Сетевые устройства для реализации «облачных вычислений»

Владимир Вычужанин (г. Одесса)

В статье рассматривается реализация стандарта 100 Гбит/с Ethernet в сетевых устройствах на основе 28-Гбит/с трансиверов, выполненных на ПЛИС Virtex-7 H580T фирмы Xilinx и оптических модулях. Устройства обладают широкой полосой передачи данных, высокой плотностью портов, низким энергопотреблением и сочетают требуемую масштабируемость с невысокой стоимостью.

Введение

«Облачные вычисления» – перспективное направление современной информатики, являющееся альтернативой локально используемому аппаратному и программному обеспечению. Идеология «облачных вычислений» заключается в переносе организации вычислений и обработки данных с локальных компьютерных центров на серверы Интернета [1]. Для реализации быстрого доступа к информации, хранящейся в «облаках», необходимо использование специальных сетевых устройств, позволяющих передавать данные на большие расстояния. Кроме того, подобные средства должны обеспечивать широкую полосу передачи данных на скорости не менее 100 Гбит/с, высокую плотность портов, низкое энергопотребление, а также сочетать требуемую масштабируемость с невысокой стоимостью сетевых устройств [2, 3].

Стандарт 100 Гбит Ethernet

Разрабатываемые быстродействующие 100-Гбит сетевые инфраструктуры для технологии «облачных вычислений» должны соответствовать требованиям стандарта IEEE Std 802.3ba 100 Gigabit Ethernet [4]. Стандарт регламентирует использование параллельной и последовательной передачи данных на скорости 100 Гбит/с, а также обеспечение максимальной скорости передачи данных 25 Гбит/с на одной несущей частоте. Кроме того, в соответствии со стандартом 100GbE, сетевые устройства для реализации «облачных вычислений» должны:

- поддерживать скорость передачи данных 100 Гбит/с на логическом MAC-уровне управления доступом к среде передачи данных;
- поддерживать только полнодуплексные режимы Ethernet уровня MAC;
- сохранять формат кадра Ethernet 802.3 уровня MAC;
- обеспечивать для интерфейса между уровнями MAC и PHY (физическим) поддержку значения BER (коэффициента битовых ошибок) не хуже 10^{-12};
- сохранять минимальный и максимальный размеры кадров стандарта IEEE 802.3;
- обеспечивать совместимость с оптическими транспортными сетями OTN.

Логический MAC-уровень управления доступом к среде передачи данных (подуровень канального, второго уровня модели OSI), согласно требованиям стандарта 100GbE, является подуровенем протокола и реализует адресацию и механизмы управления доступом к каналам передачи данных. Это позволяет нескольким терминалам или точкам доступа общаться между собой в многоточечной сети. MAC обеспечивает гибкость при взаимодействии разнотипных устройств (PHY и DTE) при передаче потока данных со скоростью 100 Гбит/с. Логический уровень реализует преобразование пакетов верхних уровней в кадры Ethernet: сегментирует, добавляет к заголовку преамбулу, MAC-адрес и контрольную последовательность FCS.

Физический уровень стандарта 100GbE состоит из трёх основных (PCS – Physical Coding Sublayer, PMA-Physical Medium Attachment, PMD – Physical Medium Dependent Sublayer) и двух необязательных подуровней [5, 6]. С уровня MAC данные попадают на подуровень согласования (RS – Reconciliation Sublayer), где последовательный поток данных преобразуется в параллельный 64-битный (64B) и через интерфейс CGMII (100 Gigabit Media Independent Interface – независимый от среды передачи 100 Гбит логический интер-

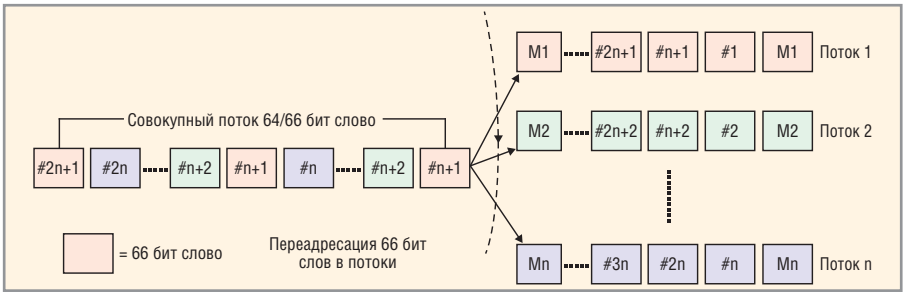

Рис. 1. Переадресация в сети 100 Гбит Ethernet 66-бит слов в индивидуальные потоки

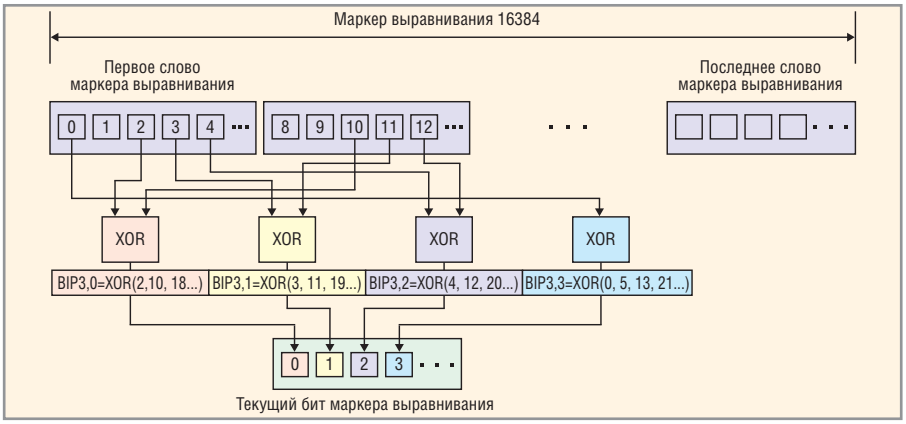

Рис. 2. Включение маркера выравнивания подуровнем PCS для полосы сдвига и полосы изменения порядка передачи данных

фейс) попадает на подуровень кодирования PCS.

Физический подуровень PCS осуществляет кодирование потока данных, поступающих в канал передачи, таким образом, чтобы они могли быть различимы приёмником и восстановлены в исходной форме. На физическом подуровне PCS часто применяет схему кодирования (скремблирования) 64B/66B, в которой 66-битное слово переадресуется карусельным образом в индивидуальные параллельные потоки, т.е. 64-битный поток данных (64B) преобразуется в 66-битный поток (66B) и разделяется на несколько потоков с меньшей скоростью (см. рис. 1). При кодировании к заголовку получения 64-битных данных добавляются дополнительные два бита синхронизации с целью формирования 66-битного блока – «01b» синхронизации заголовков пакетов данных и «10b» для управления пакетами данных. При распространении 66-битных блоков полос PCS, начиная с нулевой полосы, используется циклический механизм.

Согласно стандарту 100GbE, на физическом подуровне PCS определяются до двадцати полос передачи данных по двум направлениям (прямом – TX и обратном – RX). Так, поток со скоростью 100 Гбит/с расширяется до скорости 103,125 Гбит/с и распределяется на двадцать полос PCS по 5,15625 Гбит/с с поддержкой их интерфейсов. Для 100 Гбит Ethernet выходных битовых потоков может быть 10 или 4, с возможным их перемешиванием по определённому закону и распределением по выходным потокам.

Поскольку последовательный поток битов распределяется по индивидуальным параллельным потокам, для восстановления исходного потока в приёмнике, между параллельными потоками должна сохраняться временная синхронизация. Учитывая скорости и расстояния передачи данных, предусмотренные стандартом 100GbE, физические рассогласования потоков неизбежны. Для компенсации временных сдвигов (перекосов) используются специальные маркеры.

Маркер выравнивания имеет заголовок, контроль синхронизации («10b») и представляет собой DC-сбалансированный поток, состоящий из восьми байтов {M0, M1, M2, BIP3, M4, M5, M6, BIP7}, причём M4, M5, M6 являются побайтно обратными байтам M0, M1 и M2. Каждая полоса для подуровня

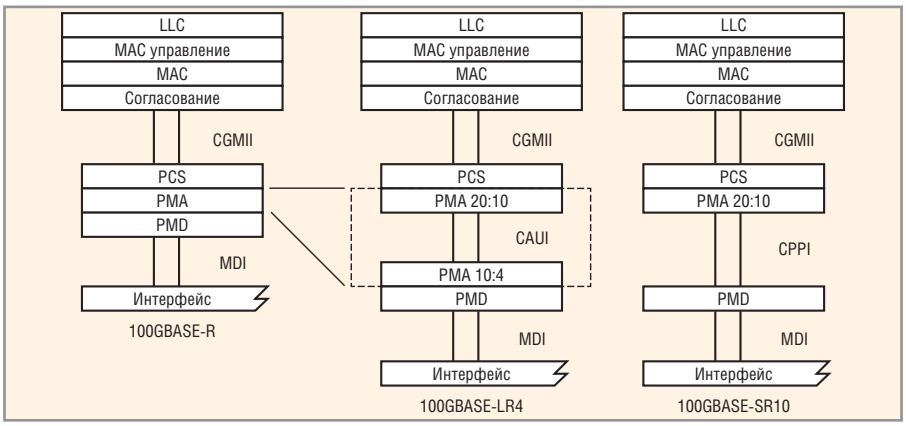

Рис. 3. Варианты архитектуры 100 Гбит Ethernet для протоколов 100GBASE-R, 100GBASE-LR4 и 100GBASE-SR10

PCS передачи данных (0, 1...19) имеет единственное байтное кодирование M0, M1, M2, позволяющее получить и расшифровать номер соответствующей полосы подуровня PCS (см. рис. 2).

Маркер выравнивания указателей вставляет 66-битный блок после кодирования 64B/66B одновременно в каждый поток данных через 16 384 кодовых 66-битных блока и удаляется в приёмнике данных при декодировании 64B/66B. Измеряя временной сдвиг поступления кодовых блоков, приёмник способен восстановить синхронность индивидуальных параллельных потоков. Передатчик, чтобы вставить маркер выравнивания, удаляет пробелы между пакетами IPG (Inter-packet gaps). С прекращением получения данных подуровнем PCS удаляются маркеры выравнивания.

Наличие периодического маркера выравнивания позволяет приёмнику нормально функционировать при значительных временных рассогласованиях сигналов между параллельными каналами передачи данных. Максимально допустимое значение перекоса в PCS составляет 180 нс для стандарта 100GbE.

После кодирования и синхронизации на подуровне PCS данные поступают на физический подуровень PMA, выполняющий функции тестирования передачи данных – генерацию тестовых последовательностей, формирование петли обратной связи данных для тестирования и т.п.

К особой группе протоколов, используемых для реализации интерфейсов физического уровня, основанной на методе блочного кодирования данных кодом 64B/66B и использующей спецификацию PMA, относится

Рис. 4. Общий вид оптического модуля CFP2

протокол 100GBASE-R физического подуровня PCS.

Протокол 100GBASE-R осуществляет кодирование/декодирование данных, полученных от интерфейса (или переданных на интерфейс) CGMII в виде кода в последовательной форме, а также распределение данных на PMA по нескольким параллельным потокам. Протокол 100GBASE-R подуровня PCS, функционально обеспечивая отображение пакетов кодом 64B/66B, распределяет пакеты на двадцать последовательных потоков PCS. Логический интерфейс CGMII обеспечивает соединение подуровня MAC с физическим уровнем PHY. В сетевых устройствах могут применяться различные варианты физического уровня PHY в ви-

Таблица 1. Основные параметры ПЛИС Virtex-7 H580T

Количество секций Slices	90 700
Число логических ячеек Logic Cells	580 480
Общее число блоков CLB	725 600
Объём блочной памяти Block RAM, Кбит	33 840
Объём распределённой памяти, Кбит	8850
Block RAM/FIFO w/ECC	940
Количество модулей CMTs (1MMCM+1PPL)	12
Максимум несимметричных I/O	600
Максимум дифф. пар I/O	288
Число аппаратных секций DSP48E1	1680
Число аппаратных модулей PCI Express Interface	2
Число приёмопередатчиков GTH 13,1 Гбит/с	48
Число приёмопередатчиков GTZ 28,05 Гбит/с	8
Объём конфигурационной памяти, Мбит	183,6

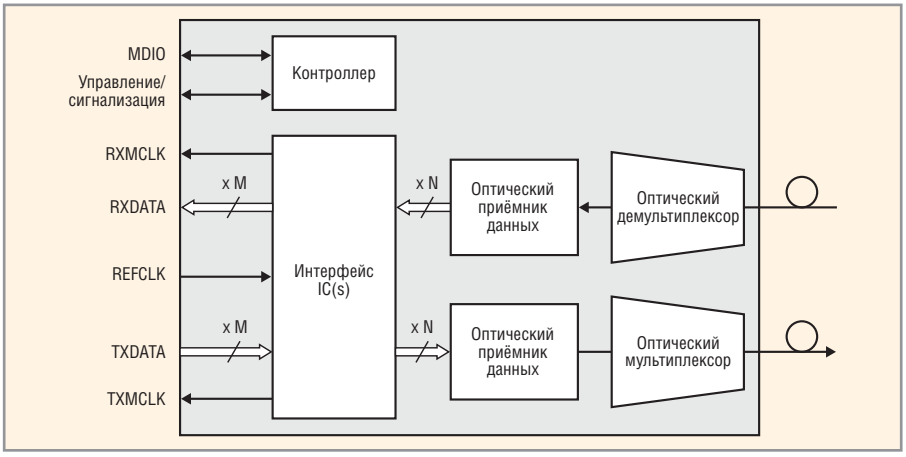

Рис. 5. Блок-схема оптического модуля CFP2

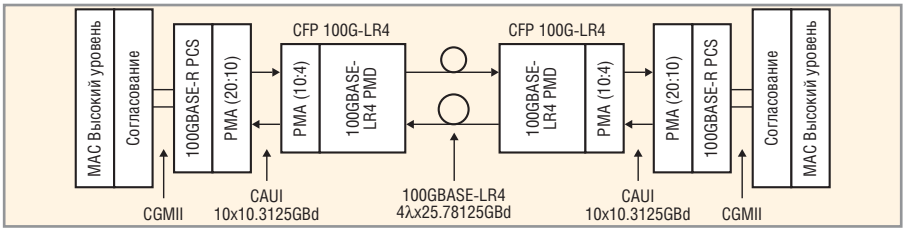

Рис. 6. Подключение оптических трансиверов CFP MSA с использование протокола 100GBASE-LR4

де модулей, стандартизированных в 802.3ba в различных MSA (Multi-source agreements) и использующих оптическое волокно (см. таблицу 1).

В качестве примера на рисунке 3 приведены три варианта архитектуры 100GbE для протоколов 100GBASE-R, 100GBASE-LR4 и 100GBASE-SR10 систем передачи сигналов. Схемы содержат: CPPI – параллельный физический интерфейс, сервисный интерфейс CAUI (gigabit per second Attachment Unit Interface) – модуль 100 Гбит/с интерфейса подключения, а также логический интерфейс CGMII.

Реализация каналов связи 100 Гбит/с

В настоящее время для одновременной передачи данных со скоростью 100 Гбит/с и более по нескольким каналам используется последовательный высокоскоростной интерфейс на основе меди. Для компенсации ухудшения качества сигнала применяют его выравнивание на сторонах передачи и приёма. Однако эффективность подобных решений ограничена расстоянием передачи: чем выше скорость, тем меньше расстояние, на которое данные могут быть переданы без ущерба для целостности сигнала. В целом подобные решения эффективны при небольших длинах медных кабелей, что в принципе не подходит для каналов связи Интернет.

Таблица 2. Стандартизованные варианты PHY

PHY	100 Гбит Ethernet
Минимум 10 км по SMF	100GBASE-LR4
Минимум 40 км по SMF	100GBASE-ER4

Реализацию стандарта IEEE Std 802.3ba 100GbE можно обеспечить высокопроизводительными сетевыми решениями для пакетной обработки трафика управления, коммутации и агрегации на основе применения оптических интерфейсов.

Известные оптические стандарты включают оптические модули форматов SFP+, CFP (C form-factor pluggable) и CFP2. Перечисленные оптические модули выгодно отличаются пропускной способностью, низкой стоимостью передачи одного бита, энергетической эффективностью, а также форм-фактором [7]. Так, оптический модуль SFP+ поддерживает скорость оптической линии связи 10 Гбит/с, а CFP – 100 Гбит/с. Хотя CFP потребляют больше энергии в расчёте на бит, чем SFP+, используемая интеграция одного волокна уменьшает сложность построения и затраты на обслуживание. Оптический модуль формата CFP2 (см. рис. 4 и 5) обладает пропускной способностью 100 Гбит/с, как и модуль CFP, но имеет в два раза меньшие размеры и энергопотребление, а также меньшую стоимость. Одним из стандартизованных оптических модулей, поддерживающих 100 Гбит Ethernet, является CFP MSA, который осуществляет первоочередные подключения оптических трансиверов с использованием протокола 100GBASE-LR4, в том числе, высокоскоростного 100 Гбит/с (см. рис. 6).

Стандарт 100GbE и ПЛИС

Реализация стандарта IEEE Std 802.3ba 100GbE для технологий «облачных вычислений», помимо решения проблемы, связанной с расстоянием передачи данных, требует применения широкополосных быстродействующих сетевых инфраструктур, способных обеспечить функционирование логического MAC и физического уровней для 100 Гбит Ethernet.

Следует также отметить, что при реализации технологии «облачных вычислений» возникает ряд аппаратных проблем, связанных с ограничением функций подсистемы хранения данных, коммутаторов, маршрутизаторов и систем ввода/вывода. Также ограничена внешняя скорость передачи данных по кабелям и другим соединениям, связывающим коммутаторы, маршрутизаторы и системы хранения данных.

Гибкость и реконфигурируемость технологии ПЛИС позволяет использовать их в системах, требующих широкого набора средств для обработки потоков ввода/вывода данных со скоростью 100 Гбит/с. Сетевые операционные преимущества подобных схемных структур вытекают из присущей им эффективной маршрутизации при обработке потоков данных 100 Гбит/с. В этом контексте технология «облачных вычислений» оказывает влияние на широкое внедрение устройств на ПЛИС для обработки высокоскоростных потоков данных.

Для обработки высокоскоростных потоков данных 100 Гбит/с фирма Xilinx предлагает использовать гетерогенные 3D-матрицы FPGA Virtex-7 H580T (см. таблицу 2), состоящие из соответствующих матриц кремния SSI (Stacked Silicon Interconnect) (см. рис. 7 и 8) [8]. FPGA Virtex-7 H580T с трёхмерной интеграцией, установленная в широкополосных быстродействующих сетевых инфраструктурах, может реализовать до 16 трансиверов с пропускной способностью 28 Гбит/с или до 72 трансиверов со скоростью 13,1 Гбит/с, а также использоваться в виде кристалла на платах N×100 Гбит/с и 400 Гбит/с. Благодаря разделению трансиверов и ядра достигается шумовая изоляция, способствующая сохранению целост-

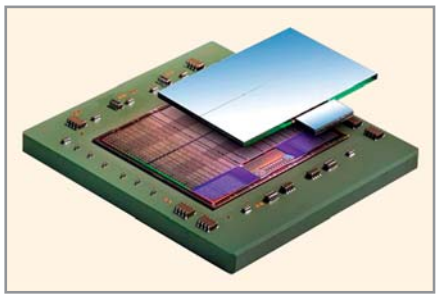

Рис. 7. Общий вид гетерогенной программируемой 3D-матрицы Virtex-7 H580T

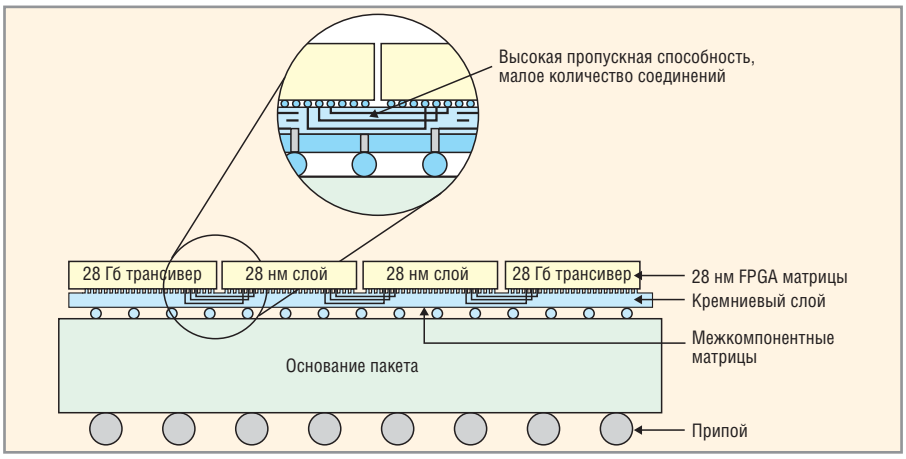

Рис. 8. Ячейка Virtex-7 H580T, выполненная по кремниевой технологии (вид сбоку)

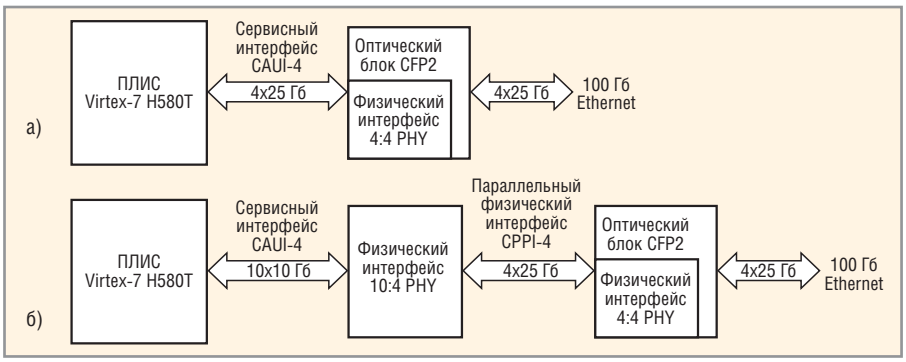

Рис. 9. Место физического уровня PHY в соединениях ПЛИС с оптическим модулем CFP2

ности обрабатываемых сигналов и увеличению ресурса системы. В устройстве FPGA Virtex-7 H580T предусмотрены дополнительные возможности отладки функций оптических транспортных сетей OTN, логического MAC-уровня, Interlaken IP и исключена необходимость использования отдельных кодируемых элементов и устройств ASSP (Application Specific Standard Product).

Гетерогенная архитектура Virtex-7 H580T, реализующая до 16 трансиверов, обеспечивает скорость 28 Гбит/с для оптического модуля формата CFP2. На скорости 100 Гбит/с предусмотрены дальний (LR – до 10 км) и сверхдальний (ER – до 40 км) режимы работы оптического модуля CFP2.

Физический уровень PHY при соединении ПЛИС с оптическим модулем поддерживает высокую мощность режима работы интерфейса CAUI-4 (см. рис. 9а) или низкую мощность режима работы CPPI-4 (см. рис. 9б). Оптический модуль CFP2 использует 10-кратный 10/11-Гбит или четырёхкратный 25/28-Гбит интерфейс. Переход на оптические модули с четырёхкратным 25/28-Гбит интерфейсом позволяет использовать совместно с ПЛИС до восьми оптических модулей 100 Гбит/с.

Структура физического подуровня PCS, подключаемого к физическому подуровню PMA, как было отмечено ранее, выполняет функцию кодирования. Программируемые FPGA-устройства с 28-Гбит поддержкой масштабируемого интерфейса SerDes (Serializer/Deserializer) могут быть использованы для реализации двухпортового блока кодирования с расширенными функциями тестирования и отладки. На рисунке 10 показано совместное подключение двухпортового блока кодирования (с расширением двух портов 100 Гбит/с) на основе Virtex-7 H580T и оптического модуля CFP2.

ПЛИС Virtex-7 H580T поддерживает:

- протокол SFI-S с 11 полосами по 11,2 Гбит/с (одна полоса – на устранение перекоса) и до 72 SerDes по 13,1 Гбит/с;
- протокол SFI-S с 5 полосами по 28 Гбит/с (одна полоса – на устранение перекоса) и до 16 SerDes по 28 Гбит/с.

Блок кодирования принимает входящие 10-кратные потоки 10/11 Гбит/с и после кодирования передаёт их четырёхкратным последовательным ин-

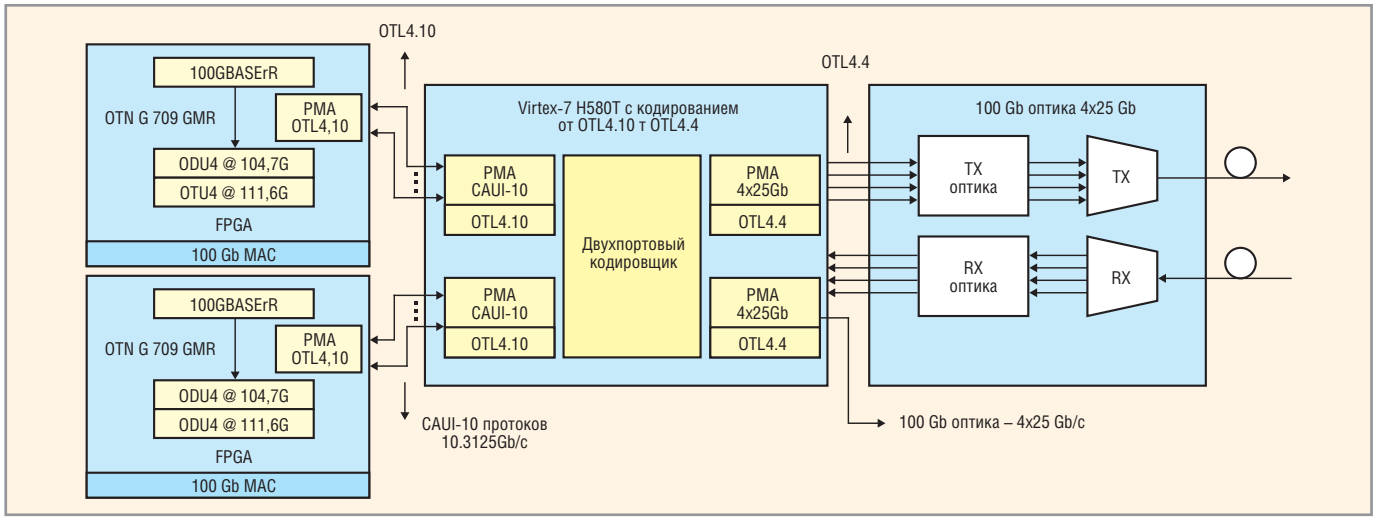

Рис. 10. Блок-схема 100-Гбит/с оптического модуля и ПЛИС

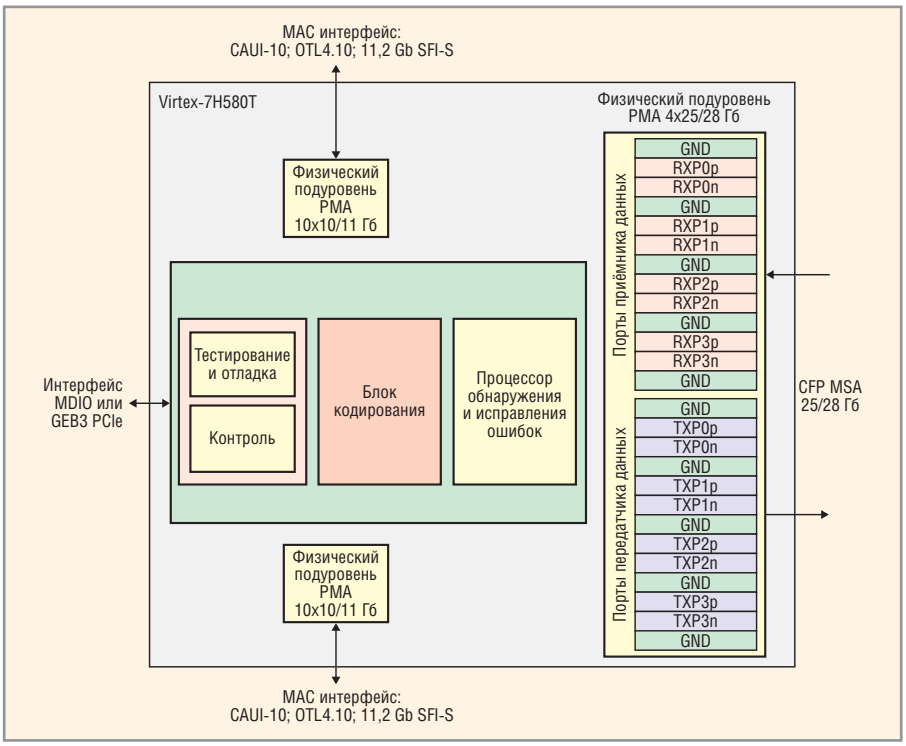

Рис. 11. Реализация блоком кодирования в ПЛИС Virtex-7 HT программируемых, расширяемых и гибких 100-Гбит/с приложений

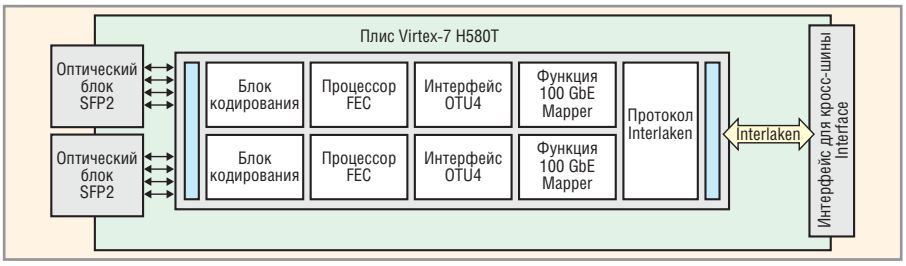

Рис. 12. OTN-транспондер 2×100 Гбит/с на Virtex-7 H580T и CFP2

терфейсом 25/28 Гбит/с с помощью подуровня PMA (20:4).

Применение ПЛИС Virtex-7 H580T позволяет реализовать 100-Гбит функции MAC-уровня, NPU, Traffic Management/QoS, а также Framer и др. (см. рис. 11).

Блок кодирования обеспечивает поддержку:

- 10×10 Гбит CAUI-интерфейса для последовательного интерфейса связи 4×25 Гбит/с;
- физического перевода интерфейса с OTL 4.10 на OTL 4.4;
- перевода 11,2 Гбит/с 10-полосного интерфейса SFI-S в 28 Гбит/с четырёхполосный интерфейс SFI-S с устранением перекоса полосы.

С целью повышения плотности портов 100-Гбит блока кодирования осуществляется подключение ASIC/FPGA/ASSP Ethernet 100 Гбит/с с использованием функций стандартов MAC или OTN. На одной ПЛИС Virtex-7 H580T можно создавать 100-Гбит транспондеры OTN, содержащие несколько оптических модулей CFP2. Применяя двух- и/или четырёхъядерный блок кодирования в ПЛИС, можно подключить к Virtex-7 H580T до восьми 4×25/28 Гбит/с оптических модулей CFP2. В качестве примера на рисунке 12 приведён 2×100 Гбит/с транспондер стандарта OTN на одной ПЛИС Virtex-7 H580T и двух оптических модулях CFP2.

ПЛИС Virtex-7 H580T, помимо физического подуровня PMA, обладает связанными синхронизацией ресурсами для поддержки интерфейсов CAUI-10×10 Гбит/с, OTL 4.10, CPPI интерфейса 4×25 Гбит/с, а также интерфейса OTL 4.4. Оптическая транспортная сеть иерархии G.709 определяет 100 Гбит Ethernet в канале оптического блока данных типа ODU4 (optical data unit), используя общие процедуры отображения GMP (generic mapping procedure). В свою очередь, ODU4 отображается на канал оптического транспортного блока OTU4 (optical transport unit). В оптическом блоке OTU4 используется в качестве интерфейса OTL4.10 или OTL 4.4. В ODU4 клиентская скорость составляет 104,79 Гбит/с, а скорость передачи данных – 111,809 Гбит/с. В блоке OTU4 интерфейс OTL 4.10 связывает более десяти полос SerDes, каждая из которых работает на скорости (255/227) × × 9 953 280 Кбит/с = 11,18 Гбит/с. Интерфейс OTL 4.4 может быть использован для блока OTU4 для связывания четырёх полос SerDes на скорости (255/227) × 24 883 200 = 27,952 Гбит/с (см. рис. 9).

Для поддержки связи 100 Гбит/с, обнаружения и исправления ошибок при совместном функционировании ПЛИС Virtex-7 H580T и оптических модулей CFP в соответствии с протоколом OIF SFI-S 1.0 используется процессор FEC на 4–20 полос масштаби-

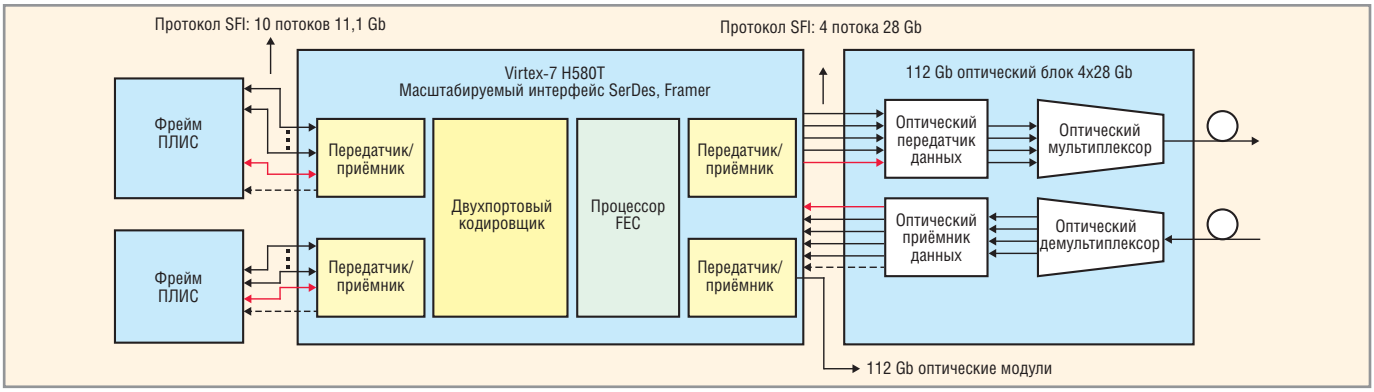

Рис. 13. Реализация протокола SFI-S в Virtex-7 H580T

руемого интерфейса (см. рис. 13). Блок кодирования в ПЛИС обеспечивает взаимодействие между устройствами, использующими различные полосы и линии фрейма SFI-S . Следует отметить, что протокол SFI-S не зависит от формата данных и может выполнять любые форматы протоколов трактов передачи и приёма данных. Протокол SFI-S используется для устранения перекоса в отдельных каналах передачи данных и позволяет упростить процедуру устранения перекоса и уменьшить сложность SerDes. Применение протокола SFI-S не даёт побочных эффектов, поскольку протокол использует дополнительную полосу SerDes.

Преимущества использования блока кодирования в Virtex-7 H580T заключаются ещё в том, что он позволяет осуществлять большие объёмы тестирования, отладки и контроля ПЛИС. Для этого в Virtex-7 H580T встроена модель генератора PRB на 13,1 Гбит/с или 28 Гбит/с для SerDes, что позволяет осуществить проверку физического подслоя PCS в различных режимах работы системы передачи данных. Кроме того, Virtex-7 H580T, имея значительное число блоков оперативной памяти (см. таблицу 2), может за несколько миллисекунд обеспечить проверку потоков получаемых данных различной длины.

Блок кодирования в ПЛИС также обеспечивает имитацию перекосов распространения сигнала. Для минимизации дрожания сигналов трансиверов в Virtex-7 H580T используется

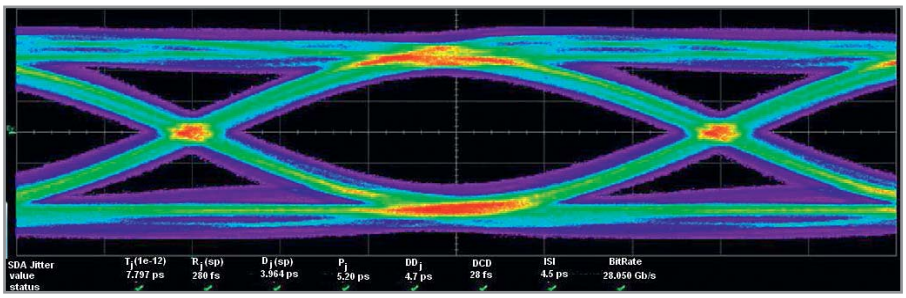

Рис. 14. Глазковая диаграмма 28-Гбит/с трансивера на Virtex-7 H580T

синхронизация с фазовой автоподстройкой частоты (PLL). Для компенсации потерь и поддержания целостности передаваемого сигнала в трансиверах реализована передача предыскажений и автоматическая адаптация в линейном эквалайзере приёмника.

Проведённые исследования 28-Гбит/с трансивера на Virtex-7 H580T показали, что полученная для него глазковая диаграмма (см. рис. 14) отражает низкий джиттер и высокое качество передаваемого сигнала.

Заключение

Для реализации стандарта 100 Гбит Ethernet в сетевых устройствах могут эффективно использоваться разработанные фирмой Xilinx на основе FPGA Virtex-7 H580T трансиверы 28 Гбит/с с низким фазовым шумом, сетевые карты N×100 Гбит/с и 400 Гбит/с и оптические модули CFP2. Такой комплект устройств обеспечивает быстрый доступ к информации, хранящейся в «облаках», обладает широкой полосой передачи данных на скорости не менее 100 Гбит/с, высокой плотностью портов, низким энергопотреблением и сочетает требуемую масштабируемость с невысокой стоимостью конечного оборудования.

Литература

1. *Zhou S.* Understanding the Evolution Dynamics of Internet Topology. Physical Review E. 2006. Vol. 74.
2. *Hewitt C.* ORGs for Scalable, Robust, Privacy-Friendly Client Cloud Computing. Massachusetts Institute of Technology. 2008. Vol. 12. № 5.
3. *Риз Дж.* Облачные вычисления. БХВ-Петербург, 2011.
4. IEEE 802.3ba-2010. IEEE Standard for Information Technology. Amendment 4: Media Access Control Parameters, Physical Layers and MaNagement Parameters for 40 Gb/s and 100 Gb/s Operation. IEEE, 22 June 2010.
5. *D'Ambrosia J.* 100 Gigabit Ethernet and Beyond. IEEE Communications Magazine. 2010.
6. *Toyoda H., Ono G., Nishimura Sh.* 100GbE PHY and MAC Layer Implementations. IEEE Commun. Mag. 2010. Vol. 50. No. 3.
7. CFP MSA Hardware Specification Revision. 2010. No. 14.
8. www.xilinx.com.