СИБРУС

Конфигурация кластера

версия 1.0, 2018 г.

Оглавление

_

1 Введение

Для обеспечения непрерывной работы ПО «СИБРУС» в высоконагруженных системах рекомендуется использовать кластерный вариант серверного ПО «СИБРУС», которое способно обслуживать системы с сотнями тысяч и миллионами пользователей.

Кластер «СИБРУС» содержит в себе следующие компоненты, на которые приходится основное потребление аппаратных ресурсов кластера.

1 Внешние компоненты.
  1.1 Базы данных MongoDB.
  1.2 Менеджер глобальной памяти и очереди сообщений Redis.

2 Компоненты серверного ПО «СИБРУС».
  2.1 Основные рабочие процессы cybrus-server.
  2.2 Медиаретрансляторы cybrus-relay.

Помимо основных серверных компонент, необходимо также развернуть компоненты мониторинга системы:

3 Компоненты мониторинга.
  3.1 Агенты мониторинга на узлах кластера.
  3.2 Система управления мониторингом вне кластера.

Существует два основных подхода к построению кластера:

1. Базовым вариантом развертывания кластера является такая конфигурация, при которой все узлы кластера однородны, и все серверные компоненты размещаются на одних и тех же узлах кластера, где они конкурируют друг с другом за аппаратные ресурсы.

   Преимуществом такого подхода является однородность и взаимозаменяемость серверных узлов кластера. Однако, аппаратные ресурсы кластера при этом могут использоваться менее эффективно, и для достижениях заданных

требований производительности может потребоваться значительно большее количество серверных узлов, чем в первом варианте развертывания.

2. Рекомендуемая конфигурация кластера, при которой различные программные компоненты системы размещаются на различных физических ресурсах (далее - «серверных узлах» или сокращенно «узлах»).

   Разные компоненты предъявляют разные требования к аппаратным ресурсам, которые при этом зависят от фактических сценариев преимущественного использования системы, и поэтому покомпонентное разбиение кластера на узлы предоставляет гибкие возможности по дальнейшему обслуживанию и расширению кластера. При этом конфигурация и количество серверных узлов различного типа могут выбираться и оптимизироваться таким образом, чтобы они были максимально ориентированы на особенности именно того серверного компонента, который будет работать на данных серверных узлах. В данном документе приводятся рекомендации по конфигурации именно такого кластера.

1 Минимальная базовая конфигурация с однородными узлами

Для минимальной конфигурации кластера на 1000-3000 пользователей возможно совместить часть узлов. Однако, при последующем расширении кластера может потребоваться остановка кластера, ручное перераспределение ролей узлов и перенастройка базы данных.

Минимальная конфигурация кластера на 1000-3000 пользователей:

1 Узлы кластера.
- 1.1 Конфигурация узла.
  - 1.1.1 Дисковая система:
    - 1.1.1.1 RAID10 8x3Тбайт HDD для основной БД и хранения файлов;
  - 1.1.2 ОЗУ 48Гбайт.
  - 1.1.3 Процессоры Intel XEON 8 ядер.
  - 1.1.4 Сеть.
    - 1.1.4.1 1Гбит NIC для основной внутренней подсети кластера;
    - 1.1.4.2 1Гбит NIC для резервной внутренней подсети кластера.

1.1.4.3 1Гбит NIC для основной сети внешних запросов;

1.1.4.4 1Гбит NIC для резервной сети внешних запросов.

1.2 Количество 3 шт.

1.2.1 Все узлы являются основными, и на них перераспределяется нагрузка в случае выхода из строя других узлов.

2 Сеть.

2.1 Внутренняя сеть кластера 1Гбит/с.

2.2 Резервная внутренняя сеть кластера 1Гбит/с.

2.3 Внешняя сеть кластера 100Мбит/с.

2.4 Резервная внешняя сеть кластера 100Мбит/с.

2 Расширяемая конфигурация со специализированными узлами

Рекомендуется выделить два сетевых контура кластера — внутренний и внешний. Во внутреннем контуре для связности узлов сервера использовать каналы не хуже, чем 1Гбит. Характеристики внешней сети зависят от возможностей датацентра и стоимости трафика, но рекомендуется канал также не менее 0.2-1 Гбит на кластер.

В целях обеспечения отказоустойчивости работы кластера рекомендуется обеспечить, как минимум, двойное резервирование по сети и питанию.

Для резервирования по сети необходимо, чтобы каждый узел подключался к внутреннему и (если применимо) к внешнему контуру двумя сетевыми портами через разные сетевые коммутаторы. Сетевые порты могут быть настроены в port trunking или port bonding.

Для резервирования по питанию необходимо, чтобы каждый узел подключался к двумя независимым источникам питания.

В качестве операционных систем рекомендуется использовать версию Debian AMD64 с самыми актуальными обновлениями безопасности.

Узлы кластера должны быть закрыты от доступа извне с использованием файервола. Рекомендуется использовать правило «белый список» - «запрещено все, что не разрешено». Список портов для доступа извне приводится в соответствующей администраторской документации.

2.1 База данных

В качестве СУБД сервер «СИБРУС» использует базу данных MongoDB. Особенностью серверного ПО «СИБРУС» является возможность распределения различных видов данных между различными экземплярами базы данных MongoDB. Тем самым обеспечивается возможность создавать такие конфигурации кластера, в которых параллельно работает несколько независимых кластеров MongoDB, каждый из которых оптимизирован на работу с определенным типом данных.

Типы данных сервера «СИБРУС», которые хранятся в базах MongoDB:

- Основная база данных сервера.
- Хранилище файлов.
- Хранилище временных данных.
- Хранилище журналов (логов).

К обработке и хранению перечисленных видов данных предъявляются разные требования, поэтому их целесообразно вынести в разные экземпляры MongoDB на разных узлах. В частности, значимыми метриками являются:

- Задержки на обработку запросов — как быстро обрабатываются запросы к базе данных и сколько запросов в единицу времени может быть обработано.
- Размер хранимых данных — какой объем данных этого типа потребуется хранить в системе.
- Пропускная способность — объемы данных, записываемых и/или читаемых в единицу времени.
- Время хранения данных — как долго система должна хранить данные этого типа.
- Критичность потери данных — насколько критичными является потеря данных этого типа и какая степень резервирования требуется для обеспечения отказоустойчивости по данным этого типа.

Ниже в Табл.1 приведены качественные показатели требований, предъявляемых к различным видам данных в системе.

	Основная база	**Хранилище файлов**	**Хранилище временных данных**	**Хранилище журналов**
Задержки на обработку запросов	Максимально низкие	Некритичное	Низкие	Некритичное
Размер хранимых данных	Средний	Большой	Малый	Средний
Пропускная способность	Высокая	Высокая	Низкая	Низкая
Время хранения данных	Вечно	Вечно	Короткий период	Средне-короткий период
Критичность потери данных	Очень высокая	Средняя	Некритичное	Некритичное

Табл.1 Качественные показатели требований к различным типам данных

База данных MongoDB имеет встроенные средства балансировки нагрузки, а также отказоустойчивого хранения и репликации данных. Кроме того, MongoDB имеет механизмы оптимизации конфигурации базы для достижения минимальных задержек на обработку запросов. Таким образом, задачи по оптимальной конфигурации базы данных MongoDB могут решаться полностью средствами MongoDB путем подбора требуемого аппаратного обеспечения и выполнению соответствующих настроек оборудования и экземпляров MongoDB.

«Горячая» отказоустойчивость обеспечивается за счет механизмов репликации данных между узлами MongoDB. При этом минимальная конфигурация должна содержать не менее 2 активных реплик, на которых хранятся данные, и один и более арбитров, которые помогают управлять репликацией и выбирать мастер-реплику. Для балансировки нагрузки и горизонтального масштабирования системы по мере роста данных и числа пользователей необходимо использовать механизмы «шардинга».

При «шардинге» данные делятся равномерно между «шардами» кластера, и в запросах участвуют только те узлы, к которым относятся данные запроса. Минимальное число «шардов» для балансировки нагрузки — 2, каждый шард, при этом, содержит набор реплик, обеспечивающих его отказоустойчивость. Таким образом комбинация «шардинг» + «репликация» требует число узлов, равное «число реплик» * «число шардов» + узлы с арбитрами реплик. Т.к. арбитры реплик занимают незначительные ресурсы, они могут использовать несколько одних и тех же узлов ,

число которых может быть константным вне зависимости от того, сколько реплик и шардов содержит весь кластер — например, достаточно 2-х узлов для обеспечения отказоустойчивости по арбитрам.

Для высоконагруженных систем, ориентированных на дальнейшее расширение и масштабирование, рекомендуется следующая базовая конфигурация узлов MongoDB.

1 Кластер MongoDB для основной базы данных.
  1.1 4 узла: 2 шарда с 2 репликами в каждом.
  1.2 Конфигурация каждого узла.
    1.2.1 Дисковая система:
      1.2.1.1 RAID10 8x1Тбайт HDD для основного хранилища;
      1.2.1.2 RAID10 4*128Гбайт SSD для быстрых индексов.
    1.2.2 ОЗУ 48Гбайт.
    1.2.3 Сеть.
      1.2.3.1 1Гбит NIC для основной внутренней подсети кластера;
      1.2.3.2 1Гбит NIC для резервной внутренней подсети кластера.
2 Хранилище файлов.
  2.1 2 узла: 2 реплики.
  2.2 Конфигурация каждого узла.
    2.2.1 Дисковая система:
      2.2.1.1 Или RAID10 24x2Тбайт HDD
      2.2.1.2 Или внешняя СХД с аналогичными емкостями.
    2.2.2 ОЗУ 48Гбайт.
    2.2.3 Сеть.
      2.2.3.1 1Гбит NIC для основной внутренней подсети кластера;
      2.2.3.2 1Гбит NIC для резервной внутренней подсети кластера.
3 Хранилище временных данных. Может быть совмещено с узлами, на которых размещаются арбитры.
  3.1 2 узла: 2 реплики или hot stand-by.
    3.1.1 Дисковая система:
      3.1.1.1 RAID10 2x1Тбайт HDD;
    3.1.2 ОЗУ 24Гбайт.
    3.1.3 Сеть.

3.1.3.1 1Гбит NIC для основной внутренней подсети кластера;

3.1.3.2 1Гбит NIC для резервной внутренней подсети кластера.

4 Хранилище журналов.

4.1 2 узла: 2 реплики.

4.1.1 Дисковая система:

4.1.1.1 RAID10 8x1Тбайт HDD;

4.1.2 ОЗУ 24Гбайт.

4.1.3 Сеть.

4.1.3.1 1Гбит NIC для основной внутренней подсети кластера;

4.1.3.2 1Гбит NIC для резервной внутренней подсети кластера.

В стартовой конфигурации для экономии числа узлов Хранилище временных данных может располагаться на тех же узлах, что и Основная база данных, а Хранилище журналов на тех же узлах, что и Файловое хранилище. При этом эти два набора узлов могут содержать взаимно арбитры друг друга. Итого, рекомендуемая минимальная конфигурация кластера MongoDB с ориентиром на последующее расширение:

1 Основной узел БД.

1.1 Конфигурация.

1.1.1 Дисковая система:

1.1.1.1 RAID10 8x1Тбайт HDD для основного хранилища;

1.1.1.2 RAID10 4*128Гбайт SSD для быстрых индексов.

1.1.2 ОЗУ 48Гбайт.

1.1.3 Сеть.

1.1.3.1 1Гбит NIC для внутренней подсети кластера;

1.1.3.2 1Гбит NIC для внешних запросов.

1.2 Количество — 4 шт.

2 Узел хранения файлов.

2.1 Конфигурация.

2.1.1 Дисковая система:

2.1.1.1 Или RAID10 24x2Тбайт HDD

2.1.1.2 Или внешняя СХД с аналогичными емкостями.

2.1.2 ОЗУ 48Гбайт.

2.1.3 Сеть.

2.1.3.1 1Гбит NIC для внутренней подсети кластера;

3 1Гбит NIC для внешних запросов.

3.1 Количество — 2 шт.

2.2 Менеджер глобальной памяти и очереди сообщений

В качестве менеджера глобальной памяти и очереди сообщений используется Redis. Для отказоустойчивой работы в кластере должно быть не менее 2 узлов Redis. Рекомендуемая конфигурация узлов Redis следующая.

1.1 Конфигурация.

1.1.1 Дисковая система:

1.1.1.1 Или RAID10 2x1Тбайт HDD.

1.1.2 ОЗУ 48Гбайт.

1.1.3 Сеть.

1.1.3.1 1Гбит NIC для основной внутренней подсети кластера;

1.1.3.2 1Гбит NIC для резервной внутренней подсети кластера.

1.2 Количество — 2 шт.

2.3 Компоненты серверного ПО «СИБРУС»

Основными компонентами серверного ПО «СИБРУС» являются основной рабочий процесс cybrus-server и медиоретранслятор cybrus-relay. В кластере может быть запущено произвольное количество процессов cybrus-server и cybrus-relay, ограничением являются только условия лицензионного соглашения.

Процессы cybrus-server и cybrus-relay, в основном, потребляют ресурсы ЦПУ и ОЗУ, и практически не потребляют дисковое пространство. Потребление сетевого трафика очень сильно зависит от сценария использования. В Табл.2 приводится качественная оценка зависимостей потребления тех или иных ресурсов процессами.

	Процесс cybrus-server	**Процесс cybrus-relay**
Число активных клиентских подключений клиент-сервер TLS	Память — высокая зависимость; Процессор — средняя зависимость; Сеть — низкая зависимость.	Не зависит
Число новых клиентских подключений клиент-сервер TLS в единицу времени	Память — средняя зависимость; Процессор — высокая зависимость; Сеть — средняя зависимость.	Не зависит
Число сообщений в единицу времени	Память — средняя зависимость; Процессор — средняя зависимость; Сеть — средняя зависимость.	Не зависит
Число сообщений в групповых чатах в единицу времени	Память — средняя зависимость; Процессор — высокая зависимость; Сеть — высокая зависимость.	Не зависит
Число активных голосовых соединений	Не зависит	Память — низкая зависимость; Процессор — средняя зависимость; Сеть — средняя зависимость.
Число активных голосовых конференций	Не зависит	Память — средняя зависимость; Процессор — средняя зависимость; Сеть — средняя зависимость.

Число активных видео соединений	Не зависит	Память — высокая зависимость; Процессор — средняя зависимость; Сеть — высокая зависимость.
Число активных голосовых конференций	Не зависит	Память — высокая зависимость; Процессор — средняя зависимость; Сеть — высокая зависимость.

Процессы cybrus-relay могут быть запущены как на тех же узлах, что и cybrus-server, так и на отдельных узлах. Причем, процессы cybrus-relay могут быть запущены вообще за пределами кластера «СИБРУС». Например, в географически распределенных системах процессы cybrus-relay могут быть запущены в разных дата-центрах таким образом, чтобы быть ближе к пулам клиентов с целью уменьшить задержки на прохождение UDP-трафика.

Рекомендуемая минимальная конфигурация узлов для работы серверных компонентов ПО «СИБРУС»:

1 Конфигурация узла.
- 1.1 Дисковая система:
  - 1.1.1 RAID10 2x1Тбайт HDD.
- 1.2 ОЗУ 48Гбайт.
- 1.3 Сеть.
  - 1.3.1 1Гбит NIC для основной внутренней подсети кластера;
  - 1.3.2 1Гбит NIC для резервной внутренней подсети кластера.
  - 1.3.3 1Гбит NIC для основной сети внешних запросов;
  - 1.3.4 1Гбит NIC для резервной сети внешних запросов.

2 Число запущенных процессов на 1 узле.
- 2.1 cybrus-server – 3 шт.;
- 2.2 cybrus-relay – 2 шт.

3 Количество узлов — 3 шт.

2.4 Компоненты мониторинга

Компоненты мониторинга являются необязательными компонентами системы, т. к. работа сервера «СИБРУС» от них не зависит. Однако, настоятельно рекомендуется настроить и использовать мониторинг системы. Компоненты мониторинга состоят из следующих составляющих:

1 Система управления мониторингом. В организации уже может использоваться своя система управления мониторингом. В случае, если такой системы нет, то рекомендуется, как минимум, поставить Zabbix.

2 Агенты мониторинга.

2.1 Стандартные агенты мониторинга, поставляемые системой управления мониторинга для сбора метрики системы: доступность, потребление ЦПУ, потребление ОЗУ, нагрузка на диски и т. п.

2.2 Внутренние агенты мониторинга компонентов серверного ПО «СИБРУС», которые работают в составе кластера и собирают метрики и статистику работы компонентов ПО «СИБРУС» внутри кластеры.

2.3 Внешние агенты мониторинга компонентов серверного ПО «СИБРУС», которые работают вне кластера и мониторят доступность служб ПО «СИБРУС» извне.

Дополнительные аппаратные ресурсы могут потребоваться только для внешних агентов мониторинга, а также для системы управления мониторингом:

1 Минимальные ресурсы для системы управления мониторингом должны оцениваться, исходя из требования используемой системы.

2 Внешние агенты мониторинга «СИБРУС» могут быть запущены на виртуальных машинах или в облачных сервисах со следующими минимальными характеристиками:

- ЦПУ 1Ггц;
- ОЗУ 1Гбайт;
- 100Мбит NIC;
- 1Гбайт HDD.