И.И. Левин

Интеллектуальные многопроцессорные системы. Опыт создания и применения

Аннотация

«НИЦ супер-ЭВМ и нейрокомпьютеров» (г. Таганрог) разрабатывает и серийно производит целый ряд интеллектуальных многопроцессорных проблемно-ориентированных вычислительных комплексов на основе программируемых логических интегральных схем (ПЛИС). Последней разработкой является универсальная реконфигурируемая вычислительная система (РВС) «Арктур», предназначенная для решения задач различных предметных областей: математической физики, цифровой обработки сигналов, а также задач искусственного интеллекта.

В настоящее время вычислительная сложность задач, использующих нейросетевые технологии, постоянно растёт, поскольку непрерывно увеличивается объём анализируемых данных, а также ужесточаются требования к скорости и качеству их обработки. Нейронные сети, обрабатывая изображения, текст или звук, демонстрируют высокую точность, если входные данные коррелируют с данными, на которых выполнялось обучение, иначе качество решения задач значительно снижается. При этом скорость появления новой необработанной на этапе обучения информации может быть достаточно высокой, в связи с чем возникает необходимость постоянного переобучения нейросетей. Соответственно, если обучение нейросетей в обычных условиях выполняется в течение нескольких месяцев, то переобучение необходимо выполнить за приемлемое время (несколько дней или даже часов) вне зависимости от объема данных, что возможно только при кратном увеличении реальной производительности интеллектуальной системы. В то же время системы, построенные на графических процессорах (GPU), при линейном масштабировании аппаратных затрат, несмотря на линейный рост декларируемой пиковой производительности, обеспечивают в лучшем случае логарифмический рост скорости машинного обучения.

В качестве альтернативы для переобучения нейронных сетей за приемлемое время вне зависимости от объема данных с требуемым качеством можно использовать РВС на базе ПЛИС. В этом плане наиболее перспективной аппаратной платформой являются реконфигурируемые вычислительные блоки (РВБ) «Арктур». В данных РВБ реализован ряд прорывных технических решений, которые позволяют реализовывать сложные задачи в едином вычислительном контуре и без разрывов выполнять вычисления за счет мощной системы информационного обмена, обеспечивая при этом необходимый уровень энергопитания и охлаждения.

Конструктив РВБ «Арктур» 3U 19” содержит 16 вертикально расположенных на кросс-плате вычислительных модулей (ВМ) по шесть ПЛИС XCVU37P фирмы Xilinx семейства UltraScale+ в каждом. РВБ «Арктур» обладает уникальной плотностью компоновки вычислительного ресурса – 96 ПЛИС высокой степени интеграции (всего более 270 млн логических ячеек) в одном блоке. Вычислительные модули расположены на кросс-плате, с помощью которой обеспечивается межмодульное информационное и управляющее взаимодействие. Уникальная кросс-плата обеспечивает переход между средами (жидкость-воздух). Внешняя область кросс-платы – различные интерфейсы для подключения периферийных устройств и 144 оптических информационных канала, используемых для межблочного информационного обмена.

В РВБ применяются ПЛИС XCVU37P фирмы Xilinx семейства UltraScale+ из линейки «HBM» (High Bandwidth Memory), в корпус которых встроены по два модуля DDR

памяти общим объемом 8 Гбайт, обеспечивающих многоканальный доступ с пиковой пропускной способностью до 460 ГБ/с. Применение данной технологии позволит существенно расширить возможности по рациональному использованию памяти в прикладных программах для решения различных задач.

РВБ «Арктур» обеспечивает уникальную реальную производительность благодаря мощной подсистеме информационного обмена, которая представляет собой множественные каналы связи между ПЛИС в пределах платы, а также между ПЛИС соседних плат. Между парой ПЛИС реализованы 24 дифференциальные линии со скоростью передачи данных до 24 Гбит/с в пределах ВМ и 24 дифференциальные линии со скоростью передачи данных до 12 Гбит/с между модулями (платами). Общая пропускная способность каналов связи ВМ – 15,6 Тбит/с, в том числе между ВМ – 9 Тбит/с.

При построении вычислительных комплексов возможна организация информационного взаимодействия между РВБ через оптические каналы. Межблочные оптические приемо-передатчики установлены на кросс-плате и обеспечивают пропускную способность до 4,5 Тбит/с.

Максимальная потребляемая мощность РВБ «Арктур» – 25 кВт. Для обеспечения охлаждения нагруженных электронных компонентов РВБ «Арктур» используется иммерсионная (погружная) система охлаждения, при которой все электронные модули блока непосредственно погружены в диэлектрический хладагент, обладающий высокой электрической прочностью и теплопроводностью, а также максимально возможной теплоемкостью при низкой вязкости.

Экспериментально подтверждена перспективность использования РВБ «Арктур» для решения различных вычислительно-трудоемких задач при использовании структурной парадигмы вычислений. Структурная парадигма позволяет организовать в процессе решения задачи конвейерную обработку данных, в том числе для таких задач как машинное или глубокое обучение. Вычислительный конвейер позволяет избежать накладных расходов на организацию управления вычислительным процессом, поскольку кроме непосредственной обработки данных в нём могут выполняться ключевые процедуры анализа результатов и формироваться решение о продолжении или завершении работы.

Проведённые исследования показали, что при решении задачи обучения нейросетевого классификатора изображений производительность одного ВМ «Арктур» соответствует производительности GPU Nvidia Tesla A100, а при увеличении числа ВМ в системе обеспечивается практически линейный рост производительности РВС. Кроме того, при равной производительности энергоэффективность РВС, созданной на базе РВБ «Арктур», будет в 2 раза выше, чем у системы Nvidia DGX, созданной на базе Tesla A100.