

СРЕДСТВА МОНИТОРИНГА И ВЕБ-ИНТЕРФЕЙС INFINIBAND НА ВК «ЗУБР»

Г.И. Артемьева, А.Ю. Озорнин, А.И. Мельников

ФГУП "РФЯЦ-ВНИИТФ им.академ.Е.И.Забабахина"

Введение

Одной из важнейших составляющих вычислительного комплекса «Зубр» [1] является коммуникационная система, которая представляет собой совокупность программных и аппаратных средств, обеспечивающих передачу данных между узлами комплекса. Сеть InfiniBand [2] в составе коммуникационной системы ВК «Зубр» используется для передачи сообщений MPI, доступа к ресурсам параллельной файловой системы Lustre и сетевой файловой системы NFS, организации бездисковых узлов и сетевой загрузки узлов, запуска задач и т.д. Существующие программные средства администрирования и диагностики сетей InfiniBand были проприетарными или не обладали необходимыми качествами (быстродействие, расширяемость, поддержка оборудования и новых версий OFED [3] и т.д.). Оригинальный пакет IBADM от компании Mellanox не поддерживает стороннее оборудование. Утилиты диагностики в СПО OFED не поддерживают работу с символьными именами узлов и уступают по функциональности пакету Mellanox IBADM. В связи с этим для ВК «Зубр» потребовалась разработка собственных программных средств.

Средства мониторинга InfiniBand

Средства мониторинга фабрики InfiniBand представляют собой программный пакет VIBADM, предназначенный для администрирования, диагностики, контроля за топологией фабрики InfiniBand и для хранения и анализа счетчиков портов устройств InfiniBand.

В основу реализации VIBADM заложен пакет утилит IBADM, поставляемый вместе с ПО IBGOLD компании Mellanox. После реинжиниринга неизменными оставлены лишь имена утилит и формат файлов топологии и журналов.

В состав пакета VIBADM входят следующие компоненты (см. Рисунок 1):

- подсистема мониторинга топологии, включающая программу `ibadmd` и набор файлов с описанием эталонной топологии сети;
- подсистема мониторинга счетчиков ошибок, включающая программу `ibmon` и ряд таблиц в базе данных `mysql`;
- подсистема разрешения имен, включающая программы `get_guid` и `ibadmd_ns`, а также таблицу в базе данных `mysql`.

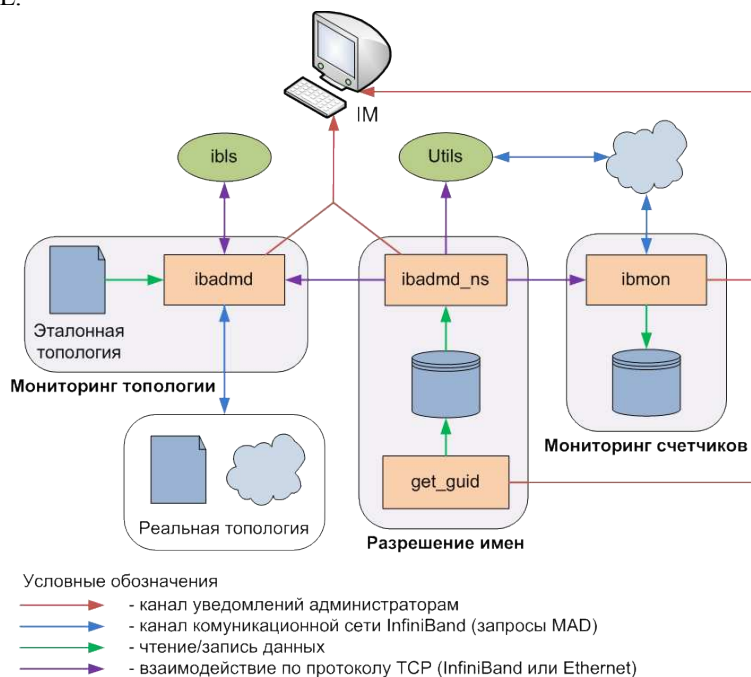


Рис. 1. Функциональная схема VIBADM

Также в состав пакета входит ряд вспомогательных утилит.

Практически все перечисленные программные средства реализованы на языке C. Настраиваемые параметры программ хранятся в едином конфигурационном файле. Предполагается, что VIBADM должен быть установлен на той же административной машине, которая используется для запуска менеджера подсетей OpenSM. При этом ряд пользовательских утилит можно установить и на других узлах, даже если они не оснащены адаптерами InfiniBand.

Доставка уведомлений администраторам осуществляется посредством системы мгновенных сообщений (IM) на основе протокола XMPP.

Мониторинг топологии

Демон ibadmd функционирует на административном узле комплекса и отслеживает изменения в сетевой топологии. Демон с заданной периодичностью сканирует фабрику InfiniBand, сравнивая реальную и эталонную топологии.

В настоящее время фиксируются следующие изменения в топологии фабрики и состоянии линков:

- подключение нового узла или кабеля;
- отсутствие узла или кабеля;
- подключение кабеля, не соответствующее эталонной схеме;
- несоответствие типа узла;
- изменение скорости или ширины полосы пропускания линка.

Эталонная топология сети хранится в локальных файлах в формате модели данных InfiniBand. Актуальность информации обеспечивают администраторы InfiniBand.

Реальная топология считывается из файла, создаваемого менеджером подсети OpenSM, либо извлекается непосредственно из фабрики InfiniBand посредством дейтаграмм MAD.

В случае обнаружения каких-либо отклонений, администраторам отправляется соответствующее уведомление.

Наряду с функцией мониторинга ibadmd выполняет функцию сетевой службы, предоставляя по запросам следующую информацию:

- список линков реальной топологии в обычном и компактном формате;
- список линков эталонной топологии в компактном формате;
- список выявленных ошибок (различий между реальной и эталонной топологиями).

Взаимодействие с ibadmd осуществляется по схеме клиент-сервер на основе протокола TCP [4].

Утилита ibls обеспечивает доступ к сетевой службе ibadmd в режиме командной строки и вывод сведений о текущей или эталонной топологии сети. Для работы ей не требуется наличие InfiniBand адаптера на узле, так как она использует Ethernet для связи с демоном ibadmd.

Мониторинг счетчиков

Утилита ibmon обеспечивает сбор состояний счетчиков фабрики, предоставляемых агентами подсистемы управления производительностью InfiniBand. Программа запускается с заданной периодичностью системной службой cron, считывает значения счетчиков со всех активных портов фабрики и записывает их в текстовый файл и в базу данных MySQL. Если база данных недоступна, утилита записывает текущее состояние счетчиков в файл-аккумулятор. Когда база данных становится доступна, утилита переносит данные из файла-аккумулятора в таблицу. Файл-аккумулятор удаляется.

Счетчики ошибок хранятся в двух таблицах. Одна таблица содержит историю счетчиков за весь период наблюдения, другая – только последнее состояние счетчиков. Получив очередные значения счетчиков порта, программа ibmon для каждого счетчика вычисляет разность между текущим и предыдущим значением. Если разность превышает заданное пороговое значение, то фиксируется ошибочная ситуация. При необходимости по завершению обработки администраторам рассылается уведомление о всех выявленных ошибках.

Вспомогательные утилиты мониторинга счетчиков:

- ibperf выводит текущее состояние счетчиков для задаваемого узла и порта;
- html_counters формирует веб-страницы для текущего и предыдущего состояния счетчиков;
- stat_db_control прореживает устаревшие наборы счетчиков в базе данных;
- get_ib_counters извлекает значения счетчиков из временного файла и выводит на консоль в удобном виде.

Разрешение имен

При разработке пакета VIBADM были приняты специальные меры для решения проблемы именования узлов в сети InfiniBand. Суть данной проблемы состоит в том, что в архитектуре InfiniBand отсутствует такое понятие, как «имя узла». В спецификации [5] определены только численные идентификаторы узлов (LID, GUID), а также атрибут NodeDesc подсистемы SM, значением которого является некоторая строка, описывающая узел в текстовом формате. Обычно NodeDesc содержит название модели устройства. Очевидно,

что отсутствие удобных идентификаторов, позволяющих быстро выявить проблемный узел, существенно осложняет задачу отождествления узлов в процессе диагностики.

Функция подсистемы разрешения имен VIBADM аналогична функции службы доменных имен (DNS). Она обеспечивает поиск уникального идентификатора (GUID) по имени узла InfiniBand, а также обратный поиск имени узла по его GUID.

Таблица соответствия имен и GUID узлов и коммутаторов хранится в базе данных MySQL на административном узле комплекса. Актуальность информации в таблице обеспечивает утилита `get_guid`, которая периодически считывает текущую топологию сети и обновляет таблицу. Для правильного определения имен коммутаторов необходимо создать файл `/etc/opensm/ib-node-name-map` и записать в него их имена и GUID-ы. В случае обнаружения новых узлов или изменения имени какого-либо узла администраторам рассылается соответствующее уведомление.

Обработку запросов на разрешение имен от клиентов подсистемы (каковыми являются `ibadmd`, `ibmon` и ряд диагностических утилит) обеспечивает демон `ibadmd_ns`. Он ищет соответствие имени узла и его GUID в таблице базы данных MySQL. Взаимодействие демона и клиентов подсистемы осуществляется по схеме клиент-сервер на основе протокола TCP.

В состав пакета VIBADM входит также ряд сервисных программ и библиотек:

- библиотека `libxmpp` обеспечивает отправку уведомлений администраторам с использованием службы мгновенных сообщений (на основе протокола XMPP);
- `ibportspeed` используется для управления портами InfiniBand (аналог программы `ibportstate` в составе OFED);
- `ibname` обеспечивает просмотр адресной информации об узлах InfiniBand (аналог утилиты `ibaddr` из OFED с возможностью задания узлов по имени).

Веб-интерфейс

Веб-интерфейс InfiniBand разработан для администраторов ВК для быстрого доступа к информации о состоянии фабрики и представляет собой набор веб-страниц.

Страница вывода текущего состояния ошибок топологии фабрики InfiniBand показана на рисунке 2. Вывод представлен в виде списка с расцветчиванием по типу ошибок. Для получения данных скриптом страницы используется утилита `ibls`.

Счетчики ошибок сети InfiniBand

Счетчики ошибок сети InfiniBand за последний период

Ошибки топологии InfiniBand

Статистика счетчиков сети InfiniBand

Топология InfiniBand

Статья по InfiniBand на русском

Документация по InfiniBand

На главную



Ошибки топологии InfiniBand

Missing Node:N0116/U1(MT26428)
Should be connected by cable from port: P1(N0116/U1/P1) to: IB1-4/P22(IB1-4/P22)

Missing Node:N0209/U1(MT26428)
Should be connected by cable from port: P1(N0209/U1/P1) to: IB1-5/P25(IB1-5/P25)

Missing cable connecting : IB1-4/P32 to N0106/U1/P1

Missing cable connecting : IB1-5/P25 to N0209/U1/P1

Wrong cable connecting : N0106/U1/P1 to IB1-4/P22 instead of IB1-4/P32

Wrong cable connecting : IB1-4/P22 to N0106/U1/P1 instead of N0116/U1/P1

Страничку сопровождает [Артемяева Г.И.](#)

Рис. 2. Ошибки топологии InfiniBand.

Страница с выводом текущей топологии фабрики InfiniBand показана на рисунке 3. Вывод представлен в виде таблицы с расцветчиванием по типу узлов и отображением текущей скорости линков между узлами. Для получения данных также используется утилита `ibls`.

Счетчики ошибок сети InfiniBand	Топология InfiniBand			
Счетчики ошибок сети InfiniBand за последний период	From port	WidthSpeed	To port	
Ошибки топологии InfiniBand	ib1-5 port 16	4x QDR	ib1-4 port 16	
Статистика счетчиков сети InfiniBand	ib1-5 port 15	4x QDR	ib1-4 port 15	
Топология InfiniBand	ib1-5 port 14	4x QDR	ib1-4 port 14	
Статьи по InfiniBand на русском	ib1-5 port 13	4x QDR	ib1-4 port 13	
Документация по InfiniBand	ib1-5 port 12	4x QDR	ib1-4 port 12	
На главную	ib1-5 port 11	4x QDR	ib1-4 port 11	
	ib1-5 port 10	4x QDR	ib1-4 port 10	
	ib1-5 port 9	4x QDR	ib1-4 port 9	
	ib1-5 port 8	4x QDR	ib1-4 port 8	
	ib1-5 port 7	4x QDR	ib1-4 port 7	
	ib1-5 port 6	4x QDR	ib1-4 port 6	
	ib1-5 port 5	4x QDR	ib1-4 port 5	
	ib1-5 port 4	4x QDR	ib1-4 port 4	
	ib1-5 port 3	4x QDR	ib1-4 port 3	
	ib1-5 port 2	4x QDR	ib1-4 port 2	
	ib1-5 port 1	4x QDR	ib1-4 port 1	
	ib1-5 port 28	4x QDR	n0202 port 1	
	ib1-5 port 27	4x QDR	sys port 1	
	ib1-5 port 24	4x QDR	n0208 port 1	
	ib1-5 port 22	4x QDR	n0206 port 1	
	ib1-5 port 23	4x QDR	n0207 port 1	
	ib1-5 port 26	4x QDR	n0210 port 1	
	ib1-5 port 20	4x QDR	n0204 port 1	
	ib1-5 port 17	4x QDR	mds port 1	
	ib1-5 port 21	4x QDR	n0205 port 1	
	ib1-5 port 18	4x QDR	n0212 port 1	
	ib1-4 port 16	4x QDR	ib1-5 port 16	
	ib1-4 port 15	4x QDR	ib1-5 port 15	
	ib1-4 port 14	4x QDR	ib1-5 port 14	
	ib1-4 port 13	4x QDR	ib1-5 port 13	
	ib1-4 port 12	4x QDR	ib1-5 port 12	

Рис. 3. Топология фабрики InfiniBand.

Страница с отображением состояния счетчиков InfiniBand из последнего полученного состояния счетчиков показана на рисунке 4. Страница генерируется утилитой `html_counters`.

Счетчики ошибок сети InfiniBand	Счетчики ошибок сети InfiniBand		
Счетчики ошибок сети InfiniBand за последний период	Посмотреть краткий вариант (без port_rcv_switch_relay_errors)		
Ошибки топологии InfiniBand	ib1-4 port 17	port_xmit_discard	10
Статистика счетчиков сети InfiniBand	ib1-4 port 18	port_xmit_discard	5
Топология InfiniBand	ib1-4 port 19	port_xmit_discard	5
Статьи по InfiniBand на русском	ib1-4 port 20	port_xmit_discard	3
Документация по InfiniBand	ib1-4 port 21	port_xmit_discard	1
На главную	ib1-4 port 23	port_xmit_discard	1
	ib1-4 port 24	port_xmit_discard	1
	ib1-4 port 25	port_xmit_discard	3
	ib1-4 port 26	port_xmit_discard	1
	ib1-4 port 27	port_rcv_switch_relay_errors	159
	ib1-4 port 28	port_xmit_discard	0

Рис. 4. Счетчики ошибок сети InfiniBand.

Страница с отображением изменения состояния счетчиков InfiniBand за последний период (разница в значениях счетчиков между двумя последними наборами). Страница также генерируется утилитой `html_counters`.

Страница работы с базой данных сохраненных состояний счетчиков InfiniBand показана на рисунке 5. Эта страница содержит инструменты для выборки из базы данных счетчиков любого сохраненного набора состояний счетчиков, осуществляет поиск по дате, предоставляет фильтр по узлам и счетчикам, а также позволяет задать порог ошибок. Кроме этого предоставляется возможность получить изменения показаний счетчиков за любой промежуток времени с теми же фильтрами.

Счетчики ошибок сети InfiniBand

Счетчики ошибок сети InfiniBand за последний период

Ошибки топологии InfiniBand

Статистика счетчиков сети InfiniBand

Топология InfiniBand

Статьи по InfiniBand на русском

Документация по InfiniBand

На главную

INFINIBAND™
Trade Association

Показания счетчиков на момент времени

Текущее время: 2014-07-03 09:05:52 Перейти: **Изменение показаний счетчиков за промежуток времени**

Поиск по дате: Порог ошибок (не меньше чем): Запись производится в середине часа

Выборка по хостам: Выборка по счетчикам:

host	port	port_rcv_errors	port_rcv_switch_relay_errors	port_xmit_discard	vl15_dropped
ib1-4	17	0	0	10	0
ib1-4	18	0	0	5	0
ib1-4	19	0	0	5	0
ib1-4	20	0	0	3	0
ib1-4	21	0	0	1	0
ib1-4	23	0	0	1	0
ib1-4	24	0	0	1	0
ib1-4	25	0	0	3	0
ib1-4	26	0	0	1	0
ib1-4	27	0	159	0	0
ib1-4	28	0	9	2	0
ib1-4	29	0	0	3	0
ib1-4	31	0	14	4	0
ib1-4	33	0	0	15	0
ib1-4	34	0	0	6	0
ib1-4	35	0	0	4	0
ib1-4	36	0	0	7	0
ib1-5	17	0	315	0	0
ib1-5	18	0	0	2	0
ib1-5	21	0	0	3	0
ib1-5	22	0	0	1	0

Рис. 5. Выборка из базы данных счетчиков InfiniBand.

Масштабируемость

Рекомендуемая частота опроса счетчиков - один час, это время можно уменьшить до времени опроса фабрики, которое составляет от одной секунды до минуты, в зависимости от количества узлов.

В пакете VIBADM нет ограничения на размер вычислительного комплекса; ограничения накладываются возможностями ПО OFED и размером базы данных MySQL.

Скорость работы веб-интерфейса зависит от количества узлов и настроек веб-сервера и php.

ЛИТЕРАТУРА:

1. Суперкомпьютер ВНИИТФ «Зубр»: сочетание вычислительной мощности, компактности и надежности. URL: <http://www.vniitf.ru/vse-kommercheskie-predlozheniya/1106-o-276>.
2. М. Кузьминский Infiniband: аппаратные средства и поддержка в Linux. Открытые системы, 2008 г. Электронный документ, URL: <http://www.osp.ru/os/2008/07/5478314>.
3. Сайт альянса OpenFabrics. URL: <http://www.openfabrics.org>.
4. Д.Э. Камер, Д.Л. Стивенс Сети TCP/IP, том 3. Разработка приложений типа клиент/сервер для Linux/POSIX.: Пер. с англ. – М.: Издательский дом «Вильямс», 2002г.
5. Infiniband Architecture Specification Release 1.1. Infiniband Trade Association. Nov. 6, 2002. Электронный документ, URL: <http://www.infinibandta.org>.