СТРАТЕГИЯ АРХИТЕКТУРНОГО ОПЕРЕЖЕНИЯ ПРИ РАЗРАБОТКЕ ОДНОКРИСТАЛЬНОГО МНОГОЯДЕРНОГО КОМПЬЮТЕРА-УСКОРИТЕЛЯ ДЛЯ ЗАДАЧ С МАССОВЫМ ПАРАЛЛЕЛИЗМОМ

С.Е. Артамонов, Ю.С. Затуливетер, В.А. Козлов, Е.А. Фищенко

1. Введение

Достижение значимых позиций на мировом компьютерном рынке – одна из приоритетных задач возрождения отечественного компьютеростроения. На первый взгляд она кажется нерешаемой, но, как далее показывается, на современном этапе развития мирового компьютеростроения вступают в действие фундаментальные рыночные факторы, которые требуют изменения базовых моделей развития индустрии массового производства компьютеров и программ.

Вступление в ВТО может способствовать снижению дискриминационных барьеров на путях доступа к современным технологиям производства СБИС и выхода на мировой высокотехнологичный компьютерный рынок с новыми разработками, но это предъявляет высочайшие требования к уровням их конкурентоспособности. Особое значение обретают проекты, направленные на формирование новых рыночных ниш высокотиражной наукоёмкой продукции в области однокристальных многопроцессорных компьютеров, которые опираются на превосходящие научные заделы и опыт практических наработок и их широкого промышленного применения.

В области высокопроизводительных и массовых компьютеров конкурентоспособность наукоёмких ноу-хау может достигаться только в рамках сложившегося на мировом рынке разделения труда с привлечением современных полупроводниковых технологий массового производства интегральных схем глубокого нанометрового диапазона 40-28-20-14-10нм.

2. Компьютерный рынок в преддверии кардинальных перемен

Компьютерный рынок – не только одна из наиболее доходных и динамично развивающихся сфер мирового рынка. Массовые компьютеры, связанные глобальными сетями, проникают во все сферы жизнедеятельности. Компьютеростроение стало ключевым фактором социально-экономического развития. Качество жизни прочно связано с уровнем технического развития компьютерных технологий, которые становятся доступными всё большему числу людей.

В первых поколениях классических микропроцессоров, которые, как известно, реализуют универсальную модель последовательных вычислений Дж. фон Неймана, повышение производительности осуществлялось не только за счёт увеличения рабочей частоты, но и на структурно-архитектурном уровне путём увеличения разрядности машинных слов и аппаратного распараллеливания алгоритмов выполнения арифметических операций.

Рис.1. Структурное насыщение микропроцессорных архитектур

Производительность в расчёте на транзистор [1] при этом быстро росла пропорционально числу транзисторов, см. рис.1. Это говорит о том, что классическая модель последовательного счёта на структурно-архитектурном уровне имела скрытые резервы внутреннего параллелизма (разрядность, параллелизм арифметических операций, специализированных устройств типа MMX, кэширование потоков данных и команд, конвейеризация операций и команд, предсказание условных переходов и др.).

Из рис. 1 видно, что максимальное значение этого показателя эффективности микропроцессорных архитектур достигнуто на первом Пентиуме (Pentium I, 3.1 млн. транзисторов). При этом изначально ограниченные внутренние резервы параллелизма последовательной модели были, в значительной мере, исчерпаны. Можно утверждать, что в диапазоне 3-25 млн. транзисторов на кристалле было достигнуто структурное насыщение микропроцессорных архитектур [1].

Компьютерная индустрия в середине 90-х вошла в начальную фазу кризиса классической модели последовательного счета. Последующие поколения одноядерных микропроцессоров в связи с исчерпанием скрытых резервов параллелизма модели последовательных вычислений быстро утрачивали компоненту наращивания производительности за счёт сверхбыстрого роста количества транзисторов на кристалле по закону Мура (удвоение за каждые 1.8-2 года). Темпы роста производительности обеспечивались, главным образом, за

счёт увеличения рабочей частоты, связанного с уменьшением размеров транзисторов. Ценой роста рабочей частоты стало непропорционально высокое потребление энергии.

Налицо растущее обесценивание возможностей прогрессирующих полупроводниковых технологий – главного двигателя компьютерного прогресса. Контраргументация о резком снижении себестоимости каждого транзистора с ростом степени интеграции не отменяет архитектурного кризиса микропроцессоров, а лишь объясняет конъюнктурную рентабельность массового производства микропроцессорных кристаллов, заполненных по большей части «безработными» транзисторами.

Компьютеростроение и компьютерный рынок приблизились к критической фазе своего развития, когда для дальнейшего прогресса необходимо принципиальное обновление компьютерных первооснов. Опережающего прогресса полупроводниковых технологий уже недостаточно.

Компьютерный рынок уже вступил в период кардинальных структурных перемен, которые неизбежно приведут к смене поколений лидеров компьютерной индустрии, существенному пересмотру приоритетов в инвестиционных процессах. Преимуществ нынешних лидеров компьютерного рынка, основанных на прежних достижениях, уже не достаточно для открытия и обустройства новых сфер массового влияния. Масштабы необходимых структурных перемен в большинстве случаев превысят адаптационные возможности лидеров, которые сильно ограничены длинными шлейфами прежних обязательств перед миллиардами клиентов. На этапе смены системообразующих принципов строить новое будущее и одновременно тащить растущий груз прошлого становится нереальным. Будущее вступает в противоречие с прошлым и требует новых правил функционирования компьютерного рынка.

3. Причины и проявления внутрикомпьютерного кризиса

Суть глубинного внутрикомпьютерного кризиса в следующем. В основе современного рынка массовых компьютеров и программ лежат два крупнейших достижения 20-го века:

- классическая модель универсальных последовательных вычислений – модель Дж. фон Неймана, которая дала старт компьютерной эпохи (конец 40-х прошлого столетия);
- микроэлектронные технологии массового производства полупроводниковых интегральных схем, которые компьютерную эпоху сделали достоянием всего человечества.

Модель фон Неймана – это свод простых логико-технических правил автоматического выполнения любых алгоритмов в последовательном режиме "команда-за-командой". Они были положены в основу первых универсальных компьютеров. Они стали основой микропроцессорной революции и до сих пор остаются единым и единственным логическим "стандартом" индустрии массовых компьютеров и программ. Главное достоинство этой модели – относительная простота и эффективность машинной реализации универсальных вычислений.

В последовательной модели вычислений в каждый момент исполняется одна команда (операция), что позволяет называть её скалярной моделью вычислений. Вычислительный процесс выглядит как последовательная траектория точечных (скалярных) событий, каждое из которых представляет исполнение одной операции. Уникальное достоинство классической модели, при этом, состоит в том, что она на инженерном уровне предлагает логически простейший и, в то же время, практически эффективный механизм управления их реализацией. Поэтому именно она легла в основу микропроцессорной революции.

Технологии полупроводниковых интегральных схем дали в виде микропроцессорных кристаллов долгосрочную материальную основу для массовой реализации компактных и недорогих универсальных компьютеров, основанных на классической модели последовательных вычислений. Темпы компьютерного прогресса стали определяться сверхвысокими скоростями развития полупроводниковых технологий массового производства интегральных схем, которые выражаются известным законом Мура: "Количество транзисторов на кристалле удваивается каждые 1.8-2 года".

Определяющей тенденцией в развитии технологий интегральных схем является уменьшение размера транзисторов и толщины проводников. Их уменьшение даёт двойной эффект. Во-первых, чем меньше транзисторы, тем быстрее они могут срабатывать и чем меньше расстояние между ними, тем скорее обмены сигналами, что позволяет увеличивать рабочие частоты, а значит и вычислительную производительность. Во-вторых, увеличивается плотность размещения транзисторов на поверхности кристалла. Быстрое увеличение количества транзисторов на кристалле (квадратичный рост с уменьшением линейных размеров транзисторов) открывает возможности для аппаратного наращивания параллелизма вычислительных устройств, что также служит ключевым фактором повышения производительности.

На начальном этапе развития микропроцессоров повышение производительности осуществлялось одновременно как за счёт увеличения рабочей частоты, так и на структурно-архитектурном уровне наращивания параллелизма. Во второй половине 90-х (после появления микропроцессора Pentium I) резервы аппаратного реализуемого параллелизма классической модели последовательного счёта были исчерпаны. Сохраняя на уровне программистов главный принцип управления – "команда-за-командой", конвейерный параллелизм (суперскалярность) также ограничен числом операций, задаваемых командой (считывание команды, операндов, вычисление, запись результата). По мере исчерпания скрытого параллелизма модели последовательного счета продолжающийся по закону Мура рост числа транзисторов в рамках классической модели перестал трансформироваться в пропорциональное прибавление производительности (рис.1). Снижение

удельной производительности каждого транзистора микропроцессоров стало самым ранним признаком проявлением начальной фазы внутрикомпьютерного кризиса, получившего название структурного насыщение микропроцессорной архитектуры [1].

В течение последующих 10 лет, вплоть до середины 00-х кризис классической модели последовательных вычислений и микропроцессорных архитектур развивался в латентной форме. Компьютерная индустрия, развиваясь за счёт расширения сфер применения массовых компьютерных устройств, игнорировала нарастающие проявления структурного кризиса. Исчерпав резервы скрытого от программистов параллелизма классической модели, она стала довольствоваться увеличением производительности новых поколений одноядерных микропроцессоров лишь за счёт наращивания рабочей частоты. А сверхбыстрый рост "избыточных" транзисторов успешно "прятали" в многоуровневых кэшах.

В соответствии с ростом рабочих частот (1-4 ГГц), достигавшемся посредством уменьшения размера транзисторов, росло быстродействие, которое позволяло ускорять выполнение всех последовательных программ (в одинаковой мере новых и старых) и обеспечивать коммерческий успех дальнейшей смены поколений микропроцессоров на рынке. Активная реклама обходила вниманием катастрофическое снижение удельной производительности в расчёте на транзистор, которое осталось совершенно неведомым как для бизнеса и потребителя, так и инвесторов.

Однако беззаботный период "гладкого" развития в рамках классической модели и "лёгких" прибылей к середине 00-х завершился. С освоением технологий 90-65нм количество транзисторов на кристалле приблизилось к миллиарду. Рабочая частота достигла 4Гц и более, но при этом энергопотребление транзисторов увеличилось настолько, что воздушное охлаждение престаёт справляться с отводом тепла. Дальнейшее наращивание производительности массовых микропроцессоров за счёт увеличения частоты стало экономически неоправданным. Кроме того, огромный и сверхбыстро растущий сектор мобильных устройств особо остро нуждается в энергоэффективных методах повышения производительности.

Тепловой барьер лишил возможности повышения производительности за счёт увеличения частоты. Стратегическая ловушка структурного насыщения микропроцессорных архитектур захлопнулась окончательно.

Чтобы хоть как-то наращивать производительность, промышленность вынужденно ответила "многоядерными" кристаллами. Стали массово реализовываться 2-4-8-ядерные микропроцессоры с простейшей архитектурой – общей для всех ядер памятью. Однако они решили проблемы наращивания производительности лишь в незначительной степени. Если 2-х ядерные давали рост производительности почти в 2 раза, то каждое новое ядро добавляло производительности все меньше и меньше. В большинстве применений максимальное количество ядер не превышает 4. Такой "несколько ядерный" параллелизм реализуется неэффективно, т.к. в этом случае проблема «узкого горла» общей для всех ядер памяти в принципе не позволяет существенно увеличивать количество ядер.

Приведём наглядный пример, который показывает масштабы кризиса структурного насыщения микропроцессорных архитектур. Технология вчерашнего дня 45нм предоставляет на кристалле около 1млрд. транзисторов. На таком кристалле можно было бы разместить более 300 ядер с эффективной архитектурой Pentium I (3.1 млн. транзисторов). Однако из-за "узкого горла" памяти все 300 ядер будут работать медленнее, чем каждое ядро в отдельности. Отсюда видна "цена" архитектурного кризиса классической модели вычислений – снижение коэффициента полезного использования транзисторов в сотни раз.

Латентный период развития кризиса структурного насыщения "внезапно" для подавляющего большинства закончился. Универсальные многоядерные микропроцессоры, которые выпускаются массовыми тиражами с середины 00-х, не устранили проблемы структурного насыщения микропроцессорных архитектур, а лишь вывели её на всеобщее обозрение.

Отсутствие долгосрочных перспектив развития такой "многоядерности" выражается в следующем:

- ограниченный параллелизм в обменах между ядрами и памятью не позволяет наращивать производительность пропорционально числу ядер, рост производительности прекращается уже на нескольких ядрах;
- прямым следствием структурного насыщение микропроцессорных архитектур стал кризис индустриальных технологий программирования, в основе которых десятилетиями оставалась модель последовательных вычислений.

Массовое производство столь бесперспективной "всего-лишь-несколько-ядерной" архитектуры стало молчаливо-безличным признанием компьютерной индустрией того свершившегося факта, что системообразующий потенциал классической модели последовательного счёта и реализующих её одноядерных микропроцессорных архитектур исчерпан, а её полноценной постнеймановской замены всё ещё нет.

4. Индустриальные проблемы внутрикомпьютерного кризиса

Начальный, относительно гладкий, потому сравнительно лёгкий, этап тридцатилетнего развития массового компьютеростроения в рамках единой и простейшей модели последовательного счёта закончился. Бизнесом собрано почти всё, что скрывалось в тонком поверхностном слое компьютерного прогресса, на который распространяется действие классической модели последовательного счёта. Более глубокие пласты тотальной компьютеризации на многие порядки более прибыльны, но требуют новых моделей и инструментов. Но их уже невозможно собрать в рамках классической компьютерной аксиоматики.

С опозданием на десятилетие компьютерный рынок вынужденно приступает к поиску индустриальных моделей параллельных вычислений, ориентированных на реализацию посредством существенно многопроцессорных архитектур (с количеством ядер сотни, тысячи, десятки тысяч и более ядер), которые оказались бы способными стать основой формирования нового, уже постнеймановского и постмикропроцессорного мэйнстрима в массовом производстве компьютеров и программ.

Для индустриализации нового мэйнстрима на смену классической – простейшей, ввиду своей скалярности, модели универсального счёта, должны прийти другие модели и архитектуры, которые регламентируют универсальные вычисления в пространстве параллельных вычислительных процессов. Фундаментальное отличие от классических последовательных траекторий скалярных вычислительных событий в том, что в пространстве параллельных процессов доминирует новое измерение, характеризующее множественность вычислительных событий, обозначаемая термином "параллелизм". Параллелизм предполагает, что в каждый момент одновременно исполняется множество операций (команд). Чем больше таких операций способна предоставить вычислительная задача, тем более высокими уровнями параллелизма должны располагать компьютерные архитектуры.

Модели параллельных вычислений и соответствующие им архитектуры начали активно разрабатываться ещё в 60-е годы. С тех пор их наработано огромное количество. Но почти все они создавались вне индустриальных требований практической реализуемости компьютеров и программ в массовых тиражах и, поэтому, не могут составить основу для нового индустриального мэйнстрима.

Массовый компьютерный рынок с одной стороны крайне динамичен в поисках и охвате новых сфер сбыта, а с другой – консервативен, поскольку отягощён колоссальной инерцией сопровождения наработанных продуктов. Целостное развитие уходящего мэйнстрима компьютеростроения обеспечивалось системообразующим потенциалом классической модели последовательных вычислений. Теперь, когда её потенциал вычерпан почти до дна, из-за отсутствия общей базовой модели параллельных вычислений, адекватной требованиям массового производства компьютеров и программ, резко вырастают риски утраты целостности компьютерного рынка. Упущенное десятилетие, в течение которого внутрикомпьютерный кризис из латентной фазы перерос в запущенную и открытую, требует экстренных мер по предотвращению стихийного распада устаревающих рыночных структур, уже не отвечающих новым вызовам.

Десятилетней задержке с активными поисками новом базовой модели трудно найти оправдание, поскольку характер моделей параллельных вычислений и соответствующих компьютерных архитектур по отношению к классике кардинально меняется. При этом научные проблемы достижения *практической эффективности* таких моделей и архитектур, отвечающие *индустриальным требованиям массового производства компьютеров и программ*, *крайне сложны* и для их решения требуется определённое время. В них доминируют уже не столько задачи совершенствования собственно технологий производства аппаратных и программных средств, сколько фундаментальные проблемы комбинаторной сложности, связанные с поиском эффективных многопроцессорных структур в условиях *математической многовариантности* структурно-динамического многообразия параллельных вычислений.

В долгосрочной перспективе продвижения на рынок моделей параллельных вычислений и архитектур, отвечающих требованиям массового производства компьютеров и программ, необходимы новые высокоэффективные многопроцессорные архитектуры и технологии их индустриального программирования. *Синергетическим эффектом* стартовой точки лавинного роста может стать любое принципиальное продвижение в части однокристальных компьютеров с высокопараллельными многопроцессорными архитектурами общего назначения.

5. Архитектурный потенциал опережения

Высокопараллельные многопроцессорные архитектуры – это та область знаний, в которых уровень конкурентоспособности изделий не может основываться только на превосходстве полупроводниковых интегральных технологий. Проблема в том, что среди огромного количества вариантов допустимого множества параллельных архитектур и процессов в них только малое их число обладает достаточной эффективностью. Необходимость поиска эффективных решений становится главной проблемой индустриальных параллельных вычислений, архитектур и технологий программирования. Лишь ничтожная часть этих структурных решений имеет практическую значимость. И пока никто не предложил регулярных методов как эти решения находить.

Отсюда особая роль наиболее удачных достижений в части эффективных параллельных архитектур. Их не так много. Одним из весьма удачных решений, является многопроцессорная архитектура отечественного компьютера ПС-2000 [2], ориентированного на решение широких классов задач с массовым параллелизмом. Это был первый в мире широкодоступный суперкомпьютер, который выпускался большой промышленной серией. Его оригинальная, масштабируемая и комплексируемая архитектура доказала свою высокую вычислительную эффективность и экономическую рентабельность во многих сферах промышленной обработки данных, а также в научно-инженерных расчётах, в системах обработки потоков данных в реального времени, а также в системах обработки больших объёмов данных. По признанию зарубежных специалистов ПС-2000 был одним из наиболее успешных отечественных суперкомпьютеров и одним из первых в мире многопроцессорных компьютеров, который выпускался серийно и использовался в многочисленных и разнообразных сферах [3,4].

Архитектуры современных многопроцессорных систем, являющиеся аналогами по сферам применимости, – однокристальные ускорители GPGPU (General Purpose Graphics Processing Units) изначально создавались и балансировались под узкие классы задач. Так, производители nVIDIA и ATI (AMD) начинали осваивать массовые рынки с изготовления графических ускорителей (видеоплат) для ПК, архитектура которых оптимизировалась под ограниченный набор алгоритмов обработки видеографики. В ходе их развития к середине 00-х сформировался новый рыночный класс однокристальных многопроцессорных компьютеров – GPU (Graphics Processing Units). Примерно в это же время IBM вывела на рынок однокристальный ускоритель Cell, с гибридной многопроцессорной архитектурой, заточенный под игровые приставки.

Так, спустя 20 лет после индустриальной премьеры ПС-2000, состоялся новый выход высокопараллельный многопроцессорных архитектур в широкие сферы применений. Конечно, масштабы промышленного тиражирования продуктов совершенно иные. ПС-2000 – выпускался большой промышленной серией в несколько сотен вычислительных комплексов. Продукция GPU и Cell – миллионные тиражи. Но это не удивительно. Современный компьютерный рынок – совершенно иная элементная база, иные масштабы сфер потребления.

Надо отметить особый вклад узкопрофильных многопроцессорных ускорителей GPU и Cell, с которыми была пройдена наиболее рисковая часть массовой рыночной инновации многопроцессорных архитектур. Уверенный бизнес ускорителей этих классов доказал, что высокопараллельные многопроцессорные архитектуры, несмотря на гораздо более высокую, в сравнении с классическими однопроцессорными компьютерами, сложность программирования, нашли своё место на компьютерном рынке и неуклонно расширяют своё присутствие.

Далее, по мере наполнения массового рынка узкопрофильных многопроцессорных ускорителей, стал формироваться спрос на недорогие ускорители для более широких классов задач. К ним относятся, прежде всего, программируемые ускорители для научно-инженерных задач высокой вычислительной сложности. Такие системы отличаются от узкопрофильных прежде всего тем, что имеют открытые для пользователей системы программирования, которые позволяют им решать свои задачи.

Так на базе рынка узкопрофильных ускорителей GPU сформировался следующий потребительский класс однокристальных многопроцессорных ускорителей общего назначения – General Purpose (GP): GPGPU.

Значительная часть ускорителей GPGPU применяется в настольных суперкомпьютерах и вычислительных серверах, центрах обработки данных, которые всё шире используются в исследовательских и проектных организациях, университетах, медицинских центрах. Такие ускорители применяются и в топовых суперкомпьютерах, обеспечивая освоение диапазона производительности более 1-20 Пфлопс и более.

В настоящее время большая часть ускорителей класса GPGPU представлена расширением архитектур, изначально специализированных на обработку графики. Одним из лидеров этого класса являются ускорители фирмы nVIDIA – Fermi (40нм, 512 ядер, пиковая производительность ~1 Тфлопс) и Kepler (28нм, 1536 ядер, ~3 Тфлопс). Однако, следует признать, что системы класса GPGPU, получаемые трансформацией из узкопрофильной многопроцессорной архитектуры GPU, изначально ориентированной на ограниченный набор алгоритмов графической видео обработки, не могут рассматриваться как окончательные решения, обеспечивающие высокую вычислительную эффективные на разнообразных классах задач. В частности одним из слабых звеньев такой архитектуры остаётся узкое горло к общей оперативной памяти, что является серьёзным ограничением в классах задач с произвольным доступом к массовым данным. Всё это снижает стратегический потенциал конкурентоспособности подходов к развитию систем класса GPGPU.

Пройдя высокорисковую стадию начального формирования на мировом рынке, потребительский класс ускорителей GPGPU уже перешёл в устойчивую стадию развития. Его характерные особенности:

- наличие устойчиво растущего спроса в разнообразных сферах применения и долгосрочных тенденций к их расширению;
- число ведущих производителей невелико (в пределах десятки: среди них nVIDIA, AMD, IBM&Sony&Toshiba, Intel);
- отсутствие однозначно лидирующей архитектуры.

Несмотря на различие представленных на рынке архитектурных подходов, ни один из них не имеет существенного превосходства по всей совокупности значимых потребительских характеристик. Так, вычислительная эффективность, измеряемая коэффициентом полезного использования ядер, в разных классах задач даёт значительный разброс реальной производительности. Отсюда разделение сфер влияния по классам задач: графические видеоплаты (nVIDIA, AMD), игровые приставки (IBM&Sony&Toshiba), научно-инженерные расчёты высокой вычислительной сложности (nVIDIA, AMD, Intel, IBM).

В рамках сложившихся сфер влияния конкуренция идёт не столько на уровне архитектурных принципов, сколько на уровне инженерно-конструкторских способов воплощения чипов в рамках своих устоявшихся архитектурных шаблонов. В нишах, где конкурируют разные игроки, вперёд на короткое время выходит чип ускорителя того производителя, который сумел опередить в темпах проектирования при переходе на очередное поколение полупроводниковой технологии.

Отсутствие явно лидирующей архитектуры свидетельствует об отсутствии качественно опережающего архитектурного решения, способного составить конкуренцию в различных сферах по большинству потребительских параметров.

В отличие от узкопрофильных многопроцессорных архитектур, заложенных в GPGPU, в основу высокопараллельной архитектуры ПС-2000 изначально положены *общие принципы* высокоэффективной обработки данных на широких классах задачах с массовым параллелизмом. Она обладает достаточным потенциалом своего развития для конкурентоспособного покрытия большинства классов задач с массовым параллелизмом, отдельные подклассы которых поделены между сегодняшними игроками.

С компьютерами класса GPGPU предстоит конкурировать новым поколениям ПС-2000. Принципиальное преимущество архитектуры ПС-2000 перед представленными на современном рынке однокристальными многопроцессорными ускорителями GPGPU, состоит в том, что функциональное ядро и архитектура ПС-2000 изначально разрабатывались и балансировались под широкие классы задач с существенно различающимися структурами. Как подтвердила обширная практика промышленных применений, именно это обеспечило высокий уровень универсализма на классах задач с массовым параллелизмом.

В то же время архитектура GPGPU nVIDIA и AMD продолжают развиваться как эклектичное расширение функций спецпроцессоров графической обработки видеоданных. Понятно, что заставить специализированные архитектуры эффективно решать задачи, структуры которых сильно отличаются от изначально реализованной, практически невозможно. Специализированное прошлое GPGPU тянет за собой излишнюю громоздкость, низкую эффективность, трудную аппаратную масштабируемость.

Архитектура ПС-2000М [5] обладает уникальным на фоне современных ускорителей свойством структурной масштабируемости, которое позволяет одновременно с увеличением количества вычислительных ядер пропорционально наращивать объём и пропускную способность встроенной распределённой памяти, произвольно адресуемой в каждом процессорном элементе (ядре), а также свойством комплексируемости как на внури-, так и на межкристальном- уровне. А это – принципиально более высокий потенциал и в использовании растущих возможностей полупроводниковых технологий, и в расширении классов эффективно решаемых задач.

Одна, вместо многих узкопрофильных, масштабируемая, комплексируемая, программно совместимая многопроцессорная архитектурная линейка, которая существенно превосходит GPGPU в эффективности (по совокупным параметрам реальная производительность в расчёте на транзистор*Ватт), даёт превосходство в охвате разнообразных классов задач. Возможность формирования однородного по аппаратно-программным средствам массового рынка высокопроизводительных компьютеров и программ даёт серьёзные преимущества в рентабельности наращивании тиражей, сопоставимых с тиражами традиционных многоядерных микропроцессоров.

Архитектура ПС-2000, доказавшая в ходе десятилетнего промышленного применения, эффективность своих принципов управления массовым параллелизмом [2], положена в основу нового проекта ПС-2000М [5] как фундамент для построения высококонкурентных однокристальных многопроцессорных компьютеров общего назначения класса GP (General Purpose).

Эффективность принципов управления и системы команд, реализованных в архитектуре ПС-2000, доказаны в ходе десятилетнего промышленного применения в разнообразных сферах промышленной обработки данных посредством сотен вычислительных комплексов на её основе.

Масштабируемая и программно комплексируемая MultiSIMD архитектура ПС-2000М не только в полной мере наследует преимущества архитектуры ПС-2000, но и дополнительно расширяет классы эффективно решаемых задач за счёт введения многозадачности и механизмов асинхронного (событийного) управления потоками данных, обеспечивающих динамическое распараллеливание в режимах программно реконфигурируемого макроконвейера [5].

Апробированная широкой практикой модель многопроцессорной архитектуры ПС-2000 не утратила своей актуальности до сих пор. Даже наоборот. Быстро прогрессирующие полупроводниковые технологии предоставляют возможности для максимального раскрытия вычислительного и экономического потенциала этой модели, что даёт долгосрочные перспективы успешной конкуренции в развитии систем класса GP на мировом уровне.

Гибридная архитектура развиваемой элементной базы на основе архитектуры ПС-2000М развивает опыт совмещённой работы в едином комплексе ПС-2000 многопроцессорного компьютера-ускорителя и управляющей универсальной мини-ЭВМ [2]. Внешняя машина своей операционной системой (ОС) обеспечивает весь цикл системного обеспечения – от тестирования и многозадачного управления вычислениями и периферией, до автоматизации программирования, включая как автономную отладку внутренних программ мультипроцессора, так и их запуск из внешних программ, выполняемых в управляющей машине. В качестве современного аналога универсальной управляющей машины на кристалле ПС-2000М будут задействованы стандартные многоядерные микропроцессоры со своей памятью и ОС.

Рис. 2. Масштабирование реализаций высокопараллельной архитектуры ПС2000М по мере развития полупроводниковых технологий.

На рис.2 приведена восходящая лестница на основе структурного масштабирования архитектуры ПС-2000М при реализации на современных и будущих полупроводниковых технологиях массового производства интегральных схем.

Масштабируемость, комплексируемость и высокая эффективность архитектуры ПС-2000М на широких классах задач с массовым параллелизмом позволит преодолеть отставание отечественного технологического производства СБИС, за счёт:

- возможности производить на имеющихся отечественных производствах микросхем младшие модели семейства (64, 128, 256, 512 ядер), обеспечивающие повышение интеллектуального характеристик высокотехнологичных изделий;
- обладания уникальным проектом масштабируемой в диапазоне 65-10 нм заказной СБИС однокристального компьютера производительностью 1-30 Тфлопс и более станет стимулом для обретения технологий глубокого нанометрового диапазона с последующим выпуском старших моделей кристаллов ПС-2000М, которые откроют возможности достижения паритета и превосходства в области высокотехнологичных изделий двойного назначения.

Однокристальный компьютер ПС-2000М предназначается для использования в широкой номенклатуре высокопроизводительных универсальных вычислительных систем и для реализации различных встраиваемых вычислительных систем и систем управления реального времени.

Выделим основные свойства архитектуры ПС-2000М, позволяющие прогнозировать ее высокую конкурентоспособность в отношении современных систем с массовым параллелизмом:

- реализация архитектуры:
  - высокое соотношение показателя «пиковая производительность/ (транзистор* стоимость* энергопотребление)»;
  - «околопиковая» (70-80% и выше) производительность на широких классах задач, обеспеченная высокой гибкостью параллельной архитектуры.
- простота аппаратных решений и гибкость управления:
  - SIMD из простых исполнительных устройств;
  - VLIW для управления (простота дешифрации, уплотнение команд организует оптимизирующий компилятор);
  - эффективная структура межпроцессорных обменов;
- пропорциональная структурная масштабируемость при переходе к новым технологиям (рис. 2):
  - по вычислительной производительности;
  - по объёму и "ширине" (пропускной способности) памяти;
  - по "ширине" (пропускной способности) внешнего интерфейса ввода/вывода;
  - по частоте (энергосбережение);

- бесшовная программно управляемая комплексируемость;
- использование сочетания параллельных и макроконвейерных структур;
- единый механизм комплексируемости на внутрикристальных и межкристальных (системных) уровнях.

В гибридной архитектуре ПС2000М в одном микрочипе объединяются универсальный микропроцессор и многопроцессорный компьютер с массовым параллелизмом, что обеспечивает высокий уровень универсализма поддерживаемых вычислений относительно классов решаемых задач.

6. Заключение

Предлагаемый к реализации проект создания масштабируемого семейства однокристальных компьютеров ПС-2000М нацелен на создание высококонкурентной отечественной ЭБ для широкого диапазона приложений и её массовое применение в составе высокопроизводительных настольных компьютеров и рабочих станций, семейств серверов, суперкомпьютеров с гибридными архитектурами производительностью петафлопного диапазона (более 10^{15} оп/сек), а также систем распределённой обработки информации и сетецентрического управления в качестве компонентов встраиваемых вычислительных систем двойного назначения, в том числе систем реального времени.

Важно подчеркнуть, что реализация предлагаемого проекта позволит организовать в России производство относительно широкой номенклатуры недорогих высокопроизводительных компьютерных систем двойного назначения, а новая технология создания прикладного программного обеспечения позволит существенно снизить издержки, связанные с программированием компьютеров с массовым параллелизмом [6]. Это обеспечит доступность высокопроизводительных вычислительных систем для широкого круга потребителей и их высокую конкурентоспособность на зарубежных рынках.

На переходе к новому этапу развития компьютеростроения глубинный внутрикомпьютерный кризис выравнивает стартовые позиции. В этом уникальный исторический шанс для новых игроков. При этом следует ясно осознавать, что пространственно-временная щель в будущее открывается лишь для тех, кто её увидит раньше других. И открыта она будет недолго.

ЛИТЕРАТУРА:

1. Затуливетер Ю.С. Компьютерные архитектуры: неожиданные повороты // Hard 'n' Soft. 1996. № 2. С. 86-94. http://zvt.hotbox.ru/p2_z1.htm.
2. Затуливетер Ю.С., Фищенко Е.А. Многопроцессорный компьютер ПС-2000 (Опыт создания и пути развития). Научное издание (Препринт). М.: Институт проблем управления РАН, 2012. 86с. URL: http://www.ipu.ru/sites/default/files/publications/16551/3477-препринт%20пс-2000__2.pdf.
3. Wolcott P., Goodman S.E. High-Speed computers of the Soviet Union // IEEE Computer. 1988. Vol. 21, No 9. P. 32-41.
4. Wolcott P., Goodman S.E. International perspectives: under the stress of reform high-performance computing in the former Soviet Union // Communications of the ACM. 1993. Vol. 36, No 10. P. 21-24.
5. Артамонов С.Е., Затуливетер Ю.С., Фищенко Е.А. Предпосылки к созданию однокристального многопроцессорного компьютера ПС-2000М производительностью 1-10 Tflops // Параллельные вычислительные технологии (ПаВТ'2011): труды международной научной конференции (Москва, 28 марта - 1 апреля 2011г.). С.402–410. Издательский центр ЮУрГУ, 2011. 730с. URL: http://omega.sp.susu.ac.ru/books/conference/PaVT2011/short/012.pdf.
6. Затуливетер Ю.С. Введение в проблему параметризованного синтеза программ для параллельных компьютеров. Научное издание (Препринт). М.: Институт проблем управления РАН. 1993. 88с.