УДК 004.75

В. А. Богатырев, д-р техн. наук, проф., e-mail: Vladimir.bogatyrev@gmail.com,
А. В. Богатырев, аспирант,
Национальный исследовательский университет
информационных технологий, механики и оптики, Санкт-Петербург

Оптимизация резервированного распределения запросов в кластерных системах реального времени

Для вычислительной системы кластерной архитектуры реального времени предложена модель оценки своевременности и безошибочности резервированного обслуживания запросов в условиях ошибок и отказов. Предложена постановка и решение задачи оптимизации, заключающейся в нахождении кратности резервирования запросов, обеспечивающей максимум вероятности того, что хотя бы в одном из принимающих запрос к резервированному выполнению узлов задержка ожидания в очереди меньше заданного предельно допустимого значения.

Ключевые слова: *надежность, своевременность, реальное время, кластер, запрос, резервирование, оптимизация*

Введение

Для информационных и управляющих систем, функционирующих в реальном времени [1, 2], требуется обеспечение высокой отказоустойчивости и функциональной надежности [3—5] при гарантии безошибочности и своевременности обслуживания критичных запросов, к которым в ряде случаев предъявляются жесткие ограничения по выдаче достоверных результатов к заданным моментам времени.

Под функциональной надежностью будем подразумевать надежность системы по выполнению функциональных запросов, при этом для систем реального времени запросы должны быть выполнены к определенным моментам времени [6, 7], т. е. должны быть выполнены условия своевременности вычислений.

Обеспечения высокой надежности, отказоустойчивости и эффективности использования ресурсов кластера можно достичь при динамическом распределении запросов [5, 8, 9], и их перераспределении [7, 10, 11] в случае перегруженности узлов или реконфигурации кластера при его деградации. При динамическом распределении запросов возможна реализация приоритетных дисциплин обслуживания критичных запросов [12, 13] в целях обеспечения своевременности их выполнения. Очевидно, что методология проектирования высоконадежных кластеров должна быть ориентирована на применение моделирования и оптимизации [14—19], что обусловливает актуальность создания соответствующих моделей и поддержки проектирования.

Обеспечение высокой функциональной надежности, помимо традиционного резервирования аппаратно-программных средств, в ряде случаев требует резервирования вычислительного процесса.

Резервирования вычислительного процесса можно достичь при распределении каждого запроса (копии запроса) на выполнение в несколько узлов системы, что потенциально позволяет повысить вероятность своевременной выдачи результатов в условиях сбоев, отказов, ошибок и пиковой загруженности узлов.

Организация и модели резервированного вычислительного процесса при обслуживании копий запросов в нескольких узлах систем, представляемых многоканальными системами массового обслуживания (СМО) с общей очередью предложены и исследованы в работах [20—22]. В этих работах показано, что при требовании своевременного обслуживания запроса хотя бы одним из выполняющих его копии узлов, резервированное обслуживание в ряде случаев позволяет снизить среднее время пребывания запросов в системе. Вместе с тем оценки важного для систем реального времени показателя — вероятности непревышения времени пребывания запросов в системе предельно допустимого значе-

ния в работах [20—22] не приведено, кроме того, в них не учитывается особенность компьютерных систем кластерной архитектуры, заключающаяся в наличии локальных очередей в каждом узле (сервере) кластера.

Для кластерных систем реального времени, консолидирующих ресурсы группы серверов, исследование возможностей повышения функциональной надежности в результате распределения запросов на резервированное выполнение группе узлов, в каждом из которых организуется локальная очередь, проведено в работе [23].

Условие успешности обслуживания запроса по [23], заключается в том, что хотя бы в одном из k узлов, принимающих запрос, не возникают отказы и ошибки, а задержка запроса в очереди меньше предельно допустимого значения t_0. В результате исследований [23] показано, что резервированное обслуживание запросов при определенных условиях потенциально позволяет существенно сократить время ожидания запроса. Это достигается в результате того, что математическое ожидание минимума из нескольких случайных величин (времен ожидания запросов в узлах) может быть значительно меньше математического ожидания каждой из этих величин. При этом показано, что резервированное выполнение запросов в кластере может увеличить вероятность того, что хотя бы в одном узле время ожидания выдачи безошибочных результатов меньше предельно допустимого значения t_0.

Эффективность резервирования запросов в кластере зависит от многих факторов, в том числе от интенсивностей входного потока запросов, отказов и ошибок, от ограничений допустимого времени ожидания t_0 и кратности резервирования вычислений. Интегрированное влияние кратности резервирования запросов на показатели эффективности кластера противоречиво. С одной стороны, его увеличение приводит к росту загрузки узлов и, как следствие, к увеличению среднего времени пребывания запросов, что в свою очередь вызывает увеличение вероятности ошибок и отказов во время нахождения запроса в узле. С другой стороны, резервирование вычислений повышает вероятность того, что хотя бы одним из k узлов, принимающих запрос к выполнению, он будет своевременно и безошибочно выполнен [23].

Разрешения указанного технического противоречия можно достичь при оптимизации резервирования запросов в кластерах реального времени, включающей:

— определение требований и критериев эффективности резервированного обслуживания запросов реального времени в кластере в условиях отказов узлов и ошибок вычислений;

— постановку и решение задачи оптимизации резервирования запросов как при заданной интенсивности входного потока запросов, так и в условиях неопределенности, связанной с вариантностью этой интенсивности.

Постановка задачи исследования

Рассмотрим вычислительную систему реального времени кластерной архитектуры, в которой в целях повышения функциональной надежности предполагается резервированное обслуживание запросов.

Кластер состоит из n одинаковых компьютерных узлов (серверов), представимых системами массового обслуживания типа M/M/1 с бесконечной очередью. В кластер с интенсивностью Λ поступает общий поток запросов, каждый из которых может быть распределен на обслуживание в любой компьютерный узел или на резервированное обслуживание в k узлов.

Исследуемая кластерная система предназначена для работы в реальном масштабе времени при воздействии сбоев, отказов и ошибок при требовании выдачи безошибочных результатов обслуживания с условием, что задержка в очереди была бы меньше заданного предельно допустимого значения.

После поступления запроса и размещения его копии в очереди k узлов, результаты обслуживания в каждом из k узлов заносятся в его буфер, выдача из которого осуществляется в определенный момент времени t, отсчитываемый с момента формирования запроса.

Запрос считается выполненным некоторым узлом успешно при условии, что к моменту выдачи результатов t рассматриваемый запрос находился в очереди узла время t_0, меньшее, чем $t - v$, где v — среднее время обслуживания запроса, причем в течение интервала $t = t_0 + v$ узел работал без сбоев, отказов и ошибок, влияющих на результаты выполнения запроса. При резервированном обслуживании запроса k узлами запрос считается безошибочно и своевременно выполнен, если хотя бы в одном из k узлов время ожидания запроса в очереди меньше t_0 и в течение времени $t_0 + v$ не было сбоев, отказов или ошибок вычислений. Время хранения результатов выполнения запроса в буфере определяется разницей реального времени ожидания и его предельно допустимым значением t_0.

Цели исследования предлагаемой работы направлены на повышение функциональной надежности кластера реального времени в результате резервированного обслуживания запросов группой k узлов в условиях отказов и ошибок вычисления.

Предлагаемые исследования предполагают построение моделей для оценки вероятности своевременности и безошибочности резервированного

обслуживания запросов в кластерах реального времени и решение задачи оптимизации, заключающейся в нахождении кратности резервирования запросов, при которой достигается максимум вероятности получения безошибочного и своевременного результата обслуживания хотя бы в одном из узлов, принимающих запрос к резервированному выполнению.

Требования эффективности резервированного распределения запросов в кластерах реального времени

Для вычислительных систем кластерной архитектуры по выполнению критичного запроса в реальном времени предъявляются следующие требования:

- готовность системы в момент поступления запроса;
- безотказность и безошибочность функционирования узлов, принявших запрос на обслуживание в период времени от поступления запроса до выдачи результатов его выполнения;
- своевременность выдачи результатов к определенному моменту времени, отсчитываемому после формирования запроса.

Соответствие системы первому требованию оценивается по коэффициенту готовности, второму требованию — по вероятности безотказной и безошибочной работы, а вместе — по коэффициенту оперативной готовности с учетом вероятности возникновения ошибок во время пребывания запроса в системе.

В качестве показателя своевременности резервированного обслуживания запросов может быть выбрана вероятность того, что время ожидания запросов в очереди хотя бы одного узла, принимающего запрос к резервированному обслуживанию, меньше предельно допустимого значения t_0 [6, 23].

Вероятность своевременности и безошибочности резервированного выполнения запроса в кластере

В кластере, содержащем n одинаковых компьютерных узлов (серверов), представимых n системами массового обслуживания типа M/M/1 с бесконечной очередью, вероятность своевременности безошибочности и надежности выполнения запроса в некотором узле кластера определим как

$$R = r(t_0)p(t), \qquad (1)$$

где $r(t_0)$ — вероятность того, что время ожидания запросов в очереди узла меньше предельно допустимого значения t_0, а $p(t)$ — вероятность того, что во время вычислений и нахождения запроса в очереди, а его результатов в буфере, отказы, сбои и ошибки не возникают.

При k-кратном резервировании выполнения запросов происходит увеличение интенсивности входного потока в k раз, при этом в случае его сбалансированного распределения между n узлами кластера с учетом модификации известной [24] формулы вероятности не превышения времени ожидания в СМО типа M/M/1 имеем:

$$r(t_0) = 1 - (v\Lambda k/n)\exp(-t_0(v^{-1} - \Lambda k/n),$$

где Λ — интенсивность суммарного потока запросов в кластер.

Вероятность безотказности и безошибочности узла во время нахождения в нем запроса на стадиях ожидания, обслуживания и хранения результатов вычислений определим как:

$$p(t) = \exp(-\lambda t) = \exp(-\lambda(t_0 + v)),$$

где λ — суммарная интенсивность сбоев, отказов и ошибок.

Вероятность своевременного получения безошибочных результатов хотя бы от одного из k узлов кластера, задействованных в резервированном выполнении запроса, определим как

$$P = 1 - (1 - R)^k, \qquad (2)$$

где R — вероятность своевременности, безошибочности и надежности выполнения запроса в некотором узле, определяемая по формуле (1).

Приведенная оценка получена для исходного состояния кластера, когда все n его узлов находятся в работоспособном состоянии. С учетом того, что в произвольный момент поступления запроса в состоянии готовности может находиться i узлов кластера и в предположении безошибочности вычислений, вероятность своевременности резервированного обслуживания запроса в кластере на основе выражения (2) найдем как

$$P = \sum_{i=1}^{n} b_i(1 - \{1 - [1 - (v\Lambda k_i/i)\exp(-t_0(v^{-1} - \Lambda k_i/i)]\}^{k_i}),$$

где b_i — вероятность готовности i узлов кластера в момент поступления запроса; k_i — кратность резервирования при готовности (работоспособности) i узлов кластера, при этом очевидно, что должно выполняться условие $k_i \leqslant i$.

Вычисление вероятностей готовности i узлов кластера b_i позволяет найти его коэффициент готовности

$$K = \sum_{i=a}^{n} b_i.$$

Причем, если для обслуживания запросов достаточно работоспособности одного узла, то $a = 1$, а при необходимости обеспечения стационарного режима обслуживания запросов значение a задается как ближайшее целое большее Λv.

С учетом возможности ошибок и отказов узлов кластера, задействованных в резервированном обслуживании запроса (при ожидании, вычислениях и хранении результата), вероятность своевременности и безошибочности резервированного обслуживания запроса в кластере на основе (1) и (2) найдем по формуле

$$P = \sum_{i=a}^{n} b_i(1 - \{1 - [1 - (\Lambda k_i/i)v\exp(-t_0(v^{-1} - \Lambda k_i/i)]\exp(-\lambda(t_0 + v))\}^{k_i}).$$

Оценка вероятности готовности i узлов в момент поступления запроса b_i может быть получена при моделировании кластера процессом размножения и гибели [24], для которого при работоспособности i узлов суммарная интенсивность отказов системы $\lambda_i = i\lambda$, а интенсивность ограниченного восстановления равна μ (λ и μ — интенсивности отказов и восстановлений одного узла).

В предположении неограниченного восстановления вероятность готовности i узлов кластера в момент поступления запроса $b_i = C_n^i K^i(1 - K)^{n-i}$, где K коэффициент готовности узла, $K = t_1/(t_1 + t_2)$, $t_1 = \lambda^{-1}$ и $t_2 = \mu^{-1}$.

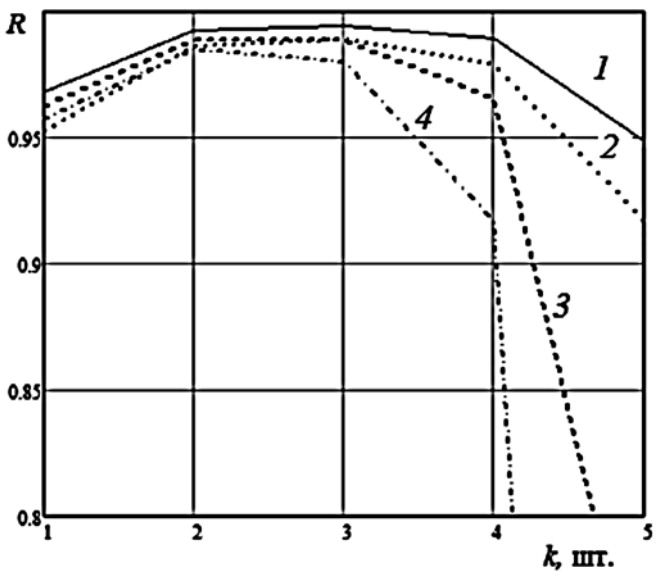

Рис. 1. Вероятность своевременности и безошибочности обслуживания запросов в зависимости от кратности их резервирования

Приведем пример расчета вероятности своевременного получения безошибочных результатов в зависимости от кратности k резервирования распределения запросов в кластере из $n = 6$ узлов при $v = 1$ с и суммарной интенсивности ошибок и отказов $\lambda = 10^{-3}$ 1/с.

Результаты расчетов для исходного состояния кластера (при готовности всех n узлов) представлены на рис. 1, на котором кривая *1* соответствует $\Lambda = 0{,}97$ 1/с, $t_0 = 2$ с; кривая *2* — $\Lambda = 0{,}97$ 1/с; $t_0 = 1{,}5$ с; кривая *3* — $\Lambda = 1{,}1$ 1/с, $t_0 = 2$ с; а кривая *4* — случаю $\Lambda = 1{,}2$ 1/с, $t_0 = 2$ с. Приведенные расчеты показывают существенное влияние на эффективность кластера кратности резервирования запросов, и следовательно, целесообразность оптимизации кратности резервирования вычислений в целях максимизации вероятности своевременного получения в реальном времени безошибочных результатов обслуживания запросов хотя бы в одном из узлов кластера.

Оптимизация резервирования запросов в кластерной системе реального времени

В результате оптимизации при заданной интенсивности потока запросов Λ требуется определить кратность резервированного обслуживания запросов k, при которой достигается максимум вероятности своевременного получения безошибочных результатов хотя бы одним из k узлов кластера, выполняющих копии запроса.

При заданной интенсивности запросов и исправности n узлов кластера критерий оптимизации задается как

$$P_1 = \max_k \{1 - [1 - [1 - (v\Lambda k/n)\exp(-t_0(v^{-1} - \Lambda k/n)]\exp(-\lambda(t_0 + v)]^k\}, \quad (3)$$

при условии стационарности процесса обслуживания $(v\Lambda k/n) < 1$.

В случае принятия решения в условиях неопределенности, когда задается вектор возможных значений интенсивностей Λ_i и вектор вероятностей соответствующих значений интенсивностей g_i, рассмотрим три варианта оптимизации (Bl, B2 и B3).

При варианте B1 оптимизацию при исправности n узлов кластера проведем соответственно критерию (3) по среднему значению интенсивности запросов:

$$\Lambda = \frac{1}{m}\sum_{i=1}^{m} g_i\Lambda_i,$$

при выполнении условий существования стационарного режима обслуживания $\max(v\Lambda_i k/n) < 1$.

Ориентация на выбор кратности резервирования по средней интенсивности запросов может привести к неоптимальности ее выбора при малых и больших интенсивностях потока запросов, отличающихся от ее среднего значения.

При варианте В2 при исправности n узлов кластера оптимизацию проведем по критерию максимума среднего значения вероятности своевременного получения безошибочных результатов хотя бы одним из k узлов кластера, при значении кратности резервирования едином для всех вариантов входного потока запросов, равном

$$P_2 = \max_k \sum_{i=1}^{m} g_i\{1 - [1 - [1 - (v\Lambda_i k/n)\exp(-t_0(v^{-1} - \Lambda_i k/n)]\exp(-\lambda(t_0 + v)]^k\}, \quad (4)$$

при условии стационарности режима обслуживания $\max(v\Lambda_i k/n) < 1$.

Ориентация на выбор кратности резервирования с учетом выполнения условий стационарности при наибольшей интенсивности потока может привести к неоптимальности ее выбора при малых интенсивностях потока запросов. Таким образом, возникает потребность адаптивного назначения кратности резервирования в зависимости от интенсивности входного потока, что, с одной стороны, потенциально должно привести к повышению своевременности безошибочного обслуживания запросов, но, с другой стороны, связано с дополнительными издержками на мониторинг и измерения текущей интенсивности входного потока.

Для варианта В3 адаптивного назначения кратности резервирования в зависимости от интенсивности входного потока целевую функцию сформируем на основе критерия Байеса как

$$P_2 = \max_k \sum_{i=1}^{m} g_i\{1 - [1 - [1 - (v\Lambda_i k_i/n)\exp(-t_0(v^{-1} - \Lambda_i k_i/n)]\exp(-\lambda(t_0 + v)]^{k_i}\}, \quad (5)$$

где n — число узлов кластера $K = \{k_1, k_2, ..., k_m\}$, при условии стационарности режима обслуживания, задаваемом для каждого варианта интенсивности входного потока

$$v\Lambda_i k_i/n < 1 \text{ при } i = 1, 2, ..., m.$$

Критерий (5) не учитывает снижения вероятности своевременности резервированного обслуживания D, связанного с диспетчеризацией и измерением текущей интенсивности запросов, требующей реализации в системе функций мониторинга и измерения нагрузки в реальном времени [25]. Определив по критерию (5) вероятность своевременности резервированного обслуживания с адаптивным назначением кратности резервирования в зависимости от измеряемой интенсивности входного потока, можно найти допустимые потери на адаптацию, при которых оптимизация по критерию (5) эффективна по сравнению с оптимизацией по критерию (3) или (4).

Приведем пример оптимизации кратности резервирования обслуживания запросов для кластера, содержащего $n = 15$ серверов при предельно допустимом времени ожидания запросов в очереди $t_0 = 2$ с, среднем времени обслуживания запросов $v = 1$ с и суммарной интенсивности ошибок и отказов одного узла $\lambda = 10^{-3}$ с$^{-1}$.

Оптимизацию проведем в условиях неопределенности входного потока, когда равновероятные варианты интенсивностей входного потока запросов представлены вектором (0,5; 1; 1,5; 2; 2,5; 3; 3,5; 4; 4,5; 5) с$^{-1}$. Результаты оптимизации представлены на рис. 2, на котором кривая *1* соответствует оптимальным значениям кратности резервирования запросов, определяемым для каждого варианта интенсивностей входного потока в соответствии с критерием оптимальности (5).

Зависимость вероятности своевременного получения безошибочных результатов, хотя бы одним из k_i узлов кластера при адаптивном задании кратности резервирования с учетом наблюдаемой интенсивности входного потока представлена кривой *2* на рис. 2. Кривая *3* на рис. 2 соответствует вероятности своевременного получения безошибочных результатов, хотя бы одним из k узлов кластера при определении значений кратности резервирования

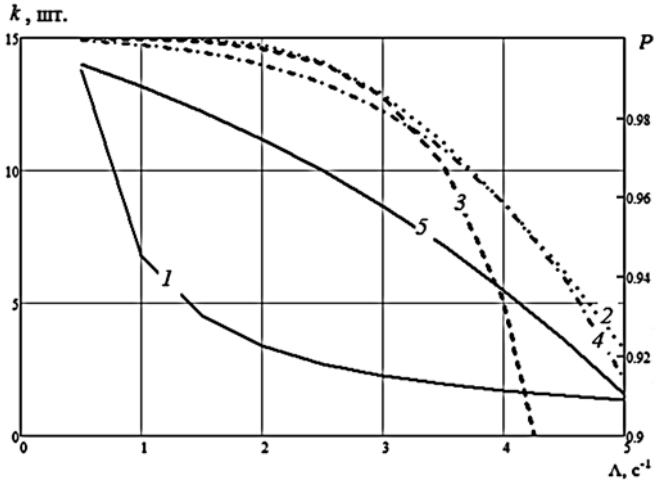

Рис. 2. Зависимость оптимальной кратности резервирования и вероятности своевременного получения безошибочных результатов при различных вариантах оптимизации

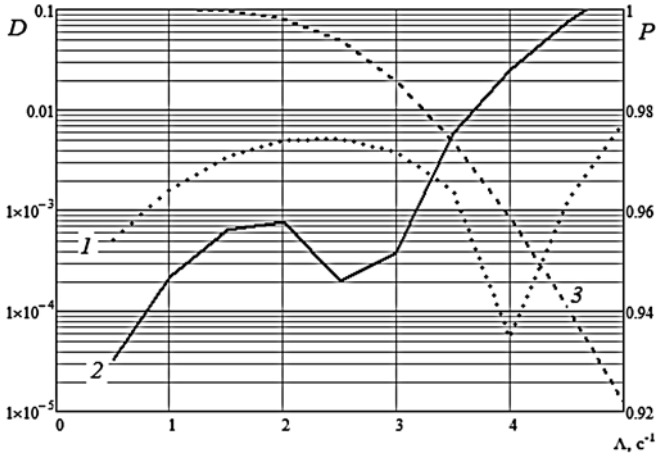

Рис. 3. Эффективность адаптивного задания кратности резервирования при изменениях интенсивности входного потока запросов

по критерию (3), а кривая *4* — по критерию (4). Кривая *5* представляет вероятность своевременности нерезервированных вычислении ($k = 1$). На рис. 3 кривыми *1* и *2* представлены потери D вероятности своевременного получения безошибочных результатов, хотя бы одним из узлов кластера при постоянной кратности резервирования, определяемой по критерию (3) и (4), относительно адаптивного назначения этой кратности при оптимизации по критерию (5).

На рис. 2 кривая *3* представляет вероятность своевременного и безошибочного обслуживания запросов, хотя бы одним из узлов кластера при адаптивном определении кратности резервирования вычислений.

Из представленных на рис. 2 и 3 зависимостей видна эффективность резервированного выполнения запросов в реальном времени, при снижении оптимальной кратности резервирования при росте интенсивности в системе реального времени. В случае значительной интенсивности запросов в результате влияния увеличения кратности резервирования на возрастание загрузки (приводящей к возможному нарушению условий стационарности) существует граница возрастания интенсивности запросов, выше которой резервированное выполнение запросов становится нецелесообразным.

Проведенные расчеты подтверждают эффективность адаптивного изменения кратности резервирования с ростом интенсивности запросов при снижении оптимальной кратности резервирования с ростом интенсивности запросов до граничного значения интенсивности запросов, когда резервирование вычислений становится нецелесообразным. Причем, при низкой интенсивности запросов оптимальная кратность резервирования вычислений наибольшая, но при этом снижается эффективность резервированных вычислений с адаптацией кратности резервирования к изменениям интенсивности.

Заключение

Для кластерных систем реального времени, консолидирующих ресурсы группы серверов с организацией локальных очередей в каждом из них, предложена оценка вероятности того, что время ожидания выдачи безошибочных результатов меньше предельно допустимого значения t_0, хотя бы в одном из узлов кластера, принимающих запрос к резервированному вычислению в условиях отказов и ошибок.

Для кластерных систем реального времени, функционирующих в условиях отказов и ошибок, показано существование оптимальной кратности резервирования запросов, обеспечивающей максимум вероятности своевременного и безошибочного их выполнения в зависимости от интенсивности запросов и допустимой задержки в очередях.

Для вычислительной системы кластерной архитектуры реального времени предложены постановка и решение задачи нахождения оптимальной кратности резервирования вычислительного процесса, при которой достигается максимум вероятности безошибочного обслуживания запроса, при времени его нахождении в очереди, меньшего предельно допустимого значения t_0, хотя бы в одном из k узлов, принимающих запрос к резервированному выполнению.

Показана высокая эффективность резервированного выполнения запросов в реальном времени при снижении оптимальной кратности резервирования в случае роста интенсивности запросов в результате влияния увеличения кратности резервирования на возрастание загрузки узлов, приводящей к существованию границы интенсивности, выше которой резервирование становится нецелесообразным.

Показана эффективность адаптивного изменения кратности резервирования в зависимости от значения интенсивности входного потока запросов с учетом существования границы интенсивности запросов, выше которой резервированное выполнение запросов нецелесообразно.

Список литературы

1. **Kopetz H.** Real-Time Systems: Design Principles for Distributed Embedded Applications. Springer, 2011. 396 p.

2. **Sorin D.** Fault Tolerant Computer Architecture. Morgan & Claypool, 2009. 103 p.

3. **Шубинский И. Б.** Функциональная надежность информационных систем: методы анализа. М.: Надежность, 2012. 296 с.

4. **Перегуда А. И., Перегуда А. А., Тимашев Д. А.** Математическая модель надежности компьютерных сетей // Надежность. 2013. № 4 (47). С. 18—30.

5. **Богатырев В. А.** К повышению надежности вычислительных систем на основе динамического распределения функций // Изв. Вузов СССР. Приборостроение. 1981. № 8. С. 62—65.

6. **Богатырев В. А., Богатырев А. В.** Функциональная надежность систем реального времени // Научно-технический вестник ИТМО. 2013. № 4. С. 150—151.

7. **Богатырев В. А., Богатырев С. В., Богатырев А. В.** Функциональная надежность вычислительных систем с перераспределением запросов // Известия вузов. Приборостроение. № 10. 2012. С. 53—57.

8. **Богатырев В. А.** Оценка вероятности безотказной работы функционально-распределенных вычислительных систем при иерархической структуре узлов // Изв. Вузов. Приборостроение. 2000. № 3. С. 67—70.

9. **Богатырев В. А.** Методы отображения и балансировки нагрузки в распределенных вычислительных системах // Информационные технологии. 1999. № 8. С. 2—5.

10. **Богатырев В. А.** Мультипроцессорные системы с динамическим перераспределением запросов через общую магистраль // Изв. вузов. Приборостроение. 1985. № 3. С. 33—38.

11. **Bogatyrev V. A., Bogatyrev S. V., Golubey I. Y.** Optimization and the Process of Task Distribution between Computer System Clusters // Automatic Control and Computer Sciences. 2012. N. 3. P. 103—111.

12. **Алиев Т. И.** Проектирование систем с приоритетами // Изв. вузов. Приборостроение. 2014. Т. 57, № 4. С. 30—35.

13. **Алиев Т. И., Муравьева-Витковская Л. А.** Приоритетные стратегии управления трафиком в мультисервисных компьютерных сетях // Изв. вузов. Приборостроение. 2011. Т. 54, № 6. С. 44—48.

14. **Гришенцев А. Ю., Коробейников А. Г.** Постановка задачи оптимизации распределенных вычислительных систем // Программные системы и вычислительные методы. 2013. № 4. С. 370—375.

15. **Гатчин Ю. А., Жаринов И. О., Коробейников А. Г.** Математические модели оценки инфраструктуры системы защиты информации на предприятии // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2 (78). С. 92—95.

16. **Богатырев В. А.** Оценка коэффициента сохранения эффективности отказоустойчивых систем из многофункциональных модулей // Методы менеджмента качества. 2001. № 9. С. 29—33.

17. **Воробьёв А. И., Колбанёв М. О., Татарникова Т. М.** Оценка вероятностно-временных характеристик процесса предоставления информационно-справочных услуг // Изв. вузов. Приборостроение. 2014. Т. 57, № 9. С. 15—18.

18. **Богатырев В. А.** Отказоустойчивость и сохранение эффективности функционирования многомагистральных распределенных вычислительных систем // Информационные технологии. 1999. № 9. С. 44—48.

19. **Богатырев В. А.** Комбинаторный метод оценки отказоустойчивости многомагистрального канала // Методы менеджмента качества. 2000. № 4. С. 30—35.

20. **Lee M. H., Dudin A. N., Klimenok V. I.** The SM/V/N queueing system with broadcasting service // Math. Probl. in Engineer. 2006. V. 2006. Article ID 98171. 18 p.

21. **Дудин А. Н., Сунь Б.** Многолинейная система MAP/PH/N с управляемым широковещательным обслуживанием ненадежными приборами // Автоматика и вычислительная техника. 2009. Т. 43, № 5. С. 32—43.

22. **Дудин А. Н., Сунь Б.** Многолинейная ненадежная система с управляемым широковещательным обслуживанием // Автоматика и телемеханика. 2009. Т. 70, № 12. С. 147—160.

23. **Богатырев В. А., Богатырев А. В.** Функциональная надежность резервированного вычислительного процесса реального времени в системах кластерной архитектуры // Автоматика и вычислительная техника. 2015. Т. 49. № 1.

24. **Вишневский В. М.** Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 512 с.

25. **Алиев Т. И., Новиков Г. И.** Метрическая теория и мониторинг компьютерных систем: состояние и проблемы // Изв. вузов. Приборостроение. 2000. Т. 43, № 3. С. 40—44.

V. A. Bogatyrev, Professor, e-mail: Vladimir.bogatyrev@gmail.com,
A. V. Bogatyrev, Post Graduate,
Saint Petersburg National Research University of Information, Technologies, Mechanics and Optics

Optimization of Redundant Routing Requests in a Clustered Real-Time Systems

The investigated cluster system is designed to work in real time when exposed to faults, failures and errors in the requirement of the accuracy and timeliness of the results of service requests with the condition that the delay in the queue is less than the specified maximum permissible value.

The aim of the research was to increase the functional reliability of the cluster real-time in the redundant service requests made by a group of κ nodes in terms of failures and errors of calculation.

To achieve the objectives of the proposed organization of the redundant distribution of queries in the cluster and the real-time model to assess the probability timeliness, and accuracy of redundant service requests. Proposed and solved the problem of optimal reservation requests in a cluster, which consists in finding the ratio of reservation requests, which guarantees the maximum probability of obtaining accurate and timely delivery of service in at least one of the nodes, receiving a request to redundant execution.

The high efficiency redundant queries in real time, while reducing the optimal ratio of redundancy in case of increase of the intensity of requests, due to the influence of increasing the ratio of the reserve to increase the load on all the nodes, leading to the existence of the border intensity, above which the reservation is not appropriate. The efficiency in the allocation of queries adaptively set the multiplicity of the reservation, pre-computed during the optimization process, depending on changes in the intensity of the input stream.

Keywords: *reliability, timeliness, real time, cluster, query, backup, optimization*

References

1. **Kopetz H.** *Real-Time Systems: Design Principles for Distributed Embedded Applications.* Springer, 2011. 396 p.
2. **Sorin D.** *Fault Tolerant Computer Architecture.* Morgan & Claypool, 2009. 103 p.
3. **Shubinskij I. B.** Funkcional'naja nadezhnost1 informacionnyh sistem: metody analiza. M.: Zhurnal *"Nadezhnost".* 2012. 296 p.
4. **Pereguda A. I., Pereguda A. A., Timashev D. A.** Matematicheskaja model nadezhnosti komp'juternyh setej. Nadezhnost'. 2013. N. 4 (47). P. 18—30.
5. **Bogatyrev V. A.** K povysheniju nadezhnosti vychislitel'nyh sistem na osnove dinamicheskogo raspredelenija funkcij. Izv. vuzov SSSR. Priborostroenie. 1981. N. 8. P. 62—65.
6. **Bogatyrev V. A., Bogatyrev A. V.** Funkcional'naja nadezhnost' sistem real'nogo vremeni. *Nauchno-tehnicheskij vestnik ITMO.* 2013. N. 4. P. 150—151.
7. **Bogatyrev V. A., Bogatyrev S. V., Bogatyrev A. V.** Funkcional'naja nadezhnost' vychislitel'nyh sistem s pereraspredeleniem zaprosov. *Izvestija vuzov. Priborostroenie.* 2012. N. 10. P. 53—57.
8. **Bogatyrev V. A.** Ocenka verojatnosti bezotkaznoj raboty funkcional'no-raspredelennyh vychislitel'nyh sistem pri ierarhicheskoj strukture uzlov. *Izv. vuzov. Priborostroenie.* 2000. N. 3. P. 67—70.
9. **Bogatyrev V. A.** Metody otobrazhenija i balansirovki nagruzki v raspredelennyh vychislitel'nyh sistemah. *Informacionnye tehnologii.* 1999. N. 8. P. 2—5.
10. **Bogatyrev V. A.** Mul'tiprocessornye sistemy s dinamicheskim pereraspredeleniem zaprosov cherez obshhuju magistral'. *Izv. vuzov. Priborostroenie.* 1985. N. 3. P. 33—38.
11. **Bogatyrev V. A., Bogatyrev S. V., Golubev I. Y.** Optimization and the Process of Task Distribution between Computer System Clusters. *Automatic Control and Computer Sciences.* 2012. N. 3. P. 103—111.
12. **Aliev T. I.** Proektirovanie sistem s prioritetami. *Izv. vysshih uchebnyh zavedenij. Priborostroenie.* 2014. V. 57, N. 4. P. 30—35.
13. **Aliev T. I., Murav'eva-Vitkovskaja L. A.** Prioritetnye strategii upravlenija trafikom v mul'tiservisnyh komp'juternyh setjah. *Izvestija vysshih uchebnyh zavedenij. Priborostroenie.* 2011. V. 54. N. 6. P. 44—48.
14. **Grishencev A. Ju., Korobejnikov A. G.** Postanovka zadachi optimizacii raspredeleOSTANOVKA jonnyh vychislitel'nyh sistem. *Programmnye sistemy i vychislitel'nye metody.* 2013. N. 4. P. 370—375.
15. **Gatchin Ju. A., Zharinov I. O., Korobejnikov A. G.** Matematicheskie modeli ocenki infrastruktury sistemy zashhity informacii na predprijatii. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki.* 2012. N 2 (78). P. 92—95.
16. **Bogatyrev V. A.** Ocenka kojefficienta sohranenija jeffektivnosti otkazoustojchivyh sistem iz mnogofunkcional'nyh modulej. *Metody menedzhmenta kachestva.* 2001. N. 9. P. 29—33.
17. **Vorob'jov A. I., Kolbanjov M. O., Tatarnikova T. M.** Ocenka verojatnostno-vremennyh harakteristik processa predostavlenija informacionno-spravochnyh uslug. *Izvestija vysshih uchebnyh zavedenij. Priborostroenie.* 2014. V. 57. N. 9. P. 15—18.
18. **Bogatyrev V. A.** Otkazoustojchivost' i sohranenie jeffektivnosti funkcionirovanija mnogomagistral'nyh raspredelennyh vychislitel'nyh sistem. *Informacionnye tehnologii.* 1999. N. 9. P. 44—48.
19. **Bogatyrev V. A.** Kombinatornyj metod ocenki otkazoustojchivosti mnogomagistral'nogo kanala. *Metody menedzhmenta kachestva.* 2000. N. 4. P. 30—35.
20. **Lee M. H., Dudin A. N., Klimenok V. I.** The SM/V/N queueing system with broadcasting service. *Math. Probl. in Engineer.* 2006. V. 2006. Article ID 98171. 18 p.
21. **Dudin A. N., Sun' B.** Mnogolinejnaja sistema MAP/PH/N s upravljaemym shirokoveshhatel'nym obsluzhivaniem nenadezhnymi priborami. *Avtomatika i vychislitel'naja tehnika.* 2009. V. 43, N. 5. P. 32—43.
22. **Dudin A. N., Sun' B.** Mnogolinejnaja nenadezhnaja sistema s upravljaemym shirokoveshhatel'nym obsluzhivaniem. *Avtomatika i telemehanika.* 2009. V. 70, N. 12. P. 147—160.
23. **Bogatyrev V. A., Bogatyrev A. V.** Functional Reliability of a Real-Time Redundant Computational Process in Cluster Architecture Systems. *Automatic Control and Computer Sciences.* 2015. V. 49, N. 1. P. 46—56. DOI 10.3103/S0146411615010022.
24. **Vishnevskij V. M.** Teoreticheskie osnovy proektirovanija komp'juternyh setej. M.: Tehnosfera, 2003. 512 p.
25. **Aliev T. I., Novikov G. I.** Metricheskaja teorija i monitoring komp'juternyh sistem: sostojanie i problemy. *Izv. vuzov. Priborostroenie.* 2000. V. 43, N. 3. P. 40—44.