ВЫЧИСЛИТЕЛЬНЫЕ ПРОЦЕССЫ С МАССОВЫМ ПАРАЛЛЕЛИЗМОМ
НОВЫЙ ПОДХОД

В.Бурцев

ПРОБЛЕМА УВЕЛИЧЕНИЯ ПРОИЗВОДИТЕЛЬНОСТИ

Требования к производительности вычислительных средств давно превзошли физические возможности одного процессора, работающего на принципе фон-Неймана. Поэтому столь остро стоит сегодня проблема реализации массовых параллельных вычислительных процессов. Современные архитектуры суперЭВМ, создаваемых на базе традиционных высокопроизводительных микропроцессоров, относятся к одному из двух классов:

- многомашинные комплексы на базе микропроцессоров с фиксированным по процессорам распределением памяти (распределенная память);
- многопроцессорные комплексы с распределяемой в процессе счета памятью (распределяемая память).

Например, объединяя неограниченное число персональных компьютеров (ПК) локальными сетями, можно достичь почти любой максимальной производительности многомашинного комплекса. Для определенных классов задач такой подход вполне оправдан – когда данные хорошо локализуются по параллельным вычислительным процессам и обмен данными между ПК не тормозит общий вычислительный процесс, производительность комплекса будет близка к максимально возможной Ппр·N, где Ппр – производительность микропроцессора ПК и N – количество ПК. Однако если задача имеет большие массивы глобальных данных, производительность многомашинного комплекса может упасть практически до нуля.

Реальная производительность комплексов (Преал) с той или иной архитектурой значительно меньше максимальной производительности. В то же время наиболее объективным параметром сравнения эффективности работы различных суперЭВМ может быть коэффициент Креал=Преал/Пмакс на представительном спектре задач. Креал существенно снижается при увеличении числа процессоров *N* или исполнительных устройств в комплексе. Причин этому две:

пространственная – недозагрузка процессоров из-за запаздывания данных при передаче их из памяти к процессору или от процессора к процессору;

временная: недозагрузка процессора из-за отсутствия данных, поставляемых другими вычислительными процессами (синхронизация по данным).

Первая причина достаточно подробно исследована в работе [1]. В ней показано, что реальная производительность многомашинного комплекса $Креал_{мм}$ пропорциональна производительности одного процессора Ппр и пропускной способности внешних запоминающих устройств и ОЗУ Е и обратно пропорционально числу процессоров N: $Преал_{мм}$~Ппр·E/N. В то же время реальная производительность многопроцессорных комплексов $Преал_{мп}$~Ппр·E. Как видно, архитектура многопроцессорных систем (распределяемая память) имеет определенные преимущества ввиду того, что не во всех задачах возможно локализовать данные при процессорах многомашинного комплекса, имеющих объем ОЗУ в *N* раз меньше, чем в комплексах с распределяемой памятью.

Однако эти качественные сравнительные оценки двух архитектур сделаны без учета задержки в системах коммутации вычислительных комплексов. Поэтому преимущества многопроцессорных комплексов справедливы только для сравнительно малых *N*. При *N* свыше 32 построение многопроцессорных комплексов существенно затруднено тем, что значительно возрастают задержки коммутатора (процессор – ОЗУ). Причем эта задержка дважды учитывается во времени выполнения каждой операции процессора. Поэтому при каждом микропроцессоре необходима сверхоперативная память (кэш). Ее использование в многопроцессорных системах решает проблему нивелирования задержки только для работы с локальными данными. При работе кэш с глобальными данными, доступными в едином адресном пространстве другим процессорам, возникает проблема когерентности работы кэш. Существующие методы обеспечения когерентности кэш при увеличении числа процессоров *N*, работающих на общую память, существенно – в десятки и сотни раз – увеличивают время доступа процессора к данным ОЗУ.

Второй причиной неэффективной работы суперЭВМ на больших задачах является временной фактор – синхронизация данных параллельных вычислительных процессов. Фактически программист, решающий задачу на суперЭВМ с большим количеством параллельных процессов, должен разработать сложнейшую программу в реальном масштабе времени – каждый параллельный процесс выполняется на выделенном для него процессоре, объеме оперативной памяти, канале передачи данных и т.д. Это, однако, крайне сложно – время реализации вычислительных процессов на аппаратуре, как правило, не известно, так как оно зависит от данных и от ситуации прохождения задачи внутри системы. Наиболее распространенный способ синхронизации процессов по данным – использовать сами микропроцессоры комплекса для слежения за появлением данных того или другого процесса. Но это приводит к дополнительному расходу процессорного времени и снижению эффективности работы всего вычислительного комплекса.

Специалисты США в своих проектах по пентафлопному суперкомпьютеру вопрос увеличения производительности решают путем сокращения длин линий связи между устройствами и увеличения числа микропроцессоров за счет совершенствования технологии изготовления элементной базы. В 2008 году США планируют перейти на технологические нормы 0,05 мкм – в три раза меньше сегодняшних (0,18–0,15 мкм). Это существенно снизит выделение энергии на одно логическое срабатывание и позволит размешать на одном кристалле 32–64 микропроцессора (однако проблема отвода тепла от кристалла обострится). Более плотная компоновка устройств и увеличение их числа, конечно, поднимет пиковую производительность новых американских суперЭВМ. Од-

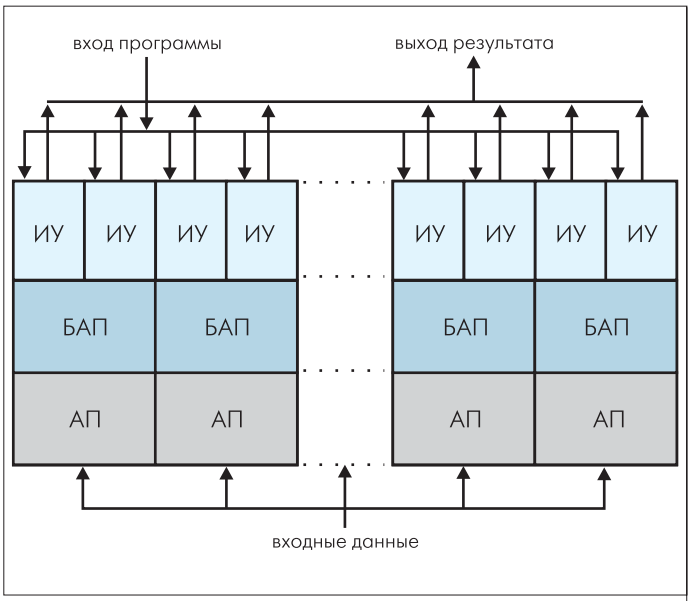

Рис.1. Структурная схема процессора

нако, оставаясь в рамках старой архитектуры, трудно ожидать существенного повышения реальной производительности высокопроизводительных вычислительных систем. Эта проблема на сегодняшний день остается не решенной.

ОТЕЧЕСТВЕННЫЙ ПРОЕКТ

Таким образом, для решения *пространственной задачи* поставки данных необходимо добиваться:

- увеличения производительности *одного* микропроцессора;
- исключения из комплекса проблемы когерентности кэш (машины Cray не имеют кэш);
- исключения из времени выполнения операций обращения к исполнительным устройствам временных задержек передачи информации между исполнительным устройством и ОЗУ;
- неучастия программиста в решении задачи распределения ресурсов вычислительных средств;
- увеличения пропускной способности между первичной и вторичной памятью.

Для решения проблемы *временных соотношений* данных необходимо:

- задачу синхронизации по данным и распараллеливания вычислительных процессов решать аппаратными средствами в процессе выполнения;
- концептуально исключить ОЗУ;
- устранить потерю процессорного времени на обработку прерываний и синхронизацию процессов;
- обеспечить работу вычислительного комплекса при минимальном составе устройств без изменения программы.

Всем этим требованиям отвечает структура микропроцессора, разработанная сотрудниками ИПИ РАН совместно с фирмой Nodal Systems Corporation. Процессор, состоящий из набора исполнительных устройств (ИУ), модулей ассоциативной памяти (АП) и их буферов (БАП), показан на рис.1 [2].

Интересная особенность приведенной структуры – в том, что функции устройства управления (организация параллельных вычислительных процессов) выполняет АП. Распределение ресурсов процессора и распараллеливание вычислительных процессов происходит автоматически, без вмешательства программиста. В то же время, если программист захочет, он может любую часть задачи запрограммировать последовательно в стиле фон-Неймана.

Сегодня уже работает макет процессора на ПЛИС APEX 20 KE фирмы Altera, разработанный на базе программ автоматизации проектирования Quartus. Макет подтвердил реализуемость новых принципов построения процессоров. Кроме того, поскольку функции устройства управления выполняет ассоциативная память, структура процессора однородна – следовательно, он технологичнее обычного процессора и может работать при наличии технологических дефектов в отдельных устройствах. По расчетам, такие процессоры можно производить на больших пластинах (wavel), сохраняя высокий коэффициент выхода годных.

Вычислительную эффективность процессора подтверждает ряд экспериментов по решению задач различных классов. Так, на рис.2 показан фрагмент решения задачи обращения матрицы ($B=A^{-1}$). Конфигурация процессора при этом – 128 исполнительных устройств и 128 модулей ассоциативной памяти. Приведены единовременно рассчитанные значения элементов обращенной матрицы. В скобках указан номер итерации для каждого элемента. Видно, что в один и тот же момент в разных точках обрабатываются различные итерации – от второй до десятой. На рис.3 показана динамика загрузки ИУ, число токенов (данных вместе с указателями) в АП и число готовых к выполнению данных в буфере АП. По оси абсцисс отложен номер такта. Рис.4 отражает процентное распределение загрузки ИУ и загрузку токенами модулей АП при конфигурации 64 ИУ и 64 модуля АП.

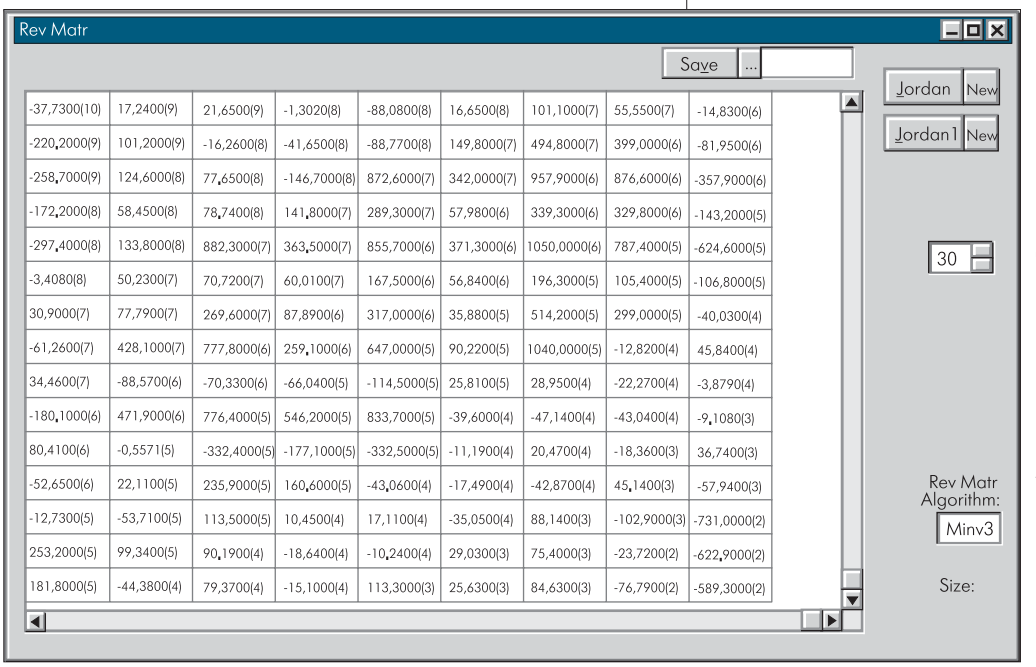

Рис.2. Ход решения задачи обращения матрицы. В скобках – номер итерации для данной ячейки

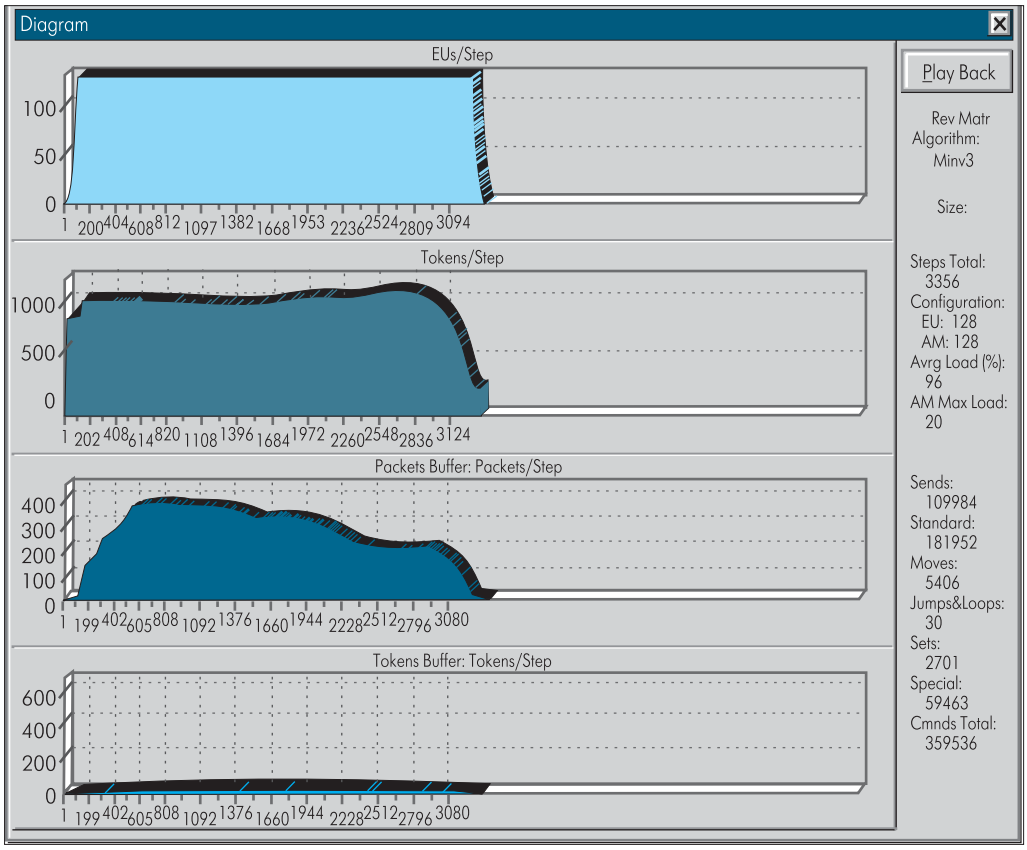

Рис.3. Динамика загрузки ИУ (верхний график), число токенов в АП (средний график) и число готовых к выполнению данных в буфере АП (нижний график) в ходе решения задачи обращения матрицы

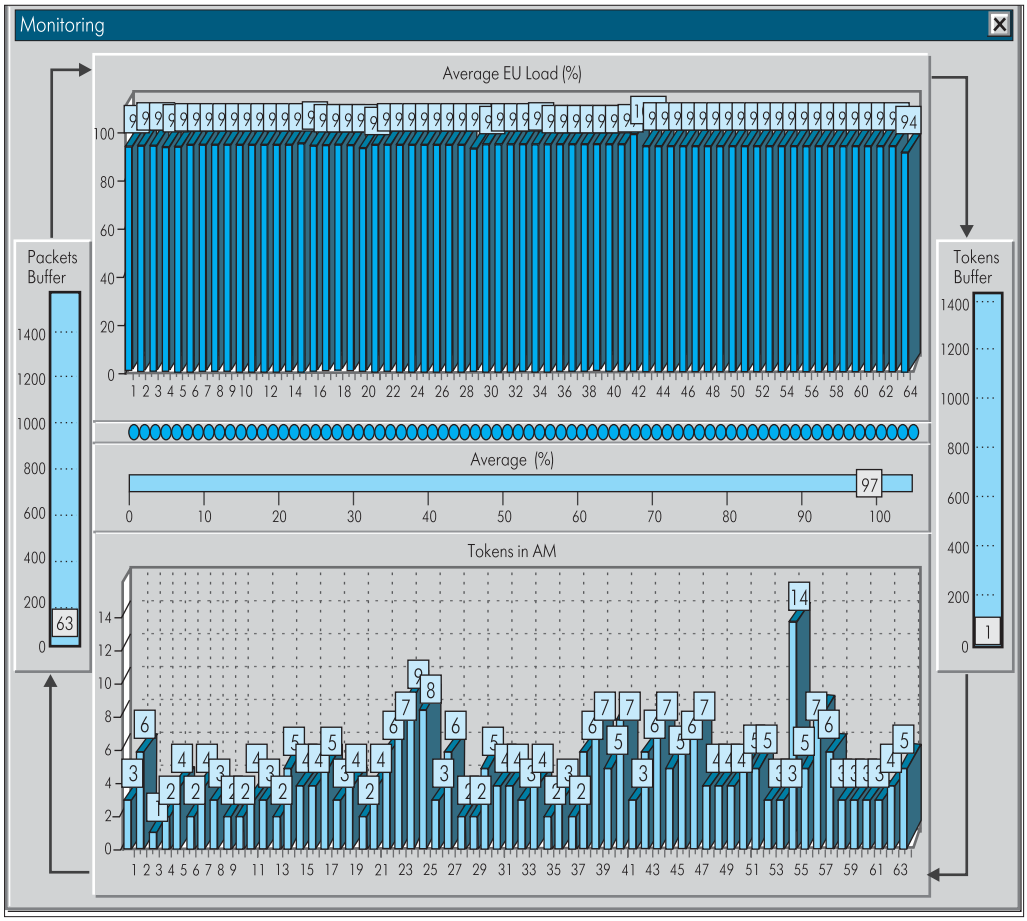

Рис.4. Процентное распределение загрузки ИУ и загрузка токенами модулей АП при конфигурации 64 ИУ и 64 модуля АП в ходе решения задачи обращения матрицы

Наличие в буфере АП достаточно большого числа готовых к выполнению данных показывает возможность увеличения параллелизма выполнения задачи за счет увеличения числа ИУ. Отметим также, что выполнение этой задачи при достаточно большом параллелизме (128 ИУ) требует не так уж много ячеек АП.

Архитектура нового процессора весьма эффективна и на задачах с неявно выраженным параллелизмом. Так, на рис.5 приведена динамика нахождения простых чисел. За счет аппаратного распараллеливания она выполняется на процессоре с достаточно большим уровнем параллелизма – 64 ИУ загружены более чем на 90%.

ВЫВОДЫ

Весьма актуальная сегодня проблема увеличения реальной производительности суперЭВМ до 10^{15} оп./с и выше не может быть полностью решена только за счет совершенствования технологии СБИС. Ее реализацию в сфере задач с массовым параллелизмом вычислительных процессов ограничивают два фактора: пространственный – обеспечение исполнительных устройств необходимыми данными и временной – синхронизация данных параллельных вычислительных процессов.

Предлагаемая новая архитектура процессора снимает эти ограничения на аппаратном уровне. Процессор может быть с успехом применен в сложнейших задачах реального масштаба времени, включая телекоммуникационные системы и создание нового поколения персональных компьютеров повышенной производительности (до 1 Tflops). История вычислительной техники еще раз подтверждает, что передовой фронт ее развития проходит через высокопроизводительные вычислительные системы – суперЭВМ.

Хочу выразить благодарность всем сотрудникам отдела 12 ИПИ РАН, как и других институтов РАН,

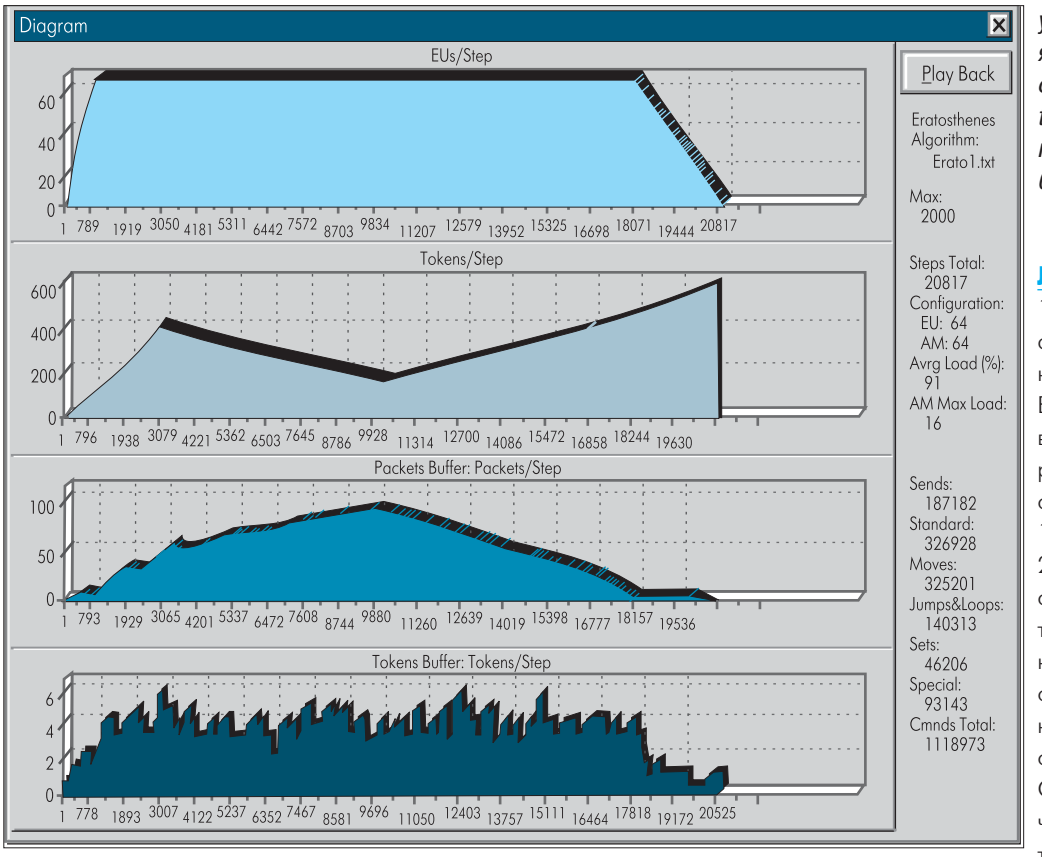

Рис.5. Динамика процесса нахождения простых чисел

участвующим в работе над настоящим проектом, и сотрудникам фирмы Nodal Systems Corporation, без помощи и моральной поддержки которых проект не был бы завершен.

ЛИТЕРАТУРА

1. **Бурцев В.С.** Новые подходы к оценке качества вычислительных средств. – В кн.: В.С. Бурцев. Параллелизм вычислительных процессов и развитие архитектуры суперЭВМ. – М.: ИВВС РАН, 1997, с. 28–40.

2. **Бурцев В.С.** Системы массового параллелизма с автоматическим распределением аппаратных средств суперЭВМ в процессе решения задачи. – Юбилейный сборник трудов институтов Отделения автоматики, вычислительной техники и автоматизации РАН. – М., 1993, т. 2, с. 5–27.