Проблемы обеспечения роста производительности отечественных суперЭВМ в период до 2020 года

В.Б. Бетелин, А.Г. Кушниренко, Г.О. Райко

Аннотация. Развитие высокопроизводительных суперкомпьютерных систем сегодня сталкивается с тремя основными проблемами: высокое энергопотребление; высокая частота случайных сбоев; высокие затраты ресурсов на организацию контрольных точек. Выход на экзафлопные производительности нельзя обеспечить путем эволюционного развития существующих технологий. В статье перечислены некоторые новые подходы к решению трех указанных основных проблем.

Ключевые слова: суперкомпьютер, производительность, Петафлопс, Экзафлопс, потребляемая мощность, надежность, контрольная точка.

В 1995 – 1996 гг. в США была развернута программа, обеспечивающая экспоненциальный рост производительности производимых в США суперЭВМ. Непосредственной целью развертывания этой программы послужила необходимость поддержки работоспособности ядерных арсеналов США в условиях Договора о запрещении ядерных испытаний в трех средах. Однако в программе были явно сформулированы и побочные цели – развертывание экономически оправданного производства суперЭВМ в США и захват значительной доли мирового рынка суперЭВМ. Программа оказалась чрезвычайно успешной: в 1996 г. была достигнута производительность 1 Тфлопс, к 2008 г. была достигнута производительность 1 Пфлопс. Американская промышленность сегодня выпускает серийно суперЭВМ производительностью от 5 до 1000 Тфлопс и доминирует на мировом рынке. Рост производительности суперЭВМ был использован американской промышленностью для повышения производительности труда, для создания новых технически сложных изделий, в том числе и военного назначения. Установленные в США суперЭВМ по суммарной производительности составляют 60% от мировой, при этом половина этой мощности используется в промышленности. Россия отстает от США, Евросоюза и Китая как по рекордной мощности установленных суперЭВМ, так и по суммарной производительности. Однако в настоящий момент это отставание не носит катастрофического характера. Опыт России в разработке микропроцессоров и коммуникационных СБИС позволяет при необходимости быстро развернуть производство отечественных суперЭВМ производительностью несколько сотен Тфлопс. В ближайшее время это положение может кардинально измениться: США планируют к 2018 г. изготовить суперЭВМ производительностью 1 Экзафлопс (1000 Пфлопс = 1 000 000 Тфлопс). Технологии разработки суперЭВМ подобной мощности пока не существует [6].

Обеспечение роста производительности суперЭВМ старыми методами сегодня оказывается невозможным. Традиционный способ построения суперЭВМ – размещение на печатной

плате максимально возможного числа микропроцессорных СБИС максимально возможной производительности (с ограничением по рассеиваемой мощности) и дальнейшее комплексирование этих плат в стойках – сталкивается с фундаментальными ограничениями.

Одним из слабых мест суперЭВМ, построенных по традиционному принципу, оказывается надежность установки в целом. Так, например, в статье [1] приводится анализ системных протоколов регистрации ошибок, зарегистрированных в период с 3 июня 2005 г. по 4 января 2006 г. на суперЭВМ IBM Blue Gene/L производительностью около 300 Тфлопс в Ливерморской Национальной Лаборатории. В указанный период суперЭВМ содержала около 200 тысяч процессоров IBM PowerPC и 1.89 Пбайт дискового пространства. В статье приведена следующая оценка. Приняв, что среднее время восстановления после сбоя составляет 10 минут, получаем среднее время наработки на отказ для суперЭВМ Blue Gene/L 5.89 часов, т.е. приблизительно случается 4 отказа в сутки. Для такого числа процессоров и такой производительности это время наработки на отказ еще приемлемо. Но с ростом числа процессоров положение начнет ухудшаться.

Современные технологии не позволяют обеспечить дальнейшее увеличение тактовой частоты при сохранении приемлемых показателей рассеиваемой мощности. Поэтому, следует ожидать, что на пути от петафлопса к экзафлопсу тактовая частота единичного микропроцессора стабилизируется на уровне нескольких гигагерц. Между тем, достижение производительности в 1 Экзафлопс требует удвоения производительности суперЭВМ каждые 12 месяцев. Тем самым нас ждет удвоение суммарного количества процессорных ядер в суперЭВМ каждые 12 месяцев, и ЭВМ производительностью 1 Экзафлопс будет содержать около миллиарда процессорных ядер. Однако количество процессорных ядер, которые можно разместить на одной микропроцессорной СБИС, ограничивается количеством выводов корпуса этой СБИС, которое не может увеличиваться экспоненциально. Таким образом, достижение требуемой производительности будет сопровождаться ростом числа микропроцессорных СБИС. По оценкам [2] число микропроцессорных СБИС в суперЭВМ будет удваиваться каждые два года. Анализ, проведенный в [3] показывает, что интенсивность отказов суперЭВМ пропорциональна числу микропроцессорных СБИС. В свою очередь, оптимистичная оценка интенсивности отказов в год в пересчете на микропроцессорную СБИС составляет 0.1 [4]. Рис. 1 показывает ожидаемый рост числа микропроцессорных СБИС в суперЭВМ и ожидаемое уменьшение наработки на отказ, вызванное этим ростом, в трех вариантах: удвоение числа микропроцессорных СБИС за 18, 24 и 30 месяцев.

Из рисунка видно, что среднее время наработки на отказ для суперЭВМ с 1 млн. микро-

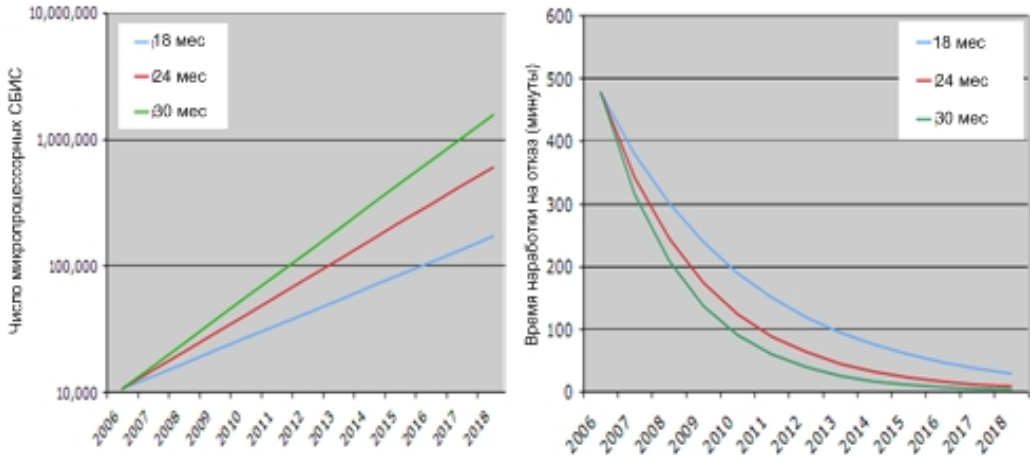

Рис. 1. Рост числа микропроцессорных СБИС (слева) и уменьшение наработки на отказ, вызванное этим ростом (справа)

процессорных СБИС будет составлять не более нескольких минут. По данным, приведенным [5], отказы модулей памяти происходят примерно с той же частотой, что и отказы микропроцессорных СБИС. Поэтому в предположении, что количество модулей памяти, приходящихся на одну микропроцессорную СБИС, составляет несколько единиц, тенденции, показанные на Рис. 1, справедливы и с учетом отказов модулей памяти.

Можно ли считать на такой суперЭВМ или нет, зависит от времени, затрачиваемого на создание контрольной точки. Чем больше это время, тем меньше доля времени, потраченного на полезную работу суперЭВМ. Для петафлопной суперЭВМ Jaguar создание контрольной точки занимает около 14 минут. Для оценок будущих суперЭВМ примем консервативное предположение, что пропускная способность подсистемы ввода/вывода и объем памяти будут расти пропорционально производительности и, таким образом, время создания контрольной точки останется постоянным. Из этого предположения вытекает пессимистический вывод: переход к экзафлопсу в старой парадигме невозможен. Действительно, на Рис. 2 изображено падение полезной загрузки суперЭВМ в условиях сделанных выше предположений. Из рисунка видно, что если количество микропроцессорных СБИС в суперЭВМ будет удваиваться каждые 30 месяцев, полезная загрузка суперЭВМ упадет до нуля в 2013 году, так как приложение будет тратить 100% времени на организацию контрольных точек и восстановление после очередного отказа.

Для предотвращения указанного катастрофического падения полезной нагрузки теоретически существует несколько стратегий.

Стратегия 1. Не увеличивать число микропроцессорных СБИС в суперЭВМ

Поскольку интенсивность отказов пропорциональна числу микропроцессорных СБИС, падения полезной нагрузки можно избежать, если сохранять число СБИС постоянным. Это, однако, означает, что

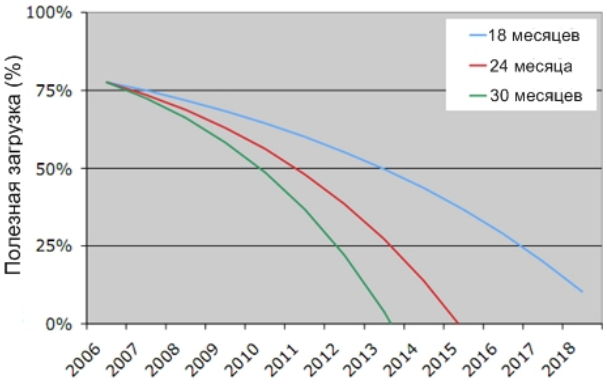

Рис. 2. Падение полезной загрузки суперЭВМ

- либо скорость роста тактовой частоты должна быть выше ожидаемой (что маловероятно, поскольку этот рост, помимо прочих факторов, сдерживается из-за соображений потребляемой мощности);
- либо должна быть выше скорость роста числа процессорных ядер (также маловероятно, поскольку обеспечить соответствующее увеличение пропускной способности памяти будет практически невозможно).

Таким образом, стратегия 1 должна быть отвергнута.

Стратегия 2. Повышение надежности СБИС

В случае если время наработки на отказ микропроцессорных СБИС будет расти пропорционально их числу в суперЭВМ, падения полезной нагрузки можно избежать. В то время как разработчики микропроцессоров не обещают подобного роста надежности СБИС, данная стратегия, по-видимому, является самой предпочтительной для заказывающих организаций, поскольку каждое ТТЗ на очередную суперЭВМ в США сохраняет величину наработки на отказ (для суперЭВМ в целом). Следование стратегии 2 потребует либо развития новых технологий микроэлектронного производства, либо радикальной переработки архитектуры микропроцессорных СБИС, а, скорее всего, и того, и другого. Работы в этом направлении уже начаты с США (проект POWER 6 компании IBM).

Стратегия 3. Уменьшение времени создания контрольных точек

Предполагая, что число микропроцессорных СБИС (и, следовательно, интенсивность отказов) ежегодно растет на 40%, полезная загрузка суперЭВМ останется на прежнем уровне, если время создания контрольных точек будет уменьшаться на 30% каждый год. Для этого требуется, чтобы пропускная способность подсистемы ввода/вывода росла пропорционально общей производительности суперЭВМ, т.е. пропускная способность должна ежегодно удваиваться. Однако до настоящего времени рост пропускной способности единичного диска был существенно ниже роста производительности единичного микропроцессора и составлял около 20% ежегодно. Это означает, что требуемое увеличение пропускной способности должно обеспечиваться за счет увеличения числа дисков в суперЭВМ. Причем скорость этого роста должна быть существенно выше, чем скорость роста числа микропроцессорных СБИС. При этом, поскольку стоимость диска несравнимо выше стоимости микропроцессорной СБИС, возрастет доля дисковой подсистемы в общей стоимости суперЭВМ и, следовательно, стоимость самой суперЭВМ существенно увеличится.

Другим фактором, ограничивающим успешность данной стратегии, является фиксированная надежность единичного диска, которая существенно ниже надежности единичной микропроцессорной СБИС.

Стратегия 4. Дублирование приложений на уровне процессов

При падении полезной загрузки суперЭВМ ниже 50% целесообразно рассмотреть вопрос о переходе к механизму дублирования приложений с организацией точек синхронизации и проверки правильности вычислений. Указанный механизм наряду с механизмом контрольных точек широко используется при организации отказоустойчивых комплексов.

Выводы

Достижение производительности суперЭВМ 1 Экзафлопс потребует развития радикально новых технологий. Россия должна решить, будет ли она принимать участие в разработке этих технологий.

Литература

1. N. Taerat, N. Naksinehaboon, C. Chandler, J. Elliott, C. Leangsuksun, G. Ostrouchov, S.L. Scott. Using Log Information to Perform Statistical Analysis on Failures Encountered by Large-Scale HPC Deployments. Материалы конференции High Availability and Performance Computing Workshop (HAPCW 2008).
2. K.Asanovic и др. The landscape of parallel computer research: a view from Berkley, Technical Report No. UCB/EECS-2006-183
3. B. Schroeder, G.A. Gibson. Understanding failures in petascale computers. Journal of Physics: Conference Series 78 (2007)
4. B. Schroeder, G. A. Gibson. A large-scale study of failures in high-performance computing systems. Proceedings of the International Conference on Dependable Systems and Networks (DSN2006).
5. B. Schroeder, E. Pinheiro, W.-D. Weber. DRAM Errors in the Wild: A Large-Scale Field Study. Proceedings of SIGMETRICS/Performance'09
6. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems. DARPA/IPTO, 2008. pp. 1 – 278.

Бетелин Владимир Борисович. Директор НИИСИ РАН. Окончил механико-математической факультет МГУ им. М.В. Ломоносова в 1970 году. Доктор физико-математических наук, профессор, академик РАН, член Президиума РАН. Автор более 100 опубликованных работ. Область научных интересов: информационные технологии.

Кушниренко Анатолий Георгиевич. Заведующий отделом НИИСИ РАН. Окончил механико-математической факультет МГУ им. М.В. Ломоносова в 1967 году. Кандидат физико-математической наук. Автор более 60 опубликованных работ. Область научных интересов: математика, информационные технологии в образовании.

Райко Глеб Олегович. Заведующий сектором НИИСИ РАН. Окончил механико-математической факультет МГУ им. М.В. Ломоносова в 1994 году. Автор более 20 опубликованных работ. Область научных интересов: системное программирование, информационные технологии.