УДК 004.052.32

ВЫЧИСЛИТЕЛЬНЫЙ КЛАСТЕР ТИПА «BEOWULF». ОСОБЕННОСТИ АРХИТЕКТУРЫ, ОРГАНИЗАЦИЯ СЕТИ, ОЦЕНКА НАДЁЖНОСТИ

РЫЖКОВА О. В.

СТАДНИК И. П.

доктор технических наук

Симферополь

Высокопроизводительные экономические расчеты нередко производятся с помощью параллельных и распределенных вычислений на кластерных системах различного типа. Одной из основных задач работы кластера является обеспечение его продолжительного надежного функционирования. Эту задачу включает три составляющие: безотказность, готовность и безопасность [1].

Одной из актуальных задач, возникающих при проектировании и модернизации кластерных систем является задача обеспечения заданного уровня надежности. При решении такой задачи возникает возможность уже на стадии проектирования оценивать уровень надежности предлагаемых схем и технологий обработки данных. Для такой оценки необходима разработка математических моделей, учитывающих особенности режима эксплуатации.

Эффективное применение сетевых кластерных технологий позволяет обеспечить не только повышение надежности функционирования компьютерной системы (КС), но и повысить ее информационную безопасность, за счет применения сетевых (распределенных территориально) кластерных систем. В этой связи важнейшую роль играет надежность программного обеспечения, используемого как для управления кластером, так и для решения конкретных прикладных задач.

Процесс оценки надежности состоит из нескольких этапов оценки показателей и характеристик отдельных компонент кластера и их совокупности в целом. На каждом этапе используются результаты анализа, выполненного с помощью теории надежности [2].

Целью данной статьи является анализ и обоснование выбора типа кластерной системы для высокопроизводительных экономических расчетов, вычислений и операций, а также разработка методики комплексной оценки надежности аппаратных средств кластерной системы и предложение обоснованных рекомендаций по повышению показателей готовности кластера для улучшения эффективности его работы.

ВЫБОР ТИПА КЛАСТЕРА ДЛЯ ПАРАЛЛЕЛЬНЫХ И РАСПРЕДЕЛЕННЫХ ЭКОНОМИЧЕСКИХ ВЫЧИСЛЕНИЙ

Двумя основными проблемами построения вычислительных систем для критически важных приложений, связанных с обработкой транзакций, управлением базами данных и обслуживанием телекоммуникаций, являются обеспечение высокой производительности и продолжительного функционирования систем. Наиболее эффективный способ достижения заданного уровня производительности – применение параллельных масштабируемых архитектур (кластерных систем) [5].

Кластер представляет собой два или больше компьютеров (называемых узлами), объединяемых при помощи сетевых технологий на базе шинной архитектуры или коммутатора и предстающих перед пользователями в качестве единого информационно-вычислительного ресурса. В качестве узлов кластера могут быть выбраны серверы, рабочие станции и даже обычные персональные компьютеры. Преимущество кластеризации – повышение работоспособности в случае сбоя какого-либо узла: при этом другой узел кластера может взять на себя нагрузку неисправного узла, и пользователи не заметят прерывания в доступе. Возможности масштабируемости кластеров позволяют многократно увеличивать производительность приложений для большего числа пользователей. Кластеризация может быть осуществлена на разных уровнях компьютерной системы, включая аппаратное обеспечение, операционные системы, программы-утилиты, системы управления и приложения. Чем больше уровней системы объединены кластерной технологией, тем выше надежность, масштабируемость и управляемость кластера.

Beowulf-кластер, как правило, является системой, состоящей из одного серверного узла (головного узла), а также одного или нескольких подчинённых узлов (вычислительных узлов), соединённых посредством стандартной компьютерной сети. Система строится с использованием стандартных аппаратных компонент, таких как ПК, запускаемых под Linux, стандартных сетевых адаптеров (например, Ethernet) и коммутаторов. Нет особого программного пакета, называемого «Beowulf». Вместо этого имеется несколько кусков программного обеспечения, которые многие пользователи нашли пригодными для построения кластеров Beowulf. Beowulf использует такие программные продукты как операционную систему Linux, системы передачи сообщений PVM, MPI, системы управления очередями заданий и другие стандартные продукты. Серверный узел контролирует весь кластер и обслуживает файлы, направляемые к клиентским узлам.

В качестве обучающей системы для проведения высокопроизводительных, трудоемких математических и экономических расчетов с большим количеством данных целесообразно использовать кластер типа Beowulf по нескольким причинам:

1. В каждом университете есть парк компьютерной техники (как правило, несколько десятков или сотен компьютеров, соединенных посредством компьютерной сети), который используется для учебных целей и занятий лишь несколько часов в сутки. В остальное время это оборудование простаивает или находится в выключенном состоянии, поэтому удобно в это время занять свободные технические ресурсы высокопроизводительными вычислениями.

2. Выгодно использовать уже готовую, функционирующую, налаженную топологию и компьютерную сеть образовательного учреждения (которая, как правило, имеет иерархическую структуру) в качестве коммутационной сети передачи данных, управления и администрирования кластера.

3. Имеющееся оборудование и компьютеры парка компьютерной техники обычно имеют различную спецификацию, характеристики, компоненты разных фирм-производителей и моделей, поэтому логично их объединить именно в кластер типа Beowulf, что является его особенностью.

Повышение надежности кластера основано на принципе предотвращения неисправностей путем снижения интенсивности отказов и сбоев за счет применения электронных схем и компонентов с высокой и сверхвысокой степенью интеграции, снижения уровня помех, облегченных режимов работы схем, обеспечение тепловых режимов их работы, а также за счет совершенствования методов сборки аппаратуры. Повышение уровня готовности предполагает подавление в определенных пределах влияния отказов и сбоев на работу системы с помощью средств контроля и коррекции ошибок, а также средств автоматического восстановления вычислительного процесса после проявления неисправности, включая аппаратурную и программную избыточность, на основе которой реализуются различные варианты отказоустойчивых архитектур. Повышение готовности есть способ борьбы за снижение времени простоя системы. Основные эксплуатационные характеристики системы существенно зависят от удобства ее обслуживания, в частности от ремонтопригодности, контролепригодности и т.д.

ОЦЕНКА НАДЕЖНОСТИ ВЫЧИСЛИТЕЛЬНОГО КЛАСТЕРА ТИПА BEOWULF ЦЕНТРА КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ ТАВРИЧЕСКОГО НАЦИОНАЛЬНОГО УНИВЕРСИТЕТА ИМ. В. И. ВЕРНАДСКОГО (ЦКТ ТНУ)

Рассмотрим схему кластера для определения компонент, которые подвержены отказам и сбоям (*рис. 1*).

Такая схема сети является одной из наиболее распространенных схем реализации кластера. Имеется центральный коммутатор, к которому подключены центральный сервер с одной стороны, и 8 внутренних подсетей с другой, в каждой из которых по 10 компьютеров, что в сумме дает 80 вычислительных узлов.

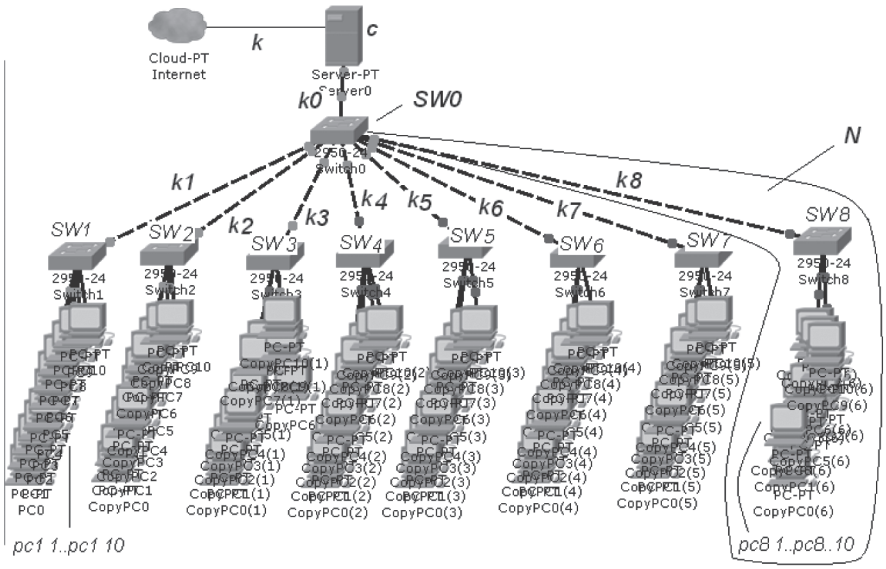

Рис. 1. Схема кластера ЦКТ ТНУ

АНАЛИЗ ПОТОКОВ ОТКАЗОВ И ИНТЕНСИВНОСТЕЙ ВОССТАНОВЛЕНИЯ ЭЛЕМЕНТОВ КЛАСТЕРА

Для такого кластера основными компонентами отказов являются (*табл. 1*)

Таблица 1

Анализ характеристик компонентов отказов

Кабель k, k0, kn, knk	UTP категории 5 (ГОСТ 15150-69)	15 лет (3 153 600 часов)	1 час
Коммутатор Sw0, Swn	DLink 2 уровня DES 1016D	89 312 часов	3 месяца (2 160 часов)
ПК	A-Line # 789600 (паспорт, соответствие ТУУ 05837085.001-97)	20 000 часов	672 часа
Сервер	A-Line # 789600 (паспорт, соответствие ТУУ 05837085.001-97)	20 000 часов	672 часа

Для упрощения дальнейших расчетов найдем параметр потока отказов λ_N и интенсивность восстановления μ_N для каждой подсети N_n. В нашем случае эти параметры для каждой подсети будут одинаковыми.

Каждая подсеть состоит из кабеля *Kn*, идущего от центрального коммутатора к коммутатору подсети *Nn*, коммутатора подсети *SWn*, 10 компьютеров в каждой из подсетей и 10 кабелей витой пары, идущим к ним от коммутатора. Объединим эти элементы в одну логическую единицу и обозначим N. Согласно [4], $\lambda = \lambda_1 + \lambda_2 + +\ldots+ \lambda_n$ интенсивность отказов независимых компонентов складываются,

Следовательно

$$\lambda_N = \lambda_{Kn} + \lambda_{SWn} + 10 * \lambda_{knk} + 10 * \lambda_{pcn}.$$

Аналогично

$$\mu_N = \mu_{Kn} + \mu_{SWn} + 10 * \mu_{knk} + 10 * \mu_{pcn}.$$

РАЗМЕЧЕННЫЙ ГРАФ ПЕРЕХОДОВ СИСТЕМЫ

Построим размеченный граф переходов системы.

Для этого определим пространство состояний {S} КС.

Состояние S0 – K C K0 SW0 NNNNNNNN: система полностью работоспособна, нет отказов или сбоев ни одного из составляющих.

S1 – KCK0SW0~~N~~NNNNNNN: отказ одной подсети (N), все остальные элементы работоспособны

S2 – KCK0SW0~~NN~~NNNNNN: отказ двух подсетей, все остальные элементы работоспособны

S3 – KCK0SW0~~NNN~~NNNNN: отказ трех подсетей, все остальные элементы работоспособны

S4 – K C K0 SW0 ~~NNNN~~NNNN: отказ четырех подсетей

S5 – K C K0 SW0 ~~NNNNN~~NNN: отказ пяти подсетей

S6 – K C K0 SW0 ~~NNNNNN~~NN

S7 – K C K0 SW0 ~~NNNNNNN~~N

Состояния {S1 – S7} образуют область состояний, где система работоспособна.

S8 – K C K0 SW0 ~~NNNNNNNN~~ – полностью неработоспособное состояние.

Если в любом из этих состояний выйдет из строя кабель K или сервер C, или кабель K0, или свитч SW0, то система также будет считаться полностью неработоспособной. Такие состояния считаем конечными.

S9 – ~~K~~ C K0 SW0 NNNNNNNN;

S10 – K ~~C~~ K0 SW0 NNNNNNNN;

S11 – K C ~~K0~~ SW0 NNNNNNNN;

S12 – K C K0 ~~SW0~~ NNNNNNNN;

S13 – ~~K~~ C K0 SW0 ~~N~~NNNNNNN;

S14 – K ~~C~~ K0 SW0 ~~N~~NNNNNNN;

S15 – K C ~~K0~~ SW0 ~~N~~NNNNNNN;

S16 – K C K0 ~~SW0~~ ~~N~~NNNNNNN;

S17 – ~~K~~ C K0 SW0 ~~NN~~NNNNNN;

S18 – K ~~C~~ K0 SW0 ~~NN~~NNNNNN;

S19 – K C ~~K0~~ SW0 ~~NN~~NNNNNN;

S20 – K C K0 ~~SW0 NN~~NNNNNN;

S21 – ~~K~~ C K0 SW0 ~~NNN~~NNNNN;

S22 – K ~~C~~ K0 SW0 ~~NNN~~NNNNN;

S23 – K C ~~K0~~ SW0 ~~NNN~~NNNNN;

S24 – K C K0 ~~SW0 NNN~~NNNNN;

S25 – ~~K~~ C K0 SW0 ~~NNNN~~NNNN;

S26 – K ~~C~~ K0 SW0 ~~NNNN~~NNNN;

S27 – K C ~~K0~~ SW0 ~~NNNN~~NNNN;

S28 – K C K0 ~~SW0 NNNN~~NNNN;

S29 – ~~K~~ C K0 SW0 ~~NNNNN~~NNN;

S30 – K C K0 ~~SW0 NNNNN~~NNN;

S31 – K C ~~K0~~ SW0 ~~NNNNN~~NNN;

S32 – K C K0 ~~SW0~~ ~~NNNNN~~NNN;
S33 – ~~K~~ C K0 SW0 ~~NNNNNN~~N;
S34 – K ~~C~~ K0 SW0 ~~NNNNNN~~N;
S35 – K C ~~K0~~ SW0 ~~NNNNNN~~N;
S36 – K C K0 ~~SW0~~ ~~NNNNNN~~N;
S37 – ~~K~~ C K0 SW0 ~~NNNNNNN~~;
S38 – K ~~C~~ K0 SW0 ~~NNNNNNN~~;
S39 – K C ~~K0~~ SW0 ~~NNNNNNN~~;
S40 – K C K0 ~~SW0~~ ~~NNNNNNN~~.

Строим размеченный граф переходов системы (*рис. 2*).

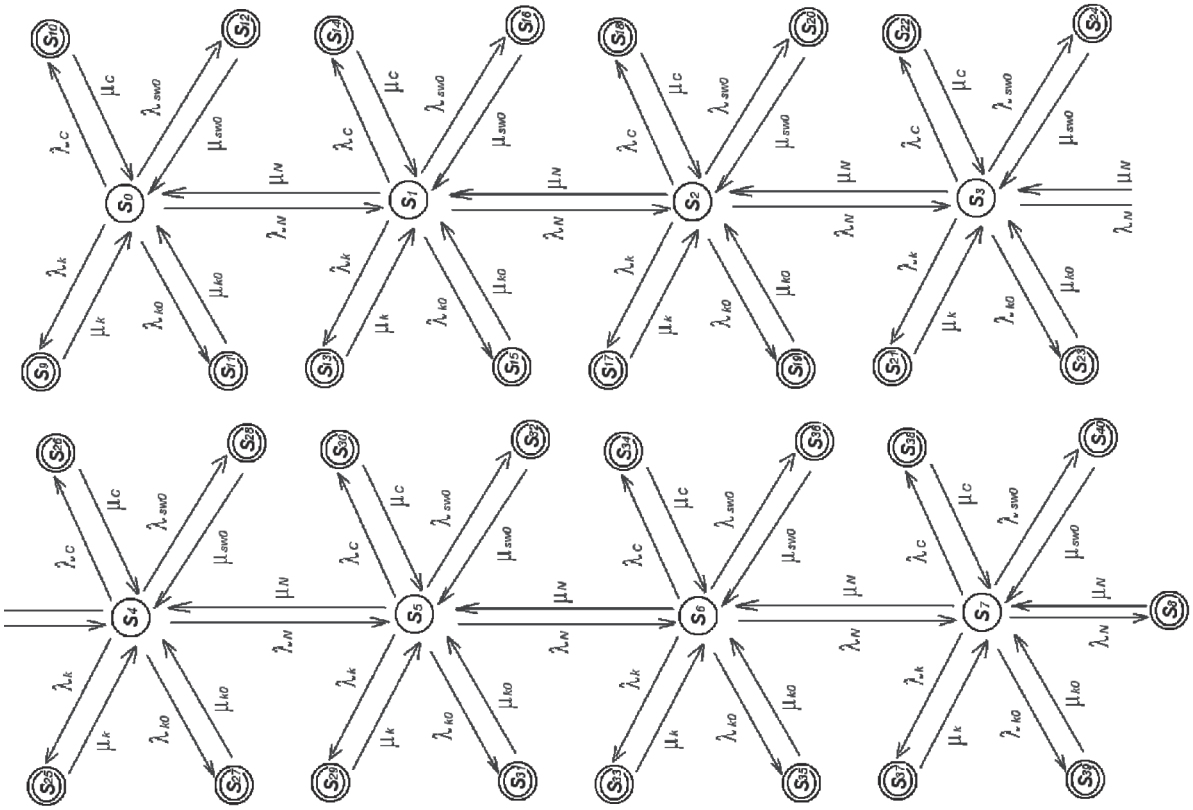

Рис. 2. Размеченный граф переходов системы

По графу составляем систему дифференциальных уравнений Колмогорова по принципу, который описан в [3]. В качестве решения этой динамической системы уравнений, используя надстройку в MathCAD, получим матрицу, элементы которой – вероятности нахождения системы в каждом из состояний в каждый момент (*рис. 3*).

Для оценки коэффициента готовности системы выберем из полученной матрицы те столбцы (область состояний), в которых система работоспособна, (вероятности состояний {S1 – S7}) и рассчитаем сумму всех этих вероятностей на каждом шаге, что и будет являться коэффициентом готовности в каждый момент времени, т. е. вероятностью в этот момент времени застать систему в работоспособном состоянии.

Представим полученные результаты графически: (*рис. 4*).

РЕКОМЕНДАЦИИ ПО УСОВЕРШЕНСТВОВАНИЮ МОДЕЛИ ОЦЕНКИ НАДЕЖНОСТИ КЛАСТЕРА

Для усовершенствования существующей модели оценки надежности кластера ЦКТ ТНУ и улучшения его

s =

	0	1	2	3	4	5	6	7	8	9	10	11
0	0	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0	0	0
1	0.1	0.26267836	0.1249828	0.1248771	0.12431764	0.12176551	0.11251376	0.08730947	0.04154704	0.00000214	$6.14624765\cdot10^{-9}$	0.00000097
2	0.2	0.40003364	0.12406691	0.12186707	0.11589557	0.10234293	0.07773848	0.04423081	0.01381082	0.00000142	$1.66543204\cdot10^{-8}$	0.00000263
3	0.3	0.53428515	0.11860337	0.11028953	0.09521916	0.07241728	0.04481657	0.01975827	0.00459143	0.00000071	$3.14800594\cdot10^{-8}$	0.00000496
4	0.4	0.65834909	0.10537983	0.08986027	0.06875347	0.04480129	0.02305577	0.00824867	0.0015266	0.00000031	$5.04216865\cdot10^{-8}$	0.00000795
5	0.5	0.76393215	0.08563081	0.06595348	0.04454464	0.02510822	0.01099124	0.00330087	0.00050764	0.00000013	$7.3020397\cdot10^{-8}$	0.00001151
6	0.6	0.84611998	0.06370973	0.04417533	0.02646894	0.01307274	0.00496428	0.00128314	0.00016882	$5.20648495\cdot10^{-8}$	$9.85989007\cdot10^{-8}$	0.00001555
7	0.7	0.90495177	0.04378569	0.02740373	0.01468327	0.00643394	0.00215395	0.00048836	0.00005615	$2.01932048\cdot10^{-8}$	0.00000013	0.00001993
8	0.8	0.94405784	0.0280966	0.01595112	0.00770766	0.00302982	0.00090611	0.00018291	0.00001868	$7.67234337\cdot10^{-9}$	0.00000016	0.00002456
9	0.9	0.96844314	0.01700518	0.00880555	0.00386782	0.00137711	0.00037194	0.00006765	0.00000621	$2.86965281\cdot10^{-9}$	0.00000019	0.00002934
10	1	0.98283893	0.00979555	0.00464952	0.00186995	0.00060803	0.00014966	0.00002477	0.00000207	$1.06012295\cdot10^{-9}$	0.00000022	0.00003422
11	1.1	0.99094714	0.00541376	0.00236426	0.00087623	0.00026206	0.00005924	0.00000899	0.00000069	$3.87744834\cdot10^{-10}$	0.00000025	0.00003915
12	1.2	0.99533195	0.00289314	0.00116408	0.00039983	0.00011066	0.00002312	0.00000324	0.00000023	$1.40662664\cdot10^{-10}$	0.00000028	0.00004411
13	1.3	0.99762021	0.00150809	0.00055742	0.00017833	0.00004593	0.00000892	0.00000116	$7.67113793\cdot10^{-8}$	$5.06857448\cdot10^{-11}$	0.00000031	0.00004908
14	1.4	0.99877693	0.00077597	0.00026052	0.00007798	0.00001878	0.00000341	0.00000042	$2.58682093\cdot10^{-8}$	$1.81661857\cdot10^{-11}$	0.00000034	0.00005407
15	1.5	0.99934456	0.00040163	0.0001192	0.00003352	0.00000758	0.00000129	0.00000015	$8.93941247\cdot10^{-9}$	$6.48750983\cdot10^{-12}$	0.00000037	0.00005906
16	1.6	0.9996149	0.00021565	0.00005353	0.00001419	0.00000302	0.00000049	$5.34386829\cdot10^{-8}$	$3.29877521\cdot10^{-9}$	$2.31606232\cdot10^{-12}$	0.00000041	0.00006404
17	1.7	0.99973923	0.00012555	0.00002364	0.00000593	0.0000012	0.00000018	$1.97617065\cdot10^{-8}$	$1.41605996\cdot10^{-9}$	$8.32967391\cdot10^{-13}$	0.00000044	0.00006903

Рис. 3. Матрица решения системы уравнений

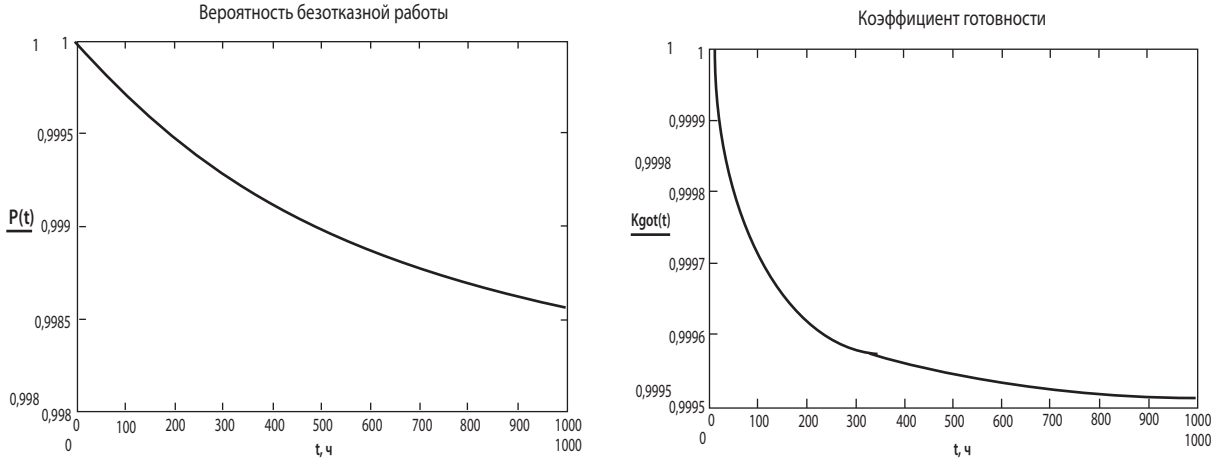

Рис. 4. Графики зависимости основных показателей надежности от времени

работы в дальнейшем, можно дать следующие рекомендации:

- учесть надежность программных средств за счет построения модели с использованием многофрагментных марковских процессов [1];
- учесть неидеальность средств контроля и улучшить систему мониторинга кластера;
- учесть периоды простоя оборудования при оценке надежности;
- для улучшения показателей надежности использовать резервирование основных элементов (центральный коммутатор, сервер, основные каналы связи.), проанализировать возможные варианты и выбрать оптимальную схему резервирования, с учетом, в том числе, финансовых расходов;
- для выбора коммутаторов, анализа их работы и опрелеления скорости коммутации между портами использовать нагрузочный тест mpitnet [4] на каждом из них.
- для анализа сетевых перегрузок кластера запустить тест mpitncf на процессорах вычислительных узлов на фоне нагрузочного теста NFS – rth [5].

ВЫВОД

При написании данной статьи были исследованы основные принципы построения кластерных систем, произведен и обоснован выбор типа кластера образовательного учреждения; проанализированы проблемы с точки зрения надежности (аппаратные отказы), возникающие при выборе и настройке компонентов кластера, а также возможные варианты их решения; сформулированы рекомендации по улучшению и усовершенствованию модели надежности кластерной системы. В рамках статьи предложен математический алгоритм оценки надежности вычислительного кластера. Решена практическая задача: на основе предложенного алгоритма реализована модель оценки надежности кластерных систем, которая позволила провести анализ кластера Центра компьютерных технологий Таврического национального университета им. В. И. Вернадского. ■

ЛИТЕРАТУРА

1. Харченко В. С., Одарущенко О. Н., Поночовный Ю. Л., Одарущенко Е. Б., Скляр В. В., Конорев Б. М., Чертков Г. Н. Технологии высокой готовности для программно-технических комплексов космических систем / Под ред. В. С. Харченко, Б. М. Конорева.– Харьков: Гос. центр регулирования качества поставок и услуг, Нац. Аэрокосмический университет «ХАИ», 2010.– 372 с.

2. Яковлев А. В. Надежность информационных систем. Лекционный материал.– Муром: Изд. ВГУМИ, 2004.– 63 с.

3. Ермаков А. А. Комплексное обеспечение надежности кластерных систем на основе математического моделирования: дис. канд. техн. наук: 05.13.01 – Москва, 2008.– 150 с.

4. Гайлунь В. Рекомендации по оценке надежности функционирования компьютеров и использованию резервных компьютеров на предприятии [Электронный ресурс].– Режим доступа: http://pda.cio-world.ru/index.php?action=article&section_id=26720&id=33594 (10.04.2011).

5. Шнитман В. З., Кузнецов С. Д. Аппаратно-программные платформы корпоративных информационных систем. [Информационно-аналитические материалы Центра Информационных Технологий. Электронный ресурс].– Режим доступа: http://citforum.ru/hardware/app_kis/contents.shtml (10.04.2011)