

Т-Платформы – перспективные исследования и разработки

Корж Антон

- Вычислитель (процессоры + ускорители)
- Память
- **Интерконнект**
- **СХД**
- Управление
- Инфраструктура
- Софт

- Новые топологии (Butterfly, Dragonfly, ...)
- Адаптивная маршрутизация
- Собственный MPI
- Цель создание систем до 100к узлов с компромиссом цена-производительность (10-15% стоимости при 100% пропускной способности ALL2ALL)

- СХД занимает 15-20% от стоимости НРС проекта.
 - По консервативным оценкам размер моделей и проблем увеличивается в 2 раза каждый год, что ведет к спросу на модули хранения
 - Мировой рынок СХД для НРС >3 \$млрд.
 - СХД наиболее медленная часть НРС и является ключевым фактором улучшения технико-экономических показателей всего НРС комплекса

- Производительность
 - Интенсивные методы (алгоритмы и параллельные команды) ->
 - Сделать больше за меньшее или тоже время
 - Уменьшить разрыв в производительности между вычислением и хранением
- Надежность
 - Быстрое время восстановления -> Минимизация простоев
 - RAID 6 и 7, борьба с скрытыми ошибками - > Обеспечение целостности и доступности к данным
- Затраты

- Дата анонса 06.2013
- Основные характеристики
 - Программный RAID 0,10,6
 - Интерфейсы IB, FC, iSCSI
 - 24-60 SAS,SATA или SSD
 - Два контроллера, Active-Passive
 - 3-7 GB/s
 - 2-8 U

SA2364, SA2368



SAI360



SA2224

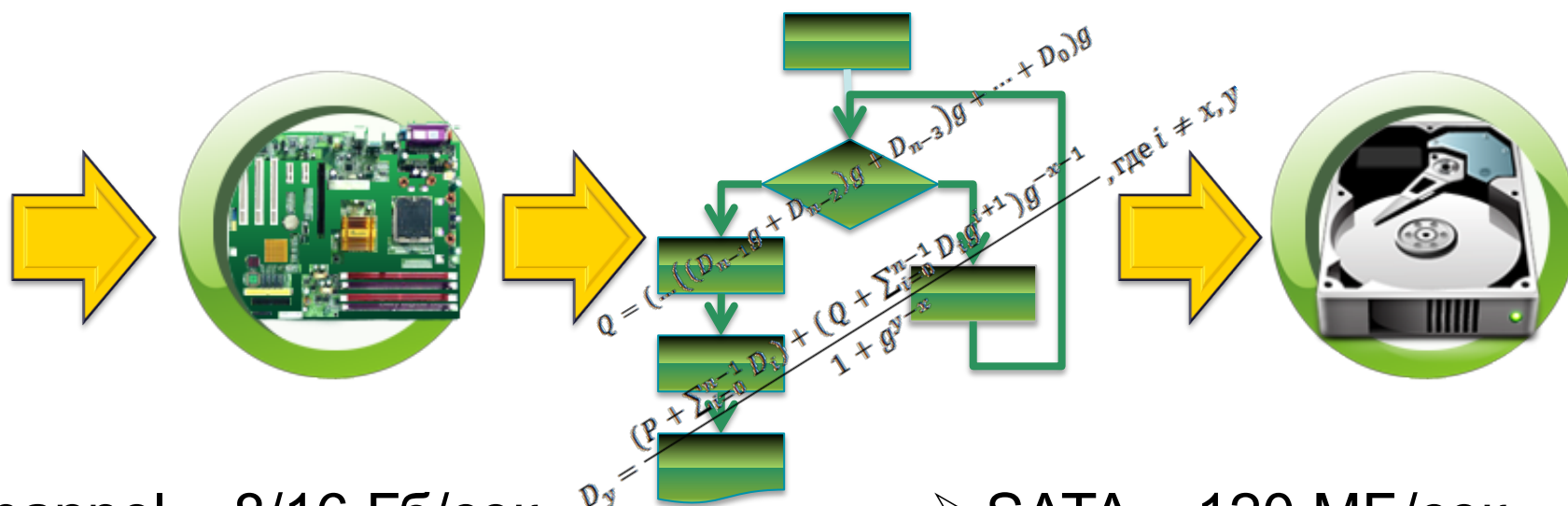


Основные характеристики

	SA2368	SA2364	SA2224
Производительность R/W RAID 6	до 7,5 GB/s	до 4,5 GB/s	до 3GB/s
Размер кэша	до 120 GB		
Тип дисков	SATA, SAS, SSD		
Количество дисков	60		24
Уровень RAID	0 6 10		
Полная отказоустойчивость	Да		
Режим работы контроллеров	Active-Passive		
Защита от скрытых ошибок	Да (во время восстановления)		
Количество дисков в RAID6	до 32		
Сырая емкость	до 240TB		
Полезная емкость	>93% от сырой		
Поддержка MPIO	VMware ESX, Windows Server, Linux, Solaris, AIX, MAC OS X		
Интерфейсы	InfiniBand, Fibre Chanel, iSCSI		
Режим Advanced Reconstruction	Да		
Размеры	8U	6U	2U

Что внутри

Программный RAID на Linux платформе с поддержкой различных интерфейсов и типов дисков



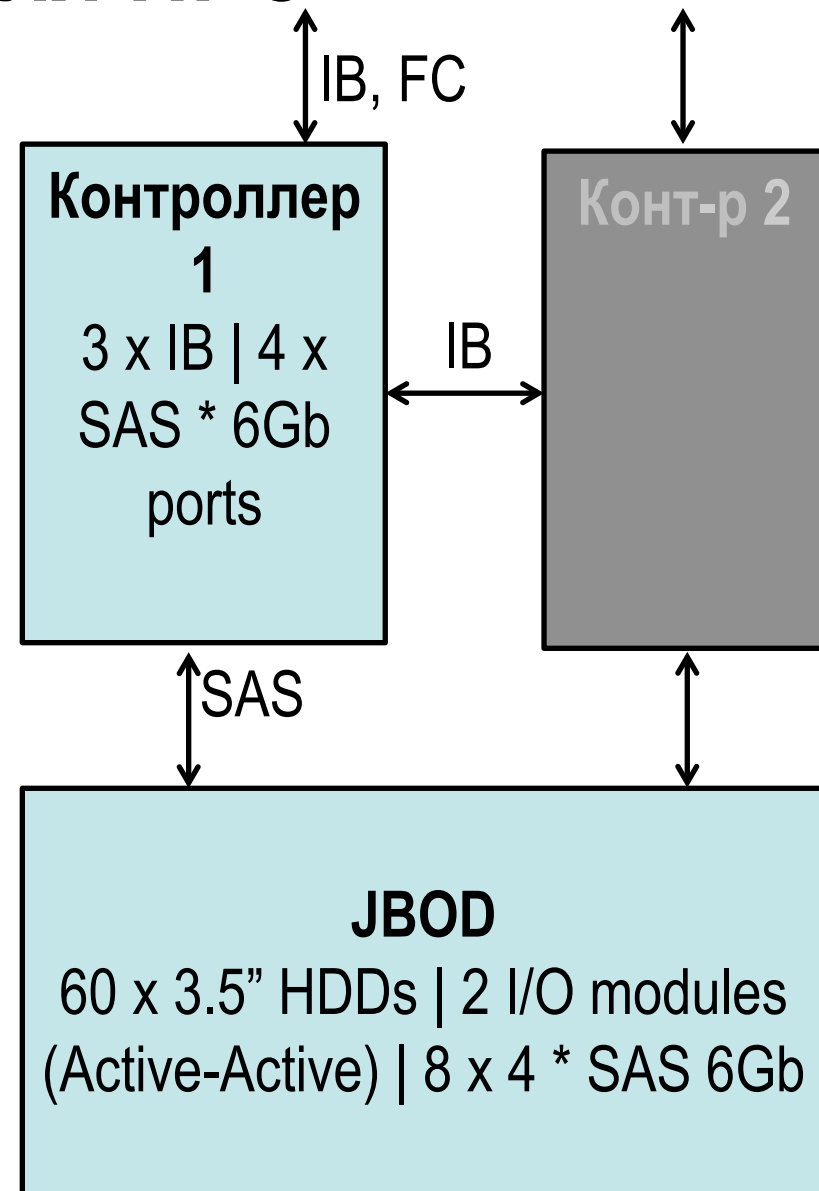
- Fibre Channel – 8/16 Гб/сек
- InfiniBand – 40/56 Гб/сек
- iSCSI – 1/10 Гб/сек

- SATA – 120 МБ/сек
- SAS – 200 МБ/сек
- SSD - 250 МБ/сек

Семейство систем T-Store SA

Архитектура блочной СХД для HPC

- T-Store SA, DDN SFA, EMC 7500 – классические «HPC» building blocks
 - Отдельные контроллеры с внешними интерфейсами IB, FC
 - Дисковые модули подключаемые по SAS
- Преимущества:
 - Большое количество внешних интерфейсных портов
 - Возможность наращивать производительность линейно, подключая дополнительные Jbod
 - Уменьшение общего количества систем при увеличении их размера



Что внутри 2

- Недостижимая конкурентами производительность RAID6 и 7.3
 - Патентованный метод вычисления, высокий параллелизм*: Запись и чтение с одинаковой скоростью
 - Минимальное время инициализации RAID
 - Стабильно высокая производительность при реконструкции
 - Настраиваемый размер кэш памяти и неограниченный размер логического тома

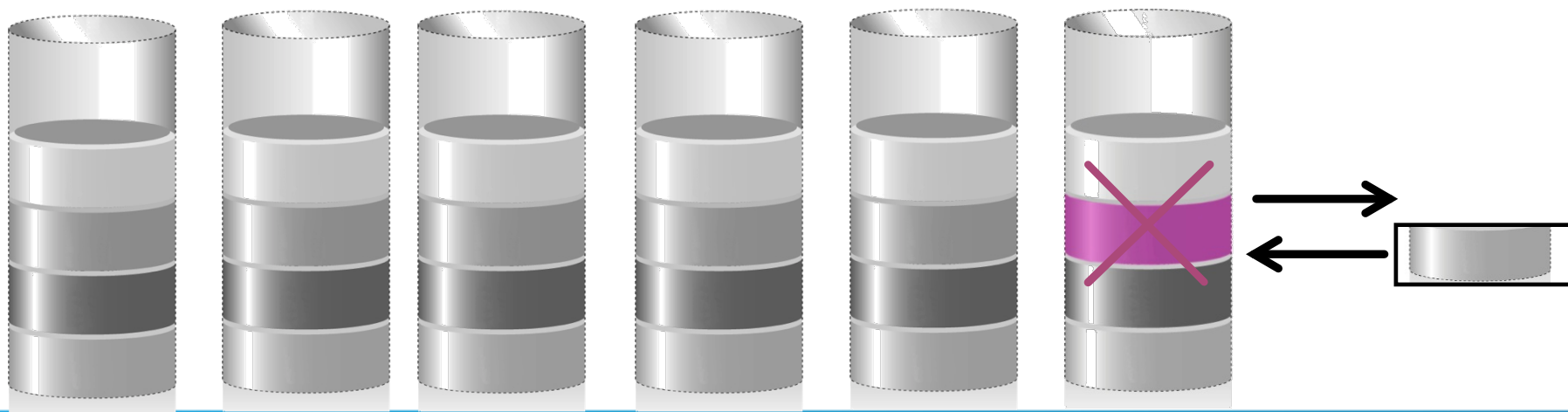
Частичная реконструкция

■ Свойства

- Каждый диск бьется на логические блоки
- В случае восстановления восстанавливается не диск целиком, а только сбойный блок

■ Преимущества

- Процесс восстановления идет быстрее в 2 048 раз!!



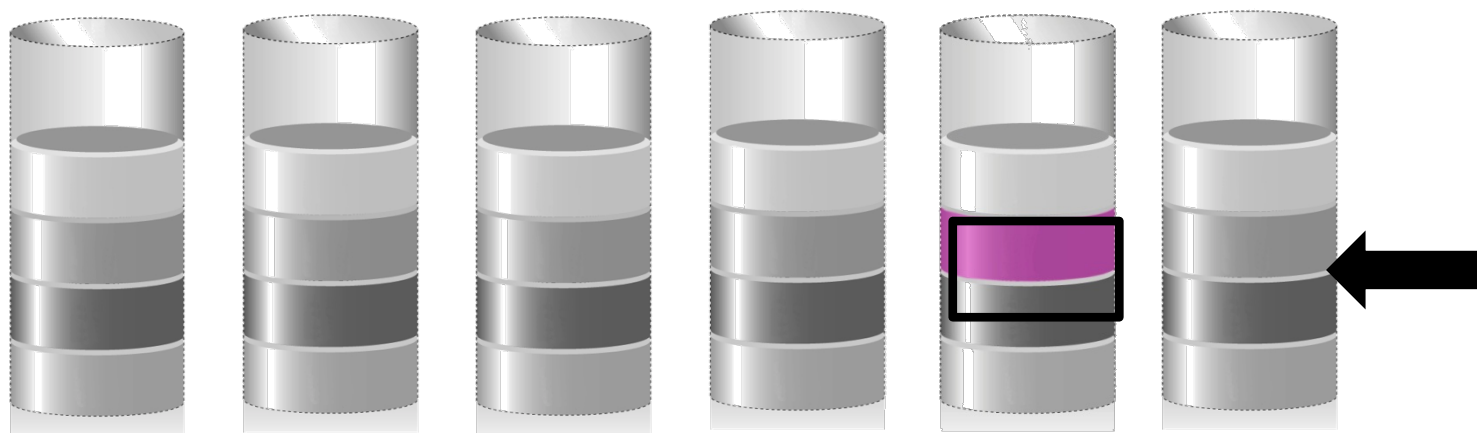
Silent Data Corruption

■ Свойства

- Алгоритм упреждающей коррекции ошибок (Silent Data Corruption), анализирует данные с целью обнаружения и исправления скрытых повреждений
- Актуально для дисков большой емкости (NL SAS, SATA)
- Выполняется при операциях чтения (не снижает производительность)

■ Преимущества

- **Надежность хранения данных на дисках большой емкости, возможность применения в системе дешевых NL SATA дисков**



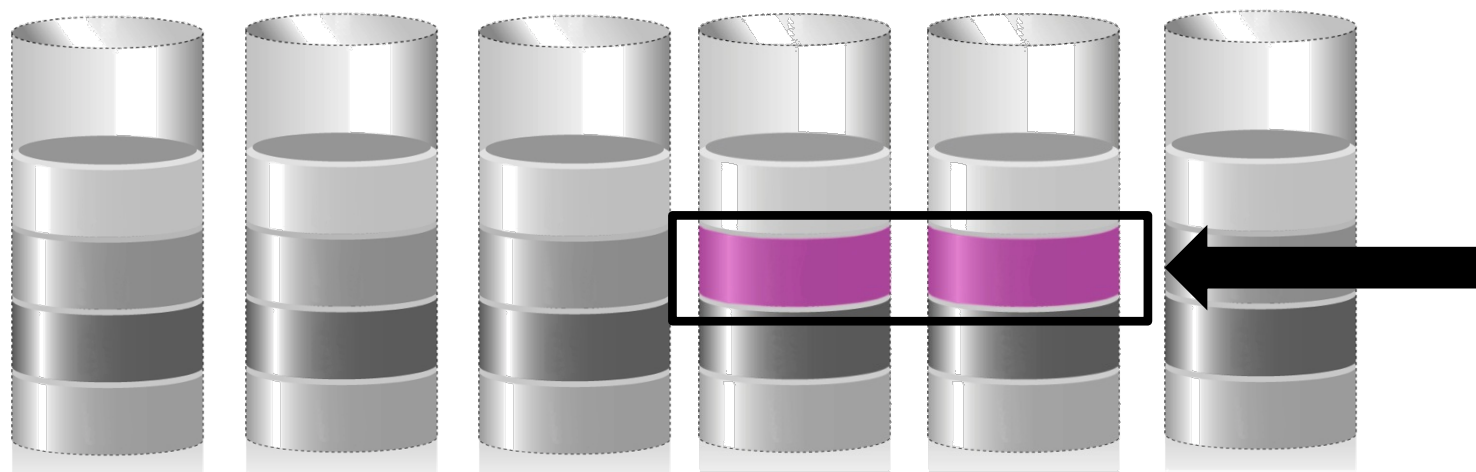
Advanced Reconstruction

■ Свойства

- Алгоритм предусматривает чтение N-2 самых быстрых диска массива и «искусственное» восстановление данных на лету с целью повышения производительности
- Актуально для потоковых операций
- Выполняется при операциях чтения

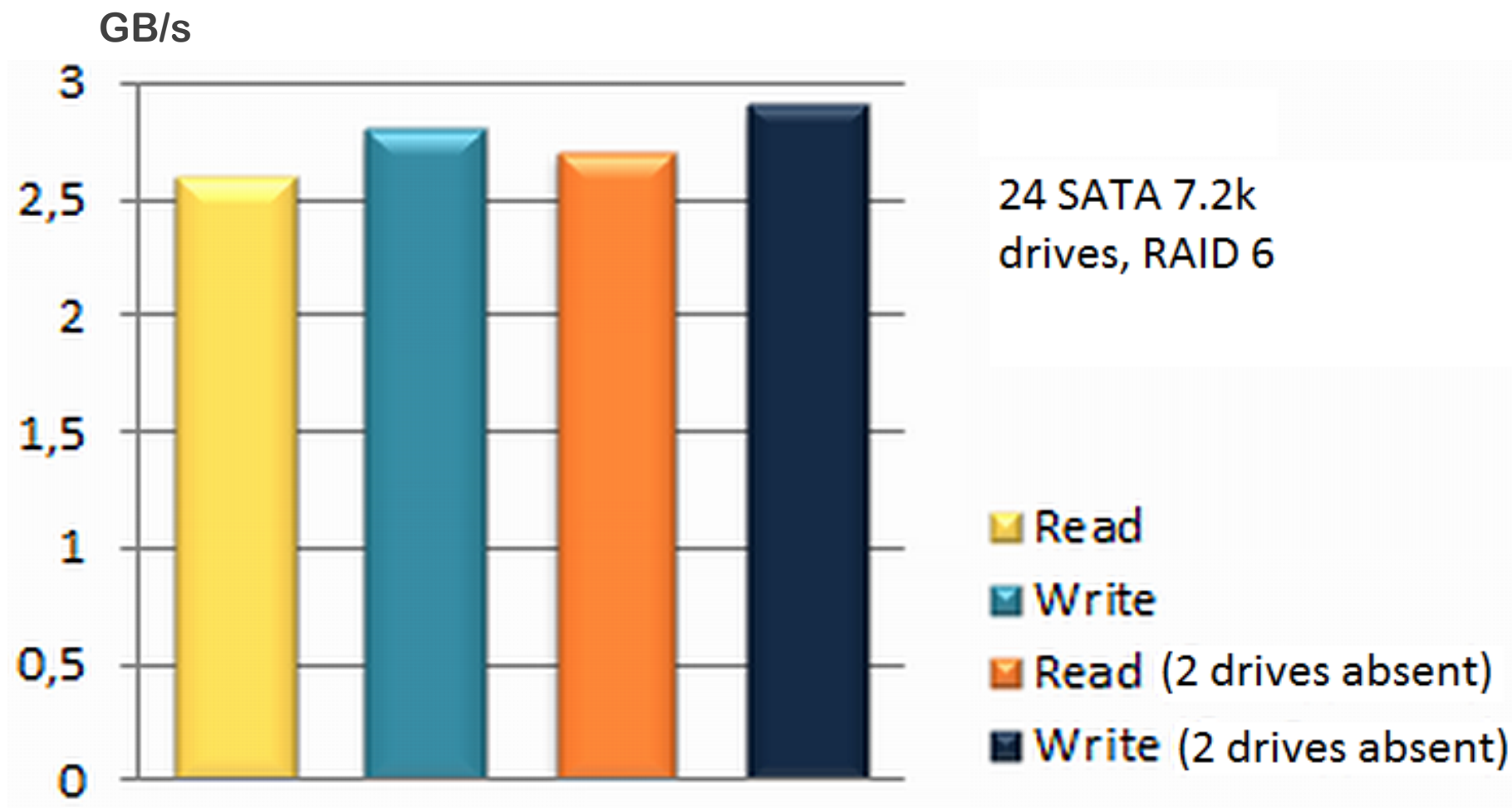
■ Преимущества

- Повышение производительности потоковых операций
- Гарантированная производительность



Уникальный функционал

Advanced Reconstruction 2



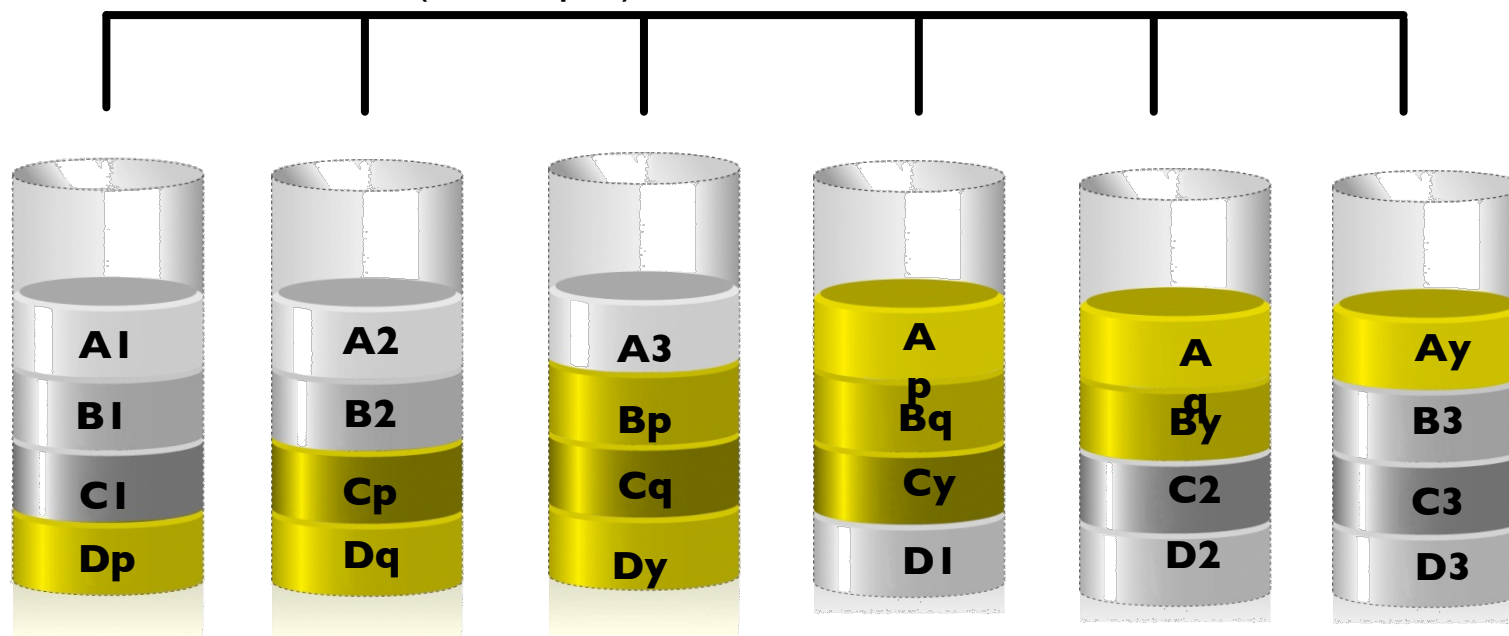
RAID 7.3

■ Свойства

- 3 контрольные суммы на каждый страйп массива данных
- Рекомендовано для массивов более 32 дисков

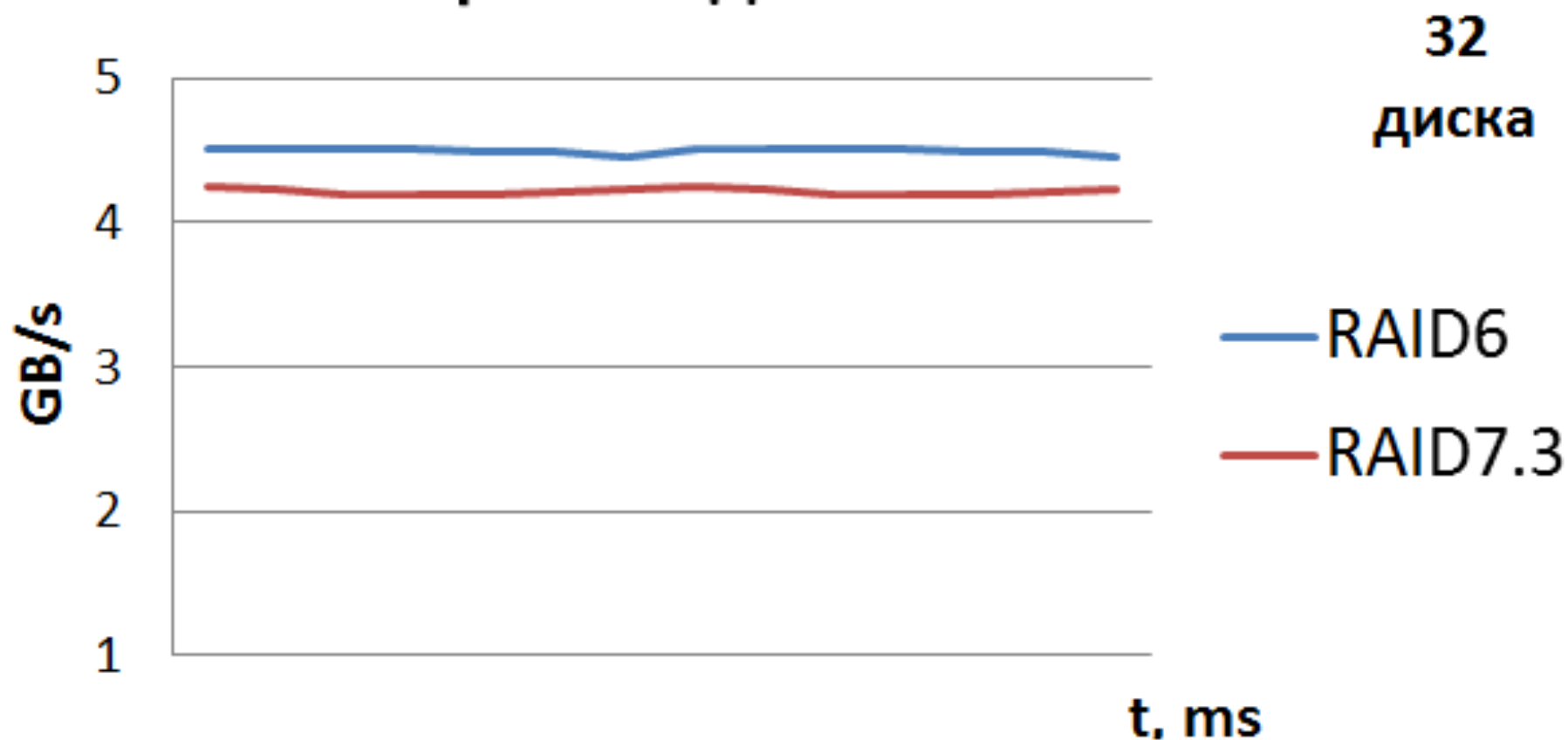
■ Преимущества

- Высочайшая надежность при максимальной производительности и минимальной стоимости (может выйти из строя 3 диска массива)
- Уникальное свойство (тендеры)



RAID 7.3

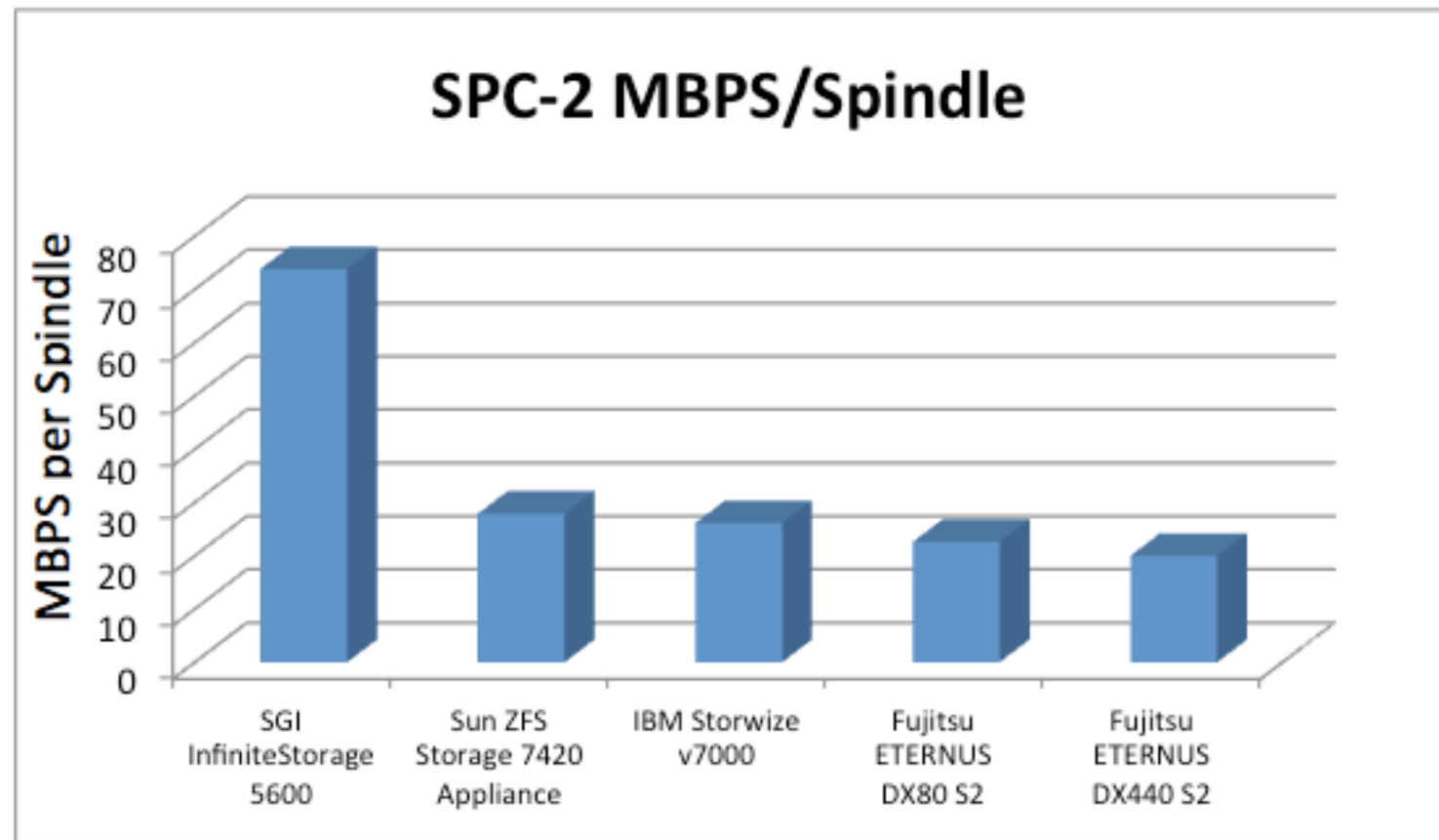
Производительность



Конкурентное сравнение

Производительность на шпиндель

- Рекорд SGI InfiniteStorage 5600 = **74** MBPS
- T-Store SA 2368 MBPS/Spindle = **110**



Надежность: время восстановления диска

- Зависит в том числе от MBPS

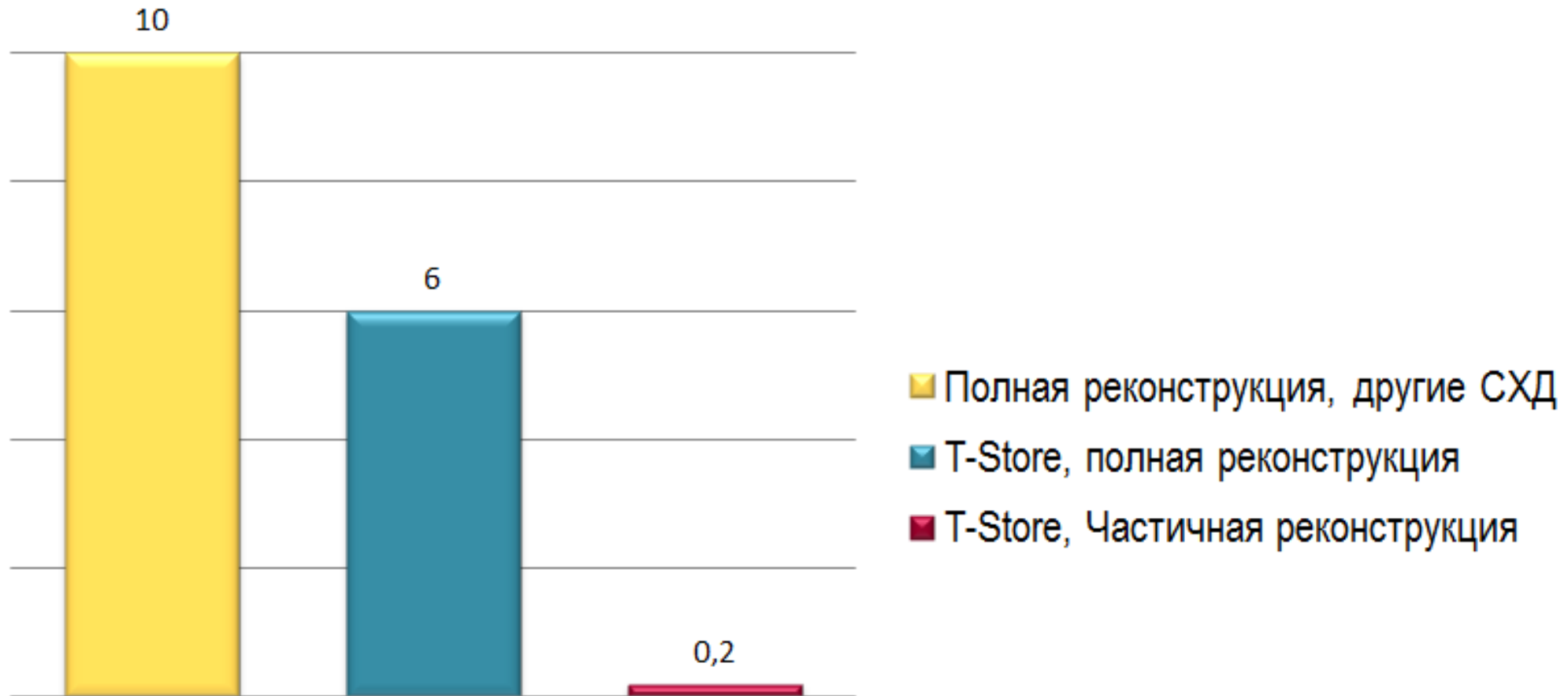
	DDN SFA 12k40	NetApp E5500	T-Store SA2368
Время восстановления 1 диска 1 ТВ, ч	8	4,5	3
Функция частичного восстановления*	-	-	да

* Позволяет восстанавливать проблемный блок в течение нескольких МИНУТ без ребилда массива

Конкурентное сравнение

Надежность 2 : Время восстановления при наличии ошибок и сбойных блоков

Кол. часов



Конкурентное сравнение

Матрица сравнения СХД для НРС

	T-Store SA2368IB	T-Store SA231215IB	DDN SFA12K-40	DDN SFA12K-40	EMC VNX 8000	NetAPP E5500
Размер, RU	8,00	12,00	45,00	45,00	28,00	8,00
Количество дисков	60,00	120,00	560,00	560,00	360,00	120,00
Интерфейсы подключения	FDR Infiniband	FDR Infiniband	FC8G or FDR IB	FC8G or FDR IB	FC8G or 16G	FC, SAS, FDR IB
Емкость (ТБ)	110,00	220,00	750,00	1120,00	360,00	180,00
Производительность ГБ/сек, параллельная файловая система	4,00	10,00	25,00	25,00	14,00	5,00
Стоимость	45000,00	80000,00	558785,00	437592,00	200000,00	62000,00
Примечание:	Посчитаны 2ТБ диски Seagate Canstellation3, производитель ность тоже дана с учетом этих дисков	q3 2014	Взята стоимость 1/4 системы для Арчера 100ГБ, 3ПБ, коэф. "сюда" = 13%	Взята стоимость 1/4 системы для Арчера 100ГБ, 5ПБ, коэф. "сюда" = 13%	контроллерный модуль 4U(2 контроллера и 2*11 внешних интерфейсных планок) каждый DAE=~3GB/s, т.е. 18GB/s на 6 DAE ??? 128GB RAM /SP	-
USD/ GB/sec	11250,00	8000,00	22351,40	17503,68	14285,71	12400,00
USD/ TB	409,09	363,64	745,05	390,71	555,56	344,44

Перспективы 2013

Модель	SA2364	SA2368	SA2224
Количество дисков	60	60	24
Режим работы контроллеров	Active-Passive		
Производительность	4.5 GB/s	8GB/s	3GB/S
Размеры	6	8	2
Функциональность	Централизованный веб интерфейс, частичное восстановление, RAID N+3, Автоматическая защита от «скрытых» ошибок при чтении		

Перспективы 2014

Модель	SA231215	SA231221
Количество дисков	120	180
Режим работы контроллеров	Active-Active	Active-Active
Производительность	15GB/s	21GB/s
Размеры	12	16
Функциональность	TBD	TBD

Спасибо!