

РЕАЛИЗАЦИЯ ФРАГМЕНТА НЕЙРОННОЙ СЕТИ НА ПЛИС XILINX С ВОЗМОЖНОСТЬЮ ИЗМЕНЕНИЯ ВЕСОВ И ФУНКЦИИ АКТИВАЦИИ В РЕАЛЬНОМ ВРЕМЕНИ

Казанцев П.А., Остапенко Г.П., Галушкин А.И.

Научный центр нейрокомпьютеров при РАСУ
105066, Москва, Новая Басманная ул, 20, тел. 263-94-30

Аннотация. Разработан прототип каскадируемого фрагмента нейронной сети с прямыми последовательными связями с возможностью выбора с точки зрения скорости изменения весовых коэффициентов или входных признаков (быстрые/медленные). Фрагмент выполнен на базе программируемой логической интегральной схемы (ПЛИС) типа FPGA Virtex-E (XCV400E-pq240) с интеграцией 570 тыс. системных вентиляей. Оработана методика проектирования нейрочипа в рамках указанной технологии в среде САПР ISE 5.2 от Xilinx.

Введение

При разработке нейросетевого алгоритма, как и любого другого вычислительного алгоритма, главными факторами оценки его эффективности являются его точность и время его выполнения. Последний фактор не всегда существен при решении простых задач, однако при ориентации на сложные задачи реального времени скорость вычисления становится критической. Использование универсальных процессоров для выполнения нейросетевого алгоритма неэффективно с точки зрения отношения производительности к стоимости, поскольку для сложных задач размеры нейронной сети велики, а разрядность входных сигналов и весов, как правило, не совпадает со стандартом универсального процессора (на данный момент это 32 бита). По этой причине возникает необходимость разработки специализированного аппаратного ускорителя нейросетевого алгоритма – нейрочипа. Ввиду узкой специализации нейросетевого алгоритма, его аппаратная реализация в виде заказной СБИС не требует организации серийного производства и достаточно дорога. По многочисленным публикациям можно сделать вывод, что наиболее эффективной аппаратной реализацией нейронной сети является реализация на программируемых логических интегральных схемах – ПЛИС (ведущие изготовители – Xilinx, Altera) [1].

Большое количество внешних выводов ПЛИС, развитая структура быстродействующих межсоединений, повышенная скорость обмена между кристаллами (до 200 МГц) в совокупности позволяют построить на основе ПЛИС законченный достаточно регулярный фрагмент нейронной сети, а затем, осуществляя каскадирование ПЛИС простым соединением указанных для аппаратного каскадирования выводов, создавать нейронные сети произвольной конфигурации и сложности.[2].

На данном этапе разработок спроектирован фрагмент нейронной сети на кристалле XCV400E-pq240 фирмы Xilinx с интеграцией 570 тыс. системных вентиляей.

Структура фрагмента нейронной сети

Ввиду того, что размещение всей нейронной сети на кристалле невозможно по причине ограниченности физических ресурсов кристалла, возникает необходимость разбивать сеть на каскадируемые фрагменты. Для реализации был принят фрагмент нейронной сети с прямыми последовательными связями, как наиболее часто встречающийся в реальных задачах, состоящий из 8-ми нейронов. Разрядность весовых коэффициентов и входных сигналов равняется 8-ми. Каждый нейрон имеет 8-мь входов. Был выбран способ вертикальный каскадирования с увеличением числа нейронов в слое. [3]

Каскадирование может быть как физическим (каждый фрагмент располагается на отдельном кристалле), так и виртуальным (фрагмент зацикливается на одном кристалле), или смешанным. На выбор одного из указанных способов влияет заданное время решения задачи. При физическом каскадировании время решения задачи меньше времени решения при виртуальном каскадировании в число раз примерно равное числу физически реализованных фрагментов.

Реализация фрагмента на кристалле

Описанный фрагмент нейронной сети занимает 86% ресурсов кристалла XCV400E-pq240 (570 тысяч эквивалентных вентиляей) фирмы Xilinx. В занимаемый фрагментом объем входит 64-ре 8-ми разрядных умножителя на константу (все работают параллельно), 56 двухвходовых сумматоров, блочная однопортовая память, реализующая функцию активации, объемом 256X16 бит на каждый нейрон (всего 8-мь блоков), блок управления режимами. Все указанные компоненты являются элементами библиотек от Xilinx [4]

Корпус кристалла имеет 158 однобитовых пользовательских ножки. В данном проекте задействованы 143 ножки: 64 входных, 64 выходных, 1 – сигнал синхронизации, 15 – для управления элементами схемы и режимом работы фрагмента. Тактовая частота работы схемы: **90 МГц**. Время вычисления выходного сигнала по входному: **70 нс**.

Функционирование и режимы работы фрагмента

Фрагмент предназначен для параллельного ввода 8-ми составляющих входного вектора, параллельного их умножения на содержимое внутренних блоков памяти умножителей, параллельного сложения 8-ми произведений и параллельной активации полученных сумм в каждом нейроне (8-мь старших бит суммы подаются на адресный порт памяти, в которой прописаны дискретные значения выбранной функции активации).

Следует отметить, что в схеме намеренно убрано различие между весами и входными данными. Перемножаемые числа отличаются лишь тем, что одно из них может меняться быстро, а другое может меняться только медленно. Это сделано для того, чтобы фрагмент был инвариантен относительно задачи, так как в некоторых задачах требуется, чтобы признаки менялись быстро, а веса медленно (распознавание, классификация). В других задачах (решение систем алгебраических уравнений) требуется быстрое изменение весов, в то время как признаки могут меняться медленно.

Константа умножителя, содержащаяся во внутренней его памяти, может быть изменена. На операцию замены необходимо 16-ть тактов. Таким образом, например, может быть реализована донастройка весов в реальное времени. Такой подход удобен в случае, если быстрая замена весов не требуется. Эта функциональная возможность реализуется подачей сигналов переключения режима и выбора нейрона, а также сигнала разрешения на запись во внутреннюю память умножителей выбранного нейрона.

Содержимое блоков однопортовой памяти функции активации также может быть изменено. Количество тактов, необходимое для модификации, определяется количеством подлежащих замене значений. С этой целью также предусмотрены сигналы переключения режима и разрешения на запись. Однако в отличие от содержимого умножителей, значения ячеек памяти изменяются во всех 8-ми блоках одинаково, так как функции активации всех нейронов одного слоя одинаковы.

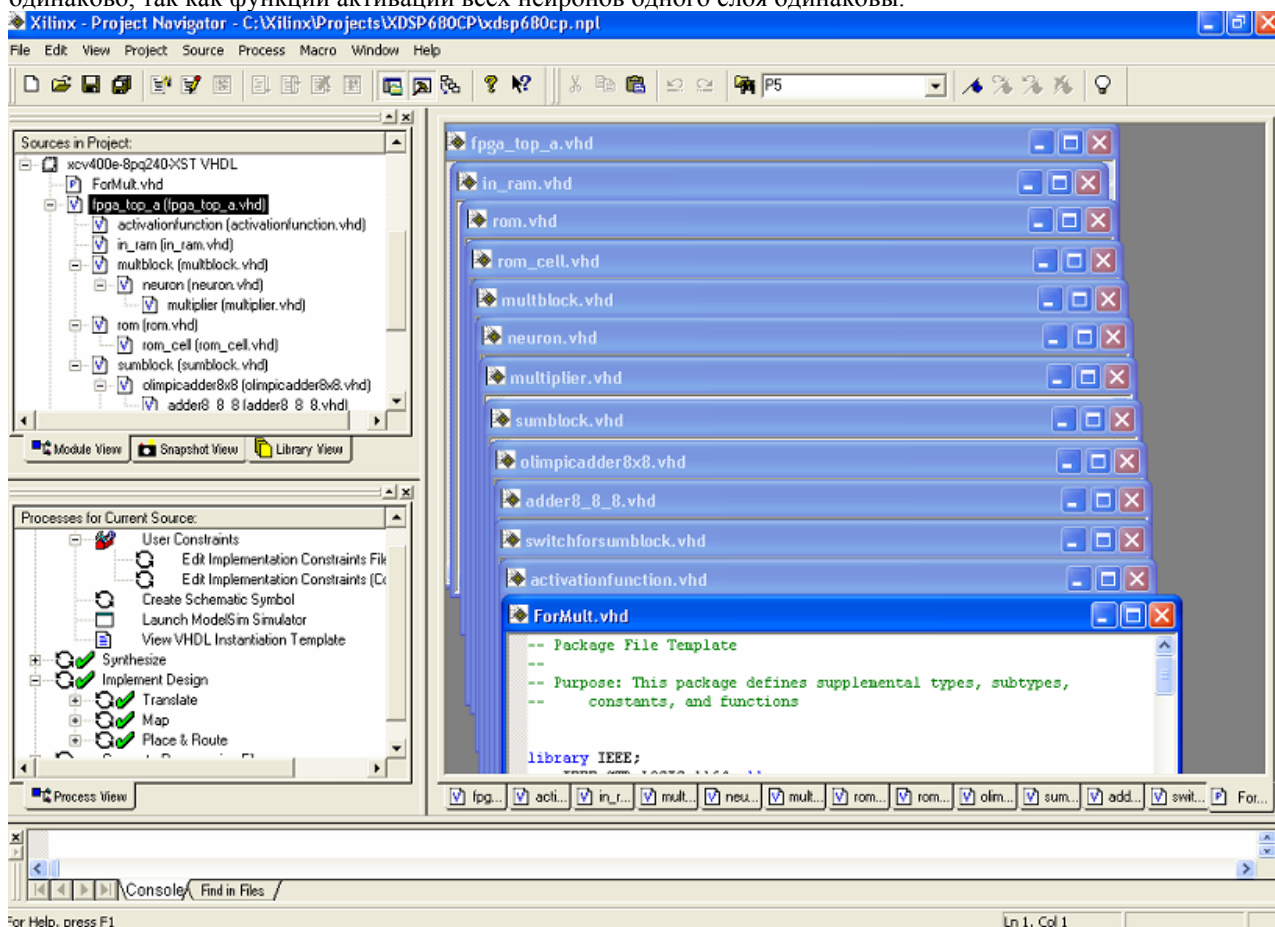


Рисунок 1. Окно проекта фрагмента нейронной сети

Разработка фрагмента в САПР Xilinx ISE 5.2.

Разработка проводилась в системе автоматического проектирования ISE 5.2 на языке VHDL (см. Рис. 1). Структура фрагмента программно разбита на основные блоки, что позволяет тестировать как отдельные компоненты, так и весь проект в целом. Временная и топологическая оптимизация проекта производилась с помощью встроенных в САПР инструментов: Xilinx Constraints Editor (временная оптимизация) и Xilinx FloorPlanner (топологическая оптимизация), а также с использованием присвоения временных и топологических атрибутов в VHDL коде компонент. Отслеживание наиболее критических задержек в схеме осуществлялось с помощью встроенного средства Timing Analyzer. С помощью этих инструментов частота работы всей схемы была увеличена примерно на 20%. Ниже на рисунке 2 приведено окно проекта в Project Navigator.

Продолжение разработок

В ближайшее время предполагается провести более детальные временные и топологические оптимизации проекта с целью повышения быстродействия фрагмента. Необходима также разработка специализированных библиотечных нейроэлементов (сумматоров, блоков функции активации), с помощью

которых предполагается увеличить производительность и функциональные возможности нейросетевого устройства. Планируется также реализация контура обучения на кристалле, для чего возможно введение режима, в котором бы фрагмент использовался в качестве фрагмента контура обучения. Однако, это является темой отдельного исследования.

На следующих стадиях исследований планируется сделать шаг в сторону реализации клеточных нейрончиков, нейрончиков с частотно-импульсной модуляцией, специализированных нейрончиков. В перспективе - выход на кристалл с интеграцией в 10 млн. системных вентилей.

Работа финансируется по гранту РФФИ (номер проекта: 03-07 90 159).

Выводы

Разработан прототип каскадируемого фрагмента нейронной сети с прямыми последовательными связями с возможностью выбора быстрых/медленных весовых коэффициентов или входные признаки. Максимальная тактовая частота работы схемы 90 МГц. Время вычисления выходного вектора по входному: 70 нс.

Разработанный фрагмент используется при решении отдельных практических задач, а также будет служить основой будущих нейросетевых систем, а освоенная методика проектирования в САПР ISE 5.2 от Xilinx послужит фундаментом для реализации нейросетевых алгоритмов, разрабатываемых в НЦН и требующих аппаратной поддержки.

Литература

1. L.M. Reyneri, "Implementation Issues of Neuro-Fuzzy Hardware: Going Towards HW/SW Design", IEEE Transactions on Neural Networks, Vol. 14, №1, January 2003, pp. 176-194.
2. Реализация нейронных сетей на ПЛИС Xilinx. Воронеж: СКАН Инжиниринг Телеком. 2000. 32 стр.
3. Галушкин А.И., Кирсанов Д.В., Цифровые нейрончики (специализированные цифровые БИС для нейрокомпьютеров)// Зарубежная радиоэлектроника 1999 №1, стр. 17-37.
4. The programmable logic databook, Xilinx Inc., 1999