ПРИБОРЫ И ТЕХНИКА ЭКСПЕРИМЕНТА, 2022, № 2, с. 37–42

ПРИМЕНЕНИЕ ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ В ЭКСПЕРИМЕНТЕ

УДК 520.374+004.052

АНАЛИЗ НАДЕЖНОСТИ ПРОГРАММНОГО ОТКАЗОУСТОЙЧИВОГО МАССИВА ПРИ ОРГАНИЗАЦИИ СИСТЕМЫ ДОЛГОВРЕМЕННОГО ХРАНЕНИЯ ДАННЫХ РАДИОИНТЕРФЕРОМЕТРИИ СО СВЕРХДЛИННЫМИ БАЗАМИ

© 2022 г. И. А. Безруков[a], А. И. Сальников[a], В. А. Яковлев[a], А. В. Вылегжанин[a,b]

[a] *Институт прикладной астрономии РАН*
Россия, 191187, Санкт-Петербург, наб. Кутузова, 10
[b] *Физико-технический институт им. А.Ф. Иоффе*
Россия, 194021, Санкт-Петербург, ул. Политехническая, 26

Поступила в редакцию 20.09.2021 г.
После доработки 25.11.2021 г.
Принята к публикации 27.11.2021 г.

Существенное увеличение скорости записи данных радиоинтерферометрии со сверхдлинными базами (р.с.д.б.) и соответственно их объема (порядка нескольких сотен терабайт) требует обеспечения надежного хранения этих данных для возможности проведения повторной обработки. В качестве одного из вариантов надежной отказоустойчивой системы долговременного хранения р.с.д.б.-данных может служить массив дисковых накопителей с избыточностью (RAID – Redundant Array of Independent Disks). В работе представлены результаты исследования надежности программных RAID в реализации файловой системы ZFS (Zettabyte File System) в вариантах raidz и raidz2 (аналоги RAID5 и RAID6) системы хранения р.с.д.б.-данных на примере стандартного серверного оборудования, переданы рекомендации по созданию такой системы для долговременного хранения этих данных в Центр корреляционной обработки РАН.

DOI: 10.31857/S0032816222020100

1. ВВЕДЕНИЕ

В 2020 г. в обсерватории “Светлое” запущен в эксплуатацию третий современный радиотелескоп РТ-13 ИПА РАН с диаметром зеркала 13 м и системой регистрации потоков данных с суммарной скоростью до 32 Гбит/с [1, 2]. Совместно с двумя аналогичными радиотелескопами в обсерваториях “Зеленчукская” и “Бадары” РТ-13 обеспечивает проведение наблюдений для определения поправок всемирного времени и суточные наблюдения, а также возможность участия в международных программах.

В связи с существенным увеличением скорости записи данных радиоинтерферометрии со сверхдлинными базами (р.с.д.б.) возрастает и объем этих данных. При этом уникальные данные р.с.д.б.-наблюдений (порядка нескольких сотен терабайт) требуют надежного хранения, с тем чтобы иметь возможность проведения повторной обработки. Главным базовым элементом любой системы хранения является дисковый накопитель. Практически любые устройства хранения, как находящиеся внутри корпуса сервера, так и подключаемые непосредственно или через сеть, строятся по принципу массива дисковых накопителей с избыточностью (RAID – Redundant Array of Independent Disks). Целью организации массива RAID является повышение надежности, увеличение скорости доступа к данным и (или) объема дисковой памяти. Кроме того, в большинстве случаев использование массива RAID обеспечивает возможность “горячей” замены отказавшего накопителя без потери каких-либо данных.

Как показывает практика создания и анализа работы Data-центров [3, 4], при планировании системы хранения данных (с.х.д.) большого объема (порядка сотен терабайт) необходимо учитывать следующее:

– задачи и тип нагрузки, создаваемой на с.х.д.;

– скорость записи/чтения массива;

– максимально возможный рабочий дисковый объем;

– вероятность разрушения массива;

– время восстановления дискового массива.

Цель данной работы – выбор наилучшего варианта программного массива RAID, построенного с использованием файловой системы ZFS (Zettabyte File System) [5] для долговременного хранения р.с.д.б.-данных объемом порядка не-

Таблица 1. Значение уровня URE для накопителей различных типов (по спецификациям производителей)

Тип накопителя	Значение URE	Объем, на который приходится одна ошибка, Тбайт
Пользовательские HDD	10^{14}	12.5
Серверные HDD, SATA	10^{15}	125
Серверные HDD, SAS	10^{16}	$1.25 \cdot 10^3$
Пользовательские SSD	10^{16}	$1.25 \cdot 10^3$
Серверные SSD	10^{17}	$12.5 \cdot 10^3$
Магнитные ленты LTO-7	10^{19}	$1250 \cdot 10^3$

скольких сотен терабайт на бюджетном серверном оборудовании.

Ожидается, что одновременно к р.с.д.б.-данным, хранящимся на с.х.д., будут обращаться до пяти сетевых клиентов, основной операцией ввода/вывода будет последовательное чтение файлов размером порядка единиц и десятков гигабайт. Производительность системы хранения данных оценивалась по скорости последовательного чтения без учета таких параметров, как задержка и число операций ввода/вывода.

2. ИНСТРУМЕНТАЛЬНЫЙ СТЕНД

Для выбора наилучшего варианта было проведено тестирование четырех конфигураций программных отказоустойчивых дисковых массивов: из десяти, восемнадцати, тридцати четырех и шестидесяти шести дисков емкостью 6 Тбайт каждый. Инструментальный стенд включал в себя сервер Dell R720 под управлением операционной системы FreeBSD 11.2-RELEASE с файловой системой ZFS. Для подключения дисков использовались три дисковые полки Supermicro CSE847 по 44 диска (всего 132 диска с форм-фактором 3.5" каждый).

Конфигурация сервера включала в себя процессор Intel Xeon E5-2643 v1 и 96 Гбайт оперативной памяти DDR3. Дисковые полки соединялись последовательно и подключались через контроллер HBA SAS 3.0, обеспечивающий суммарную скорость чтения/записи до 48 Гбит/с.

По результатам тестирования оценивались следующие параметры: доступный объем дисковой памяти, скорость локальных и сетевых операций чтения/записи, вероятность разрушения массива RAID, время восстановления массива и нагрузка на центральный процессор в процессе работы с.х.д.

Следует отметить, что исследования отказоустойчивости отдельных дисков различного объема, как и дисковых массивов RAID, проводятся производителями дисков и фирмами, обслуживающими и предоставляющими ресурс в Data-центрах с большим количеством дисков (порядка нескольких сотен тысяч единиц) [3]. Таким образом, в настоящее время мы имеем достаточно представительную статистику отказов дисков различных производителей.

3. ВЕРОЯТНОСТЬ РАЗРУШЕНИЯ ПРОГРАММНЫХ RAID-МАССИВОВ

Основными причинами отказов жестких дисков являются функциональные (аппаратные) и скрытые ошибки [4, 6]. Функциональный отказ обнаруживается на аппаратном уровне обслуживающим накопитель контроллером и при правильном построении системы не ведет к потере данных. Частота функциональных отказов определяет надежность диска как устройства и выражается через среднее время наработки на отказ. Чаще вместо среднего времени наработки на отказ (MTBF – Mean time between failures) используют такой показатель, как AFR (Annualized Failure Rate) – средний процент отказов в год, AFR следует интерпретировать как вероятность выхода из строя диска в течение одного года.

Под скрытыми ошибками данных понимают ошибки, которые не обнаруживаются электроникой дискового накопителя (диска) в процессе работы. Параметром, характеризующим такие ошибки, является URE (Unrecoverable Read Error) – невосстановимые ошибки чтения, указываемые производителем в спецификации диска.

Параметр URE означает, что в процессе чтения с диска указанного количества бит велика вероятность получить одну невосстановимую ошибку чтения, несмотря на то что до этого события возможные ошибки чтения были восстановлены по кодам коррекции ошибок (ECC – error-correcting code). Типовые значение уровня URE для дисков различных классов представлены в табл. 1.

Невосстановимые ошибки чтения явно не проявляются, но их последствия могут быть катастрофическими и в итоге способны привести к полной потере данных. Расчет вероятности разрушения массивов RAID5 и RAID6 для дисков емкостью от 1 до 20 Тбайт и числа дисков в массиве от 3 до 66 проведен по приведенным ниже формулам. Принципы расчета описаны в работе [7].

Разрушение массива RAID5 в процессе восстановления после отказа диска может произойти в следующих случаях:

– отказ одного из оставшихся в массиве дисков (P_{afr});

– возникновение во время восстановления невосстановимой ошибки чтения (P_{ure}).

При этом вероятность разрушения массива RAID5 будет равна сумме вероятностей совместных событий:

$$P_{raid5} = (P_{afr} + P_{ure} - P_{afr}P_{ure}). \quad (1)$$

Здесь

$$P_{ure} = 1 - \left(1 - \frac{1}{\mathrm{URE}}\right)^{n}, \quad (2)$$

где n, бит – объем данных в массиве из N неповрежденных дисков;

$$P_{afr} = 1 - \left(1 - \frac{1}{Der}\right)^{h}, \quad (3)$$

где h – время восстановления массива, определяемое как $h = T_{disk}N$ (T_{disk} – время восстановления одного диска, N – количество неповрежденных дисков в массиве), Der – вероятность разрушения диска в пересчете на час, рассчитываемая по формуле

$$Der = \frac{1}{1 - \sqrt[8766]{1 - \mathrm{AFR}}}. \quad (4)$$

Разрушение дискового массива RAID6 в процессе восстановления после отказа одного диска может произойти в следующих случаях:

– отказ одного диска и одновременно одна невосстановимая ошибка чтения URE;

– отказ двух дисков.

С учетом этих отказов вероятность разрушения массива может быть определена как:

$$P_{raid6} = (P_1 + P_2 - P_1P_2), \quad (5)$$

где $P_1 = P_{afr1}P_{ure2}$ и $P_2 = P_{afr1}P_{afr2}$; P_{ure2} рассчитывается по формуле (2) при числе дисков $n(N–2)$ в массиве; P_{afr1} и P_{afr2} рассчитываются по формуле (3) при числе дисков соответственно $h(N–1)$ и $h(N–2)$ в массиве.

Рассчитанные по формулам (1), (5) вероятности разрушения массивов RAID5 и RAID6 в зависимости от объема диска и их количества в массиве представлены на рис. 1. При расчетах значение AFR принималось равным 0.1, что соответствовало максимально возможной вероятности отказа диска, а URE – 10^{15} – значению URE для серверных HDD SATA из табл. 1.

Вероятности разрушения массива в процессе восстановления при выходе из строя одного диска объемом 6 Тбайт приведены ниже:

Количество дисков в группе	10	18	34	66
RAID5, %	35.5	56.2	80.0	95.8
RAID6, %	0.2	0.6	1.8	4.3

Из приведенных данных следует, что даже минимальная вероятность разрушения массива RAID5 из десяти дисков составляет 35.5%. Отметим, что многие производители оборудования начиная с 2007 г. дают рекомендацию не использовать массивы RAID5 при создании отказоустойчивых с.х.д. Наши расчеты подтверждают этот вывод.

Для массивов RAID6 вероятность разрушения существенно меньше и не превышает порядка 4.3%. С учетом этого обстоятельства дальнейшие оценки параметров проводились только для массива RAID6, близкого в реализации к ZFS raidz2.

4. СРАВНЕНИЕ ДИСКОВЫХ МАССИВОВ

4.1. Сравнение дисковых массивов по объему дисковой памяти

В практике создания массивов с избыточностью при выборе количества дисков в массиве обычно используют величину, равную степени двойки, и добавляют необходимое количество дисков четности: один диск – для ZFS-массива (пула в терминах ZFS) типа raidz1 и два диска – для пула типа raidz2.

Было рассмотрено несколько возможных конфигураций пулов ZFS с $N \times$ raidz2 для с.х.д. из трех дисковых полок Supermicro CSE847, по 44 диска объемом 6 Тбайт каждый с форм-фактором 3.5". В каждом варианте конфигурации отдельные отказоустойчивые группы дисков raidz2 (ZFS vdev) были объединены в общий пул с помощью технологии stripe (аналог RAID60).

Результаты сравнения различных конфигураций пулов ZFS с.х.д. по таким параметрам, как емкость массива, количество задействованных дисков и дисков с избыточными данными, обеспечивающих отказоустойчивость, приведены в табл. 2.

Максимальный полезный объем дисковой памяти (620 Тбит) имеет конфигурацию пула из двух групп, по 66 дисков каждая. Эта конфигурация использует наименьшее количество дисков четности (4), но плюсы такой конфигурации нивелируются самым высоким значением вероятности разрушения массива. Наиболее интересной, на наш взгляд, выглядит конфигурация из семи групп, по 18 дисков каждая. Полезная емкость такого хранилища будет равна 592 Тбайт при 14 дисках четности.

4.2. Сравнение ZFS-пулов по скорости операций ввода/вывода

При выборе наилучшего варианта конфигурации ZFS-пула следует также учитывать скорость операций чтения/записи. В табл. 3, 4 приведены оценки скорости операций чтения/записи как для локального, так и для сетевого режима при

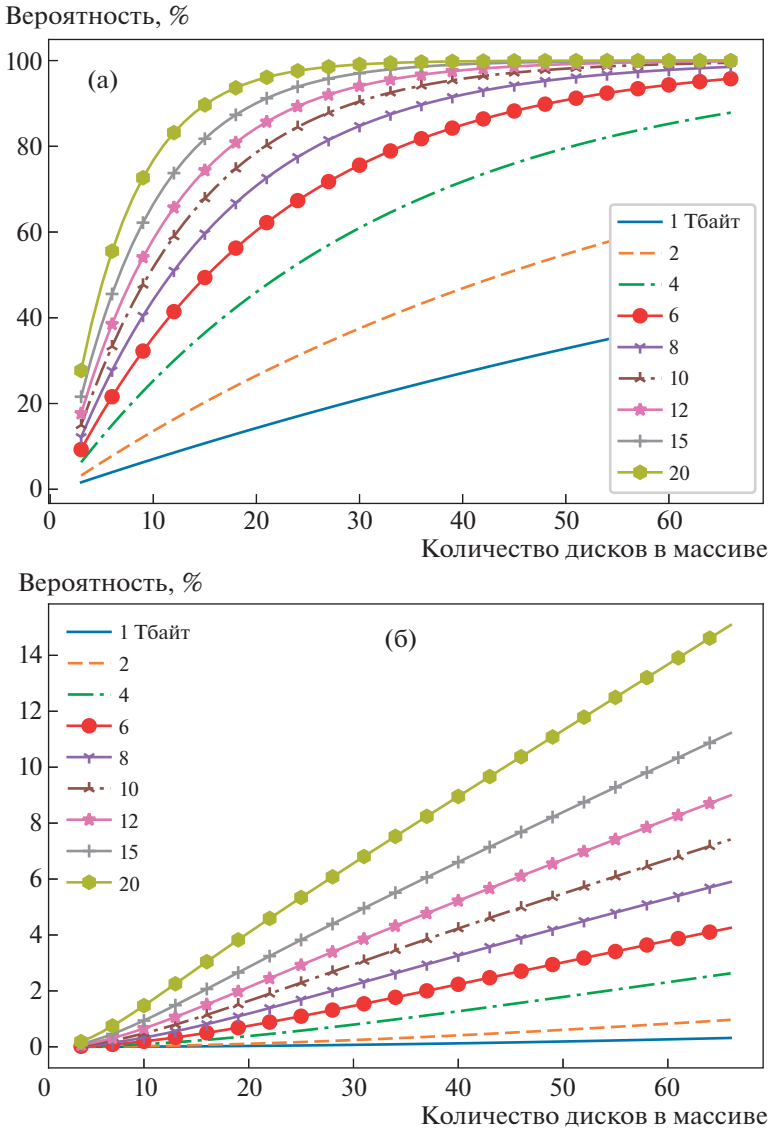

Рис. 1. Вероятность разрушения RAID5 (**а**) и RAID6 (**б**) в зависимости от количества дисков в массиве и объема одного диска.

одновременном обращении к с.х.д. пяти клиентов. Тестирование скорости дисковых операций проводилось утилитой fio.

Анализ приведенных оценок показывает, что наиболее приемлемым вариантом, с учетом накладных расходов на обеспечение избыточности

Таблица 2. Основные параметры для различных конфигураций ZFS-пула $N \times$ raidz2

Конфигурация пула	Полная дисковая емкость с.х.д., Тбайт	Доступная дисковая емкость с.х.д., Тбайт	Количество дисков в с.х.д.	Количество дисков четности
2 × 66d	720	620	132	4
7 × 18d	688	592	126	14
13 × 10d	710	524	130	26
3 × 34d	557	479	102	6

Таблица 3. Скорость локальных операций чтения/записи

Конфигурация пула	Скорость чтения, Мбайт/с	Скорость записи, Мбайт/с	Одновременные запись и чтение	
			Чтение, Мбайт/с	Запись, Мбайт/с
2 × 66d	230	687	174	386
7 × 18d	265	762	176	636
13 × 10d	257	821	212	804
3 × 34d	197	666	146	382

Таблица 4. Скорость сетевых операций чтения/записи при одновременном подключении к с.х.д. пяти клиентов

Конфигурация пула	Скорость чтения, Мбайт/с	Скорость записи, Мбайт/с	Одновременные запись и чтение	
			Чтение, Мбайт/с	Запись, Мбайт/с
2 × 66d	532	463	42	368
7 × 18d	769	425	42	356
13 × 10d	1091	494	41	332
3 × 34d	529	414	40	355

и полезной доступной емкости хранилища, также является вариант ZFS-пула из семи групп по 18 дисков.

5. ВРЕМЯ ВОССТАНОВЛЕНИЯ ZFS-ПУЛА

Теоретические оценки времени восстановления пула в нормальное состояние из режима Degraded Mode (состояние, при котором в ZFS-пуле raidz2 отказали один или два диска, но потерь данных нет) для трех наиболее вероятных значений скорости восстановления массива приведены ниже:

Скорость восстановления массива, Мбайт/с	15	30	60
Время восстановления массива, ч	117	58.3	29.1

При расчете принят пул из дисков объемом 6 Тбайт при заполнении 100%. Скорости чтения диска задавались аналогичными тем, которые представлены в исследованиях фирмы IBM [7].

Проведенные на экспериментальном стенде исследования показали приблизительно такие же, что и в работе [7], значения скорости чтения и времени восстановления ZFS-пула raidz2 для дисков объемом 6 Тбайт при заполнении на 70%.

Для массива из десяти дисков по 6 Тбайт были проведены две серии тестов: для 4.03 и 3.50 Тбайт восстановленных данных. Время восстановления соответственно составило 37.8 и 35.8 ч, скорость восстановления – 31.1 и 28.5 Мбайт/с. Для массива из восемнадцати дисков по 6 Тбайт (4.23 Тбайт восстановленных данных) время восстановления составило 40.4 ч, скорость восстановления – 30.5 Мбайт/с. Таким образом, эксперимент показал, что время восстановления массива не зависит от количества дисков в массиве, а в основном определяется загруженностью массива, т.е. числом и скоростью операций ввода/вывода.

6. ЗАГРУЗКА CPU

Одним из параметров, используемых при сравнении трех вариантов ZFS-пулов, являлась утилизация (загрузка) сервера CPU (Central Processing Unit) с.х.д. Так как ZFS является программным RAID, то загрузка CPU прямо пропорциональна количеству дисков в массиве. В табл. 5 представлены результаты измерений загрузки CPU для всех испытанных конфигураций отказоустойчивых массивов, измерения осуществлялись штатной утилитой FreeBSD vmstat. В качестве нагрузки на массив в режиме одновременной записи/чтения использовалась утилита fio.

Анализ результатов измерений загрузки CPU при тестировании утилитами fio и vmstat на ZFS-

Таблица 5. Загрузка CPU для различных конфигураций ZFS-пула

Конфигурация	Средняя загрузка CPU, %	Среднеквадратичное отклонение
1 × 10d	10	5
13 × 10d	49	16
1 × 18d	27	15
7 × 18d	51	19
1 × 34d	27	9
3 × 34d	45	11
1 × 66d	27	6
2 × 66d	54	13

Таблица 6. Итоговое сравнение характеристик различных конфигураций отказоустойчивых программных ZFS-пулов raidz2

Конфигурация пула	Баллы за вероятность разрушения (разд. 3)	Баллы за емкость с.х.д. (разд. 4.1)	Баллы за скорость операций ввода/вывода (разд. 4.2)	Сумма баллов
13 × 10d	3	1	3	7
7 × 18d	2	2	2	6
3 × 34d	1	0	0	1
2 × 66d	0	3	1	4

пулах из групп raidz2 по 10, 18, 34 и 66 дисков показывает, что максимальная загрузка CPU не превышает 54% (среднеквадратичное отклонение порядка 13%), и она существенно не влияет на скорость обработки операций ввода/вывода.

7. ВЫВОДЫ

В табл. 6 представлены итоговые результаты тестов, описанных в разд. 3, 4.1 и 4.2. Результаты тестов, описанных в разд. 5 и 6, практически идентичны для всех вариантов конфигураций ZFS-пула и не оказывают существенного влияния на процесс работы с.х.д. Каждый тест оценивался по шкале от 0 до 3 – чем лучше результат конфигурации ZFS-пула в тесте, тем больше баллов присуждалось.

В результате итогового сравнения варианты конфигурации ZFS-пула из трех групп по 34 диска и двух групп по 66 дисков набрали наименьшее количество баллов и далее не будут рассматриваться.

Вариант конфигурации ZFS-пула из 13 дисковых групп, по 10 дисков каждая (13 × 10), получил по 3 балла за минимальный шанс разрушения в процессе восстановления и за самую высокую скорость дисковых операций ввода/вывода и 1 балл – за получившуюся емкость с.х.д.

Вариант ZFS-пула из семи дисковых групп, по 18 дисков каждая (7 × 18), получил на 1 итоговый балл меньше, за счет того что во всех тестах показал средний результат. Данный вариант конфигурации позволяет дополнительно получить 68 Тбайт данных при сравнимых значениях вероятности разрушения ZFS-пула и скорости дисковых операций, что, на наш взгляд, является достаточным преимуществом при создании долговременного хранилища р.с.д.б.-данных.

Таким образом, наиболее подходящим вариантом отказоустойчивого массива для долговременного хранения р.с.д.б.-данных является ZFS-пул из семи групп raidz2, по 18 дисков объемом 6 Тбайт каждый.

СПИСОК ЛИТЕРАТУРЫ

1. *Ипатов А.В.* // Успехи физ. наук. 2013. Т. 183. № 7. С. 769. https://doi.org/10.3367/UFNr.0183.201307i.0769
2. *Безруков И.А., Сальников А.И., Яковлев В.А., Вылегжанин А.В.* // Труды ИПА РАН. 2015. Вып. 32. С. 3.
3. Backblaze Hard Drive Stats for 2020 [Электронный ресурс]. URL: https://www.backblaze.com/blog/backblaze-hard-drive-stats-for-2020/ (дата обращения 20.04.2021).
4. *Schroeder B., Gibson G.A.* // ACM Transactions on Storage (TOS). 2007. V. 3. Issue 3. P. 8. https://doi.org/10.1145/1288783.1288785
5. RAID-Z Storage Pool Configuration [Электронный ресурс] // Oracle Solaris ZFS Administration Guide URL: https://docs.oracle.com/cd/E19253-01/819-5461/gamtu/index.html (дата обращения 20.04.2021).
6. *Pinheiro E., Weber W.D., Barroso L.A.* Failure Trends in a Large Disk Drive Population. [Электронный ресурс] URL: https://www.usenix.org/legacy/events/fast07/tech/full_papers/pinheiro/pinheiro_old.pdf
7. Re-Evaluating RAID-5 and RAID-6 for slower larger drives [Электронный ресурс]. URL: https://www.ibm.com/support/pages/re-evaluating-raid-5-and-raid-6-slower-larger-drives (дата обращения 20.04.2021).