

UPPER MID-RANGE · ПРОГРАММНО-ОПРЕДЕЛЯЕМАЯ СХД

Два контроллера. Оба в работе.

Отказ одного узла не останавливает доступ.
Клиент видит каждый том на обоих контроллерах.

до 14,4 ПБ

объём хранения

iSCSI · iSER · FC

Ethernet и Fibre Channel

RAID 5 / 6 / 7

одна, две или три чётности

СХД из двух контроллеров

КУБ — программно-определяемая СХД класса Upper Mid-range.
Отдаёт блочные тома серверам и виртуализации.
Оба контроллера обслуживают ввод-вывод одновременно.

до 14,4 ПБ

Двухпортовые диски NVMe

Путь к каждому диску есть с обоих контроллеров.

Каждый том доступен по нескольким путям

Multipath по Ethernet или Fibre Channel на оба узла.

Консоль и API

Управление из браузера и автоматизация через REST.

Что уже есть

ДОСТУП**Symmetric Active-Active**

Оба контроллера работают одновременно, без простоя и потери скорости.

ДОСТУП**iSCSI / iSER / FC**

Блочный доступ по Ethernet и Fibre Channel к любой инфраструктуре.

УПРАВЛЕНИЕ**Web и REST API**

Консоль в браузере и полная автоматизация по API.

НОСИТЕЛИ**Двухпортовые диски NVMe**

Высокая скорость и резервирование путей к диску.

ЗАЩИТА**RAID 5 / 6 / 7**

Одна, две или три диска чётности.

ТОМА**Толстые тома**

Ёмкость выделяется сразу.
Производительность предсказуема.

ЗАЩИТА**Hot-spare**

Отказавший диск меняется из резерва без администратора.

МАСШТАБ**До 14,4 ПБ**

Объём хранения системы.

Что появится дальше

Толстые тома остаются. К ним добавляются тонкие тома, снимоты и NVMe-oF.

КОНЕЦ 2026

Тонкие тома

Место на Сторе занимаетса по мере записи, а не целиком в момент создания тома.

КОНЕЦ 2026

Снимоты

Снимок состояния сразу. Откат к нему за секунды.

КОНЕЦ 2026

NVMe-oF

Поддержка NVMe over Fabrics: доступ к томам по сети на протоколе NVMe.

Два слоя на каждом контроллере



Стор, RAID, резерв

Стор собирается из RAID-групп. Потолок системы — до 14,4 ПБ. Конкретный Стор занимает столько, сколько в него поставлено дисков.

5

переживёт потерю одного диска

Одна чётность. Страйп задаётся на группу.

6

переживёт потерю двух дисков

Две чётности.

7

переживёт потерю трёх дисков

Три чётности.

Толстый том

Заявленный размер сразу вычитается из Стора. Производительность предсказуема.

Hot-spare

Резерв лежит вне групп. После отказа диск подменяется, идёт пересчёт чётности, группа снова обслуживается с обоих контроллеров.

На чём измеряли

Два одинаковых контроллера, Node 1 и Node 2. Оба онлайн.

КАЖДЫЙ УЗЕЛ

Intel Xeon Gold 6548Y+

32 ядра. Emerald Rapids. Память 512 ГБ.

СЕТЬ

200 Гбит/с на фронт и 100 Гбит/с на интерконнект

Mellanox ConnectX-6 Dx. К клиентам 200 Гбит/с, между контроллерами 100 Гбит/с.

ДИСКИ

20 двухпортовых Samsung PM1733

U.2, по пути на каждый контроллер. 10 дисков по 3,84 ТБ и 10 дисков по 1,92 ТБ. Сырой объём 57,6 ТБ.

Как был собран замер

СТОП

Две группы RAID 6 по 10 дисков

Страйп 16 КиБ. Ёмкость 41,9 ТиБ, занято 37,7 ТиБ. Потолок продукта — до 14,4 ПБ.

10 дисков по 3,49 ТиБ и 10 дисков по 1,75 ТиБ. 28 толстых томов по 1,35 ТиБ.

НАГРУЗКА

Случайные 4К, 70/30, один час

FIO, 28 томов, по два активных пути на том. Прочитано 48,1 ТиБ, записано 20,6 ТиБ.

Таких прогонов было несколько, цифры совпали. Итог следующего слайда — с инициатора по всем 28 томам.

СЛУЧАЙНОЕ 4К · 70/30 · ЧАС

5,1 млн IOPS

Соотношение 3,59 к 1,54 млн — те самые 70/30. Несколько прогонов дали те же цифры.

	Чтение	Запись
IOPS	3,59 млн	1,54 млн
Поток	13,7 ГиБ/с	5,86 ГиБ/с
Средняя задержка	1,28 мс	1,68 мс
p50	1,21 мс	1,50 мс
p99	3,56 мс	4,55 мс
p99.9	5,21 мс	7,11 мс

Нагрузка легла на оба узла

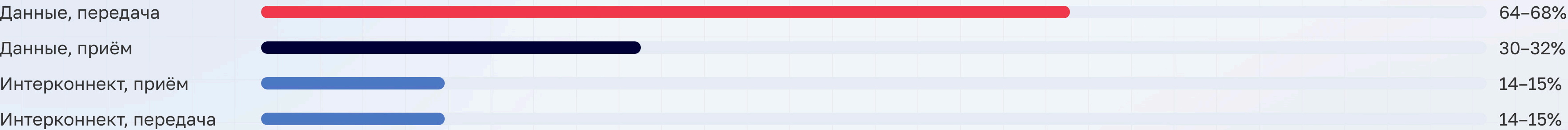
Счётчики block device по 28 томам, плато в течение часа. Это задержки на обработку IO команд в нашем стеке. Эти цифры не включают задержку клиента и транспортной инфраструктуры.

	Чтение	Запись	Задержка чтения	Задержка записи
Node 1	2,0–2,1 млн	0,85–0,9 млн	0,13 мс	0,40 мс
Node 2	1,8–2,0 млн	0,80–0,88 млн	0,13 мс	0,40 мс

Задержка на контроллере ниже клиентской: в 1,3–1,7 мс на FIO сидят очередь, сеть и путь до инициатора. Ядра подсистемы передачи данных обоих узлов — 90–98% весь час.

Порт 100 Гбит/с не упёрся в потолок

Доля линка на плато. Передача по данным — это чтение клиента. Картина на обоих контроллерах одинаковая.



Около двух третей порта данных уходит клиенту на чтение, около трети приходит как запись. Интерконнект занят слабо: пара синхронизируется, основной поток идёт наружу.

РЕШЕНИЕ РАЗРАБОТАНО КОМПАНИЕЙ

NOVATOR

Upper Mid-range СХД КУБ — разработка компании «НОВАТОР».