УДК 004.338

СРЕДНЕЕ ВРЕМЯ ДО ПОТЕРИ ДАННЫХ ДВУХДИСКОВОГО МАССИВА

Рахман П.А.

ФГБОУ ВПО «Уфимский государственный нефтяной технический университет», Филиал в г. Стерлитамаке, Россия, e-mail: pavelar@yandex.ru

Рассматриваются системы хранения данных на базе отказоустойчивого двухдискового массива RAID-1, которые широко используются на практике и имеют приемлемую аппаратную избыточность. Также рассматривается модель надежности двухдискового массива RAID-1 на базе цепей Маркова, учитывающая конечного время замены неисправного диска, различные интенсивности отказов дисков при нормальной работе и при синхронизации данных после замены неисправного диска, и вероятность ошибки чтения данных при репликации данных. Также представлены математическая модель надежности, методика расчета среднего времени до потери данных, методика оценки параметров надежности диска и контроллера массива, и пример расчета.

Ключевые слова: Избыточный массив недорогих дисков, отказоустойчивая система хранения данных, среднее время до потери данных, цепь Маркова с непрерывным временем

MEAN TIME TO DATA LOSS OF DUAL-DISK ARRAY

Rahman P.A.

Ufa State Petroleum Technological University, Sterlitamak branch, Russian Federation, e-mail: pavelar@yandex.ru

This paper deals with data storage systems based on fault-tolerant dual-disk RAID-1, which are widely used as high-reliable data storage systems and have acceptable overhead expenses in hardware implementation. Advanced reliability model of dual-disk RAID-1 array based on Markov chains, which takes into consideration finite time of disk replacement after disk failure, different disk failure rate in array's normal and rebuild states, and probability of read errors during array rebuild procedure, are also overviewed in this paper. Mathematical solution of reliability model, calculation formula for mean time to data loss, estimation of disk and array reliability parameters and MTTDL calculation example are also provided.

Keywords: Redundant array of inexpensive disks (RAID), Fault-tolerant disk system (FTDS), Mean time to data loss (MTTDL), Continuous-time Markov chain (CTMC)

Введение

В последние три десятилетия наблюдается бурное развитие информационных технологий и их внедрение в самые различные сферы деятельности человека, и информация, представленная в электронном виде, стала ключевой частью жизни и работы не только организаций, но и каждого отдельного человека. Более того, сохранность и доступность информации для ее пользователей, как правило, имеет критическую важность, а потеря данных нередко может приводить к катастрофическим последствиям. В такой ситуации анализ показателей надежности дисковых массивов имеет достаточно высокую актуальность, особенно для предприятий среднего и крупного масштабов, поскольку такой анализ также позволяет оценивать риски потери данных и принимать соответствующие решения, и при необходимости внедрять дополнительные технические средства.

В настоящее время существует множество вариантов построения дисковых хранилищ с применением одного или нескольких дисковых массивов по той или иной технологии RAID (Redundant Array of Inexpensive Disks), причем как классических (RAID-0, RAID-1, RAID-5, RAID-6), так и каскадных (RAID-10, RAID-50, RAID-60, RAID-51, RAID-61), матричных и других специализированных видов массивов.

С целью достижения высокой отказоустойчивости (особенно для баз данных), как правило, применяются RAID-1 массивы (также известные как «зеркало»), в котором все диски хранят одни и те же данные, и массив сохраняет работоспособность до тех пор, пока хотя бы один диск работоспособен. В силу высоких накладных расходов (при любом количестве дисков полезная емкость массива всегда равна емкости одного диска), на практике, как правило, используют двухдисковый RAID-1 массив.

Что касается моделей надежности, то с одной стороны имеется ряд академических учебников по теории надежности [1, 2], в которых рассматриваются обобщенные модели надежности технических систем, но нет конкретных примеров по современным системам хранения данных, в частности, избыточным дисковым массивам. С другой стороны имеется специализированная литература [3], посвященная надежности вычислительных машин, систем и сетей, в ко-

торых рассматриваются дисковые массивы, но приведенные модели надежности слишком упрощены и дают завышенные значения для показателей надежности.

Соответственно, в рамках научных исследований автора в области надежности систем [4-10] возникла научная задача разработки специализированной модели надежности для двухдискового массива RAID-1, для последующего использования полученных результатов при проектировании систем хранения данных для промышленных предприятий.

Базовая модель надежности двухдискового массива

Рассмотрим сначала известную упрощенную модель надежности двухдискового массива на базе модели дублированной системы с двумя с независимыми элементами.

Введем следующее множество состояний двухдискового массива RAID-1 и условий переходов из одного состояния в другое:

Состояние 0 (online) – оба диска исправны, данные массива доступны. Из этого состояния массив может с интенсивностью $2\lambda_D$ (отказ любого из исправных дисков) перейти в состояние 1.

Состояние 1 (degraded) – один диск исправен, другой диск отказал. Из этого состояния массив может с интенсивностью λ_D (отказ оставшегося диска) перейти в состояние 2, либо с интенсивностью μ_D (замена отказавшего диска и репликация данных с оставшегося диска) в состояние 0.

Состояние 2 (offline) – оба диска отказали, и массив разрушен.

где, λ_D – интенсивность отказов дисков в исправном состоянии.

μ_R – интенсивность замены диска и репликации данных.

Ниже на рис. 1 приведена марковская цепь, отражающая множество состояний системы и условия переходов:

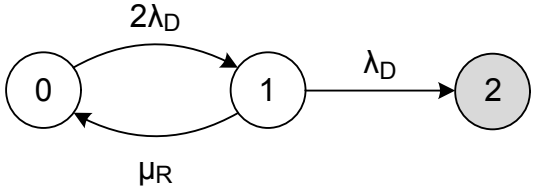

Рис. 1. Базовая модель надежности двухдискового массива RAID-1

Соответственно, система дифференциальных уравнений Колмогорова-Чепмена для этой цепи выглядит следующим образом:

$$\begin{cases} P_0(0)=1; \quad P_1(0)=0; \quad P_2(0)=0; \\ P_0(t)+P_1(t)+P_2(t)=1; \\ \dfrac{dP_0(t)}{dt}=-2\lambda_D P_0(t)+\mu_R P_1(t); \\ \dfrac{dP_1(t)}{dt}=2\lambda_D P_0(t)-(\mu_R+\lambda_D)P_1(t); \\ \dfrac{dP_2(t)}{dt}=\lambda_R P_1(t); \end{cases} \quad (1)$$

Тогда, учитывая, что состояние 0 является начальным, а состояние 2 – финальным, при котором массив разрушается, и теряются данные, мы имеем следующую формулу для расчета среднего времени наработки массива до потери данных:

$$T_{DL}=\int_0^\infty (P_0(t)+P_1(t))dt=\frac{\mu_R+3\lambda_D}{2\lambda_D^2}. \quad (2)$$

Усовершенствованная модель надежности двухдискового массива

Теперь рассмотрим предлагаемую автором модель надежности двухдискового массив RAID-1 с учетом конечного времени обнаружения и замены вышедшего из строя диска, конечного времени репликации данных (процедура rebuild) на замененном диске, возможности отказа как оставшегося диска, так реплицируемого диска, а также возможности срыва процедуры репликации из-за ошибки чтения данных с оставшегося диска.

Введем следующее множество состояний двухдискового массива RAID-1 и условий переходов из одного состояния в другое:

Состояние 0 (online) – оба диска исправны, данные массива доступны. Из этого состояния массив может с интенсивностью $2\lambda_D$ (отказ любого из исправных дисков) перейти в состояние 1.

Состояние 1 (degraded) – один диск исправен, другой диск отказал и ожидает замены, данные массива доступны. Из этого состояния массив может с интенсивностью λ_D (отказ исправного диска) перейти в состояние 2, либо с интенсивностью μ_D (замена отказавшего диска) в состояние 3.

Состояние 2 (offline 2) – оба диска отказали, и массив разрушен.

Состояние 3 (rebuild) – один диск исправен, другой диск заменен, на замененном диске идет репликация данных с исправного диска, данные массива доступны. Из этого состояния массив может с интенсивностью μ_R (завершение репликации данных на замененном диске) перейти в состояние 0, либо с интенсивностью λ_R (отказ реплицируемого диска) в состояние 1, либо с интенсивно-

стью λ_D (отказ исправного диска) в состояние 4, либо с интенсивностью ε_D (критическая ошибка чтения данных исправного диска в процессе репликации) в состояние 5.

Состояние 4 (offline 1) – один из ранее отказавших дисков заменен, но данные на него не успели реплицироваться, так как другой диск, с которого выполнялась репликация данных, отказал, и массив разрушен.

Состояние 5 (offline 0) – оба диска исправны, но произошла ошибка при репликации данных на замененный диск, и массив разрушен.

где, λ_D – интенсивность отказов дисков в исправном состоянии.

μ_D – интенсивность замены отказавшего диска.

λ_R – интенсивность отказов при репликации или восстановлении данных на замененный диск (большой объем операций записи).

μ_R – интенсивность восстановления или репликации данных.

ε_D – интенсивность ошибок чтения данных исправного диска при репликации данных на другой диск (большой объем операций чтения).

Ниже на рис. 2 приведена марковская цепь, отражающая множество состояний системы и условия переходов.

Соответственно, система дифференциальных уравнений Колмогорова-Чепмена для этой цепи выглядит следующим образом:

$$\begin{cases} P_0(0)=1; \quad P_1(0)=0; \quad P_2(0)=0; \\ P_3(0)=0; \quad P_4(0)=0; \quad P_5(0)=0; \\ P_0(t)+P_1(t)+P_2(t)+P_3(t)+P_4(t)+P_5(t)=1; \\ \dfrac{dP_0(t)}{dt}=-2\lambda_D P_0(t)+\mu_R P_3(t); \\ \dfrac{dP_1(t)}{dt}=2\lambda_D P_0(t)-(\lambda_D+\mu_D)P_1(t)+\lambda_R P_3(t); \\ \dfrac{dP_2(t)}{dt}=\lambda_D P_1(t); \\ \dfrac{dP_3(t)}{dt}=\mu_D P_1(t)-(\lambda_R+\mu_R+\lambda_D+\varepsilon_D)P_3(t); \\ \dfrac{dP_4(t)}{dt}=\lambda_D P_3(t); \\ \dfrac{dP_5(t)}{dt}=\varepsilon_D P_3(t); \end{cases} \quad (3)$$

Тогда, учитывая, что состояние 0 является начальным, а состояния 2, 4 и 5 – финальными, при котором массив разрушается, и теряются данные, мы имеем следующую формулу для расчета среднего времени наработки до потери данных:

$$T_{DL}=\int_0^\infty (P_0(t)+P_1(t)+P_3(t))dt= \frac{(\mu_D+3\lambda_D)(\mu_R+\lambda_D+\varepsilon_D)+\lambda_D(3\lambda_R+2\mu_D)}{2\lambda_D(\lambda_D(\lambda_R+\mu_R)+(\lambda_D+\mu_D)(\lambda_D+\varepsilon_D))}. \quad (4)$$

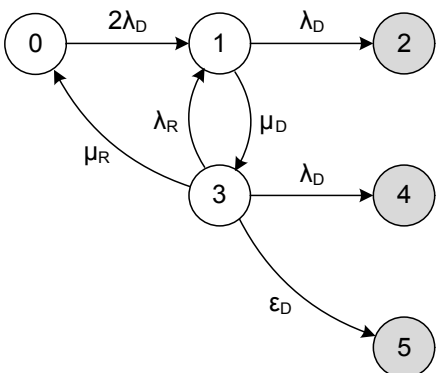

Рис. 2. Усовершенствованная модель надежности двухдискового массива RAID-1

Оценка исходных параметров надежности дисков и массива. Интенсивность отказов дисков λ_D можно оценить на основе параметра MTTF (Mean Time To Failure), предоставленного производителем дисков или полученного из практического опыта эксплуатации. Следует отметить, что производители часто завышают MTTF, указывая более миллиона часов. Практика же показывает, что MTTF диска лежит в пределах 50-300 тысяч часов. Что касается интенсивности отказов в режиме репликации (восстановления) данных λ_R, то в силу большого объема операций записи интенсивность отказов реплицируемого диска выше базовой интенсивности. Мы будем упрощенно полагать, что интенсивность реплицируемого диска втрое выше:

$$\lambda_D=1/\mathrm{MTTF}_{\mathrm{disk}}; \quad \lambda_R=3/\mathrm{MTTF}_{\mathrm{disk}}. \quad (5)$$

Интенсивность замены диска зависит от того, происходит ли замена автоматически за счет применения дополнительных дисков (помимо основных дисков в массиве) и технологии горячего резерва, или же обнаружения и замена диска осуществляется специалистами. В первом случае замена может занимать несколько минут, во втором – несколько часов. Соответственно, обобщая оба случая можно сказать, что интенсивность замены определяется параметром MTWS (Mean Time Waiting for Spare):

$$\mu_D=1/\mathrm{MTWS}_{\mathrm{disk}}. \quad (6)$$

Интенсивность репликации данных μ_R для массивов RAID-1 зависит от емкости диска V (в байтах), средней скорости записи v_{WR} на диск (в байт/сек) и средней скорости чтения v_{RD} данных (в байт/сек), и может быть оценена следующим образом:

$$\mu_R=\frac{3600\, v_{RD}v_{WR}}{V(v_{RD}+v_{WR})}. \quad (7)$$

Например, для диска емкости 1012 байтов, скорости записи $v_{RD} = 80 \cdot 10^6$ байт/сек и скорости чтения $v_{WR} = 50 \cdot 10^6$ байт/сек, интенсивность репликации данных составит $\mu_R \sim 1/9$ час-1 (в среднем репликация данных длится 9 часов).

Интенсивность ошибок чтения ε_D диска можно определить на основе параметра P_{UER} (вероятности невосстанавливаемой ошибки чтения бита), предоставленного производителем дисков или полученного из практического опыта эксплуатации, емкости диска V (в байтах) и среднего времени репликации данных, равного $1/\mu_R$ (в часах). Для дисков персональных компьютеров P_{UER} составляет ~10-14, для дисков серверных систем ~10-15.

Тогда, учитывая, что при репликации данных в массиве RAID-1 требуется считывать весь диск размером 8V битов, то вероятность ошибки чтения $Q = 1 - (1 - P_{UER})^{8V}$. С другой стороны полагая, что время наработки на ошибку – экспоненциально распределенная случайная величина с параметром ε_D, и регенерация длится в течение $1/\mu_R$ часов, имеем равенство $Q = 1 - e^{-\varepsilon/\mu}$. Тогда, из двух тождеств получаем $\varepsilon_D = -8V\mu_R \ln(1 - P_{UER})$. Тогда, учитывая, что P_{UER} очень малая величина, и $\ln(1 - P_{UER}) \sim -P_{UER}$, окончательно получаем:

$$\varepsilon_D = 8V\mu_R P_{UER}. \quad (8)$$

Например, для диска емкости V = 10^{12} байтов, интенсивности репликации данных μ_R = 1/9 час$^{-1}$ и вероятности невосстанавливаемой ошибки чтения бита $P_{UER} = 10^{-14}$, интенсивность ошибок чтения составит $\varepsilon_D \approx 1/112$ час$^{-1}$.

Пример расчета

Имеется массив RAID-1 с двумя дисками емкостью V = 10^{12} байтов. Среднее время наработки до отказа диска составляет $MTTF_{disk} = 120000$ часов. Интенсивность отказов реплицируемого диска втрое выше. Вероятность невосстанавливаемой ошибки чтения бита $P_{UER} = 10^{-14}$. Средняя скорость чтения данных $v_{RD} = 80 \cdot 10^6$ байт/сек. Средняя скорость записи данных $v_{WR} = 50 \cdot 10^6$ байт/сек. Среднее время замены дисков $MTWS_{disk} = 8$ часов.

Оценим сначала исходные параметры надежности по формулам 5-8.

Интенсивность отказов диска:

$\lambda_D = 1/MTTF_{disk} = 1/120000$ час$^{-1}$.

Интенсивность отказов реплицируемого диска: $\lambda_R = 3/MTTF_{disk} = 3/120000$ час$^{-1}$.

Интенсивность замены дисков:

$\mu_D = 1/MTWS_{disk} = 1/8$ час$^{-1}$.

Интенсивность репликации данных в массиве: $\mu_R = \frac{3600\, v_{read} v_{write}}{V(v_{read} + v_{write})} \sim 1/9$ час$^{-1}$.

Интенсивность ошибок чтения при репликации: $\varepsilon_D = 8V\mu_R P_{UER} \sim 1/112$ час$^{-1}$.

Рассчитаем среднее время наработки до потери данных дискового массива по известной упрощенной модели (формула 2):

$$T_{DL} = \frac{\mu_R + 3\lambda_D}{2\lambda_D^2} \approx 800180000 \text{ часов.}$$

Теперь рассчитаем среднее время наработки до потери данных дискового массива по предложенной автором модели (формула 4):

$$T_{DL} = \frac{(\mu_D + 3\lambda_D)(\mu_R + \lambda_D + \varepsilon_D) + \lambda_D(3\lambda_R + 2\mu_D)}{2\lambda_D(\lambda_D(\lambda_R + \mu_R) + (\lambda_D + \mu_D)(\lambda_D + \varepsilon_D))} \approx$$

805522 часа.

Нетрудно заметить, что специализированная модель, учитывающая ряд дополнительных параметров надежности дисков и массива, дает значительно более низкую и реалистичную оценку среднего времени наработки массива RAID-1 до потери данных, нежели чем известная упрощенная модель.

Заключение

Таким образом, в рамках данной статьи рассмотрены двухдисковый массив RAID-1, известная упрощенная модельная надежности и предложенная автором специализированная модель надежности для расчета среднего времени наработки массива до потери данных. Также рассмотрены методика оценки исходных параметров надежности дисков и массива, и приведен пример расчета среднего времени наработки.

Полученные научные результаты использовались автором при проектировании систем хранения данных для НИУ МЭИ (ТУ), Балаковской АЭС, ОАО «Красный Пролетарий» и ряда других предприятий.

Список литературы

1. Черкесов Г.Н. Надежность аппаратно-программных комплексов. СПб.: Питер, 2005.

2. Половко А.М., Гуров С.В. Основы теории надежности. 2-е изд. СПб.: БХВ-Петербург, 2006.

3. Martin L. Shooman. Reliability of computer systems and networks. John Wiley & Sons Inc., 2002.

4. Каяшев А.И., Рахман П.А., Шарипов М.И. Анализ показателей надежности двухуровневых магистральных сетей // Вестник Уфимского государственного авиационного технического университета. 2014. Т. 18. № 2 (63). С. 197-207.

5. Каяшев А.И., Рахман П.А., Шарипов М.И. Анализ показателей надежности локальных компьютерных сетей // Вестник Уфимского государственного авиационного технического университета. 2013. Т. 17. № 5 (58). С. 140-149.

6. Каяшев А.И., Рахман П.А., Шарипов М.И. Анализ показателей надежности избыточных дисковых массивов // Вестник Уфимского государственного авиационного технического университета. 2013. Т. 17. № 2 (55). С. 163-170.

7. Рахман П.А., Каяшев А.И., Шарипов М.И. Марковская цепь гибели и размножения в моделях надежности технических систем // Вестник Уфимского государственного авиационного технического университета. 2015. Т. 19. № 1. С. 140-154.

8. Рахман П.А., Каяшев А.И., Шарипов М.И. Модель надежности отказоустойчивой пограничной маршрутизации с двумя интернет-провайдерами // Вестник Уфимского государственного авиационного технического университета. 2015. Т. 19. № 1. С. 131-139.

9. Рахман П.А., Каяшев А.И., Шарипов М.И. Модель надежности отказоустойчивых систем хранения данных // Вестник Уфимского государственного авиационного технического университета. 2015. Т. 19. № 1. С. 155-166.

10. Рахман П.А., Шарипов М.И. Модель надежности двухузлового кластера приложений высокой готовности в системах управления предприятием // Экономика и менеджмент систем управления, 2015. № 3 (17). С. 85-102.