СЛЕДУЮЩЕЕ ПОКОЛЕНИЕ УНИВЕРСАЛЬНЫХ КОМПАКТНЫХ СУПЕРЭВМ ТЕРАФЛОПНОЙ ПРОИЗВОДИТЕЛЬНОСТИ

В. Н. Лашманов, Н. А. Дмитриев, Е. Н. Кандюрин, А. Г. Селякин, А. Ю. Ушаков, А. В. Шатохин, С. Н. Шлыков, Д. Р. Ярулин

ФГУП «РФЯЦ-ВНИИЭФ», г. Саров Нижегородской обл.

Введение

С 2010 года ФГУП РФЯЦ-ВНИИЭФ в рамках президентского Проекта «Развитие суперкомьютеров и грид-технологий» ведет исследования и разработки в области компактных суперЭВМ (далее − КС-ЭВМ) [1]. На данный момент предприятиям-участникам поставлено 28 КС-ЭВМ − аппаратно-программных комплексов производительностью 1 Тфлоп/с (далее − АПК-1). Аппаратные решения защищены патентом на полезную модель № 99213 [2].

В ходе эксплуатации АПК-1 был выявлен ряд недостатков, связанных, прежде всего, с работой систем охлаждения и электропитания. Кроме того, пользователями были высказаны пожелания и предложения по доработке конструктива и расширению функционала АПК-1. С целью устранения недостатков и учета предложений пользователей, а также с учетом выхода на рынок новых, более производительных микропроцессоров возникла потребность в разработке и реализации технических решений по модернизации АПК-1 до уровня АПК-1М.

В основе модернизации АПК-1 лежит использование новых более производительных микропроцессоров Interlagos компании AMD (или Sandy Bridge компании Intel), а также разработка унифицированного конструктива для АПК-1М. Использование этих микропроцессоров в АПК-1М позволяет, при сохранении пиковой производительности компактной суперЭВМ не менее 1 Тфлоп/с, обеспечить улучшение ряда ее характеристик (потребляемой мощности, габаритов, веса, стоимости, надежности, ремонтопригодности и т. д.).

1. Основные направления модернизации АПК-1

В качестве основных методов оптимизации технических решений, применяемых в АПК-1, были выделены следующие:

− актуализация применяемых базовых элементов вычислительной подсистемы АПК-1 (микропроцессоры, материнские платы и т. п.);

− выработка конструктивных изменений, направленных на повышение надежности, ремонтопригодности и эргономики АПК-1;

− устранение замечаний пользователей, связанных с особенностями эксплуатации АПК-1.

В основе АПК-1 лежит использование четырехпроцессорных материнских плат H8QG. Материнские платы оснащены 12-ядерными микропроцессорами AMD Opteron 6168, тактовая частота этих микропроцессоров составляет 1,9 ГГц. Производительность одного микропроцессора AMD Opteron 6168 составляет 91,2 Гфлоп/с, одной материнской платы соответственно – 364,8 Гфлоп/с. Использование трех таких материнских плат в АПК-1 обеспечивает достижение пиковой производительности 1094,4 Гфлоп/с. В ходе проектирования АПК-1 эти выкладки легли в основу выбора центрального микропроцессорного элемента, и выбор был сделан в пользу AMD Opteron 6168. Сравнение выполнялось с серверными микропроцессорами Intel, но на момент разработки АПК-1 подходящего микропроцессора среди микропроцессоров Intel не оказалось. Здесь нужно уточнить, что в ходе сравнения микропроцессоров Intel и AMD анализу также подвергались и результирующие технические характеристики АПК-1, получающиеся при использовании того или иного микропроцессора, ведь выбор микропроцессора практически однозначно определит и количество вычислительных узлов в системе, и топологию коммуникационной среды, скажется на таких показателях, как суммарная потребляемая мощность, уровень акустического шума, габариты, вес.

С момента разработки АПК-1 ведущие производители микропроцессоров выпустили новые модели своих изделий, и в ходе модернизации АПК-1 это было учтено.

Во второй половине 2011 года компания AMD представила на рынке новые микропроцессоры серии Opteron 6200 с кодовым именем Interlagos. Это 16-ядерные микропроцессоры с архитектурой Bulldozer. Данные микропроцессоры совместимы с материнскими платами H8QG. В ходе проектирования АПК-1 уже решен ряд конструктивных вопросов, связанных с применением данной материнской платы. Использование этих наработок позволило существенно сократить сроки разработки и усилия, затрачиваемые на разработку АПК-1М. Это является ключевым моментом, определяющим использование данного процессора в АПК-1М. В случае использования микропроцессоров Interlagos для обеспечения производительности 1 Тфлоп/с достаточно установить два вычислительных узла в АПК-1М вместо трех вычислительных узлов АПК-1.

Важной особенностью технических и конструктивных решений АПК-1М является улучшение ремонтопригодности и эргономики. Это потребовало радикальной переработки корпуса АПК-1, повышения технологичности процесса изготовления корпуса. Как будет показано ниже, использование двух вычислительных узлов при реализации АПК-1М относительно варианта с тремя вычислительными узлами в АПК-1 существенно повысило ремонтопригодность, упростило конструктивные решения. Более удачная компоновка элементов АПК-1М упростила доступ к портам ввода-вывода вычислительных узлов и т. д.

Пожелания и предложения пользователей АПК-1 также были учтены при планировании путей модернизации АПК-1. В первую очередь необходимо отметить проблемы АПК-1, связанные с эксплуатацией коммерческого программного обеспечения, использующего реализации MPI, не обладающие поддержкой бескоммутаторной схемы коммуникационной среды межпроцессорных обменов [3]. В АПК-1М используется топология коммуникационной среды «точка-точка». Такая топология поддерживается любой реализацией коммуникационной программного обеспечения. Кроме того, в АПК-1М для обеспечения масштабирования предусмотрена возможность опциональной установки 8-портового коммутатора InfiniBand. Это будет востребовано пользователями, желающими использовать несколько образцов АПК-1М в единой системе, а также интегрировать АПК-1М в существующую вычислительную инфраструктуру предприятия. Важным замечанием пользователей АПК-1 является отсутствие возможности выполнения визуализации сложных графических объектов, это связано с тем, что в базовую комплектацию АПК-1 изначально не были включены высокопроизводительные графические адаптеры. На данный момент для визуализации как в серийные АПК-1, так и в разрабатываемые АПК-1М предполагается устанавливать видеокарту GeForce GT 430 Low Profile. Внешний вид видеокарты представлен на рис. 1.

PCI Express 2,0×16
Low profile
GeForce GT430
DDR31 Gb
OpenGL 4,01×HDMI, 1×D-SUB, 1×DVI
2560×1600

Рис. 1. Внешний вид видеокарты GeForce GT 430

Такая видеокарта не окажет существенного влияния на потребляемую мощность АПК-1(М), не потребует дополнительных конструктивных и схемотехнических изменений. Далее будет показано, каким образом в АПК-1М предусмотрена возможность установки более мощных графических ускорителей.

Далее рассмотрены технические решения АПК-1М.

2. Функциональная схема

На рис. 2 представлена функциональная схема АПК-1М. В варианте использования процессоров AMD Interlagos вычислительная подсистема АПК-1М строится на базе двух вычислительных модулей (четырех сокетных материнских плат Supermicro H8QG).

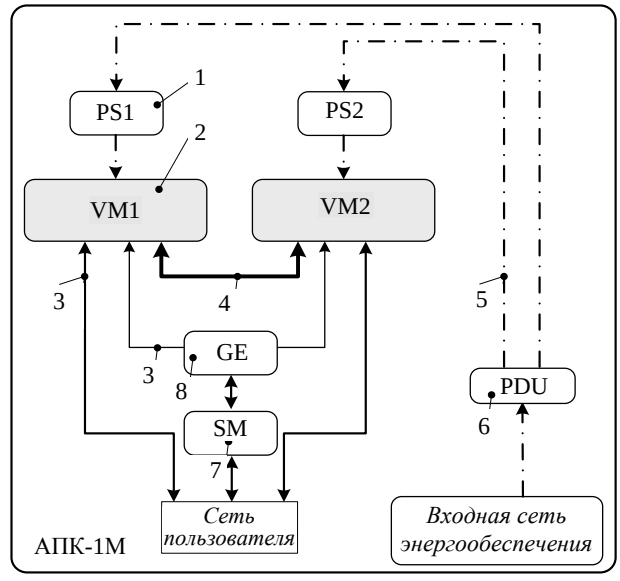

Рис. 2. Функциональная схема АПК-1М: 1 – блок питания вычислительного модуля (2 шт.); 2 – вычислительный модуль (2 шт.); 3 – канал управляющей сети (5 шт.); 4 – канал коммуникационной сети (1 шт.); 5 – линия электропитания (2 шт.); 6 – распределительная коробка питания (1 шт.); 7 – сервисный модуль (1 шт.); 8 – внутренний коммутатор Gigabit Ethernet (1 шт.)

Использование двух вычислительных модулей значительно упрощает конструкцию, а значит, производство и обслуживание АПК-1М.

Технические характеристики АПК-1М при использовании процессоров Interlagos (2,1 ГГц) представлены в таблице.

Технические характеристики АПК-1М

Теоретическая пиковая производительность	1,075 Тфлоп/с
Объем оперативной памяти	до 1024 Гбайт
Емкость дисковой памяти	до 36 Тбайт
Габариты (В × Ш × Г)	650 × 240 × 550
Среда межпроцессорных обменов	InfiniBand 2xQDR «точка-точка» + возможность масштабирования
Потребляемая мощность	до 2,5 кВт
Стоимость	от 1,2 млн. руб

3. Коммуникационная подсистема

Бескоммутаторная топология коммуникационной среды АПК-1 вызывает ряд сложностей при работе с коммерческим программным обеспечением,

использующим стандартные реализации коммуникационного программного обеспечения. Однако при использовании прямого соединения материнских плат «точка-точка» такие проблемы не возникают. На рис. 3 представлена схема коммуникационной среды АПК-1М.

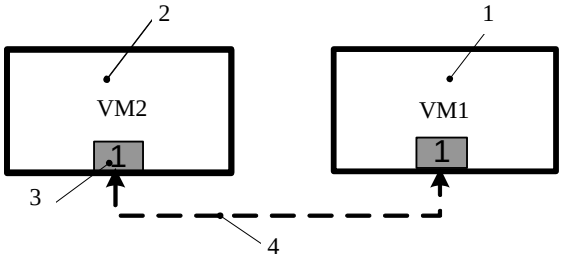

Рис. 3. Коммуникационная среда АПК-1М: 1 − вычислительный модуль 1 (1 шт.); 2 − вычислительный модуль 2 (1 шт.); 3 − адаптер InfiniBand QDR (2 шт.); 4 − канал коммуникационной сети InfiniBand (1 шт.)

Для реализации коммуникационной среды АПК-1М использованы однопортовые адаптеры InfiniBand QDR HCA фирмы Mellanox.

В конструкции АПК-1М предусмотрена возможность опциональной установки 8-портового QDR коммутатора InfiniBand Mellanox IS5022. Такое решение предназначено для обеспечения масштабируемости АПК-1М либо облегченного включения АПК-1М в существующую инфраструктуру заказчика.

4. Подсистема охлаждения

В АПК-1М наибольшей модернизации подверглась система охлаждения. Для обеспечения требуемого уровня шума с сохранением высокой эффективности отвода тепла от центральных процессоров система охлаждения центральных процессоров осталась жидкостной. В целях повышения надежности был предпринят ряд мер.

Минимизация количества резьбовых и вращающихся соединений трактов СЖО. Как известно, чем больше соединений, тем больше мест вероятной протечки контура СЖО. Контур СЖО одной материнской платы, содержащей четыре процессора, в АПК-1 имеет: 16 резьбовых соединений, восемь вращающихся соединений, восемь соединений гибких труб с металлическими фитингами типа «елочка». Преобразование СЖО подразумевает использование моноблока, охватывающего по два процессора. На рис. 4 представлен внешний вид 3D модели такого моноблока. Из рисунка очевидно, что применение моноблока в АПК-1М значительно повысило надежность тракта СЖО и упростило производство серийных образцов АПК-1М. При этом конструкция моноблока не затрудняет физический доступ к памяти (для замены).

Следующим направлением модернизации является повышение эффективности воздушного охлаждения оперативной памяти. На рис. 5 представлен снимок распределения температур, сделанный тепловизором, одной из крайних плат АПК-1 (без боковой крышки). Вентиляторы системы охлаждения расположены слева и осуществляют прокачку воздуха слева направо.

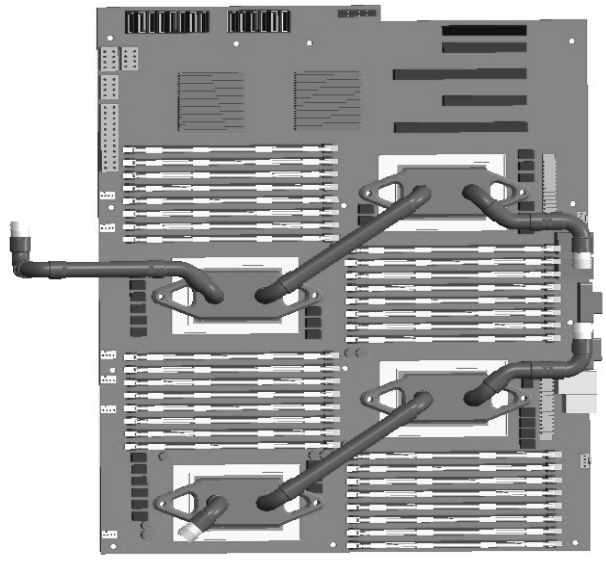

Рис. 4. Моноблок АПК-1М. Сравнительные характеристики СЖО АПК-1 и АПК-1М (для одной материнской платы). АПК-1М: − четыре резьбовых соединения; АПК-1: − 16 резьбовых соединений; − восемь вращающихся соединений; − восемь соединений гибких труб с металлическими фитингами типа «елочка»

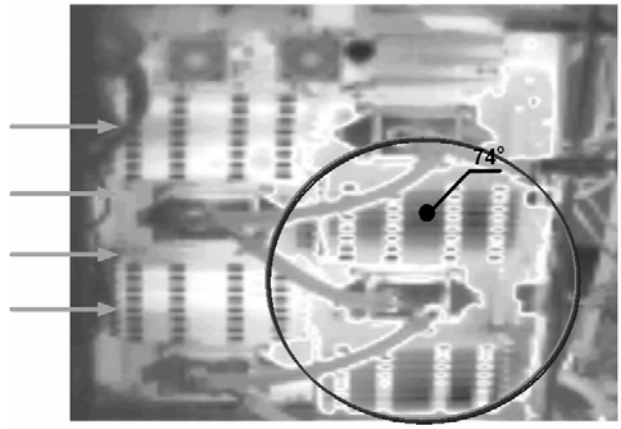

Рис. 5. Снимок тепловизора

Необходимо отметить, что данный снимок сделан при значительной загрузке АПК-1 и при открытой боковой крышке корпуса. При закрытой боковой крышке память нагревается меньше, так как крышка способствует правильной ориентации воздушного потока. Как видно из рисунка, при использовании реализованной схемы охлаждения памяти, воздушный поток, создаваемый вентиляторами, используется не совсем эффективно – часть воздуха растекается по пути наименьшего сопротивления, что приводит к тому, что модули памяти, находящиеся дальше от вентиляторов, нагреваются больше (до 70 °С), чем модули, расположенные ближе к вентиляторам. Для

повышения эффективности охлаждения оперативной памяти и концентрации воздушного потока предлагается использовать дополнительные направляющие для воздушного потока – воздуховоды. Модель воздуховода представлена на рис. 6.

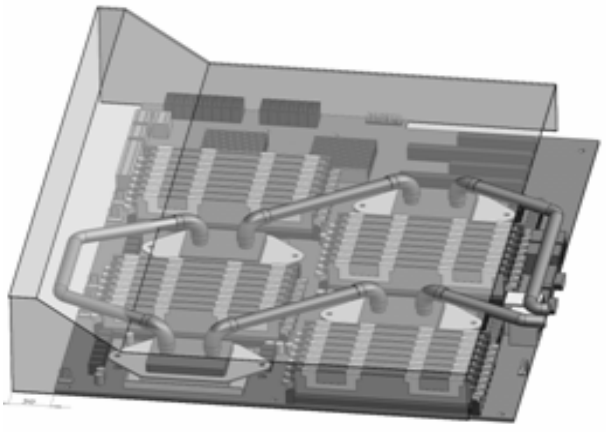

Рис. 6. Модель воздуховода

Основной задачей воздуховода является оптимальное распределение воздушного потока, создаваемого вентиляторами, для более эффективного охлаждения оперативной памяти, расположенной вдали от вентиляторов.

Еще одним важным нововведением в систему охлаждения АПК-1М является использование резервуара с переменным объемом. Резервуар переменного объема предназначен для компенсации перепадов давления, вызванного температурным расширением жидкости в тракте СЖО. Резервуар переменного объема – инновационная разработка РФЯЦ-ВНИИЭФ. В настоящий момент проведены все процедуры по регистрации данного технического решения в патентной службе РФ. Особенностью данного технического решения является то, что переменный объем бака дает возможность сохранить оптимальный уровень давления внутри СЖО, при этом не используются механизмы (клапаны), связывающие тракт СЖО с атмосферой, что позволяет сохранить тракт СЖО герметичным. На рис. 7 представлена модель бака СЖО с переменным объемом.

В основе функционирования бака переменного объема − использование гибкой мембраны, ко-торая герметично отделяет тракт СЖО от окружающей среды, при этом предоставляет запас теплового расширения для жидкости и воздуха в тракте СЖО.

Изменение претерпела и компоновка элементов системы СЖО. Радиатор (в АПК-1М достаточно использовать один радиатор, в отличие от АПК-1, где их использовалось два), предназначенный для охлаждения жидкости, и блок вентиляторов, устанавливаемый на него, вынесены в верхний отсек корпуса АПК-1М (см. рис. 14). Данное решение явилось следствием следующих улучшений:

− упрощен слив охлаждающей жидкости из тракта СЖО (радиатор находится значительно выше, чем в АПК-1);

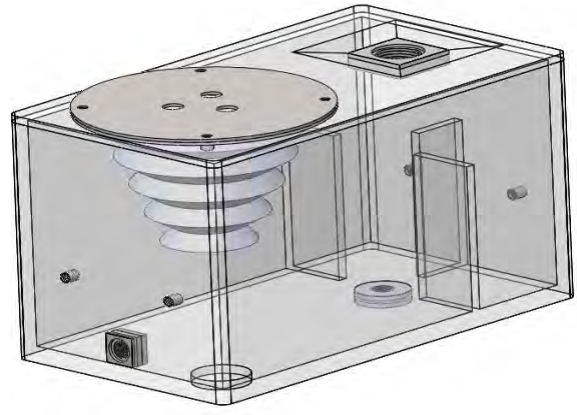

Рис. 7. Модель резервуара переменного объема

− уменьшена длина корпуса (в АПК-1 радиаторы размещались в задней части корпуса, что влекло за собой увеличение его длины);

− порты ввода-вывода вычислительных модулей, для удобства доступа, выведены на заднюю панель корпуса (размещение радиаторов в АПК-1 не позволяло вывести порты ввода-вывода вычислительных модулей АПК-1 непосредственно на заднюю панель корпуса).

В основе действия применяемого в АПК-1 датчика потока положено механическое перемещение ротора, интенсивность вращения которого свидетельствует о скорости тока жидкости в системе. Основным минусом при использовании датчика потока такого типа является то, что механическое перемещение ротора может быть легко заблокировано пузырьком воздуха либо незначительным загрязнением. При этом останов ротора датчика потока приводит к автоматическому аварийному выключению всей машины.

5. Сервисная подсистема

Сервисная подсистема также претерпит ряд изменений [4]. В АПК-1М планируется использование нового типа сервисного модуля. Новый тип сервисного модуля представляет собой панельный PC промышленного исполнения. На рис. 8 представлен внешний вид сервисного модуля. Кроме LCD дисплея данный сервисный модуль также обладает двумя интегрированными контроллерами Gigabit Ethernet, что очень важно относительно применения данного устройства в АПК-1М. Дело в том, что один из Ethernet-контроллеров должен быть использован для связи с внутренней (управляющей) сетью Gigabit Ethernet, а второй − для подключения внешних пользователей. В АПК-1 эта проблема решалась использованием дополнительного адаптера USB-Ethernet, что снижало надежность работы внешнего сетевого подключения к сервисному модулю.

Использование нового типа сервисного модуля влечет за собой изменения в системе мониторинга. В состав системы мониторинга включен модуль локального графического отображения информации на

VIA VIPRO VP7806;
VIA Nano 1.3 ГГц;
DDR2 2 ГБ;
6.5" color TFT LCD panel;
640 x 480;
HDD 2,5'';
2xGE
Сенсорный экран пятипроводной резистивный

Рис. 8. Внешний вид сервисного модуля VIA VIPRO VP7806

сенсорном LCD дисплее сервисного модуля с возможностью управления непосредственно с него.

Для примера на рис. 9−11 представлены иллюстрации работы графического модуля локального отображения информации.

Управляющая сеть представляет собой внутреннюю сеть Gigabit Ethernet. Управляющая сеть Gigabit Ethernet объединяет материнские платы АПК-1М, обеспечивая возможность их взаимодействия, осуществляя связь по каналу с пропускной способностью 1 Гбит/с и выполняя функции сетевой подсистемы мониторинга. В качестве коммутатора используется пятипортовый коммутатор Gigabit Ethernet D-Link DGS-1005D/GE. Схема управляющей сети АПК-1М представлена на рис. 12.

Рис. 9. Стартовый экран и основное меню

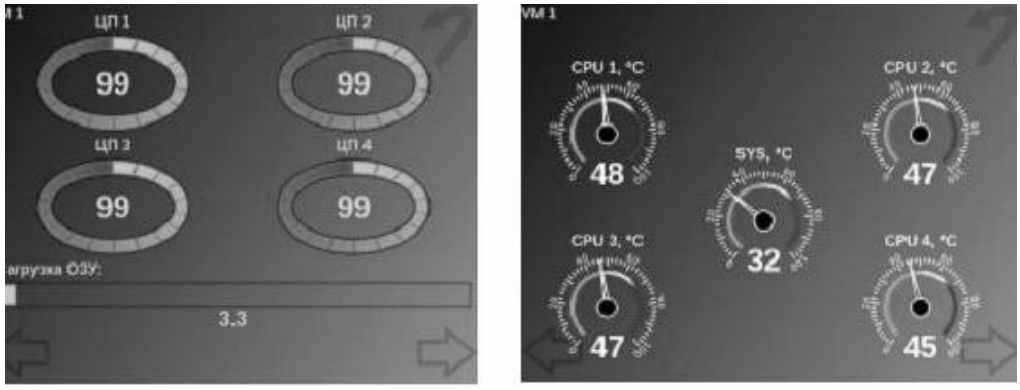

Рис. 10. Экран ресурсов и датчиков температуры

Рис. 11. Экран вентиляторов

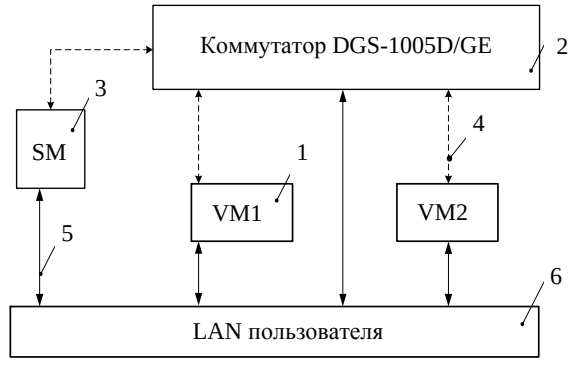

Рис. 12. Схема управляющей сети АПК-1М: 1 – вычислительный модуль (2 шт.); 2 – коммутатор Gigabit Ethernet DGS-1005D/GE (1 шт.); 3 – сервисный модуль (1 шт.); 4 – канал управляющей сети Gigabit Ethernet (3 шт.); 5 – канал пользователя (4 шт.); 6 – LAN пользователя (1 шт.)

Встроенные порты LAN1 Gigabit Ethernet контроллеров, расположенные на вычислительных модулях VM1, VM2 и сервисном модуле, подключаются к коммутатору. Так организуется внутренняя сеть АПК-1М, которую, в частности, использует система мониторинга. К встроенным портам LAN2 вычислительных модулей VM1, VM2, сервисного модуля, а также к коммутатору подключается LAN пользователей.

6. Корпус

Работы по модернизации корпуса велись в направлении уменьшения габаритов, снижения веса, улучшения ремонтопригодности и эргономики АПК-1М. Большинство технических решений, применяемых при разработке корпуса, направлено на создание наиболее универсального корпуса, который при незначительных изменениях можно было бы использовать для различных двухплатных модификаций АПК-1М (с использованием микропроцессоров Intel либо графических ускорителей). В ходе работ над АПК-1М был разработан *трехмерный* макет корпуса АПК-1М. На рис. 13 представлен внешний вид АПК-1М.

Рис. 13. Внешний вид АПК-1М

Как видно из рис. 13, корпус АПК-1М значительно компактней корпуса АПК-1. На передней панели кроме кнопок управления расположен LCD дисплей сервисного модуля, на нем отображается информация о текущем состоянии АПК-1М (температурные датчики, контроль ресурсов и т. п.). Кроме того, в отличие от АПК-1, где для замены диска требуется демонтировать боковые стенки корпуса АПК-1 и крепления жестких дисков, в конструкции АПК-1М предусмотрена возможность использования hot-swap дисков для их «горячей» замены.

На рис. 14, 15 представлено расположение комплектующих внутри корпуса АПК-1М.

Рис. 14. Компоновка АПК-1М, вид справа (ne2)

Рис. 15. Компоновка АПК-1М, вид слева (ne1)

Важно отметить, что корпус разработан таким образом, чтобы в узел ne1 можно было установить любые адаптеры, включая графические ускорители большого формата. Кроме того, на рис. 15 видно, что корпусе предусмотрена возможность опциональной установки восьмипортового коммутатора Infiniband. Еще одним важным изменением корпуса АПК-1М относительно АПК-1 является то, что порты ввода-вывода материнских плат для удобства эксплуатации выведены непосредственно на заднюю панель, а не спрятаны внутри корпуса, как в АПК-1.

7. Направления развития АПК-1М

Как показано выше, разрабатываемая для АПК-1М аппаратная конструкция может стать базовой основой для различных модификаций КС-ЭВМ. Далее

представлено описание возможных модификаций АПК-1М для использования в качестве основы при построении КС-ЭВМ с микропроцессорами Intel и графическими ускорителями.

Использование микропроцессоров Intel в АПК-1М

Выше в докладе представлено описание конструктивных и схемотехнических решений модернизированного образца КС-ЭВМ (АПК-1М), построенного с использованием микропроцессоров AMD 6200 Interlagos.

Для сохранения лидирующего положения РФЯЦ-ВНИИЭФ в российском сегменте рынка компактных суперЭВМ предлагается модификация АПК-1М на платформе Intel [5]. Для создания такой модификации КС-ЭВМ могут быть использованы микропроцессоры Sandy-Bridge, выход четырехпроцессорных версий которых Intel намечает на начало 2012 года. В случае применения этих микропроцессоров пиковая производительность свыше 1 Тфлоп/с достигается путем использования двух четырехпроцессорных материнских плат, как и в случае АПК-1М − AMD. Для того чтобы выработать модификацию АПК-1М, использующую другой тип микропроцессоров, требуется минимальная доработка, которая будет касаться только вычислительной подсистемы. Большинство базовых подсистем: корпус, сервисная подсистема, коммуникационная подсистема, подсистема управления и подсистема жидкостного охлаждения − претерпят минимальные изменения.

Возможность создания гибридных суперЭВМ на базе АПК-1М

Учитывая то, что использование гибридных технологий при проведении высокопроизводительных вычислений все чаще находит применение в серьезных коммерческих продуктах [6], следует рассматривать и такую возможность, как включение графических ускорителей в состав разрабатываемых аппаратных средств.

Для установки ускорителей в материнской плате необходимо наличие слотов расширения PCI-Ex16 v2.0. В рамках платформы АПК-1М предусмотрена возможность использования соответствующих материнских плат.

Потребляемая мощность одного ускорителя, как правило, составляет до 300 Вт [7]. Поэтому подсистема электропитания должна иметь выделенные линии электропитания для питания графических ускорителей, с соответствующим сечением проводов. В платформе АПК-1М используются блоки питания и платы распределения питания, в которых уже предусмотрена возможность подключения графических ускорителей.

Корпус АПК-1М предусматривает возможность установки полноразмерных адаптеров (например, графических ускорителей) на одну из двух материнских плат.

Таким образом, технические решения, примененные в АПК-1М, в перспективе могут стать платформой для построения не только универсальных компактных суперЭВМ, но и визуализационных, и гибридных машин, включающих в себя GPU.

Заключение

В данном докладе представлено описание модернизации конструктивных и схемотехнических решений АПК-1, направленной на повышение надежности, улучшения ремонтопригодности и эргономики, снижение стоимости. В АПК-1М снижены габариты и потребляемая мощность, максимально учтены замечания пользователей АПК-1. Важно отметить, что конструктивные решения, состав и структура подсистем АПК-1М разработаны таким образом, чтобы позволить АПК-1М стать базой для реализации дополнительных модификаций КС-ЭВМ, как с микропроцессорами Intel, так и с графическими ускорителями. Таким образом, существенно повышена функциональность АПК-1 и его потенциальная полезность потребителю.

Литература

1. Стрюков В. Н., Бартенев Ю. Г., Басалов В. Г. и др. Универсальная компактная суперЭВМ: Доклад на XII международный семинар «Супервычисления и математическое моделирование», 2010.

2. Пат. на полезную модель № 99213 от 10 ноября 2010 года. Компактная суперЭВМ.

3. Жуков Д. А., Вялухин В. М., Басалов В. Г. Особенности реализации коммуникационного программного обеспечения КС-ЭВМ при использовании бескоммутаторной технологии: Доклад на XII международный семинар «Супервычисления и математическое моделирование», 2010.

4. Дмитриев Н. А., Стрюков В. Н., Лашманов В. Н. и др. Сервисная подсистема универсальной компактной суперЭВМ: Доклад на XII международный семинар «Супервычисления и математическое моделирование», 2010.

5. [Электронный ресурс]. Режим доступа: http://www.intel.ru

6. [Электронный ресурс]. Режим доступа http://www.ansys.msk.ru

7. [Электронный ресурс]. Режим доступа http://www.nvidia.ru