УДК 007.52 **Технические науки**

***Лазарева Наталия Борисовна**, старший преподаватель,*
Тихоокеанский государственный университет, г. Хабаровск

ОРГАНИЗАЦИЯ ОТКАЗОУСТОЙЧИВОГО (НА) КЛАСТЕРА

Аннотация: в статье рассматривается способ обеспечения высокой доступности ИТ- инфраструктуры путем создания кластерного решения.

Ключевые слова: кластер, инфраструктура, отказоустойчивость, высокая доступность.

Abstract: this article describes highly available IT-infrastructure solution based on cluster technology.

Keywords: cluster, infrastructure, fault tolerance, high availability.

Современные концепции развития IT-инфраструктуры предприятий любого уровня в настоящее время подразумевают кластеризацию составляющих ее элементов. Кластер [1] - связанная совокупность нескольких вычислительных систем, работающих совместно для выполнения общих приложений, и представляющихся пользователю единой системой.

История создания кластеров начинается с 70-х годов прошлого века, основы которых были заложены группой разработчиков стандарта TCP/IP в лаборатории Xerox PARC. Дальнейшее развитие технологии проводилось компаниями Sun Microsistems, Datapoint, DEC, IBM. Создание кластеров в это время могли позволить себе только очень крупные IT-концерны и государственные структуры (NASA). Целью создаваемых кластерных решений этого времени было увеличение быстродействия и мощности вычислительных ресурсов. Это направление кластеризации в дальнейшем получило название «High Performance» (HP). Бесспорным лидером этого направления является

компания Google, количество серверов которой составляет более 1 000 000 экземпляров.

Удешевление аппаратной части серверного оборудования и активный рост количества персональных компьютеров привели к «компьютеризации» банков, компаний и предприятий всех уровней. Это потребовало создания систем высокой надежности. Кластеры, создаваемые для этих целей получили название High Availability (HA).

В рамках реализации проекта ЦАС КИТ в эксплуатацию вводится новый программно-аппаратный комплекс, являющийся логическим продолжением имеющегося комплекса РСДУ, но на обновленной и расширенной программно-аппаратной части. Данный кластер имеет много «фирменных» решений «от производителя», обеспечивающих высокую производительность создающих условия эффективной и надежность в эксплуатации.

При всех рекламируемых плюсов этих решений существует несколько не афишируемых минусов: зависимость от конкретного производителя (невозможность заменить составляющие на другого при поломке/модернизации); завышенная начальная цена (за «brand»); платная «поддержка» системы в дальнейшем. Стоимость возможного простоя этих систем превышает стоимость затрат создания. Поэтому одной из задач при эксплуатации является необходимость того, чтобы ни одна составная часть такого кластера не стала «точкой отказа» системы в целом (NSPF — No Single Point Failure).

Разработка физической и логической схем организации отказоустойчивого (HA) кластера без дополнительных вложений и является темой данной статьи.

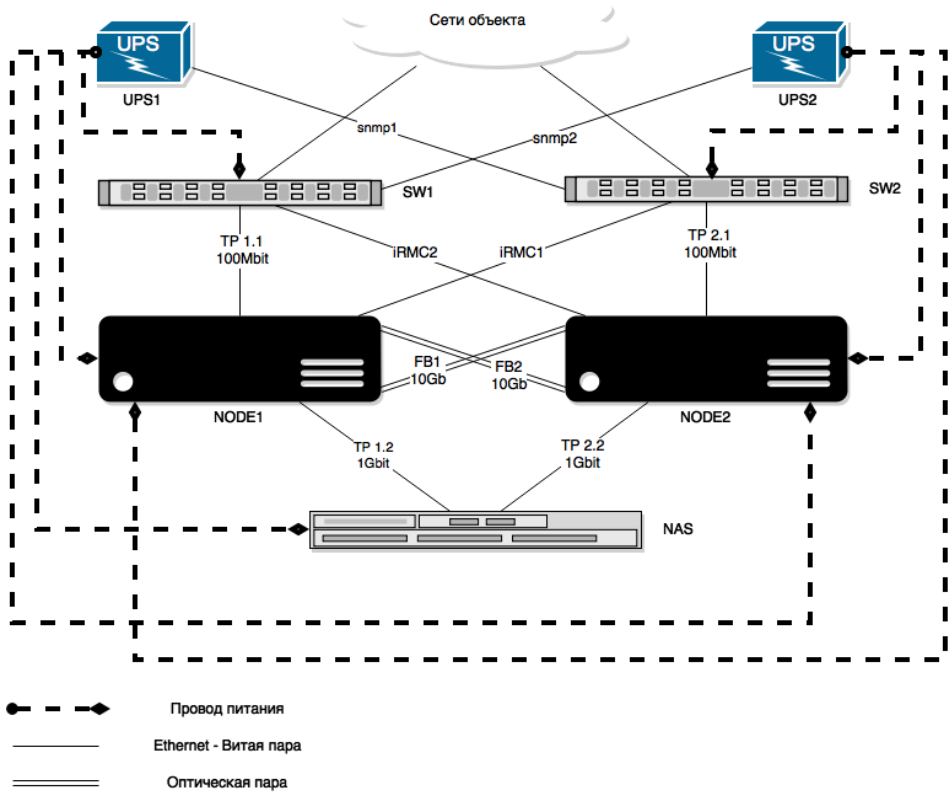

Рис. 1. Физическая схема организации кластера

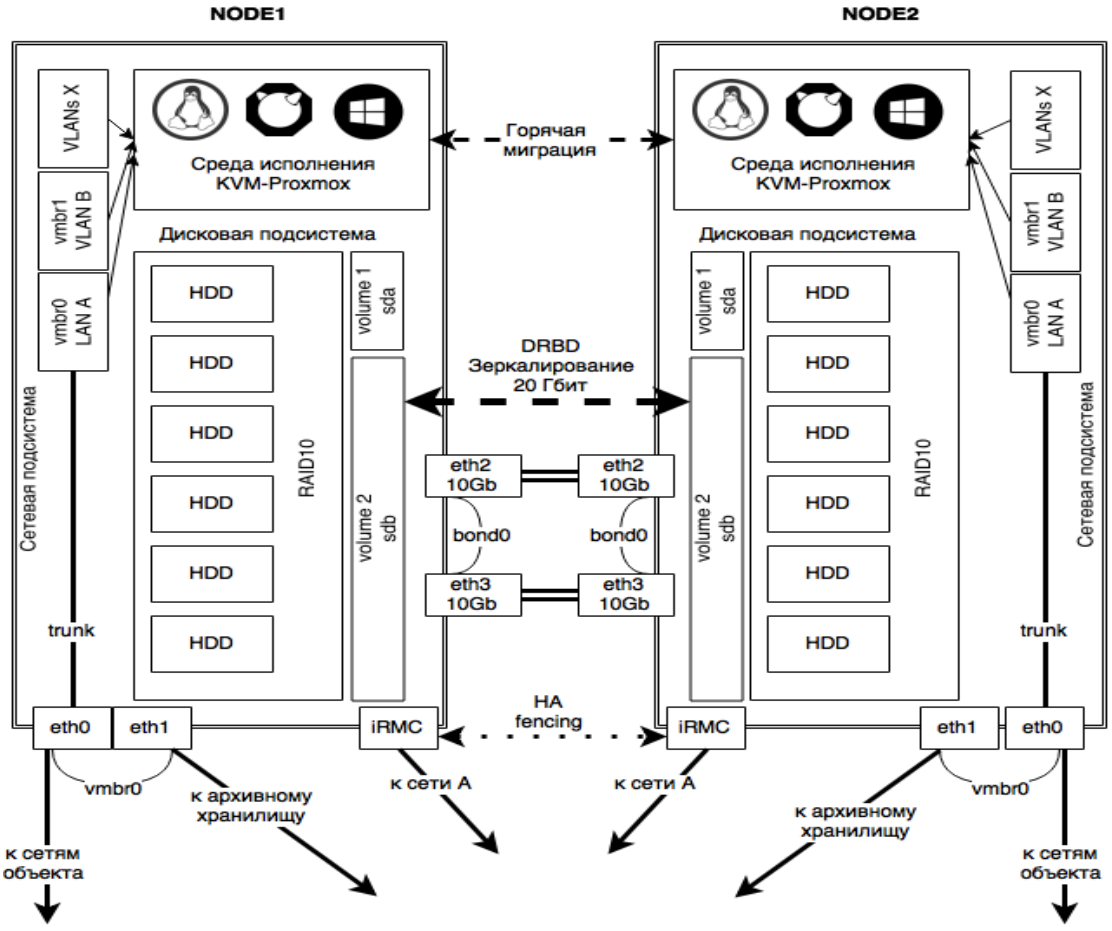

Рис. 2. Логическая схема организации кластера

При разработке предусмотрены шесть групп мер обеспечения надежности.

1. Отказ носителя. При отказе одного из жестких дисков работа комплекса не прерывается. Производится физическая замена диска из имеющегося ЗИП. Кроме физического извлечения поврежденного диска и установки на его место нового никаких действий не требуется. Восстановление RAID массива осуществляется контроллером автоматически. Допускается потеря с сохранением работоспособности от 2 до 6 жестких дисков из комплекса.

2. Отказ оптической связи. Физическая связь между нодами кластера осуществляется двумя независимыми прямыми оптическими кроссами. В штатном режиме суммарная скорость обмена между нодами составляет сумму скоростей кросов и равна 20 гбитам. При этом допускается без остановки рабочих задач физическое повреждение или извлечение одного оптического кабеля из SFP модуля, физическое извлечение любого SFP модуля.

3. Отказ ethernet связи. Отказ сетевого подключения к eth0 (первый сетевой порт) сервера может произойти по причине:

- Случайное извлечение или повреждение патч-корда.
- Отказ порта на стороне коммутатора или сервера.
- Отключение коммутатора по питанию или сброс ПО коммутатора.

Восстановление работоспособности в данном случае осуществляется путем переноса рабочей нагрузки (виртуальные машины) на ноду подключенную к рабочему линку. Это возможно осуществить благодаря тому, что каждый коммутатор обеспечивает гарантированный доступ к своей сети A или B и в обычном состоянии передает эту сеть на второй коммутатор vlan-ом. В случае отказа любого из них или A или B сеть все равно будет доступна. Управляющий интерфейс proxmox подключен сразу к двум сетям сетям и при этой ситуации управление кластером сохраняется по одной из сетей.

4. Отказ операционной системы. В случае некритичного отказа ОС Proxmox [3] или случайной её перезагрузки система HA посредством технологии fencing и модулей iRMC [4] серверов осуществит перезапуск виртуальных машин на ноде, оставшейся в работе. При этом зависшая система будет отключена и вновь включена по питанию, то есть будет предпринята попытка автоматического восстановления работы оборудования. В случае с виртуальными машинами, которые не используют режим HA, виртуальные машины должны запуститься автоматически возобновлении работы ноды. В случае критического отказа ОС ноды, повреждение файловой системы, ошибочные параметры и прочее виртуальные машины с HA перенесутся на рабочую ноду автоматически. Виртуальные машины без HA необходимо перенести вручную.

5. Отказ ПО виртуальных машин. Для минимизации простоя в работе комплекса при отказе ПО, ОС и прочих параметров внутри контейнеров виртуальных машин, можно осуществлять ежедневное архивирование методом онлайнового мгновенного снимка - образа виртуальных носителей и конфигураций контейнеров виртуальных машин. Это означает, что архивация выполняется прозрачно, без прерывания работы виртуальной нагрузки, без существенного оказания воздействия на её производительность. Благодаря независимым каналам связи с дисковым архивным хранилищем каждой ноды индивидуально, архивирование не воздействует на рабочую сеть и может осуществляться одновременно обеими нодами кластера на максимальной скорости. В зависимости от реального объема может сохраняться от двух и более версий архивов виртуальных машин. Система автоматически ведет учет числа выполненных архивов и затирает наиболее старый в случае успешного создания нового. Система может уведомлять на электронной почте об успешности или сбое в результате архивирования персонал, ответственный за работу комплекса. Восстановление может осуществляться как в режиме затирания исходной виртуальной машины, так и в режиме дублирования, в том

числе на другом сервере proxmox, не входящем в кластер при его доступе к хранилищу архивов.

6. Отказ электропитания. Для предотвращения отказа по электропитанию каждый сервер имеет два блока питания, которые подключены к двум разным ИБП. В свою очередь ИБП могут быть подключены к системе электроснабжения имеющей два независимых питающих ввода (основной и резервный). Коммутацию между вводами осуществляет АВР. Отключение любого ИБП на обслуживание по очереди не приводит к полному отказу системы.

Характеристики нод кластера (размер дискового пространства и оперативной памяти, количество процессоров) подбирается исходя из потребностей виртуальных машин, размещаемых на кластере. Предложенная схема позволяет создать отказоустойчивый кластер на открытых решениях, лишенный ориентированности на какого-либо производителя.

Библиографический список:

1. https://www.ixbt.com/cpu/clustering.shtml (дата обращения: 30.01.2021).

2. Официальная документация Proxmox https://pve.proxmox.com/wiki/ (дата обращения: 30.01.2021).

3. https://ru.wikipedia.org/wiki/Кластер_(группа_компьютеров) (дата обращения: 30.01.2021).

4. Официальная документация Fujitsu Remote Management iRMC https://sp.ts.fujitsu.com/dmsp/Publications/public/ds-irmc-s5-en.pdf (дата обращения: 30.01.202).