Настройка отказоустойчивого кластера с помощью утилиты live-cluster

Статья актуальная для версий IVA MCU 5.1 и выше. Для более ранних версий используйте следующую статью.

Введение

Общие положения

Есть две схемы развертывания продукта:

- один головной сервер
- один головной сервер + N медиа серверов

Отказоустойчивость требуется обеспечивать только для головного сервера, т.к. в случае падения одного из медиа серверов конференции расположенные на нём будут перенесены на оставшиеся медиа сервера средствами ПО ИВКС расположенного на головном сервере. Соотвественно, всё что будет описано ниже относится только к головным серверам.

Отказоустойчивость головных ИВКС обеспечивается на нескольких уровнях:

1. На аппаратном уровне резервирование осуществляется посредством использования двух идентичных серверов с резервированными блоками питания.
2. На аппаратном уровне реализовано резервирование дискового массива (hardware raid)
3. На сетевом уровне реализовано резервирование всех сетевых интерфейсов по технологии Linux Ethernet Bonding. При этом предполагается что сетевые

интерфейсы подключены к разным коммутаторам, за счёт чего достигается резервирование сетевой инфраструктуры. Подробнее про варианты резервирования смотри в разделе "Способы резервирования сети".

4. На прикладном уровне для головного сервера реализован кластер высокой доступности под управлением программного обеспечения Cluster Resource Manager. Кластер работает в режиме холодного резервирования (active/standby) обеспечивает резервирование следующих компонентов:
    - Базы данных PostgreSQL с репликацией данных на резервный узел в режиме реального времени.
    - Файлового хранилища с репликацией данных в режиме реального времени (DRBD). (опционально, если заказчик НЕ предоставил доступ к своему файловому хранилищу, если предоставил то резервирование файлового хранилища выполняется средствами заказчика)
    - Виртуального IP адреса, привязанного к активному узлу кластера и используемому для доступа пользователей к ИВКС. При выходе из строя активного узла кластера, в задачи Cluster Resource Manager также входит перенос виртуального адреса на резервный узел кластера.

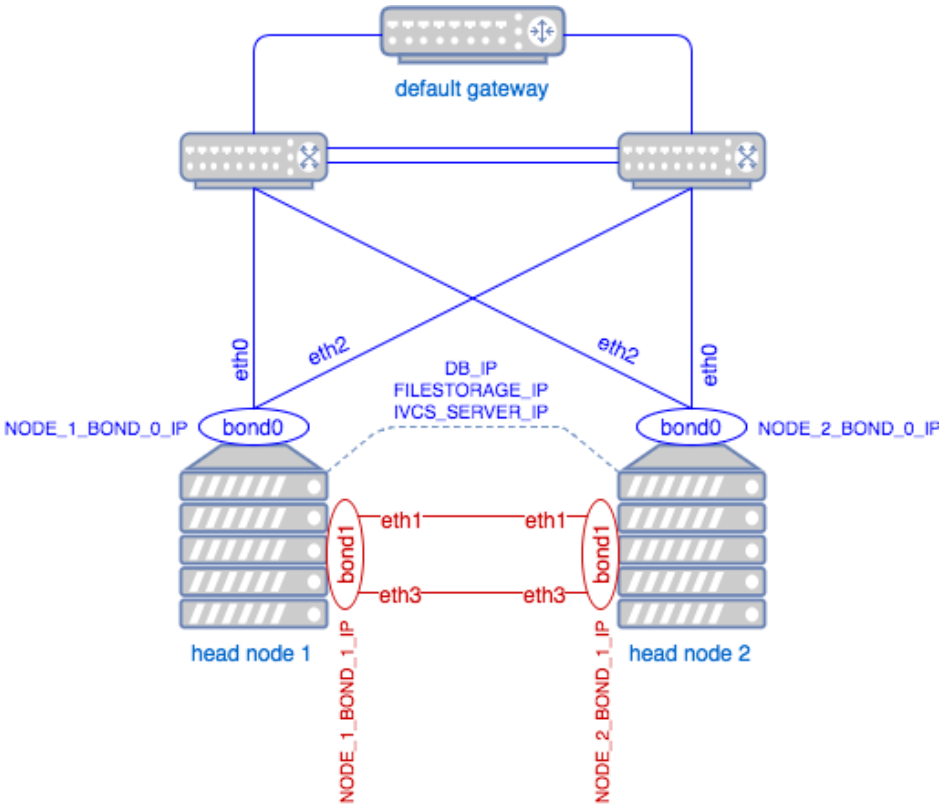
Терминология

- Cluster resource manager (CRM) - програмное обеспечение Pacemaker.
- Cluster resource - абстрактный сервис работающий под управлением CRM. В решении ИВКС используются следующие ресурсы:
    - База данных PostgreSQL
    - Виртуальный IP адрес
    - Сервис бизнес логики ivcs-server
    - Сетевая файловая система Samba
- Cluster resource agent - реализация стандартизированного интерфейса управления ресурсом, выполняющая трансляцию стандартного набора операций (start, stop, monitor и т.д.) в набор действий, специфических для конкретного ресурса или приложения. Как правило реализуется в виде shell скрипта.
- Cluster information base (CIB) - распределенная база данных, содержащая описание узлов кластера, ресурсов, их параметров и текущего состояния. В задачи CRM входит поддержание CIB в консистентном состоянии на всех узлах, входящих в кластер.

Перед началом работ

Выбор способа резервирования сети

Полное резервирование

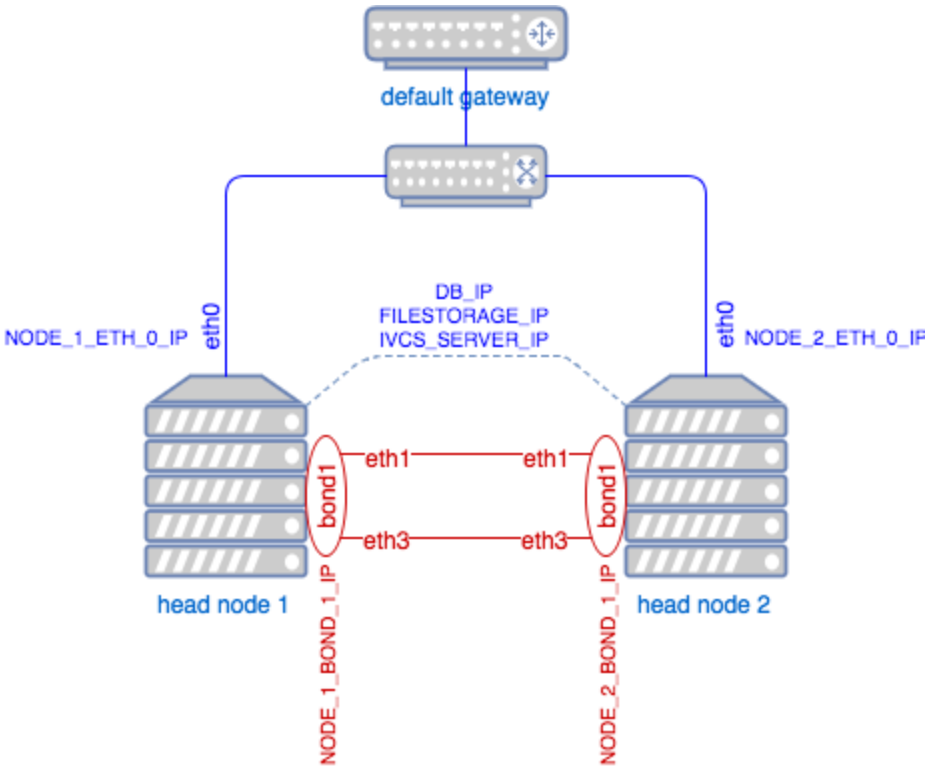


Это рекомендуемая схема. Оба сервера входящие в состав кластера имею по 4 сетевых интерфейса.

Интерфейсы eth0 и eth2, объединенные по технологии Linux Ethernet Bonding, используются для подключения ИВКС к сети заказчика. Через эти интерфейсы осуществляется доступ пользователей к сервису ИВКС, а также удаленное управление решением со стороны системных администраторов и обслуживающего персонала. Также эти два интерфейса используются CRM для проверки доступности узлов кластера, обмена информацией о состоянии узлов и ресурсов, работающих под управлением кластера.

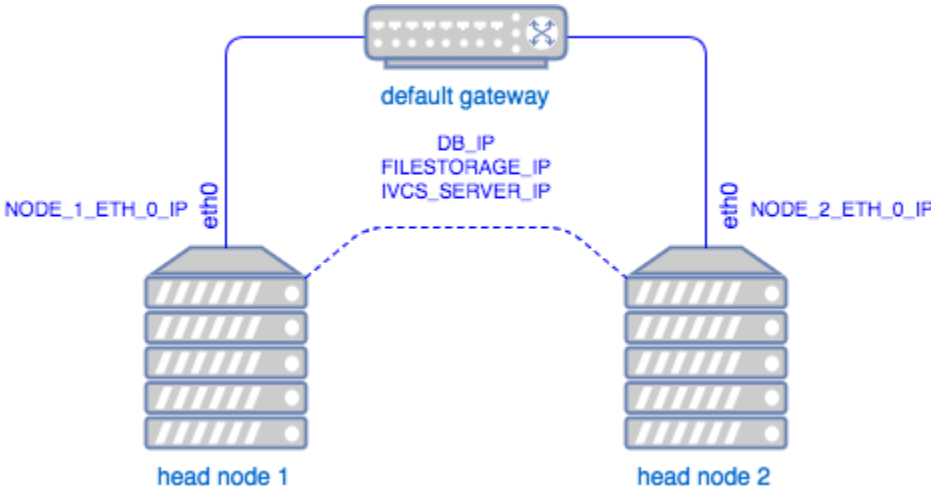
Интерфейсы eth1 и eth3, объединенные по технологии Linux Ethernet Bonding, используются CRM в качестве резервного канала связи для проверки доступности узлов кластера, обмена информацией о состоянии узлов и ресурсов, работающих под управлением кластера. Дополнительно через эти интерфейсы осуществляется резервное копирование базы данных PostgreSQL и репликация данных с активного узла базы данных на резервный узел.

Частичное резервирование



В этом случае предполагается, что развернутое решение ИВКС может стать недоступным из вне при потери связи с единственным свитчом. Связь между узлами кластера сохраняется за счёт резервированного канала связи bond1. Используется в случае если у заказчика есть только один свитч.

Без резервирования



Как не трудно догадаться из названия - резервирования сети в этом случае нет. Данная схема может использоваться в следующих случаях:

- внутреннее тестирование сценариев падения без тестирования сценария пропадания связи между узлами.
- тестовая демонстрация заказчику без демонстрации случая пропадания связи между узлами.

Основная проблема данной схемы - split-brain, когда при пропадании сетевой связности между узлами каждый узел будет считать себя активным.

Выбор файлового хранилища

Оптимальным вариантом является использование файлового хранилища заказчика. Поддерживаются следующие виды хранилищ:

- NAS. В этом случае файловое хранилище доступно IVA MCU как сетевая файловая система. Поддерживаемые протоколы CIFS, NFS
- SAN. В этом случае файловое хранилище на стороне IVA MCU представлено блочным устройством. Доступ осуществляется по одному из протоколов iSCSI, FCoE и т.д.

В случае отсутствия у заказчика файлового хранилища надо использовать локальное на основе технологии DRBD. Перед использованием его надо предварительно настроить.

Общие моменты

Предполагается, что перед настройкой отказоустойчивого кластера выполнено следующее:

1. Выбран способ резервирования сети. См. раздел выше.
2. Выбран тип файлового хранилища. См. раздел выше.
3. В инфраструктуре заказчика выделено следующее:
    1. два сервера под головные сервера IVA MCU
    2. N серверов под медиа сервера IVA MCU (если требуются отдельные медиа сервера).
    3. IP адреса под каждый из серверов выше.
    4. Три дополнительных IP адреса, которое будут мигрировать между головными серверами.
    5. Разрешен доступ по SSH (порт 22, TCP) между всеми узлами кластера.
4. На головных серверах развёрнут IVA MCU
5. Головные сервера синхронизированы по времени
6. На медиа серверах развёрнут IVA MCU media (если требуются отдельные медиа сервера)
7. В случае использования внешнего файлового хранилища, представленного блочным устройством на нём создана файловая система ext4 (sudo mkfs.ext4 DEVICE_PATH)

8. В /etc/iptables-config/ipv4/filter.d/0000-default на всех серверах добавлена запись -A INPUT -p tcp -m tcp --dport 5432 -m comment --comment "DB" -j ACCEPT и рестартовать сервис командой sudo systemctl restart iptables-config
9. Примечание. Утилита при работе меняет хостнеймы машин на ivcs-main-1, ivcs-main-2, ivcs-media-0, ivcs-media-1 и так далее. Если есть необходимость именовать машины как-то иначе, настоятельно рекомендуется ПОСЛЕ установки кластера, при помощи команды hostnamectl, установить необходимые хостнеймы и в файле /etc/hosts НА ВСЕХ МАШИНАХ прописать соответствия IP к нашим хостнеймам, для того, чтобы каждой машины имена ivcs-main-1 и т.д. резолвились в локальные IP адреса соответствующих машин

Настройка кластера

На **всех** серверах (в том числе и на медиа) выполняем команду:

```
sudo live-configure unlock-configurator
```

На любом из головных серверов выполняем команду (при написании команды, смотри примечания в данной статье):

```
sudo live-cluster configure \
    --head-node-1-ip HEAD_NODE_1_IP_1 \
    --head-node-1-secondary-ip HEAD_NODE_1_IP_2 \
    --head-node-2-ip HEAD_NODE_2_IP_1 \
    --head-node-2-secondary-ip HEAD_NODE_2_IP_2 \
    --public-ip PUBLIC_IP \
    --public-fqdn PUBLIC_FQDN \
    --database-ip DATABASE_IP \
    --filestorage-ip FILESTORAGE_IP \
    --filestorage-device FILESTORAGE_DEVICE \
    --filestorage-username FILESTORAGE_USERNAME \
    --filestorage-password FILESTORAGE_PASSWORD \
    --license-server-instance-id LICENSE_SERVER_INSTANCE_ID \
    --license-file LICENSE_FILE \
    --external-mail-server-hostname EXTERNAL_MAIL_SERVER_HOSTNAME \
    --external-mail-server-username EXTERNAL_MAIL_SERVER_USERNAME \
    --external-mail-server-password EXTERNAL_MAIL_SERVER_PASSWORD \
    --external-mail-disable \
    --smsc-host SMSC_HOST \
    --smsc-port SMSC_PORT \
    --smsc-username SMSC_USERNAME \
    --smsc-password SMSC_PASSWORD \
    --smsc-sender-address SMSC_SERNDER_ADDRESS \
    --smsc-disable \
    --ssl-certificate-file SSL_CERTIFICATE_FILE \
    --ssl-private-key-file SSL_PRIVATE_KEY_FILE \
    --outgoing-sip-proxy OUTGOING_SIP_PROXY \
    --default-sip-from-header DEFAULT_SIP_FROM_HEADER \
    --default-h323-from-header DEFAULT_H323_FROM_HEADER \
    --media-node-1-ip MEDIA_NODE_1_IP \
    ... \
    --media-node-n-ip MEDIA_NODE_N_IP
```

где

- HEAD_NODE_1_IP_1 - IP адрес первого узла головного сервера
- HEAD_NODE_1_IP_2 - второй IP адрес первого узла головного сервера (указывается в случае частичного или полного резервирования сети)
- HEAD_NODE_2_IP_1 - IP адрес второго узла головного сервера
- HEAD_NODE_2_IP_2 - второй IP адрес второго узла головного сервера (указывается в случае частичного или полного резервирования сети)
- PUBLIC_IP - плавающий IP адрес по которому будут доступны веб интерфейс, SIP/H323 сигнализация
- PUBLIC_FQDN - FQDN по которому будет доступен веб интерфейс (опционально, если не указывать будет использоваться PUBLIC_IP)
- DATABASE_IP - плавающий IP адрес по которому будет доступен текущий master базы данных
- FILESTORAGE_IP - плавающий IP адрес внутреннего файлового хранилища
- FILESTORAGE_DEVICE - имя блочного устройства используемого в качестве внешнего файлового хранилища или URI сетевого файлового хранилища (NAS)
- FILESTORAGE_USERNAME - имя пользователя для авторизации в сетевом файловом хранилище (необходимо, если файловое хранилище требует авторизацию)
- FILESTORAGE_PASSWORD - пароль пользователя для авторизации в сетевом файловом хранилище (необходимо, если файловое хранилище требует авторизацию)
- LICENSE_SERVER_INSTANCE_ID - server instance id из файла лицензии (опционально, на сервере будет принудительно настроен указанный server instance id)
- LICENSE_FILE - путь до файла с лицензией (опционально)
- EXTERNAL_MAIL_SERVER_HOSTNAME - адрес внешнего почтового сервера (опционально)
- EXTERNAL_MAIL_SERVER_HOSTNAME - имя пользователя для аутентификации на внешнем почтовом сервере (опционально)
- EXTERNAL_MAIL_SERVER_USERNAME - пароль пользователя для аутентификации на внешнем почтовом сервере (опционально)
- --external-mail-disable - флаг отключающий интеграцию с внешним почтовым сервером (опционально, нужен если требуется отключить ранее включенную интеграцию)
- SMSC_HOST - адрес SMSC шлюза (опционально)
- SMSC_PORT - порт SMSC шлюза (опционально)
- SMSC_USERNAME - имя пользователя для аутентификации на внешнем SMSC шлюзе (опционально)
- SMSC_PASSWORD - пароль пользователя для аутентификации на внешнем SMSC шлюзе (опционально)
- SMSC_SENDER_ADDRESS - адрес отправителя в исходящих SMS (опционально)
- --smsc-disable - флаг отключающий интеграцию с SMSC (опционально, нужен если требуется отключить ранее включенную интеграцию)
- SSL_CERTIFICATE_FILE - путь до SSL сертификата (опционально)

- SSL_PRIVATE_KEY_FILE - путь до приватного ключа (опционально)
- OUTGOING_SIP_PROXY - адрес SIP прокси для исходящих звонков (опционально, по умолчанию sip:${public-fqdn}:5060)
- DEFAULT_SIP_FROM_HEADER - значение заголовка From для исходящих SIP звонков (опционально, по умолчанию "IVCS #<CONFERENCE_ID>" <sip:<CONFERENCE_ID>@${public_fqdn}>)
- DEFAULT_H323_FROM_HEADER - значение заголовка From для исходящих H323 звонков (опционально, по умолчанию "IVCS #<CONFERENCE_ID>" <h323:<CONFERENCE_ID>@${public_fqdn}>)
- MEDIA_NODE_1_IP - IP адрес первого медиа сервера (необходимо, если есть отдельные медиа сервера)
- MEDIA_NODE_N_IP - IP адрес N'ого медиа сервера (необходимо, если есть отдельные медиа сервера)

С помощью команды "sudo live-cluster status" убеждаемся, что все ресурсы запустились. Её вывод должен быть примерно следующим:

```
Stack: corosync
Current DC: ivcs-main-1 (version 1.1.16-94ff4df) - partition with quorum
Last updated: Thu Jun 13 16:37:17 2019
Last change: Tue Jun 11 16:20:25 2019 by root via cibadmin on ivcs-main-1

2 nodes configured
14 resources configured

Online: [ ivcs-main-1 ivcs-main-2 ]

Full list of resources:

 Resource Group: db-group
     db-ip      (ocf::heartbeat:IPaddr2):      Started ivcs-main-2
 Resource Group: filestorage-group
     filestorage-fs     (ocf::heartbeat:Filesystem):   Started ivcs-main-1
     filestorage-ip     (ocf::heartbeat:IPaddr2):      Started ivcs-main-1
     samba      (systemd:smbd):Started ivcs-main-1
 Resource Group: ivcs-server-group
     ivcs-server-ip     (ocf::heartbeat:IPaddr2):      Started ivcs-main-2
     ivcs-server        (systemd:ivcs-server): Started ivcs-main-2
     opensips   (systemd:opensips):    Started ivcs-main-2
     gnugk      (systemd:gnugk):       Started ivcs-main-2
 Master/Slave Set: ivcs-db-ms [ivcs-db]
     Masters: [ ivcs-main-2 ]
     Slaves: [ ivcs-main-1 ]
 Clone Set: diskspace-clone [diskspace]
     Started: [ ivcs-main-1 ivcs-main-2 ]
 Clone Set: monitor-clone [monitor]
     Started: [ ivcs-main-1 ivcs-main-2 ]
```

После завершения настройки на **всех** серверах (в том числе и на медиа) выполняем команду:

```
sudo live-configure lock-configurator
```

В случае успешного выполнения будет сделано следующее:

- У домена по умолчанию в качестве веб адреса будет указано значение --public-fqdn
- Изменены следующие системные настройки:
  - "Прокси исходящего SIP звонка" на "sip:${значение --public-fqdn}:5060"
  - "SIP-header" на "IVCS #<CONFERENCE_ID>" <sip:<CONFERENCE_ID>@${значение --public-fqdn}>
  - "H323-header" на "IVCS #<CONFERENCE_ID>" <h323:<CONFERENCE_ID>@${значение --public-fqdn}>
- Установлена лицензия с указанным server instance id (при указании соответствующих параметров)
- Установлены SSL сертификат и приватный ключ в nginx, opensips и gnugk (при указании соответствующих параметров)
- Настроена отправка почты через внешний почтовый сервер (при указании соответствующих параметров)
- Настроена отправка SMS через указанный SMSC (при указании соответствующих параметров)
- Подключено внешнее файловое хранилище

Далее в интерфейсе администрирования, в разделе медиа сервера надо сделать следующее:

1. Удалить медиа сервер с адресом 127.0.0.1
2. Добавить медиа сервера с адресами указанными в качестве значений параметров --head-node-1-ip и --head-node-2-ip (если планируется использовать медиа сервера на головных серверах)
3. Добавить медиа сервера с адресами указанными в качестве значений параметров --media-node-n-ip

Прочие полезные команды

Посмотреть статус кластера

```
sudo live-cluster status
```

Остановить все кластеризованные ресурсы

```
sudo live-cluster stop-resources
```

Запустить все кластеризованные ресурсы

```
sudo live-cluster start-resources
```

Примечания

1. Команда "live-cluster configure" является идемпотентной, т.е. приводит кластер в указанное состояние, а стало быть может быть запущена много раз без последствий.
2. В целях безопасности действие команды "sudo live-configure unlock-configurator" ограничено по времени одним часом или временем до ближайшей перезагрузки.
3. **Рекомендуется экранировать значения всех параметров передаваемых в live-cluster, т.е. писать например**

```
4. --license-file 'путь до файла лицензии с пробелом'
   --smsc-sender-address 'адрес отправителя SMS с пробелом'
```

вместо

```
--license-file путь до файла лицензии с пробелом
--smsc-sender-address адрес отправителя SMS с пробелом
```

Настройка таймаутов переключения серверов

Файл

/etc/corosync/corosync.conf

```
...
totem {
    ...
    # How long before declaring a token lost (ms)
    token: 3000
    # How many token retransmits before forming a new configuration
    token_retransmits_before_loss_const: 10
    ....
}
```

Если в течении 0,75 от времени token не будет связи, будет попытка реконфгурации

но до начала будет <token_retransmits_before_loss_const> попыток соединения каждая попытка будет в течении token/(token_retransmits_before_loss_const+0,2)

после этого начнется запуск сервисов на оставшейсе ноде.

token/(token_retransmits_before_loss_const+0,2) > должно быть более 30ms

Итого время переключения примерно равно 0,75*token + token