Использование средств GPGPU для ускорения процесса построения карт диспаратности

А.Н.Волкович

В работе рассматриваются проблемы увеличения производительности методов восстановления трехмерных моделей. Обсуждаются аспекты использования многоядерных и многопроцессорных систем. Также рассматривается архитектура новейших вычислительных систем, использующих GPU как многоядерных высокоскоростных вычислителей, а так же рассмотрены аспекты их использования для увеличения производительности методов стереообработки.

Введение

Построение объемной модели на основе стерео изображений традиционно была, и остается одним из наиболее актуальных направлений в развитии компьютерного зрения. Последние исследования в этой области значительно продвинули область знания в вопросах качества и адекватности построений. К сожалению на современном этапе исследований требования к производительности значительно превышают возможности элементной базы – алгоритмы стереовосстановления обычно требуют от нескольких секунд до нескольких минут машинного времени, для построения единственной карты диспаратности. Однако существует значительное количество актуальных приложений, таких как задачи навигации и виртуальной реальности, которые требуют построения карт диспаратности с частотой близкой или эквивалентной стандартному видео. Кроме того обработка больших изображений (таких как аэрофотоснимки и т.п.) существующими методами требуют неприемлемо больших временных затрат.

1. Использование параллельных систем с целью увеличения эффективности методов обработки

На сегодняшний день достигнут предел увеличения вычислительной мощности одноядерных вычислительных процессоров, базирующийся на увеличении тактовой частоты и архитектурных инновациях. Для решения данных задач требуется концентрация вычислительных мощностей, а так же решение задач оптимизации вычислений. Ведущие изготовители микроэлектронных компонентов для сохранения тенденций роста производительности, переходят на разработку многоядерных процессоров с новой архитектурой, обеспечивающих распараллеливание обработки данных. Появление многоядерных процессоров является качественным скачком на пути создания эффективных супервычислителей, обладающих существенно более высокими показателями производительность/ стоимость, по сравнению с существующими высокопроизводительными системами вычислений на базе суперЭВМ и кластерных систем. Использование многоядерных процессоров предоставляет гибкие возможности в части варьирования конфигураций и масштабирования мощности вычислительных систем от персональных компьютеров, рабочих станций, серверов до кластерных систем.

Идея распараллеливания вычислений базируется на том, что большинство задач может быть разделено на набор меньших задач, которые могут быть решены одновременно.

Производя декомпозиционный анализ процесса построения объемных моделей на основе стерео изображений, оптимальным для параллельной реализации определен этап построения плотной карты диспаратности. Данный этап представляет собой совокупность однотипных опреаций по сравнению областей изображений и/или строк. Данный шаг возможно разделить на независимые блоки, которые будут обрабатываться в различных вычислительных узлах.

2. Построение карт диспаратности на «классических» параллельных системах

С целью определения целесообразности использования параллельных методов были проведены тестовые сравнения процесса построения плотных карт диспаратности в параллельном и последовательном режимах на многоядерных вычислительных системах.

Для проведения эксперимента производилась обработка двух пар изображений: стереоизображения реальной местности; калибровочная стереопара (тестовая пара стереоизображений куба в сфере).

Для получения информации об отношении скорости расчетов в последовательном и параллельном режимах произведено измерение времени выполнения последовательно реализованного алгоритма, а затем и параллельно реализованной версии. С целью получения наиболее объективных данных проведено десять замеров и вычислено среднее значение скорости выполнения.

Эксперименты показали, что параллельная реализация алгоритма позволяет увеличить его производительность на двуядерных или двуконвеерных системах на 30-40% по сравнению с их последовательной реализацией. Однако двуядерная архитектура не позволяет достичь скорость обработки близкой к частоте видео. Таким образом возникает необходимость использования более сложных параллельных вычислительных систем, главным недостатком которых являются: чрезвычайно высокая стоимость; сложность инфраструктуры; большие потери на межузловой коммуникации и громоздкость системы.

3. Высокопроизводительные вычислительные системы на базе GPU

На современном этапе в качестве альтернативы могут выступать системы использующие графические процессоры в качестве высокопроизводительных вычислителей.

Изначально GPU не могли использоваться для вычислений. Однако рост требований представляемых перед графическими ускорителями стал толчком в увеличению производительности и совершенствованию архитектуры.

В последнее время, в ходе своего развития, программируемые графические процессоры превратились в полноценную вычислительную единицу. Обладая многоядерной архитектурой и высокопропускной памятью, современные GPU представляют высокопроизводительные ресурсы и для графической и для неграфической обработки.

Процессор типа G80 (NVidia GeForce 8800) является многоядерным и многопоточным высокопроизводительным микропроцессором. По своим функциональным характеристикам и вычислительной мощности он может рассматриваться как графический процессор и как универсальный процессор для эффективной реализации неграфических приложений, требующих интенсивных вычислений. Как графический процессор он полностью реализует функции классического графического конвейера, устраняя недостатки предшествующих моделей GPU. Как универсальный процессор на операциях с плавающей точкой он превосходит по критерию производительность-стоимость все существующие традиционные и многоядерным CPU и GPU. Базовыми инновациями G80 являются:

унифицированная архитектура массива ядерных потоковых процессоров с плавающей точкой. пригодных для исполнения как графических конвейерных операций (геометрических преобразований,

обработки вершин и пикселов), реализуемых единообразно на потоковых процессорах, так и неграфических вычислений;

технология NVIDIA GigaThread Technology, широкомасштабная многопоточная архитектура, поддерживающая исполнение тысячи независимых, параллельно исполняемых тредов(потоков команд), обеспечивающая высокую эффективность обработки потоковых данных и использования вычислительного потенциала нового поколения многоядерных GPU. Для сравнения современные многоядерные CPU поддерживают работу на порядок и даже на два порядка меньше количества нитей.

При реализации на G80 неграфических вычислений наиболее значимыми компонентами являются массив унифицированных потоковых процессоров, доступные им ресурсы памяти, коммуникационные и управляющие средства. На рис.1 приведена блок-схема унифицированного массива процессоров G80.

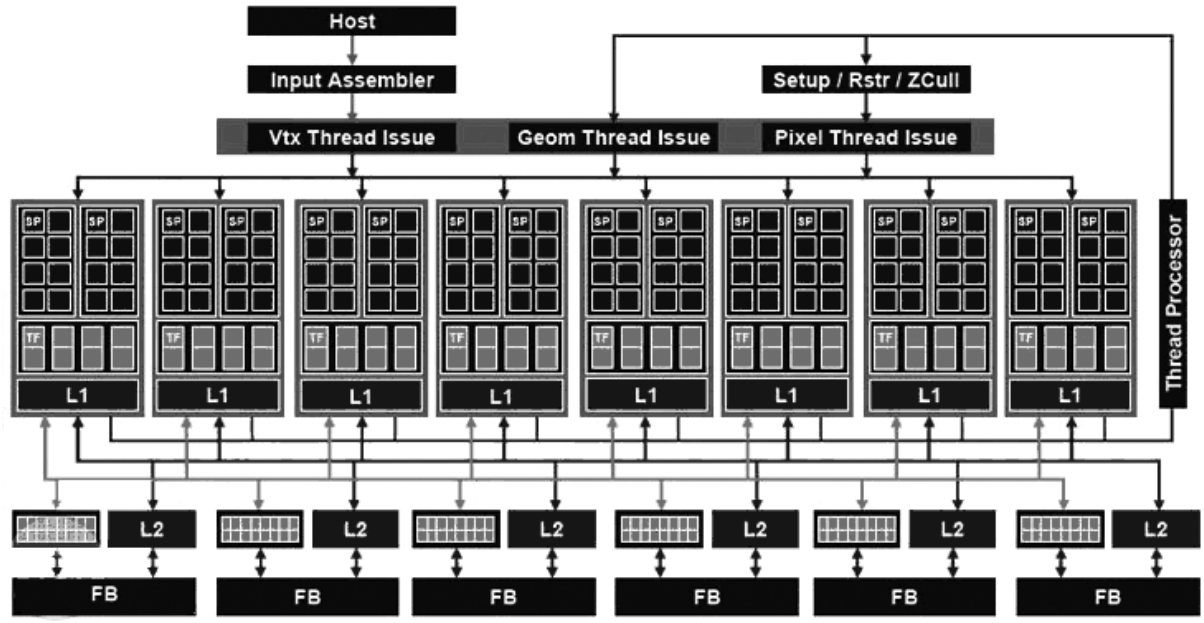

Рис. 1 Архитектура G80

G80 содержит 128 потоковых SP процессоров в виде 8 групп(кластеров, мультипроцессоров) по 16 SP. Они являются унифицированными скалярными процессорами с плавающей точкой, обрабатывающие не только графические, но и другие виды данных. Объединение SP в кластеры позволяет наиболее эффективно использовать ими аппаратные ресурсы G80: регистры 32бит, локальную разделяемую внутрикристальную память по 16 KB на кластер, разделяемую кэш память (64KB) для чтения констант из внешней памяти G80, разделяемую кэш память текстур. Разделяемые ресурсы памяти внутри кластеров позволяют обеспечить синхронизацию и коммуникацию междунитями (потоками команд), работающими внутри кластера.

G80 обладает мощной параллельной архитектурой. Каждый потоковый процессор, на основе механизмов управления работой нитей, способен динамически переназначаться для исполнения конвейерных графических или других вычислительных операций, обеспечивая таким образом пиковую загрузку ресурсов GPU и максимальную сбалансированную гибкость при обработке задач.

G80 содержит 128 потоковых SP процессоров в виде 8 групп(кластеров, мультипроцессоров) по 16 SP. Они являются унифицированными скалярными процессорами с плавающей точкой, обрабатывающие не только графические, но и другие виды данных. Объединение SP в кластеры позволяет наиболее эффективно использовать ими аппаратные ресурсы G80: регистры 32бит, локальную разделяемую внутрикристальную память по 16 KB на кластер, разделяемую кэш память (64KB) для чтения констант из внешней памяти G80, разделяемую кэш память текстур.

Разделяемые ресурсы памяти внутри кластеров позволяют обеспечить синхронизацию и коммуникацию между нитями (потоками команд), работающими внутри кластера. G80 обладает мощной параллельной архитектурой. Каждый потоковый процессор, на основе механизмов управления работой нитей, способен динамически переназначаться для исполнения конвейерных графических или других вычислительных операций, обеспечивая таким образом пиковую загрузку ресурсов GPU и максимальную сбалансированную гибкость при обработке задач.

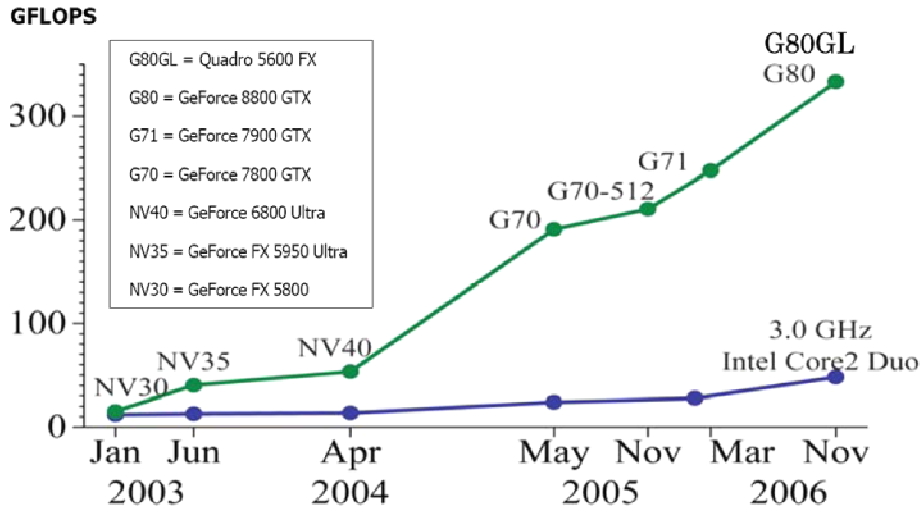

Рис. 2 Сравнение развития вычислительных мощностей GPU и CPU (по данным NVidia)

Тенденция роста вычислительной мощности GPU, проиллюстрированная на рисунке позволяет говорить о возможности использования систем для больших вычислений. В мировой практике уже существуют прецеденты использования вычислений на GPU, а также попытки сравнения с эффективностью фактических вычислений на CPU. Так Исследователи в Антверпенском университете (Бельгия) создан высокопроизводительный компьютер на базе четырех видеокарт NVIDIA GeForce 9800 GX2 (8 GPU). По результатам сравнения вычислительных мощностей объявлено его вычислительная эквивалентность кластеру из 300 ПК с Intel Core 2 Duo 2.4GHz.

4. Использование GPU для построения карт диспаратности

На сегодняшний день в мировой практике были предприняты попытки реализации алгоритмов построения карт диспаратности на GPU. Исследователями был реализован алгоритм динамического программирования с коррекцией карты за счет второго прохода. В свою очередь пред и пост обработки производились на центральном процессоре с тактовой частотой 3 GHz.

Данный алгоритм был выбран по причине его хорошей распараллеливаемости в отличие от глобальных алгоритмов и одновременной возможности получения удовлетворительных результатов. Результаты работы показывают возможность построения 16 уровневой карты диспаратности для рисунка 320x240 точек с частотой 42 кадра в секунду (0.023 секунды на обработку одного кадра).

Также было произведено сравнение вычислений карт диспаратности по указанному алгоритму для изображений различного разрешения и с различным количеством уровней:

Таблица 1. Сравнительные характеристики вычислений диспаратности на CPU и GPU

Размер	Уровней	GPU	CPU
640x480	16	0.079	15.2
	32	0.131	29.1
	48	0.183	42.4
320x240	16	0.023	3.61
	32	0.042	6.78
	48	0.054	9.63

Базируясь на приведенных данных возможно предположить, что использование одного GPU для построения карты диспаратности изображений SD-разрешения (720x576) позволит достичь частоты около восьми кадров в секунду, а HD-разрешения (1920x1080) порядка одного.

Заключение

Анализируя вычислительные мощности графических ускорителей заявляемые производителями, базируясь на имеющихся в мировой практике опытах использования GPU в качестве высокопроизводительных систем, а также на заявляемой возможности использования нескольких GPU-вычислителей, возможно создать систему оперативного восстановления объемных моделей на основе стереизображений. Текущая производительность GPU-систем и возможность использования нескольких ускорителей в одной системе позволяют судить о возможности достичь скорость обработки близкую или эквивалентную частоте стандартного видео для изображений SDTV и HDTV разрешений.

Кроме того использование GPU-вычислений позволит ускорить построение карт диспаратности при помощи алгоритмов, которые не предполагают быстрого выполнения, но обладают, на современном этапе критически долгое выполнение, затрудняющее их использование.

В свою очередь значительно меньшая стоимость оборудования, простота монтажа, а также компактность размещения позволит использовать данные системы в качестве программно-аппаратных комплексов стереовидения.

Литература

1. Christopher Zach, David Gallup, Jan-Michael Frahm. Fast Gain-Adaptive KLT Tracking on the GPU . University of North Carolina Chapel Hill, NC 2008 7 p.
2. Liang Wang, Miao Liao, Minglun Gong, Ruigang Yang, David Nister. High-quality Real-time Stereo using Adaptive Cost Aggregation and Dynamic Programming. Abstracts of University of Kentucky 2008, 8 p.
3. NVIDIA CUDA Homepage. http://www.nvidia.ru/object/cuda.html
4. Воробьев А., Медведев А. NVIDIA GeForce 8800 GTX (G80). http://www.ixbt.com/video2/g80-part1.shtml
5. Аляутдинов М.А., Троепольская Г.В. Использование современных многоядерных процессоров в нейрокомпьютерах для решения задач математической физики Нейрокомпьютеры: разработка, применение, № 9, 2007 г. С 71-80.
6. A.N. Volkovich, D.V. Zhuk, A.V. Tuzikov. Construction of three-dimensional models from images using parallel systems. Proceedings of the 9th International Conference “Pattern Recognition and Information Processing”, 22-24 May, 2007, Minsk, Belarus, vol. 2, 232-235.
7. А.Н. Волкович, Д.В. Жук, А.В. Тузиков. Методы построения трехмерных моделей местности и их реализация для параллельных систем. Доклады 5-й международной конференции "Обработка информации и управление в чрезвычайных и экстремальных ситуациях", 24 – 26 октября, Минск, Беларусь, 2006, 100-104.