ОПТИМИЗАЦИЯ ПАРАМЕТРОВ РАБОТЫ КОНТРОЛЛЕРА PCI EXPRESS АДАПТЕРНОГО БЛОКА СИСТЕМЫ МЕЖПРОЦЕССОРНЫХ ОБМЕНОВ СМПО-10GA-AD

М. П. Авдеев, С. В. Дыдыкин, Ю. А. Малых, В. С. Попов

ФГУП «РФЯЦ-ВНИИЭФ», г. Саров Нижегородской обл.

На протяжении нескольких лет, в отделе 0828 разрабатывается отечественная система межпроцессорного обмена (СМПО), предназначенная для построения вычислительных кластеров петафлопсного класса [1]. Прототип СМПО-10G разработан на базе программируемой логической интегральной схемы (ПЛИС). На основе отработанной в ПЛИС RTL модели разработан кристалл сверхбольшой интегральной схемы (СБИС).

Одним из ключевых элементов СМПО является адаптерный блок СМПО-10GA-AD, представляющий собой многослойную печатную плату, выполненную в конструктиве платы расширения для вычислительного узла. Интерфейсом взаимодействия между адаптерным блоком и вычислительным узлом является PCI Express2.0x8. Для связи со смежными вычислительными узлами адаптерный блок СМПО-10GA-AD имеет четыре высокопроизводительных последовательных канала.

На рис. 1 представлена упрощенная структурная схема адаптерного блока СМПО. Адаптерный блок имеет в своем составе:

– 4 высокоскоростных канала производительностью 40 Гбит/с каждый;

– блок коммуникационного управления, в котором реализованы сетевой и транспортный уровень;

– контроллер PCI Express – обеспечивающий взаимосвязь блока коммуникационного управления и ядра PCI Express.

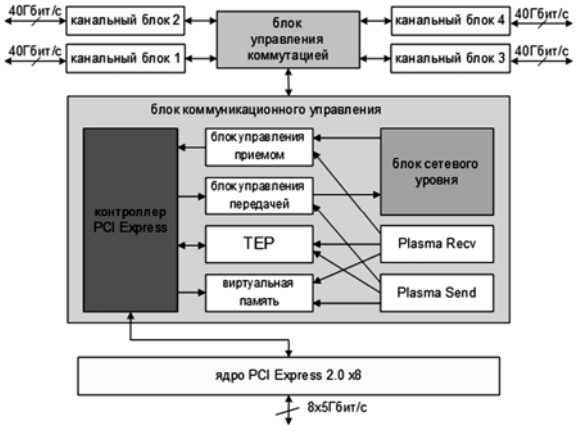

Рис. 1. Структурная схема адаптерного блока

Эффективность работы контроллера PCI Express во многом определяет характеристики производительности системы.

На рис. 2 представлены значения пиковой и теоретической производительности различных интерфейсов адаптерного блока.

Под пиковой производительностью будем понимать производительность интерфейса без учета применяемой кодировки (символьная производительность).

Под теоретической производительностью будем понимать максимальную производительность, которую может достичь интерфейс с учетом используемой кодировки.

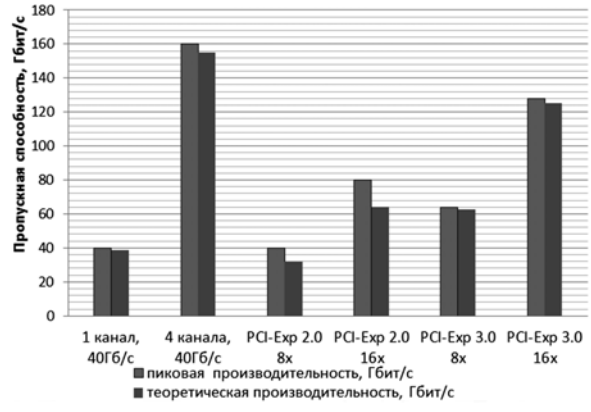

Рис. 2. Прописная способность различных интерфейсов

Для применяемого высокопроизводительного канала пиковая производительность составляет 40 Гбит/с, а теоретическая 39 Гбит/с, соответственно 4 таких канала позволяют агрегировать пропускную способность 157 Гбит/с.

PCI Express 2.0 ×8 имеет пиковую производительность 40 Гбит/с, а теоретическую 32 Гбит/с. Стандарт PCI Express 2.0 несет большие накладные расходы, равные 20 % [2], применение нового стандарта PCI Express 3.0 позволит повысить эффективность передачи данных.

Сравним теоретическую производительность 4-х высокоскоростных каналов, равную 160 Гбит/с, с теоретической производительностью PCI Express 2.0x8, равную 32 Гбит/с. PCI Express имеет в 5-ть раз меньшую пропускную способность, из чего можно сделать вывод, что дальнейшее увеличение скорости канала не приведет к существенному росту производительности СМПО.Для увеличения производительности СМПО необходимо либо применение стандарта PCI Express 3.0 или оптимизация существующего контроллера PCI Express 2.0.

Недостатки реализации контроллера PCI Express:

– фиксированные значения размера кадра для запросов записи и чтения. Max_Payload_Size = 256, Max_Read_Request;

– поддержка прерываний типа INTA;

– поддержка 32/64 разрядных транзакций с регистрами BAR.

Предлагаемые пути оптимизации контроллера PCI Express:

– снижение накладных расходов при работе с DMA или регистрами BAR;

– поддержка многовекторных прерываний.

Проведем оценку накладных расходов контроллера PCI Express. Накладные расходы представляют собой отношение количества полезных переданных данных к общему количеству переданных данных в одном кадре сообщения.

На рис. 3 представлен формат кадра PCI Express [2]. Кадр состоит из следующих полей:

– поле Start, размер 1 байт;

– поле Sequence, размер 2 байта;

– поле Header, размер 12 или 16 байт в зависимости от типа кадра;

– поле Payload, размер 0-4096 байт;

– поле ECRC, размер 4 байта;

– поле LCRC, размер 4 байта;

– поле End, размер 1 байт.

При передаче «коротких сообщений» накладные расходы составляют 15–40 %. Таким образом, уменьшение накладных расходов возможно за счет увеличения полезной нагрузки кадра PCI Express.

При работе с DMA, или регистрами BAR формируются запросы на чтение/запись. Стандарт PCI Express ограничивает количество запрашиваемых/передаваемых данных параметрами Max_Read_Request, Max_ Payload_Size соответственно. В свою очередь, значения указанных параметров зависят от конечного устройства и вычислительного сервера, с которым работает устройство. Максимальное значение параметров может достигать 4096 байт. Значения, с которыми может работать устройство и сервер, считываются из конфигурационного пространства на этапе инициализации драйвера устройства.

Использование фиксированных значений Max_Read_Request, Max_Payload_Size не является оптимальным с точки зрения достижения максимальной пропускной способности. ПриMax_Read_Request = 512, накладные расходы составляют 4 %, а Max_Payload_Size = 256 составляют 7 %.

Снижение накладных расходов позволит эффективней использовать пропускную способность PCI Express, так как появится возможность формировать запросы на большее количество данных.

Стандарт PCI Express [2] ограничивает количество прерываний типа INTA. Согласно стандарту такое прерывание может быть только одно, что влечет за

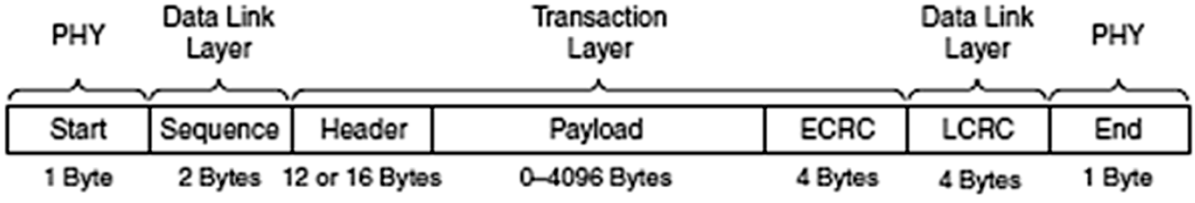

Рис. 3. Формат кадраблока работы с DMA

Размер служебных полей кадра составляет 20–28 байт, а возможный размер передаваемых полезных данных составляет 0–4096 байт. На рис. 4 приведен график накладных расходов в зависимости от размера передаваемых полезных данных.

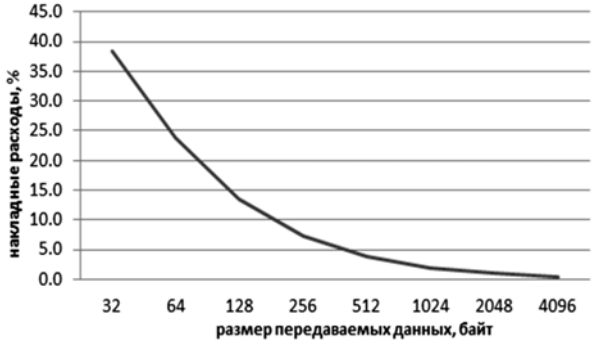

Рис. 4. График зависимости накладных расходов от количества передаваемых данных

собой дополнительные расходы на установление источника прерываний путем считывания регистра статуса, в котором хранится информация об источнике прерывания.

Использование двух и более векторов прерываний позволит ускорить обработку прерываний, так как не будет необходимости считывать регистр статуса, и обработчики прерываний смогут работать одновременно. Использование двух и более векторов прерываний возможно при использовании прерываний типа MSI/MSIX.

Одним из основных достоинств MSI/MSIX прерываний является тот факт, что в много процессорных и многоядерных системах обработчики прерываний MSI/MSIX могут выполняться на разных ядрах одновременно и независимо.

Прерывания MSIX поддерживают до 2048 векторов. Предлагается использовать два вектора прерываний типа MSIX, один для передаваемого потока данных, другой для принимаемого потока данных.

За счет разделения обработчиков прерываний можно ускорить их обработку и, как следствие, улучшить производительность контроллера PCI Express при работе в дуплексном режиме.

Используемое ядро PCI Express не предусматривает готового контроллера MSI-X прерываний, поэтому данный блок реализован самостоятельно. Разработанный блок является универсальным и способен работать с ядрами PCI Express других производителей. Структурная схема модуля представлена на рис. 5.

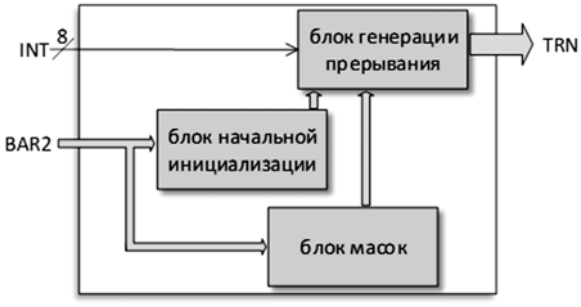

Рис. 5. Структурная схема модуля генерации MSI-X прерываний

Блок генерации MSI-X прерываний состоит из блока начальной инициализации, блока масок и блока генерации прерываний. Блок начальной инициализации работает при инициализации драйвера. Блок масок позволяет производить маскирование прерываний, блок генерации формирует запрос типа сообщение, при получении которого сервер формирует соответствующее прерывание.

Аппаратные затраты на оптимизацию, см. таблицу, составили единицы % от общего количества блоков ПЛИС, из чего можно сделать вывод, что проведенная оптимизация является не «дорогой» и может быть реализована в дальнейших проектах.

Аппаратные затраты на оптимизацию

	Исходный вариант, шт./%	**Оптимизированный вариант, шт.**	**Δ, шт./%**
Slice LUT	38305/24	39345/26	1040/2
Slice Register	40544/13	45047/14	4053/1
Block RAM	170/14	194/16	24/2

В процессе оптимизации контроллера PCI Express создан блок генерации MSIX прерываний, позволяющий драйверу отслеживать и обрабатывать одновременно до восьми независимых прерываний, что позволило увеличить пропускную способность в дуплексном режиме.

Добавлена возможность считывать и применять параметры Max_Read_Request_Size, Max_Payload_Size из конфигурационного пространства, обеспечивающая гибкость настройки контроллера PCI Express.

Добавлена поддержка работы с 128 разрядными транзакциями на операциях с регистрами BAR, результатом чего стало увеличение производительности системы, за счет снижения числа затратных транзакций на операциях с BAR.

По итогам проведенной работы, с использованием оптимизированного контроллера PCI Express получены следующие результаты.

На тесте Ping Pong:

1. задержка передачи в одном направлении:
   - до оптимизации – **5,92** мкс;
   - после оптимизации – **5,77** мкс.
2. максимальная пропускная способность в одном направлении:
   - до оптимизации – **1707** Мбайт/с;
   - после оптимизации – **1807** Мбайт/с.

На тесте Send Recv:

1. задержка передачи в двух направлениях:
   - до оптимизации – **6,2** мкс;
   - после оптимизации – **5,93** мкс.
2. максимальная пропускная способность в двух направлениях:
   - до оптимизации – **2242** Мбайт/с;
   - после оптимизации – **2894** Мбайт/с.

При аппаратных затратах не более 2% оптимизация контроллера PCIExpress позволила увеличить производительность системы:

- 6 % при передаче сообщений в одном направлении;
- 22 % при работе в дуплексном режиме.

Литература

1. Холостов А. А. Масштабируемая система межпроцессорных обменов 10 G // [Электронный ресурс] – Национальный суперкомпьютерный форум, 2013. Режим доступа: www.nscf.ru.

2. PCI Express Base Specification Rev. 2.0, 2006. [Электронный ресурс] режим доступа – http://pcisig.com.