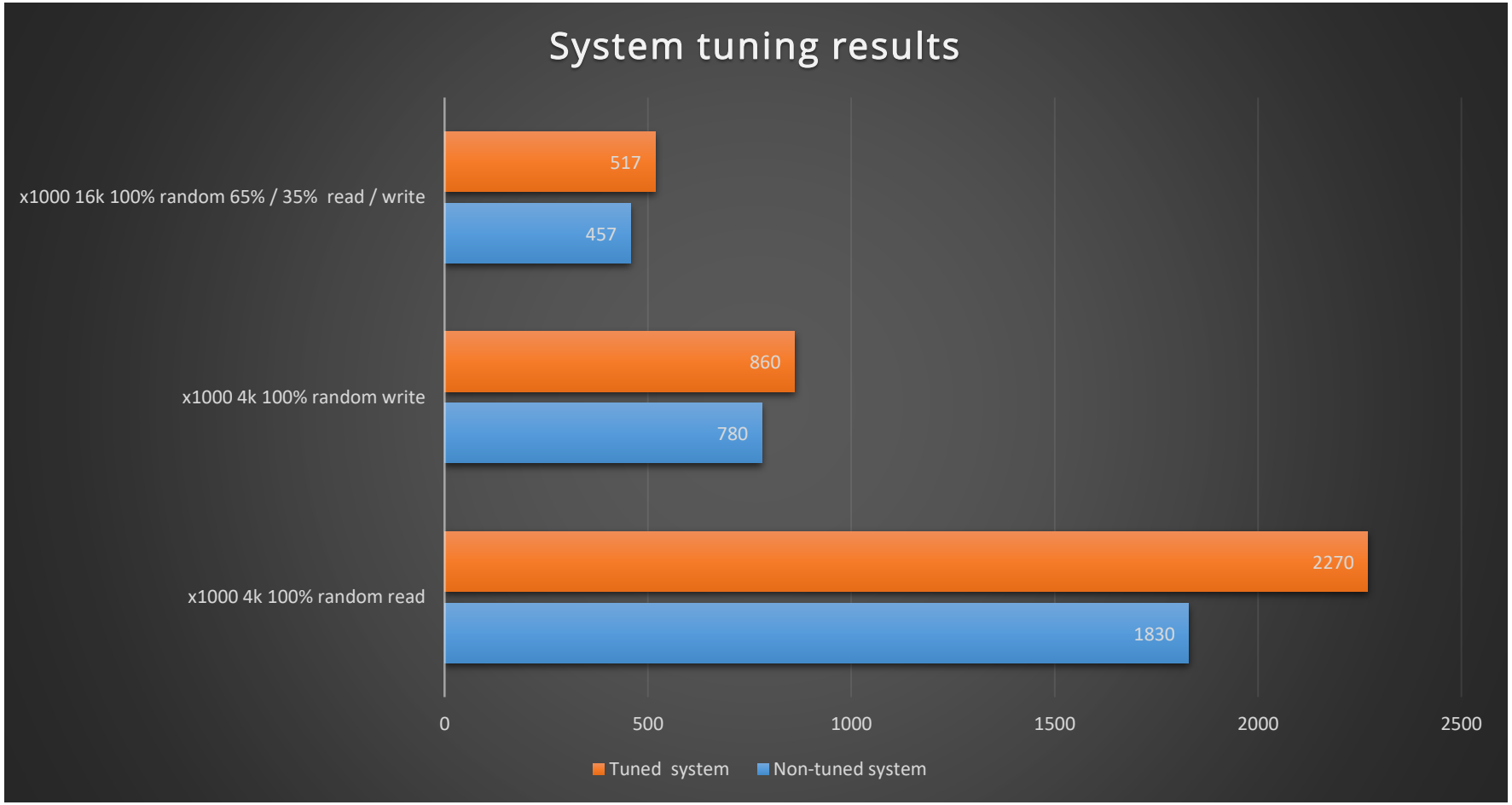
Disable HT
Disable C-States
Disable Node Interleaving

OS

tuned-adm profile latency-performance
cpupower frequency-set --min 2100000
echo “performance” > /sys/devices/system/cpu/cpu/cpufreq/scaling_governor
SMP affinity doesn’t work!

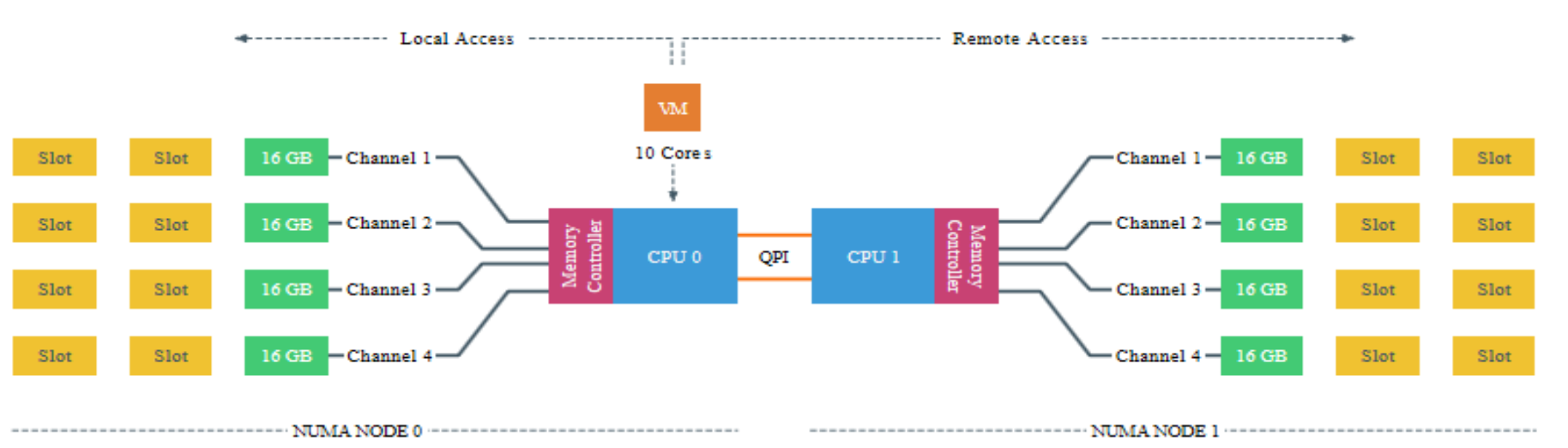
RAIDIX

Результаты настройки



RAIDIX

Numa

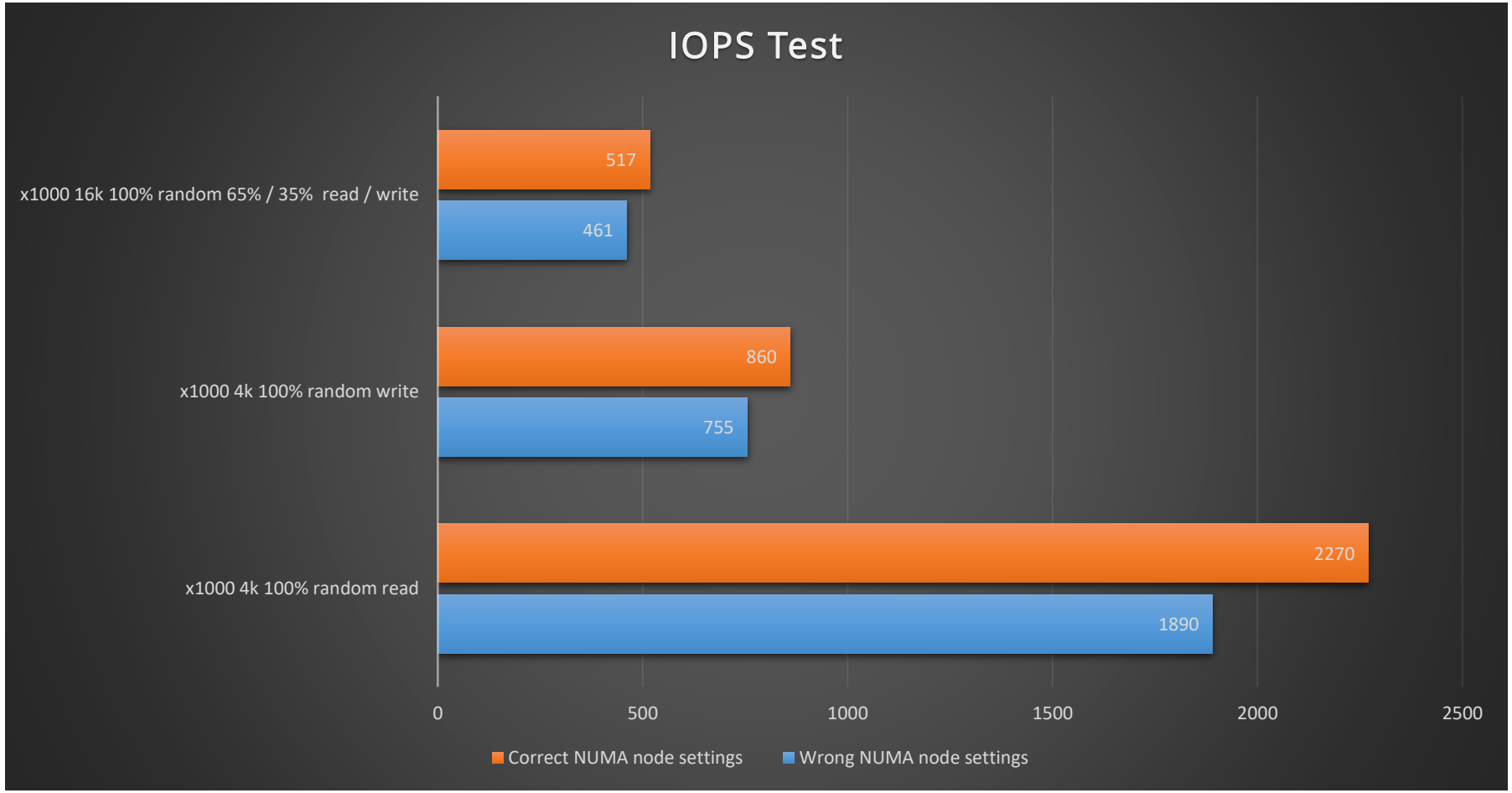


lspci
cat /sys/class/pci_bus/<>/device/numa_node

Fio settings
 numa_cpu_nodes=
 numa_mem_policy=

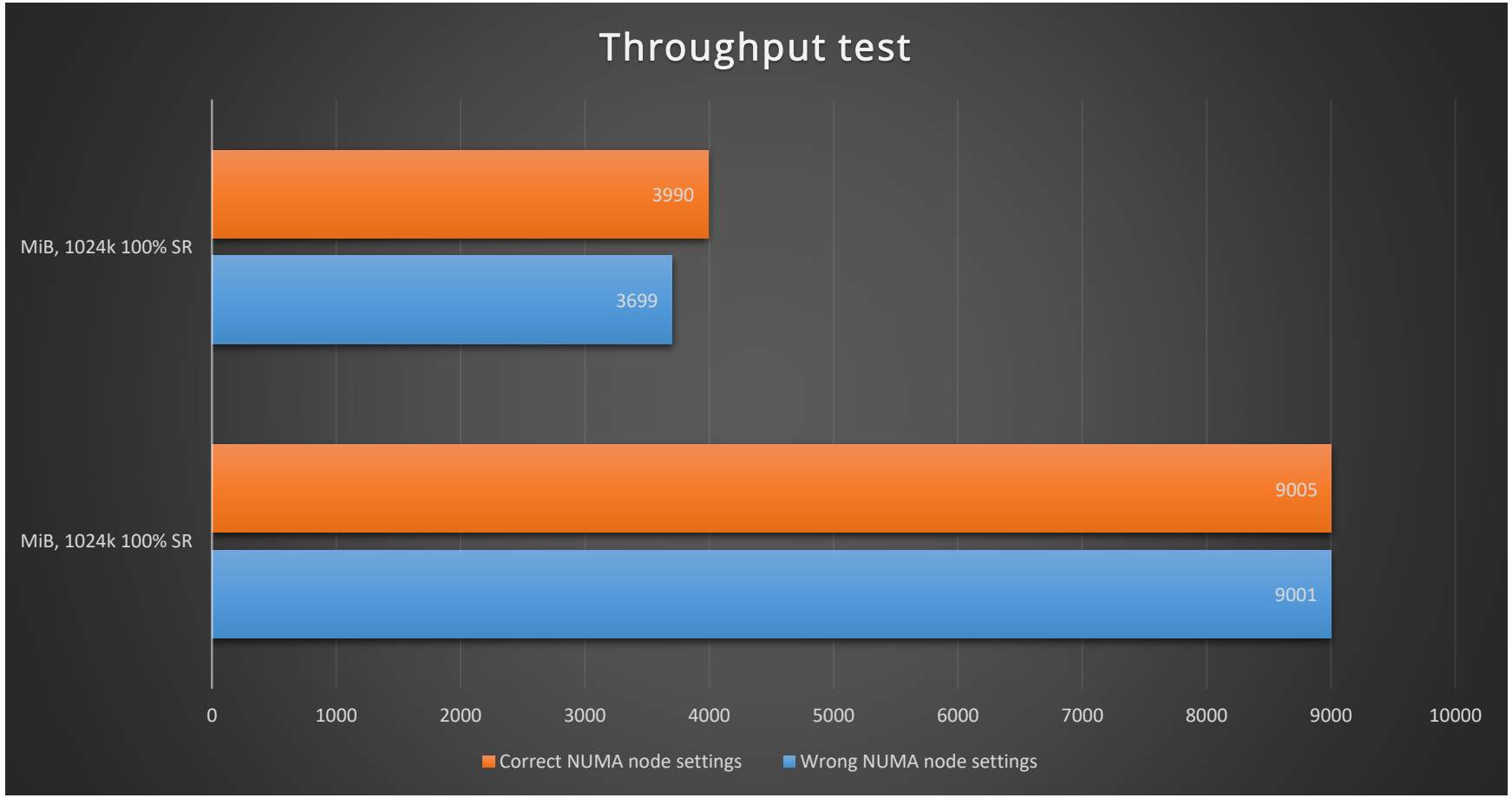
RAIDIX

Влияние Numa



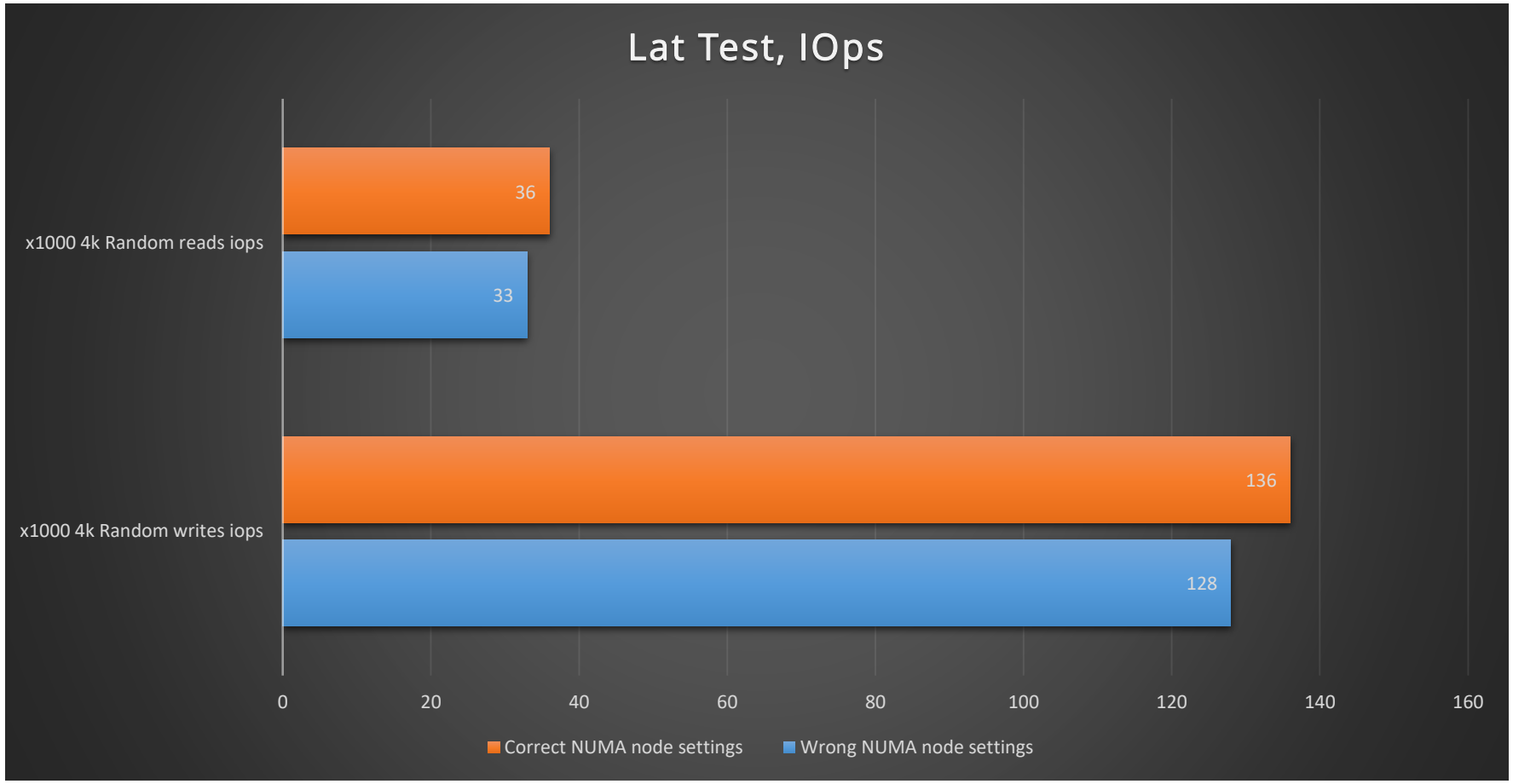
RAIDIX

Влияние Numa



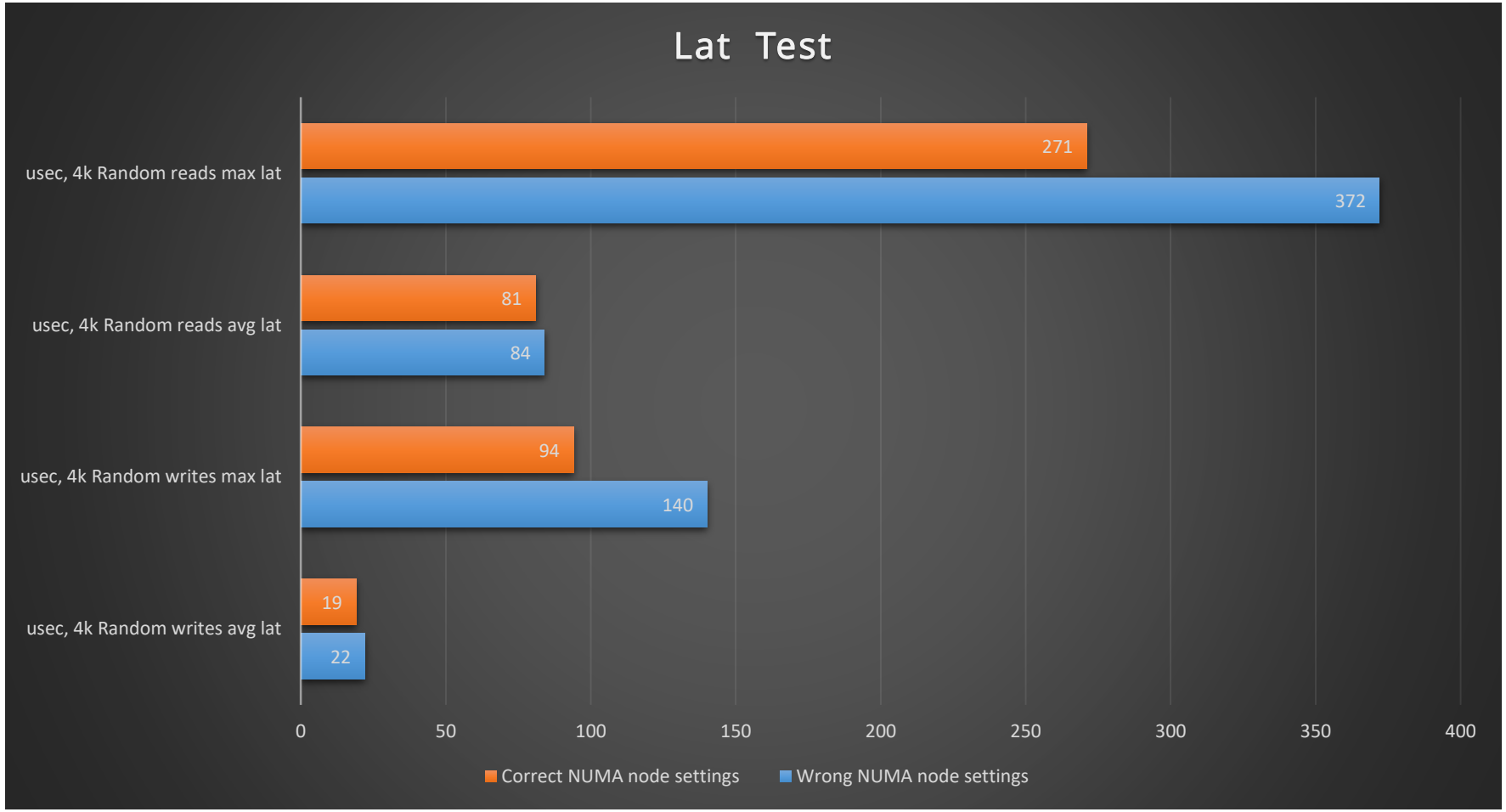
RAIDIX

Влияние Numa



RAIDIX

Влияние Numa



RAIDIX

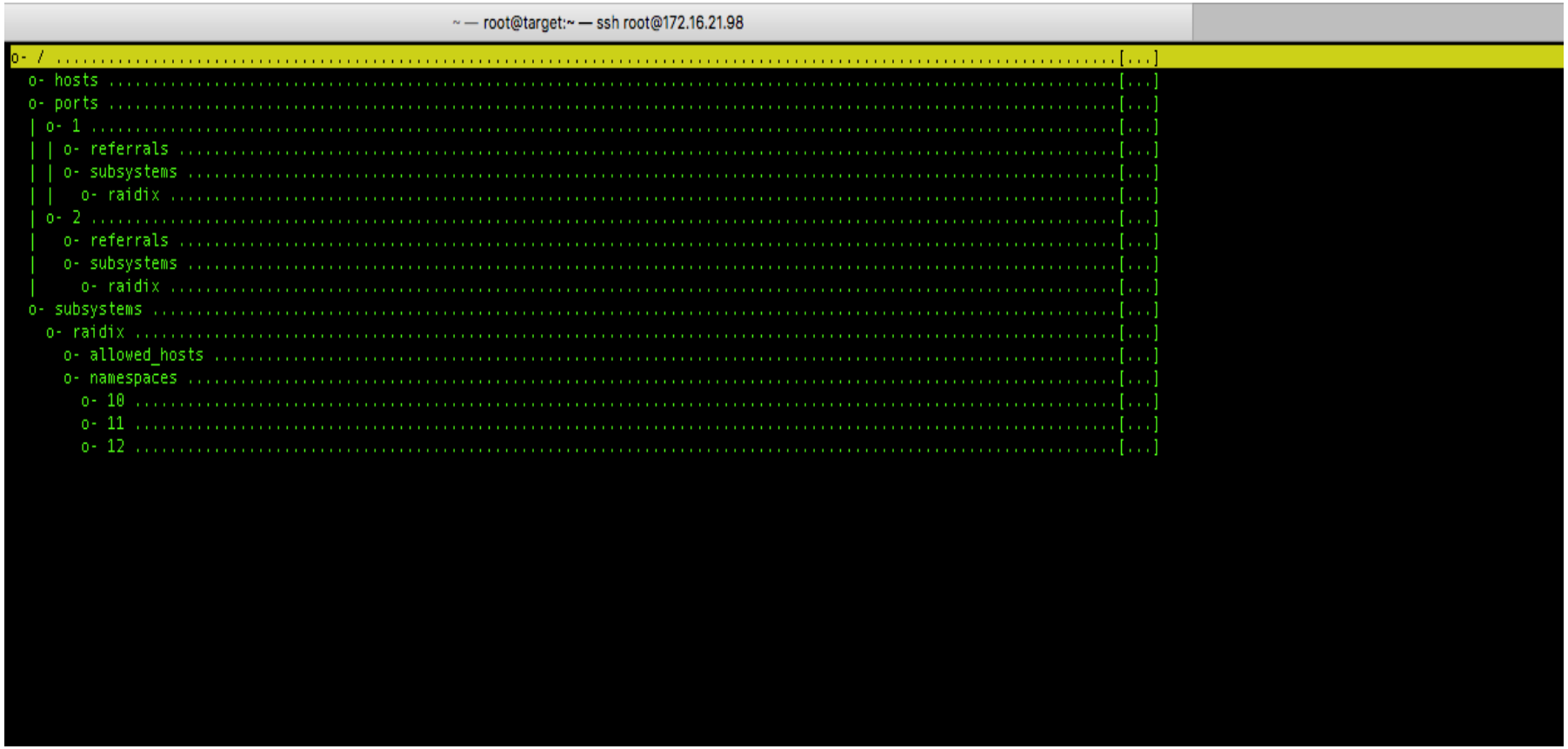
Компоненты таргета NVMf

- NVMe target core: определяет и управляет сущностями NVMe
- Применение команд NVMe admin
- Применение команд NVMe I/O
- Поддержка NVMe over Fabrics: отвечает за обслуживание команд Fabrics (connect, property get/set).
- Сервис обнаружения (discovery) NVMe over Fabrics: отвечает за передачу Discovery log страницы через специальный Discovery-контроллер.

- Поддерживает минимум необходимых команд NVMe:
  - READ, WRITE, FLUSH + admin command
- Начиная с ядра 4.10 поддерживается команда write-zeroes
- Поддержка DSM (aka discard)
- Поддержка ACLs
- Persistent reservations не поддерживаются
- Controller failover???

RAIDIX

Компоненты таргета NVMf



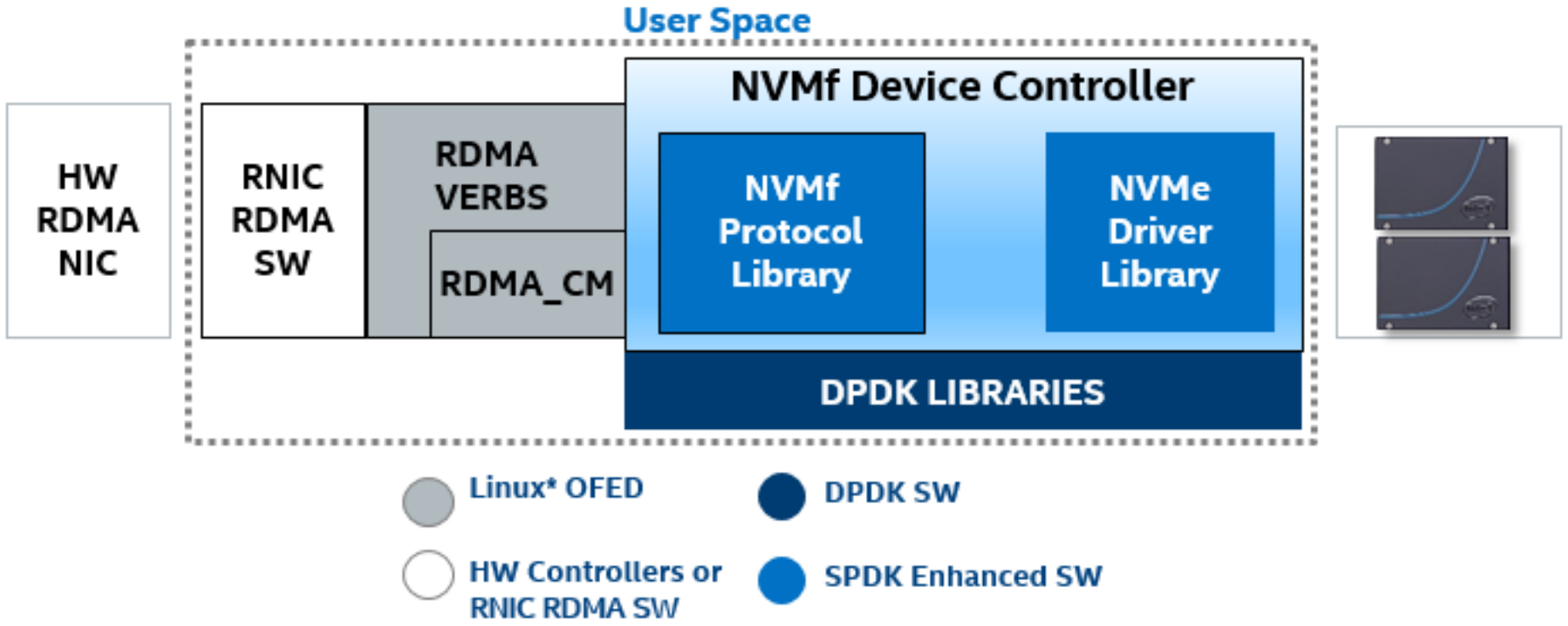
RAIDIX

Компоненты клиента

```
nvme discover -t rdma -a 30.3.1.1 -s 1023

                Discovery Log Number of Records 1, Generation counter 2
                =====Discovery Log Entry 0======
                trtype:  rdma
                adrfam:  ipv4
                subtype: nvme subsystem
                treq:    not specified
                portid:  2
                trsvcid: 1023

                subnqn:  raidix
                traddr:  30.3.1.1

                rdma_prtype: not specified
                rdma_qptype: connected
                rdma_cms:    rdma-cm
                rdma_pkey: 0x0000

nvme connect -t rdma -n raidix -a 30.3.1.1 -s 1023
```

RAIDIX

Intel SPDK



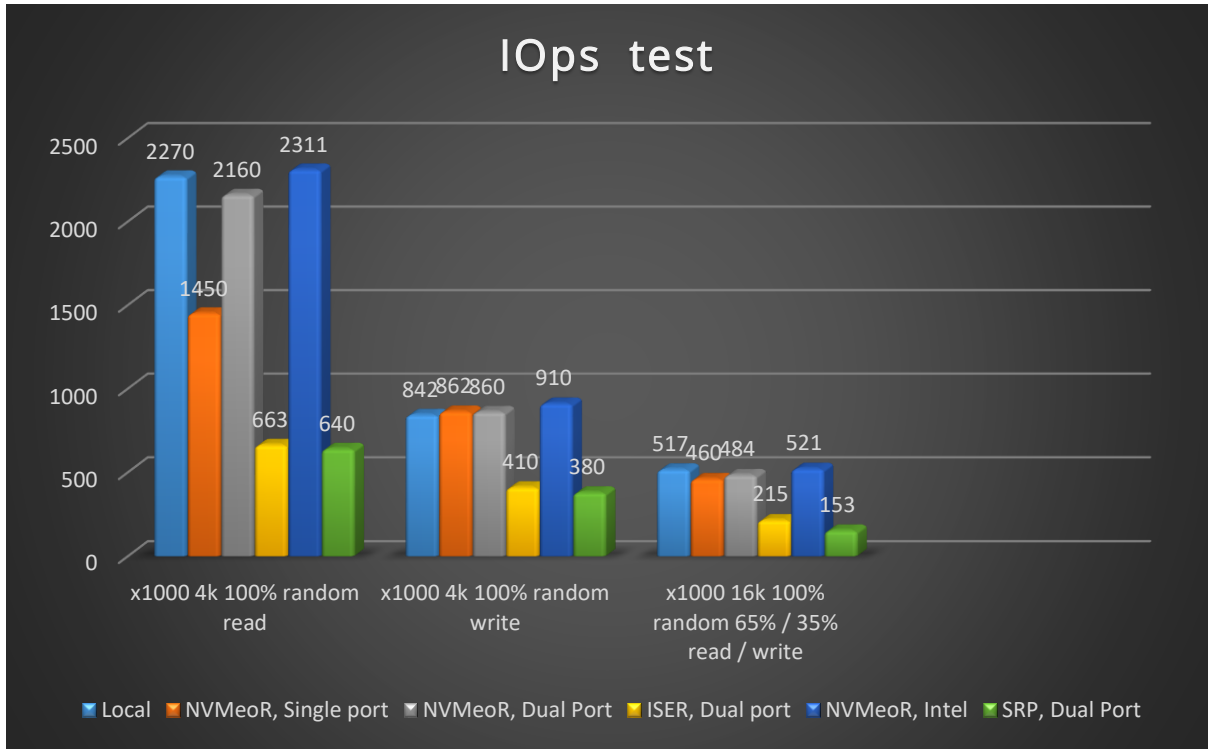
RAIDIX

Компоненты таргета Intel NVMf

```
# NVMf Target Configuration File
#
# Please write all parameters using ASCII.
# The parameter must be quoted if it includes whitespace.
#
# Configuration syntax:
# Leading whitespace is ignored.
# Lines starting with '#' are comments.
# Lines ending with '\' are concatenated with the next line.
# Bracketed ([]) names define sections

[Global]
  # Users can restrict work items to only run on certain cores by
  #  specifying a ReactorMask.  Default ReactorMask mask is defined as
  #  -c option in the 'ealargs' setting at beginning of file nvmf_tgt.c.
  #ReactorMask 0x00FF

  # Tracepoint group mask for spdk trace buffers
  # Default: 0x0 (all tracepoint groups disabled)
  # Set to 0xFFFFFFFFFFFFFFFF to enable all tracepoint groups.
  #TpointGroupMask 0x0

  # syslog facility
  LogFacility "local7"

[Rpc]
  # Defines whether to enable configuration via RPC.
  # Default is disabled.  Note that the RPC interface is not
  # authenticated, so users should be careful about enabling
  # RPC in non-trusted environments.
  Enable No

# Users may change this section to create a different number or size of
#  malloc LUNs.
# This will generate 8 LUNs with a malloc-allocated backend.
# Each LUN will be size 64MB and these will be named
# Malloc0 through Malloc7.  Not all LUNs defined here are necessarily
#  used below.
[Malloc]
  NumberOfLuns 8
  LunSizeInMB 64

# Define NVMf protocol global options
[Nvmf]
  # Set the maximum number of submission and completion queues per session.
  # Setting this to '8', for example, allows for 8 submission and 8 completion queues
  # per session.
  MaxQueuesPerSession 4

  # Set the maximum number of outstanding I/O per queue.
  #MaxQueueDepth 128

  # Set the maximum in-capsule data size. Must be a multiple of 16.
  #InCapsuleDataSize 4096

  # Set the maximum I/O size. Must be a multiple of 4096.
  #MaxIOSize 131072

  # Set the global acceptor lcore ID, lcores are numbered starting at 0.
  #AcceptorCore 0

  # Set how often the acceptor polls for incoming connections. The acceptor is also
  # responsible for polling existing connections that have gone idle. 0 means continuously
  # poll. Units in microseconds.
  AcceptorPollRate  1000

# Define an NVMf Subsystem.
# - NQN is required and must be unique.
# - Core may be set or not. If set, the specified subsystem will run on
#   it, otherwise each subsystem will use a round-robin method to allocate
#   core from available cores.  lcores are numbered starting at 0.
# - Mode may be either "Direct" or "Virtual". Direct means that physical
#   devices attached to the target will be presented to hosts as if they
"nvmf.conf" 108L, 3932C                                                    16,1          Наверху
```

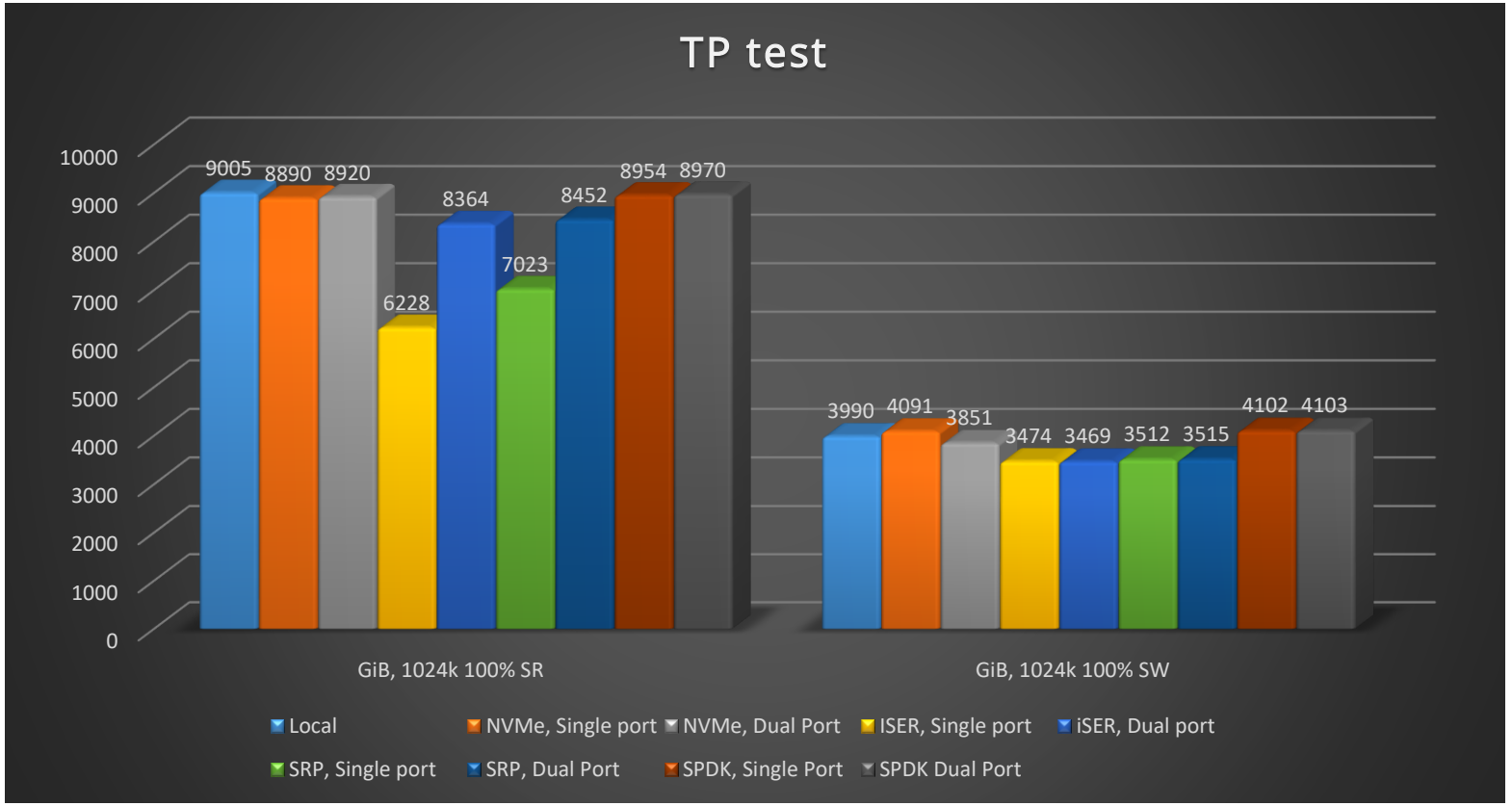
RAIDIX

Результаты тестирования производительности



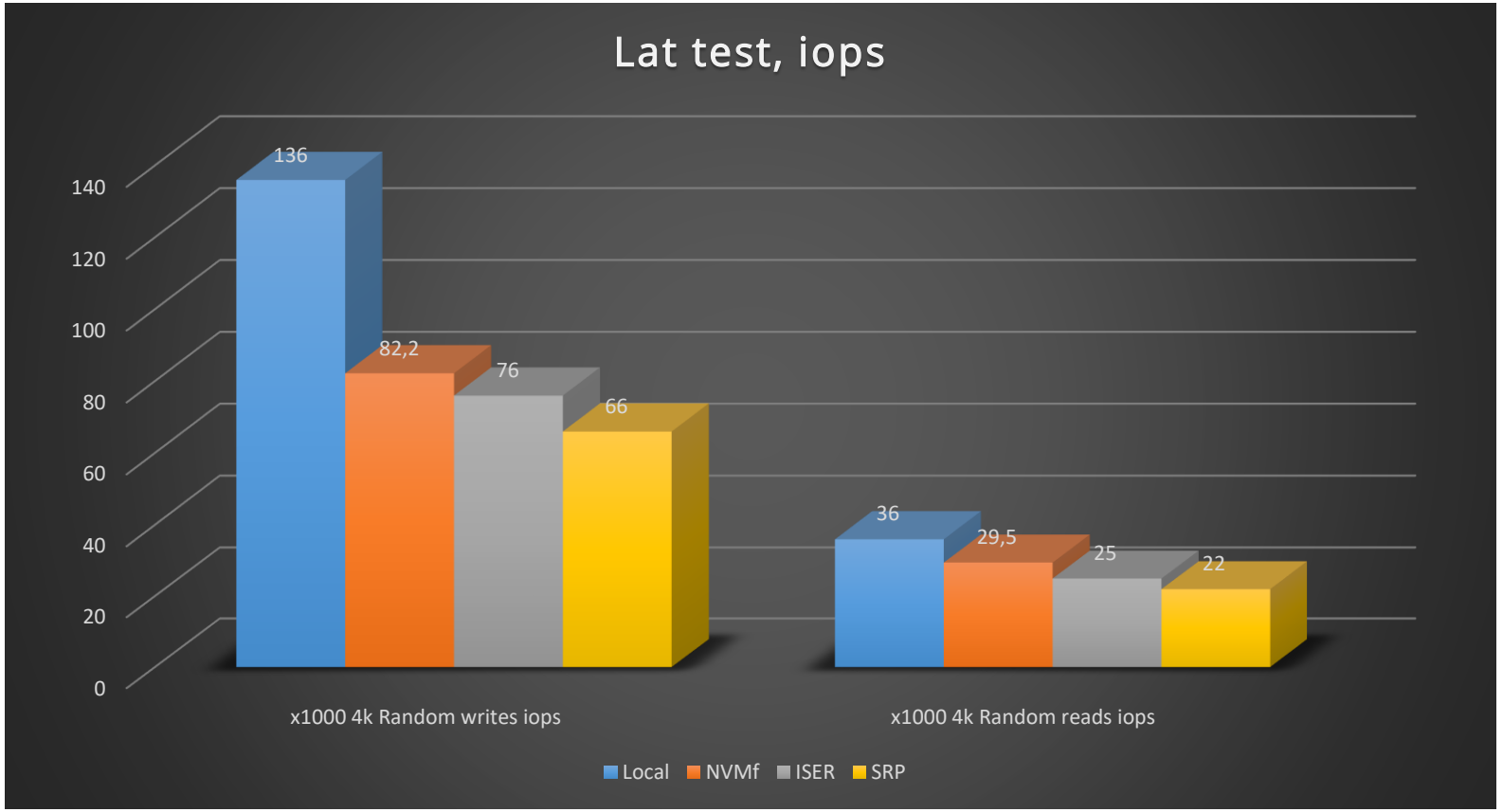
RAIDIX

Результаты тестирования производительности



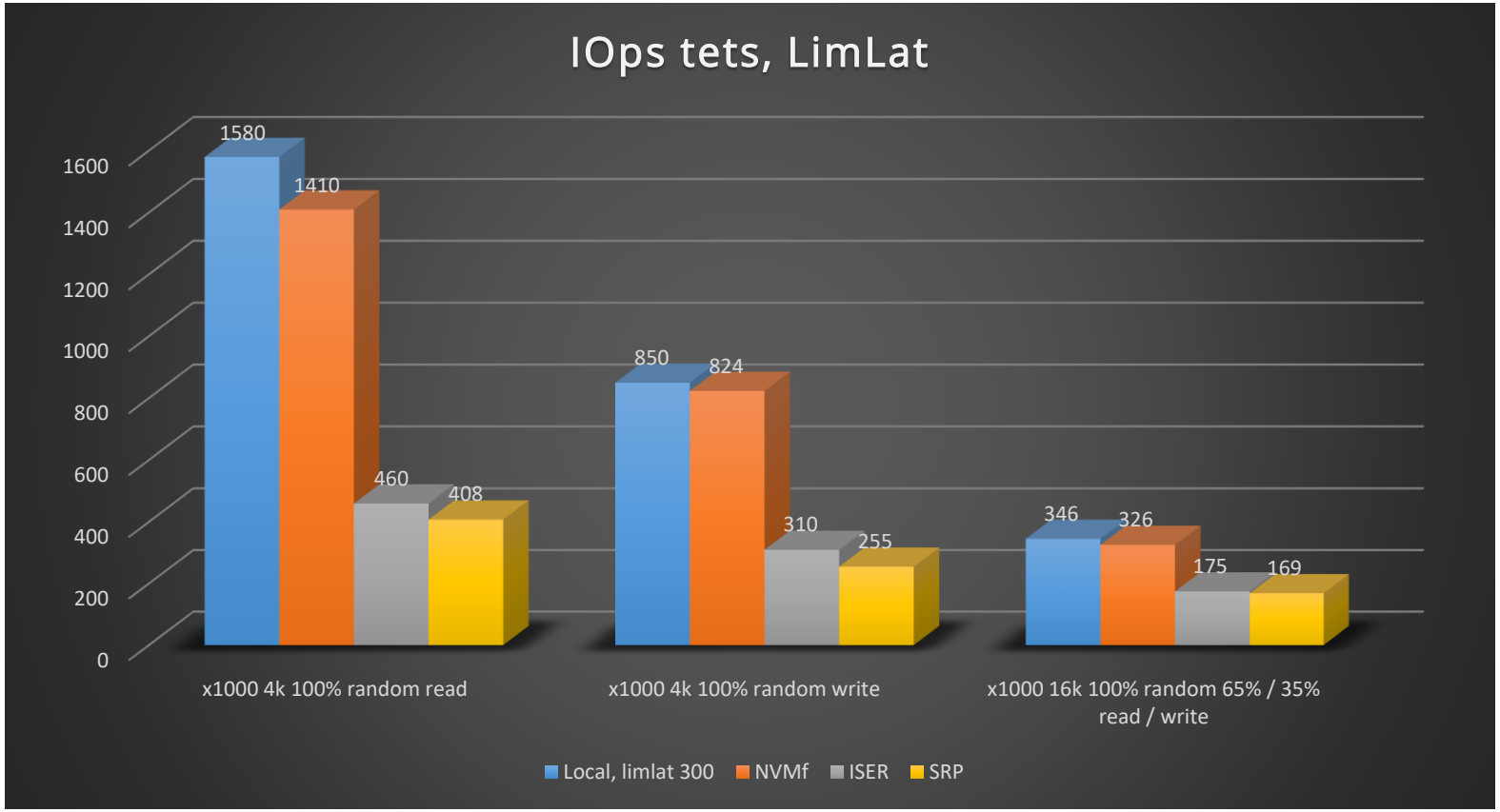
RAIDIX

Результаты тестирования производительности



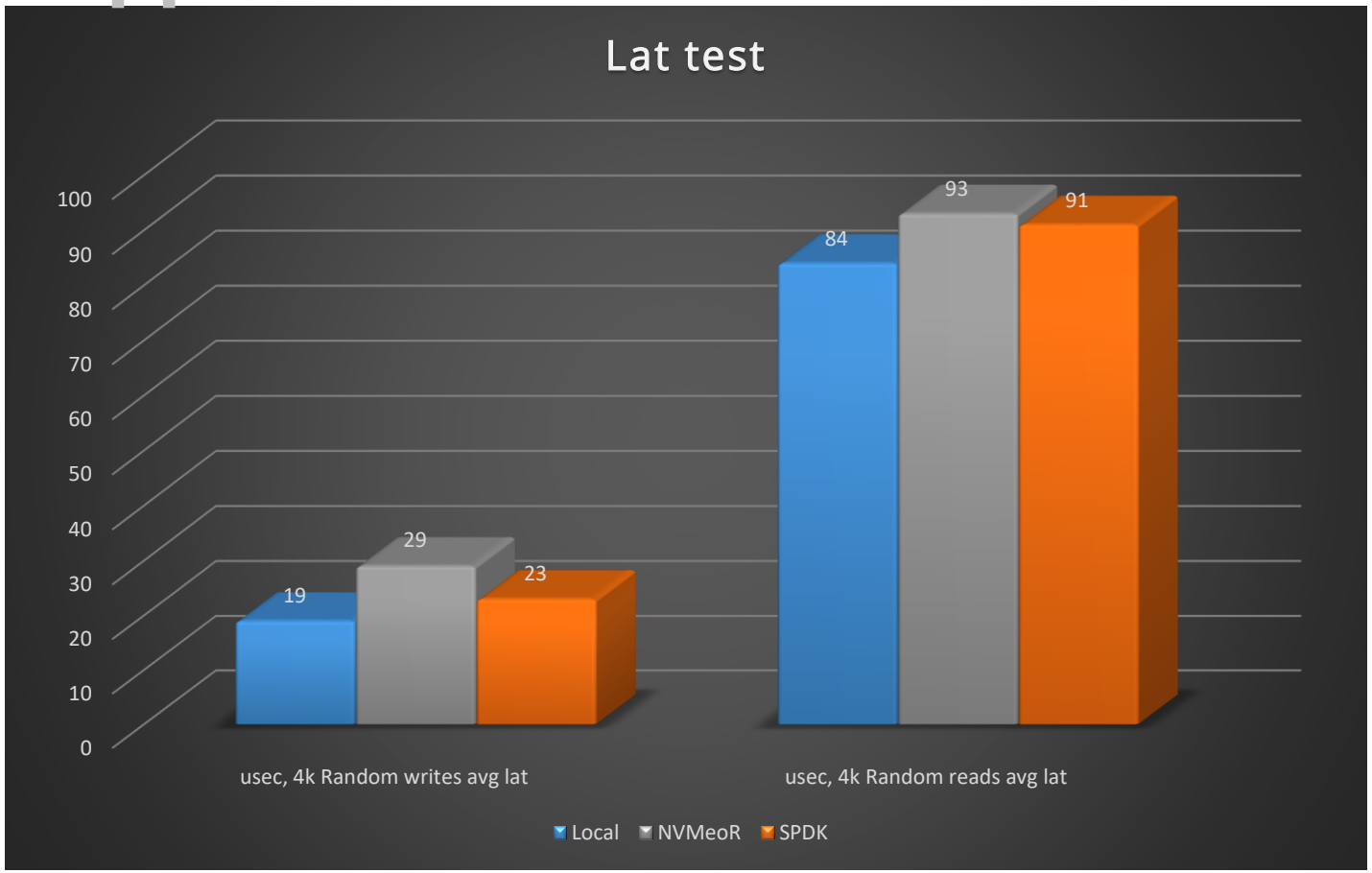
RAIDIX

Результаты тестирования производительности



RAIDIX

Результаты тестирования производительности



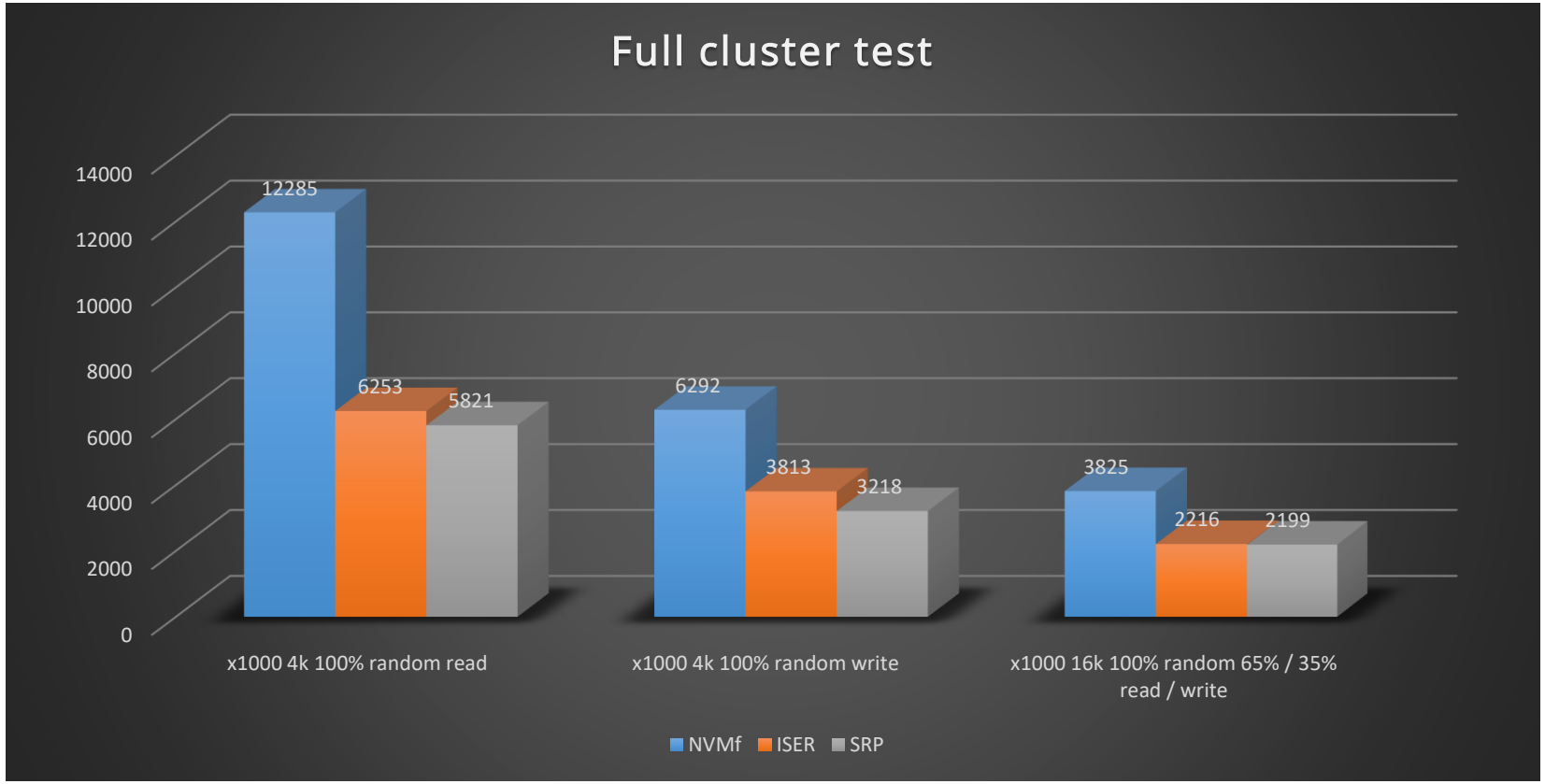
RAIDIX

Результаты тестирования по кластеру

- 12 узлов
- 24 накопителя NVMe
- Ограниченные задержки
- Все накопители распределяются между узлами
- Дополнительный сетевой трафик генерируется с помощью dd

RAIDIX

Результаты тестирования производительности



RAIDIX

Зачем же нам NVMe over Fabric?

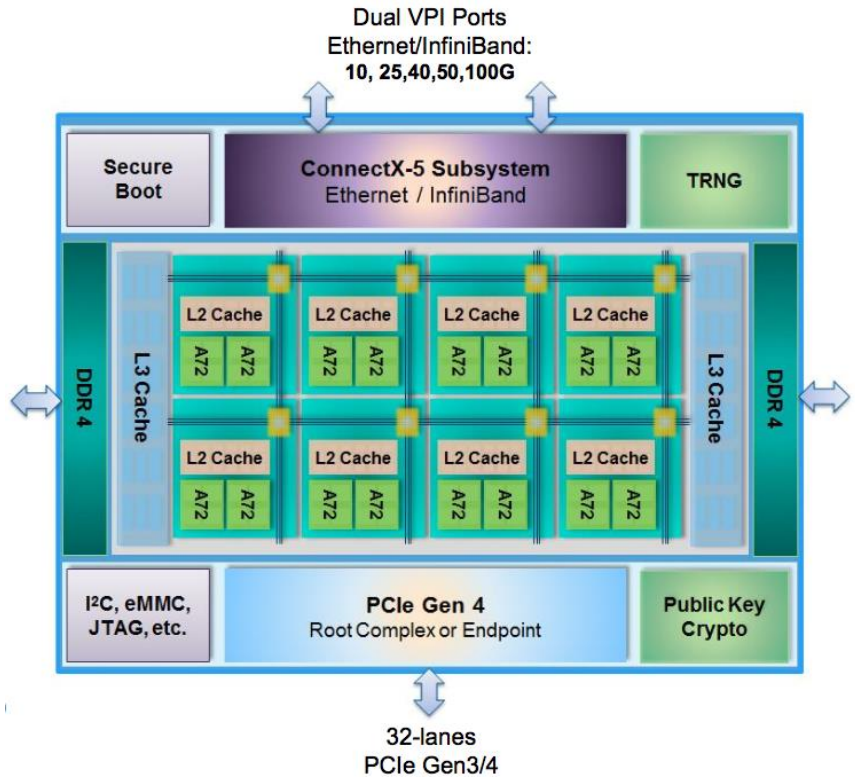
Enterprise
- Гиперконвергентные решения
- Кластеры СУБД
- **Распределенные СХД**
- **“Ускорители”**

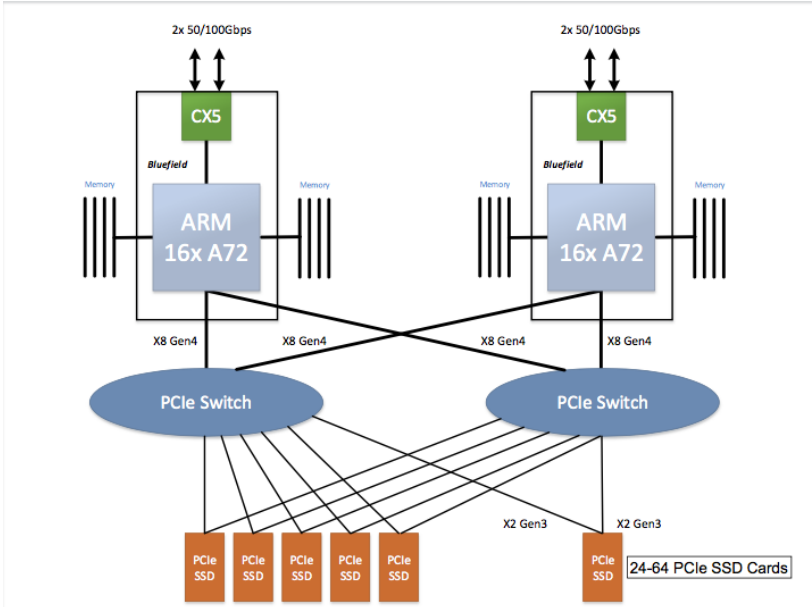
Высокопроизводительная аналитика

HPC
- **BurstBuffer**
- Визуализация и симуляция

RAIDIX

Оборудование





RAIDIX

Результаты

- NVMe over fabrics демонстрирует лучшие результаты по сравнению с другими протоколами хранения на базе RDMA
- При этом ожидаемых результатов я не достиг
- Дополнительные задержки выше, чем показывают тесты от вендора.

RAIDIX