УДК 519.223.42

ПОДХОД К ОЦЕНКЕ НАДЕЖНОСТИ КЛАСТЕРНЫХ СТРУКТУР

М. Б. КУПЕРМАН
Д. Е. АВЕРЬЯНОВ

ЗАО «ИНФОРМСВЯЗЬ ХОЛДИНГ», г.Москва

Приведены модели и получены соотношения для оценки надежности кластерных систем. Показана важность учета надежности переключателя резерва при моделировании резервируемых структур. Рассмотрен «феномен» превосходства дублирования над троированием.

Ключевые слова: кластерные системы, высокая готовность, отказоустойчивость, марковские процессы, моделирование, наработка отказ, коэффициент готовности, надежность переключателя.

Новые условия обеспечения непрерывности бизнеса постоянно ужесточают требования к надежности IT-инфраструктуры, особенно вычислительных платформ и систем хранения данных (СХД). Основным методом достижения высоких значений готовности таких систем является резервирование, так как только оно способно обеспечить значения коэффициента готовности в районе «пяти девяток» – 0, 99999. В качестве типовых решений применяются отказоустойчивые системы (Fault Tolerance, FT) и системы высокой готовности (High Availability, HA). Наибольшую популярность на сегодняшний день имеют HA-кластеры.

Кластер – это группа серверов, которые связаны между собою и функционируют как один узел обработки информации. Спектр предлагаемых кластерных решений весьма обширен, ниже будут рассмотрены конфигурации применительно к классу HA-кластеров, основной целью которых является обеспечение отказоустойчивости. Однако, учитывая, что на практике часто используются смешанные конфигурации HA-кластеров, например с задачами балансировки нагрузки (Load Balancer cluster) и обеспечения высокой производительности (High Performance, HP), рассматриваемые подходы к оценке надежности могут быть распространены на такие смешанные системы. Кроме того, предложенные модели могут быть использованы при расчетах других резервируемых слабосвязанных многоузловых систем.

Одной из проблем оценки надежности отказоустойчивых систем, включая FT-системы, HA-кластеры и СХД, является проектная оценка надежности. Практика показывает, что часто используются упрощенный подход к такой оценке и получение явно завышенных величин показателей надежности. Ниже будут рассмотрены традиционные модели и предложены меры по увеличению их адекватности для расчета рассматриваемого класса систем.

Марковские процессы

Статические модели, построенные на методах, использующих основные формулы теории вероятности, комбинаторики, других логико-вероятностных методов, используемых, главным образом, для описания последовательно-параллельных структур, не позволяют учитывать изменения в характеристиках и процессах в зависимости от уже происходящих событий, отказов. Поэтому выбор модели надежности для описания резервированной структуры обусловлен ориентацией на класс динамического моделирования.

Моделирование кластерных структур марковскими процессами позволяет отразить в модели и учесть изменения процессов и отказов элементов во времени, временные условия осуществления других событий. Марковский процесс обладает ха-

рактерными свойствами, определенными, в первую очередь, экспоненциальными распределениями времени пребывания в каждом состоянии. Для применения экспоненциального закона распределения, при котором вероятность отсутствия отказов за t:

$$P(t) = e^{-\lambda t},$$

необходимо и достаточно соблюдение условия существования простейшего потока отказов. Простейший поток характеризуется следующими свойствами:

1) стационарностью, которая означает, что вероятностные характеристики потока для любого интервала времени зависят только от протяженности этого интервала, но не зависят от момента, когда он начался;

2) ординарностью, т.е. появление в один и тот же момент времени более одного отказа невозможно (дискретность времени);

3) отсутствием последствия, которое означает, что вероятность появления события в потоке, начиная с некоторого произвольного момента времени, не зависит от всей предыстории реализации этого потока.

Предполагается, что система восстанавливаемая. Ниже будут рассмотрены варианты с одной или несколькими ремонтными бригадами (параллельное восстановление узлов). Распределение времени наработки на отказ подчиняется экспоненциальному распределению с интенсивностью отказов:

$$\lambda = 1/\text{To},$$

где To – средняя наработка между отказами, а распределение времени восстановления подчиняется экспоненциальному распределению с интенсивностью восстановления;

$$\mu = 1/\text{TB},$$

где TB – среднее время восстановления.

Граф состояний такой системы представлен на рис. 1.

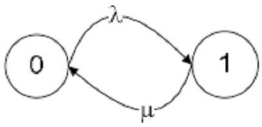

Рис. 1. Граф состояний одиночного узла (сервера):
состояния системы: 0 – работоспособное, 1 – неработоспособное

В качестве показателя надежности (ПН) отказоустойчивых структур, как правило, используется коэффициент готовности (Кг), определяемый формулой:

$$\text{Кг} = \frac{\text{наработка на отказ}}{\text{наработка на отказ} + \text{среднее время восстановления}}.$$

В случае необходимости получения оценки надежности на начальном, относительно непродолжительном интервале времени, необходимо использовать нестационарный Кг. Однако в большинстве случаев достаточно определить стационарный коэффициент надежности – вероятность того, что восстанавливаемый объект окажется работоспособным в произвольно выбранный момент времени в установившемся процессе.

Расчет надежности дублированной группы

При расчете надежности сетевого кластера, как правило, рассматривается дублированная группа узлов. При этом отказом считается выход из строя обоих узлов. В качестве параметров модели используются интенсивности отказов и восстановления, приведенные на рис. 1. Рассмотрим модель дублированной группы с идентичными узлами, приведенную на рис. 2.

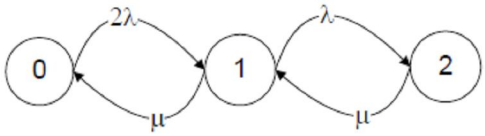

Рис. 2. Граф состояний кластера с двумя узлами и идеальной системой контроля

Отказ дублированной группы наступает тогда, когда во время восстановления одного из узлов откажет второй узел. Возможные состояния:

0 – оба узла исправны;

1 – отказ в одном узле;

2 – отказ в обоих узлах.

Таким образом, состояния исправности системы 0, 1, отказа 2.

В случае отказа одного из элементов группы, отказавший узел ремонтируется (заменяется) без остановки системы и после восстановления через случайный промежуток времени, распределенный по экспоненциальному закону с параметром μ, включается в состав дублированной группы $\mu=1/T_в$, где $T_в$ – среднее время восстановления. Одновременно может восстанавливаться один узел.

Кратко рассмотрим общий подход к расчету марковских цепей. На основе подготовленного графа состояний модели для каждого состояния k составляем дифференциальные уравнения вида:

$$p^{©}{}_k(t) = -p_k(t)\sum_{i\in e(k)}\Lambda_{ki} + \sum_{i\in E(k)}\Lambda_{ik}p_i(t), \qquad (1)$$

где запись i∈A означает, что суммирование ведется по всем таким состояниям i, которые относятся к множеству A; E(k) – множество тех состояний, из которых возможен непосредственный переход в некоторое состояние k; e(k) – множество состояний, в которые возможен непосредственный переход из состояния k. Через Λ_{ij} обозначена интенсивность перехода из состояния i в состояние j, а через $p_i(t)$ – вероятность пребывания системы в i-м состоянии в момент t. Если граф переходов содержит n различных состояний, то в результате может быть составлено n различных дифференциальных уравнений. Для определения коэффициента готовности необходимо взять n-1 уравнения и одно дополнительное:

$$\sum_{i=1}^{n} p_i(t) = 1. \qquad (2)$$

В результате расчета модели графа состояний рис. 2 в соответствии с (1) и (2) получим формулу для расчета коэффициента готовности:

$$Кг = \frac{2\lambda\mu + \mu^2}{2\lambda^2 + 2\lambda\mu + \mu^2}. \qquad (3)$$

Отметим, что типовой набор моделей дублированной группы узлов, как правило, включает следующие варианты: наличие одной (ограниченное восстановление) или нескольких ремонтных бригад и нагруженный \ ненагруженный режим резервирования. Выражение (3) определяет систему с нагруженным резервом и ограниченным восстановлением.

Для ненагруженного режима резервирования с ограниченным восстановлением в соответствии с (1) и (2) получим формулу расчета коэффициента готовности:

$$Кг = \frac{\lambda\mu + \mu^2}{\lambda^2 + \lambda\mu + \mu^2}.$$

Аналогично для дублированной системы с нагруженным резервом и неограниченным восстановлением (две ремонтные бригады) получим.

$$Кг = \frac{2\lambda\mu + \mu^2}{\lambda^2 + 2\lambda\mu + \mu^2},$$

и с ненагруженным резервом и неограниченным восстановлением:

$$Кг = \frac{2\lambda\mu + 2\mu^2}{\lambda^2 + 2\lambda\mu + 2\mu^2}.$$

Приведенные выше формулы для расчета дублированной группы сетевых узлов являются наиболее распространенными. Однако, при подстановке значений параметров λ и μ модели в (3) получаются явно завышенные значения показателей надежности, не отражающие, как правило, реальную надежность системы. При исходных данных интенсивности отказов λ = 0,00005 1/ч (наработка на отказ составляет 20 000 часов) и интенсивности восстановления μ = 0,25 1/ч (4 часа восстановления) получим из (3) значение Кг=0,999 999 92 (семь девяток). Подчеркнем, что взятая наработка – 20 тыс. часов – является нижней планкой MTBF (MeanTime Before Failure, средняя наработка на отказ) серверных платформ, обычно для серверов приводятся значения 50-100 тыс. часов и, следовательно, получаются еще более «хорошие» результаты.

Представленная на рис. 2 модель могла бы быть применена, например, для систем класса Stratus Continuum, в которых каждые два физических процессора объединяются парами и одновременно выполняют одну и ту же команду. При этом схема сравнения в каждом такте проверяет, что оба процессора пары вычислили тот же самый результат. Если результаты в паре различаются, то принимается решение о сбое, а пользователь использует результаты другой пары. Даже при такой организации вычислительного процесса FT-системы Stratus Continuum, обеспечивающие непрерывную готовность (Continuous Availability), заявленный коэффициент готовности составляет 99,999% (пять девяток – время недоступности системы 5 минут в год). При этом четыре процессора выполняют единственную команду с потактовой синхронизацией и сравнением результата. В кластерных системах отсутствует птактовая синхронизация и мажоритарный контроль.

Не смотря на такие фантастические расчетные значения, рассмотренная выше модель для расчета кластерных структур является типовой и очень удобной для подтверждения «высокой» надежности проектируемых отказоустойчивых систем. Вместе с тем, эта модель является очень упрощенной и не адекватна реальным процессам. Она не учитывает ряд факторов, существенно влияющих на надежность кластера. Не учитывается надежность внешних по отношению к узлам кластера элементам, включая коммуникационную среду, например элементы СКС и ЛВС, связывающие их. Но основной причиной неадекватности является исключение из рассмотрения дополнительных состояний системы.

Состояние необнаруженного отказа

В каждом элементе могут быть скрытые отказы. Модель, приведенная на рис. 2, не учитывает вероятность обнаружения отказа, надежность «переключателя» резерва, задержку при переключении резервов и другие, учет которых возможен при введении дополнительных параметров модели.

В дополнение к множеству состояний традиционной модели, представленной на рис. 2, в модернизированной модели, приведенной на рис. 3, добавляется состояние 3 – не обнаруженно средствами внутреннего (внутрикластерного) контроля отказа. Таким образом, состояниями отказа системы являются 2 и 3.

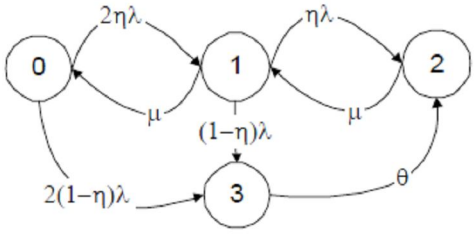

Рис. 3. Граф состояний кластера с двумя узлами и неидеальной системой контроля

Отметим, что для учета ненагруженности резерва и наличия нескольких ремонтных бригад применяются подходы к выводу зависимостей Кг, аналогичные приведенным выше при расчете подобных модификаций модели графа на рис. 2.

С позиции контролируемости кластер представляется как дублированная структура с непрерывным неполным контролем (внутренними средствами кластера), заданным η, и периодическим – внешним полным контролем работоспособности узлов, заданным θ, причем отказ узла с вероятностью η обнаруживается мгновенно, а с вероятностью 1-η обнаружение отказа задерживается на время 1/θ (в среднем). Время задержки обнаружения скрытых отказов имеет экспоненциальное распределение с параметром θ.

Путем расчета полученной модели (граф рис. 3) для той же наработки на отказ и времени восстановления определим значение коэффициента готовности. Предположим, что обнаружение «не обнаруженного» внутренними средствами кластера отказа составляет 15 мин., например, за это время клиенты, убедившись в отсутствии сервисов, начинают звонить в техническую поддержку, вынуждая администраторов вручную убедиться в работоспособности кластера. При подстановке значений получим, что уже при одном необнаруженном отказе на 100 отказов (η=0,99), обнаруженных системой управления, имеет место резкое снижение значения коэффициента готовности: с семи «девяток» до пяти «девяток». При необнаруженном каждом десятом Кг составит уже менее пяти «девяток», а при необнаруженном каждом втором – менее четырех.

Дублирование или троирование

Как было показано выше, при расчете дублированных структур без учета состояния необнаруженного отказа получаются завышенные значения показателей надежности. При увеличении кратности резервирования увеличиваются и значения показателей надежности, при этом для Кг получаются еще более астрономические цифры. Однако, при использовании модели с неидеальным контролем, например схемы троированных узлов, приведенной на рис. 4, ситуация меняется кардинально.

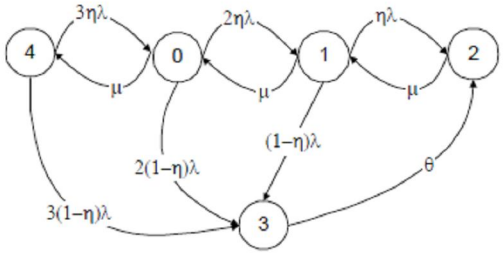

Рис. 4. Граф состояний троированной системы с неидеальной системой контроля

Введение состояния необнаруженного отказа позволяет наблюдать не только количественное, но и качественное изменение зависимости значений показателя надежности. Имеет место «феномен», когда метод повышения надежности путем повышения кратности резервирования, т.е. добавлением нового резервного узла, перестает действовать с учетом в модели фактора неполноты контроля функционирования коэффициентом полноты контроля η. На рис. 5 приведены результаты расчетов Кг, выполненные по представленным выше моделям дублирования и троирования с неидеальной системой контроля (представлены на рис. 3 и 4 соответственно).

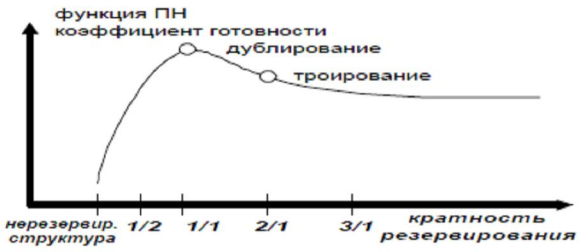

Рис. 5. Зависимость Кг от кратности резервирования при $\eta \leq 0{,}999$

Схемы двухузловой организации кластера могут показывать лучшие ПН, чем более избыточные: трех-, четырех- и т.д. узловые. Так, при равных системных параметрах λ,μ,θ и $\eta \leq 0{,}999$ двухузловый кластер обеспечивает лучший Кг по сравнению с аналогичной моделью трехузлового кластера. При одном необнаруженном на одну тысячу обнаруженных отказов ($\eta = 0{,}999$) значение Кг двухузлового кластера уже превосходит значение трехузлового. При ухудшении степени контролируемости, т.е. при уменьшении η, преимущества двухузловых конфигураций еще более высоки по сравнению с трехузловыми.

Это объясняется тем, что определяющую роль при малых значениях интенсивности отказов играет составляющая $m(1-\eta)\lambda$, где $m = 1, 2$ для схемы дублирования (рис. 3) и $m = 1, 2, 3$ – для троирования (рис. 4). То есть для резервируемых систем, претендующих на высокие показатели надежности, доминирующий вклад в результирующее значение Кг вносит вероятность обнаружения отказа, а не добавление нового резервного узла.

Выводы

Представленные методические рекомендации по оценке надежности кластерных структур нашли применение при проектировании центров обработки данных для банковского сектора и внутренних войск МВД РФ. Предложенные модели были апробированы и обеспечили существенное повышение адекватности моделирования, по сравнению с традиционными.

Литература

1. Ю. К. Беляев, В. А. Богатырев, В. В. Болотин и др. / под ред. И. А. Ушакова. Надежность технических систем: справочник. – М.: Радио и связь, 1985.

2. Шпаковский Г.И., Верхотуров А.Е., Серикова Н.В. Руководство по работе на вычислительном кластере: учеб. пособие. – Мн.: БГУ, 2004.

MODELS ARE RESULTED AND PARITIES ARE RECEIVED FOR CLUSTER RELIABILITY ESTIMATION

M. B. KUPERMAN
D. E. AVERJANOV

JSG «INFROMSVYAZ HOLDING», Moskow

Importance of the account of reliability of the switch of a reserve is shown at modeling of reserved structures. "Phenomenon" of the superiority of duplication over thrice-repeated is considered.

Key words: klasternye systems, high readiness, fault tolerance, марковские processes, modeling, an operating time refusal, readiness factor, reliability of the switch.