МОДЕЛИРОВАНИЕ КЛАСТЕРА ВЫСОКОЙ ГОТОВНОСТИ

И.В. Комаров

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия ВО, д. 39
<mur@linux-ink.ru>

УДК 681.3
И. В. Комаров. **Моделирование кластера высокой готовности** // Труды СПИИРАН, Вып. 2, т. 2. — СПб.: Наука, 2005.

Аннотация. *Рассмотрены основные цели, решаемые системами высокой готовности. Как пример приведена архитектура и принципы работы кластера с различным количеством узлов. Проведено экспериментальное исследование зависимости критерия готовности от архитектуры кластера. Рассмотрено изменение этой зависимости при различных порядковых отношениях между исходными характеристиками.* — Библ. 7 назв.

UDC 681.3
I. V. Komarov. **Modeling High-Availability Cluster** // SPIIRAS Proceedings. Issue 2, vol. 2. — SPb.: Nauka, 2005.

Abstract. *The basic purposes decided by High-Availability systems are considered. The architecture and principles of work of a cluster with various amount of nodes is shown. As a result of experiment, dependence model of availability criterion from cluster architecture is found. Also change of dependence is considered at various ordinal relations between initial characteristics.* — Bibl. 7 items.

1. Введение

Повышение надежности компьютерных систем и сервисов является одной из наиболее востребованных задач, при этом стоимость системы часто становится определяющим фактором при проектировании таких систем. Поэтому, наряду с традиционными, достаточно дорогими отказоустойчивыми системами (Fault Tolerance Systems), в которых резервирование компонентов реализуется на аппаратном уровне, в последнее время особой популярностью стали пользоваться кластеры высокой готовности (High Availability Clusters), как относительно недорогие системы, обеспечивающие высокую надежность выполнения задач.
Технология создания кластеров высокой готовности обычно использует избыточность основных компонентов и специализированное программное обеспечение (ПО), обеспечивающее своевременное переключение сервисов при возникновении сбоя в работе. Максимально отказоустойчивая система не должна иметь активного элемента, отказ которого может привести к потере функциональности системы. В подобных случаях система обозначается как NSPF (No Single Point of Failure, — англ., отсутствие единой точки отказа) [6].

При построении систем высокой готовности, главная цель — обеспечение минимального времени простоя. Для достижения этой цели система высокой готовности должна характеризоваться: максимальной надежностью компонентов, отказоустойчивостью, допустимостью замены компонент без останова комплекса и удобством в обслуживании.

Обеспечение заданной надежности достигается использованием электронных компонентов высокой и сверхвысокой интеграции, поддержанием нормальных режимов работы. Отказоустойчивость осуществляется путем использования специализированных компонентов, а также с помощью технологий кластеризации. Благодаря кластеризации достигается такая схема функционирования, когда при отказе одного из компьютеров задачи перераспределяются между

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

исправно функционирующими узлами кластера. Причем одной из важнейших задач производителей кластерного ПО является обеспечение минимального времени восстановления системы после сбоя. Удобство в обслуживании, которое служит уменьшению плановых простоев (например, замены вышедшего из строя оборудования) является одним из важнейших параметров систем высокой готовности. Если система не допускает замены компонентов без выключения комплекса, то ее готовность уменьшается. Пренебрежение любым из указанных факторов, может привести к потере функциональности системы.

В настоящее время наиболее применяемы следующие два типа архитектуры кластеров высокой готовности [6] (рис. 1,2):

- архитектура без разделения ресурсов (рис. 1);
- архитектура с общей дисковой системой (рис. 2).

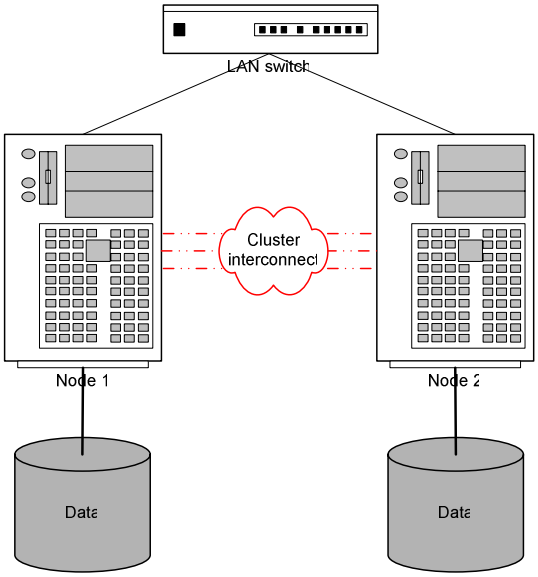

Рис. 1. Архитектура без разделения ресурсов.

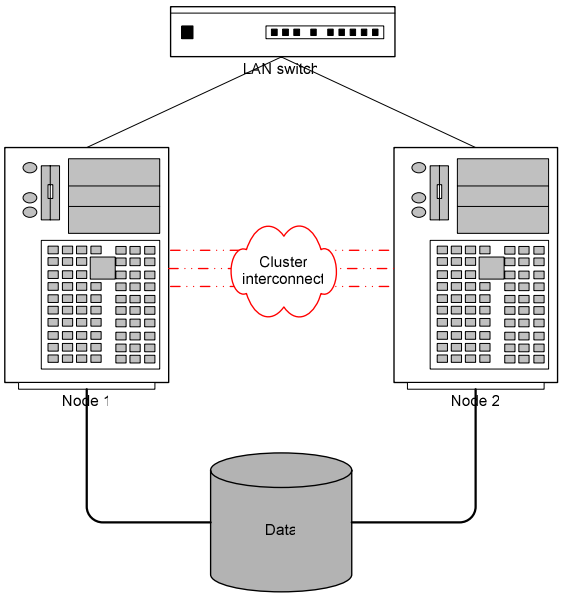

Рис. 2. Архитектура с общей дисковой системой.

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Архитектура без разделения ресурсов не использует общей системы хранения данных, каждый узел имеет свои дисковые накопители, которые не используются совместно узлами кластерной системы. Фактически на аппаратном уровне разделяются только коммуникационные каналы. Если задача позволяет логически разделить данные для того, чтобы запрос из некого подмножества запросов можно было бы обработать с использованием части данных, то система без разделения ресурсов может оказаться более эффективным решением.

Архитектура с общей дисковой системой (рис. 2) используется для построения кластерных систем высокой готовности, ориентированных на обработку больших объемов данных. Такая система состоит из узлов кластера и общей системы хранения данных, доступной для любого узла. При работе с задачами обработки данных, архитектура с общими дисками является более эффективной, так как в этом случае не нужно держать несколько копий данных и, в то же время, при выходе из строя узла данные могут быть мгновенно доступны для других узлов. Кроме того, использование архитектуры с общей дисковой системой эффективно тогда, когда в качестве задач, работающих на кластере, используются уже готовые приложения.

С точки зрения распределения программных ресурсов между узлами при построении кластера высокой готовности могут использоваться различные технологии. Так, при выполнении одной задачи на нескольких узлах одновременно (модель “активный-активный”) к отказоустойчивости добавляется высокая производительность. При этом в случае отказа одного из узлов, его часть задачи мигрирует или распределяется между рабочими узлами. Однако довольно часто требуется обеспечение отказоустойчивого функционирования уже готовых программных решений. К сожалению, модель “активный-активный” в таком случае не работает. Для подобных ситуаций используется модель “активный-пассивный”, в которой обеспечивается миграция задач со сбойного узла на работающие [6].

2. Архитектура и принципы работы

Рассмотрим модель двухузлового кластера с резервированием основных компонентов и связей, работающего в режиме “активный-пассивный” (рис. 3).

Комплекс состоит из двух узлов, один из которых находится в рабочем режиме, а второй в режиме ожидания. Два массива с общими данными, к которым оба узла имеют доступ, синхронизируют данные в процессе работы, что позволяет в случае отказа одного продолжить работу со вторым без перезагрузки узла. Внутренние диски каждого из узлов также дублированы. У каждого из узлов резервированы те компоненты и связи (пути доступа), на которые приходится основная нагрузка в процессе работы: внутренние диски с ОС и кластерным ПО, дисковые контроллеры и пути к массивам с данными, межкластерные соединения, сетевые контроллеры и пути доступа во внешнюю сеть. Межкластерные соединения служат для обмена служебной информацией между узлами (определением состояния узлов и синхронизацией настроек).

При отказе дублированного компонента или связи, узел “на ходу” переключается на резерв с минимальным временем простоя. В случае сбоя недублированного компонента или отказа ОС, пассивный узел берет на себя выполнение задачи, переключает соединения с пользователями и становится активным. После устранения отказа первый узел становится пассивным и переходит в режим

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ожидания. Сформулируем ключевые принципы, используемые при проектировании кластера высокой готовности:

1. организация системы хранения данных и синхронизация данных в процессе работы;
2. межузловой мониторинг и синхронизация депозитария ресурсов кластера;
3. миграция сервисов (задач) между узлами;
4. выбор кратности резервирования.

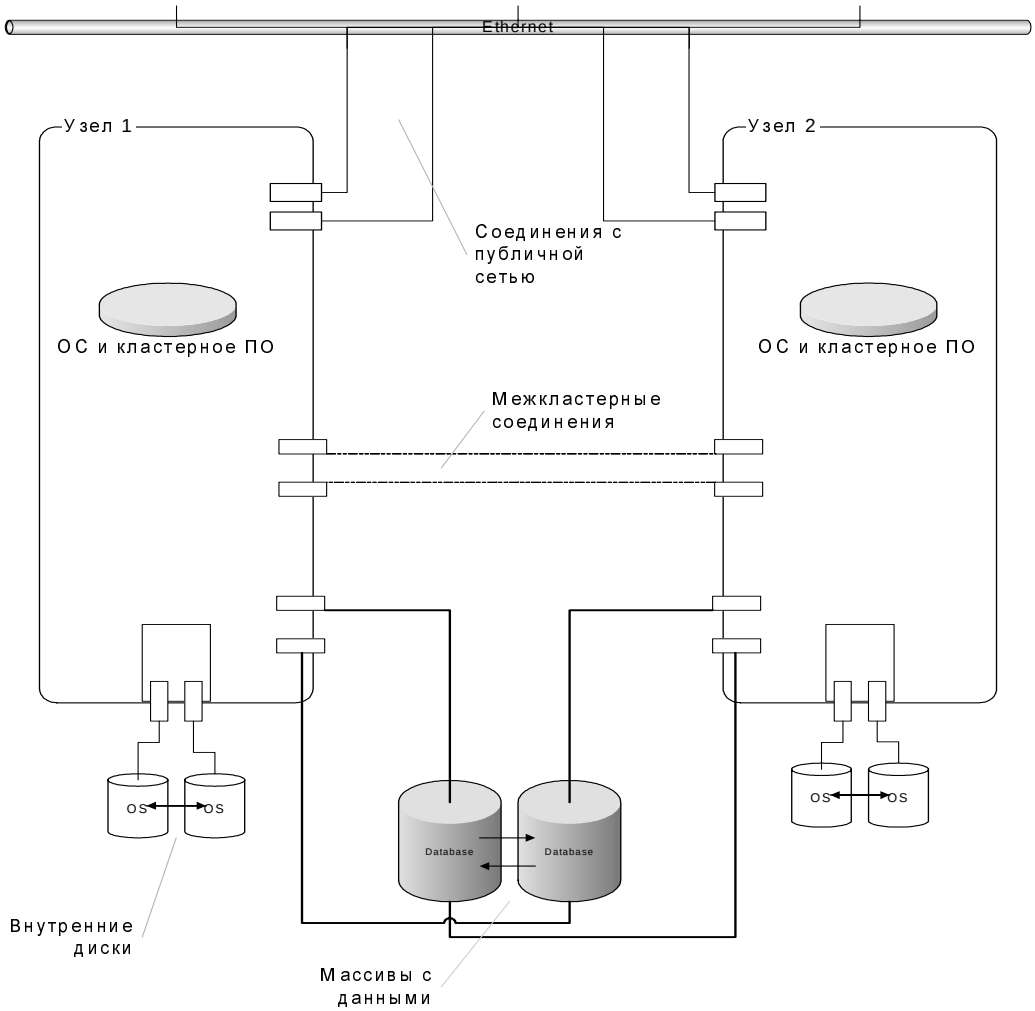

Рис. 3. Архитектура двухузлового кластера с общей дисковой системой.

Рассмотрим подробнее каждый из принципов проектирования.

1. При построении общей дисковой системы могут использоваться технологии доступа к данным SCSI или FC-AL (Fibre Channel Arbitrated Loop). В случае SCSI соединение “узел1-массив-узел2” строится на общей SCSI-шине, что обеспечивает возможность доступа к данным на массиве с обоих узлов. Учитывая принцип резервирования, таких массивов должно быть два или больше, таким образом добавляется еще одна общая SCSI-шина и получившиеся соединения условно можно представить как “узел1{SCSI-контроллер1}-массив1-узел2{ SCSI-контроллер1}” и “узел1{SCSI-контроллер2}-массив2-узел2{ SCSI-контроллер2}”. На уровне операционной системы драйвер метаустройств осуществляет зеркалирование массивов 1 и 2, что позволяет данным синхронизироваться в процес-

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

се изменения мгновенно (например, чтение происходит с массива 1, а запись на массив 1 и массив 2 одновременно). В случае использования технологии Fibre Channel, соединения узлов с массивами строятся через оптические развязки (FC-AL хабы или коммутаторы) и в принципе мало чем отличаются от обычных сетевых соединений. Зеркалирование данных на массивах может осуществляться как средствами ОС (как в случае с общей SCSI-шиной), так и средствами самих массивов (современные интеллектуальные системы хранения данных это позволяют). При нескольких путях доступа к массивам можно применять принцип выравнивания нагрузки потока данных, то есть обмен данных между узлом и массивом будет проходить через оба канала одновременно.

2. Межузловой мониторинг является основной частью кластерного ПО и отвечает за определение состояния комплекса: опрос узлов, определение их доступности и арбитраж в спорных ситуациях (например, при одновременной попытке захвата ресурса несколькими узлами). Синхронизация депозитария ресурсов необходима для своевременного обновления системной информации и конфигурации сервисов кластера на всех узлах. Как правило, межузловой мониторинг реализуется как часть ядра, а синхронизация депозитария ресурсов как набор сервисов и приложений. Для связи и обмена данными между узлами используются межкластерные соединения

3. Автоматическая миграция сервисов с одного узла на другой происходит в том случае, когда система межузлового мониторинга принимает решение о неисправности или неполной функциональности узла. Решение формируется либо на основании диагностической информации, либо при отсутствии связи с узлом (в случае выхода из строя узла или отсутствия связи через межкластерные соединения). В некоторых случаях перемещение сервисов между узлами может быть инициировано оператором, например, в случае сервисных работ на узле. Такая миграция называется принудительной [5].

4. Коэффициент (порядок, кратность) резервирования компонента (комплекса) характеризует степень его избыточности. Кратность резерва — это отношение числа резервных элементов объекта к числу резервируемых ими основных элементов, выраженное несокращенной дробью [6]. Резервирование с целой кратностью имеет место, когда основной элемент резервируется одним или более резервными элементами. Резервирование, кратность которого равна единице, называется дублированием. Так в модели рис. 3 коэффициент резервирования дисковых контроллеров для узла - 1, а для комплекса – 3. Это значит, что при выходе из строя двух контроллеров на первом узле он потеряет функциональность, а при выходе из строя всех четырех контроллеров весь комплекс прекратит работу. Выбор порядка резервирования при проектировании кластера является важной задачей, при решении которой нужно руководствоваться тем, что для наиболее нагруженных и наименее надежных компонентов следует обеспечивать большую кратность резервирования [4]. Заметим, что в данном случае речь идет о структурном резервировании, другие типы резервирования (временное, информационное, функциональное и нагрузочное) не рассматриваются.

На основе двухузловой модели кластера можно построить трехузловую модель путем добавления дополнительного узла, межкластерных соединений и связей с дисковыми массивами и внешней (публичной сетью). Это не единственно возможная, но часто применяемая на практике конфигурация. Кластеры с большим числом узлов могут строиться путем объединения этих базовых моде-

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

лей через межкластерные соединения, за счет чего достигается резервирование комплекса более высокого порядка.

Надо отметить, что описанная выше архитектура двухузловой модели кластера является наиболее применяемой в связи с тем, что получение полнофункционального комплекса достигается при относительно низких затратах на оборудование и обслуживание. В некоторых случаях возможны видоизменения модели, например внесение дополнительных контроллеров и путей доступа к данным, добавление оптических развязок и коммутаторов, увеличение количества межкластерных соединений и.т.д., однако в целом принцип построения остается неизменным.

3. Определение критерия готовности кластера

Увеличение порядка резервирования не всегда ведет к росту готовности и, следовательно, эффективности кластера. Существуют дополнительные аспекты, связанные с увеличением количества узлов и связей в кластере высокой готовности. Во-первых, время восстановления узла в кластере и время синхронизации депозитария ресурсов при увеличении количества узлов в кластере возрастает. Во-вторых, увеличивается время диагностики целостности кластерной системы. И, наконец, растет стоимость построения и обслуживания комплекса, а также его сложность. Таким образом, пока неясно насколько и до какой степени эффективно наращивать кластерный комплекс и при какой конфигурации получается наилучший результат. Именно это является предметом настоящего исследования.

Прежде всего, необходимо отметить что кластер - многопараметрический объект с большим числом степеней свободы и многие его характеристики выпадают из поля зрения. Поэтому для оценки качества модели мы попробуем определить основные характеристики, важные при проектировании и использовании кластера.

Для начала рассмотрим основные элементы комплекса, по которым производиться резервирование (фактически эти элементы определяют конфигурацию системы, как кластера). С точки зрения структуры, кластер характеризуется количеством: узлов, связей с внешней сетью, путей доступа к данным, межкластерных соединений, массивов с данными. На первый взгляд может показаться, что некоторые из этих характеристик зависимы, хотя на самом деле это не так. Например, при построении двухузлового кластера резервирование межкластерных соединений может иметь коэффициент больше 2 (например 5), а резервирование связей с внешней сетью - коэффициент 4 (или любой другой). Кроме того, может использоваться более двух массивов с данными, а количество путей от узлов к массивам может быть увеличено (за счет добавления оптических развязок или иным способом).

Помимо описанных количественных характеристик, определяющих структуру кластера, необходимо ввести дополнительные, определяющие качество функционирования, а именно: среднее время бесперебойной работы комплекса в год (MTBF - англ., Mean Time Before Failure), среднее время восстановления узла при сбое (MTTR - англ., Mean Time To Repair) и стоимость кластерного комплекса.

Таким образом, оценка качества модели кластерного комплекса, которую нужно получить, зависит от структурных особенностей (резервирование элементов), надежности работы (время работы и время восстановления) и стоимости построения и обслуживания. Количественную оценку эффективности модели,

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

учитывающую указанные факторы, будем называть критерием готовности кластера.

Для расчета критерия готовности использовалась Оболочка Системы Поддержки Принятия Решений (ОСППР) АСПИД-3W, предназначенная для всестороннего оценивания сложных объектов в условиях неопределенности [2]. На основе результатов расчетов (оценки сводных показателей) строились графики зависимости критерия готовности от типа архитектуры. В эксперименте исследовалась выборка из семи моделей: 2-х, 3-х, 4-х, 6-и, 8-и, 16-и и 32-х узловые кластеры. Для каждой из моделей был произведен расчет исходных характеристик.

Структурные характеристики рассчитывались исходя из архитектуры кластера и топологии соединений. При выборе среднего значения времени бесперебойной работы кластера использовались как экспериментальные данные, так и данные из различных источников и публикаций [1], [3], [5].

Среднее время восстановления узла рассматривалось как комплексная величина, состоящая из времени диагностирования неисправности оборудования и кластерного ПО ($15s + N \times 1s$, где N — количество узлов, s - секунды), времени восстановления системы (20 s), времени перезапуска сервисов или перезагрузки операционной системы (20 s.) и временем регистрации в кластере ($10s + N \times 1s$). Эти значения приблизительно соответствуют среднему времени восстановления узла в системе Sun Cluster 3.0 при отсутствии активных сервисов.

Стоимость кластерного комплекса оценивалась как сумма условных стоимостей элементов комплекса и стоимостей работ по установке и обслуживанию ПО на каждом из узлов. Элементы и соединения кластера, а также стоимость работ по обслуживанию оценивались в условных единицах, отношения между которыми примерно соответствуют реальным ценам.

Итоговый расчет исходных характеристик для каждой из исследуемых моделей представлен в табл. 1.

Таблица 1. Исходные характеристики

Модель кластера	Количество					MTBF	MTTR	Стоимость
	узлов	связей с внешней сетью	путей к данным	межкластерных соединений	массивов			
2 узла	2	4	4	2	2	3000	69	8.6
3 узла	3	6	6	2	2	4000	71	12.1
4 узла	4	8	8	4	4	6000	73	17.6
6 узлов	6	12	12	4	4	6000	77	24.2
8 узлов	8	16	16	8	8	7000	81	35.2
16 узлов	16	32	32	16	16	8000	97	70.4
32 узла	32	64	64	32	32	8000	129	140.8

Вычислительный эксперимент проводился в два этапа: на первом этапе расчет граничных (минимально и максимально возможных) значений исходных характеристик производился ОСППР АСПИД-3W, а на втором - методом экспертных оценок.

Для каждого из этапов были рассмотрены различные порядковые отношения между весовыми показателями характеристик:

- степень взаимовлияния характеристик не определена (все характеристики равнозначимы);

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- основной вес имеют стоимость комплекса и время бесперебойной работы, время восстановления узла вторично (cost=MTBF, MTBF > MTTR), что является условиями для оптимизации по стоимости и надежности;
- основной вес имеют время бесперебойной работы комплекса и время восстановления узла (MTBF=MTTR), что соответствует оптимизации по надежности и времени восстановления;
- основной вес имеет время бесперебойной работы комплекса (MTBF>MTTR,MTBF >cost), что характерно для критически важных систем (Mission Critical Systems).

Результаты зависимости сводной оценки (критерия готовности) от количества узлов (типа модели) для обоих этапов эксперимента отражены на графиках (рис. 4, 5).

Рис. 4. График зависимости критерия готовности от количества узлов, граничные значения исходных характеристик рассчитываются программой.

Из рис. 4 и 5 видно, что весовые коэффициенты, задающие степени влияния отдельных показателей на критерий готовности, не изменяют характер кривой. Однако при различных условиях, которые задаются порядковыми отношениями весов, точки излома и угол наклона кривой различны.

Рассмотрим графики на рис. 5. В случае, когда время бесперебойной работы имеет доминирующее значение по сравнению со временем восстановления узла (два верхних графика), наиболее эффективными конфигурациями становятся 8 и 16 узловые модели. Кроме того, хорошо виден прогиб кривой для 6

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

узлового кластера и малый прогиб для 3 узлового. Это можно объяснить меньшим уровнем резервирования базовой 3 узловой модели (коэффициент резервирования соединений — 3, а массивов с данными — 2) при большей стоимости и большем времени восстановления. Таким образом, можно предположить, что при выбранной технологии построения базовой 3 узловой модели все архитектуры, содержащие ее как элемент (6, 12 и.т.д.), будут иметь меньший критерий готовности, чем модели, построенные на резервированных парах. Интересно отметить, что для различных порядковых отношений между весовыми показателями характеристик наибольшим значением критерия готовности обладают кластерные модели с 4, 8 и 16 узлами. График, где отношения между весовыми показателями не задан (непрерывная линия), наиболее обще отражает изменения критерия готовности, так как результирующая оценка оптимизирована по всем характеристикам. Надо также отметить тот факт, что на всех графиках для модели с количеством узлов больше 16 критерий готовности падает.

Рис. 5. График зависимости критерия готовности от количества узлов, граничные значения исходных характеристик рассчитываются методом экспертных оценок.

4. Заключение

Полученные результаты позволяют сделать следующие выводы:

- на всех этапах эксперимента характер кривой, отражающей зависимость критерия готовности от модели кластера неизменен;

Труды СПИИРАН. 2005. Вып. 2. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2005. Issue 2. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- для разных порядковых отношений весовых коэффициентов наибольший критерий готовности имеют соответствующие модели;
- наивысший критерий готовности имеют модели с 4, 8 и 16 узлами;
- 3 узловая модель либо неправильно спроектирована, либо малоэффективна;
- по расчетным данным для сбалансированности по параметрам стоимость - время бесперебойной работы нужен 8 или 16 узловой кластер, а для сбалансированности по всем параметрам — 4 или 8 узловой;
- при количестве узлов в модели больше 16 критерий готовности падает, при любых порядковых отношениях между характеристиками;
- для исследования поведения кластеров с большим количеством узлов нужна большая выборка моделей и, возможно, дополнительные характеристики.

Предлагаемая оценка готовности и ряд характеристик требует дальнейшего как экспериментального, так и теоретического обоснования. Уточнение анализа и рассмотрение дополнительных характеристик, таких, как вероятность восстановления узла после сбоя и интенсивность отказов [7], а также увеличение исследуемой выборки моделей с различными типами архитектур является предметом последующих исследований.

Литература

[1] *Ira Pramanick, Cluster Software Development Group.* Modeling Sun Cluster Availability // Sun BluePrint Online, December 2002.

[2] *Хованов К.Н.* Свидетельство об официальной регистрации программы для ЭВМ №960087. Программа для ЭВМ "Анализ и Синтез Показателей при Информационном Дефиците. АСПИД-3W" // РосАПО. М., 22.09.1996.

[3] *Michel R. Lyu, Veena B. Mendiratta.* Software Fault Tolerance in a Clustered Architecture: Techniques and Reliability Modeling // Hong Kong, 2003.

[4] *Таненбаум Э.* Распределенные системы принципы и парадигмы // СПб: Питер, 2003

[5] *Sun Cluster Software Development Group.* Sun Cluster 3.1 Documentation // Sun Microsystems, 2003.

[6] *Савяк В.* Эффективные кластерные решения. // <http://www.ixbt.com/cpu/clustering.shtml>, 11.04.2002

[7] *Ramakumar R.* Reliability Engineering, The Electrical Engineering Handbook // CRC Press LLC, 2000.