УДК 681.32

УСТАНОВКА «СКИФ-АВРОРА» В ЮЖНО-УРАЛЬСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ

А.А. Московский[1], Е.А. Дружинин[2]
[1]Институт программных систем им. А.К. Айламазяна РАН, Переславль-Залесский;
[2]ЗАО «РСК СКИФ», Переславль-Залесский

Приводятся основные характеристики и технические решения по опытной установке СКИФ ряда 4 ~ «СКИФ-Аврора» - в Южно-Уральском государственном университете.

Введение

Платформа «СКИФ-Аврора» изначально разрабатывалась как основа для высокопроизводительных систем большого масштаба. Целый ряд технических решений, использованных в СКИФ ряда 4, сдвигает баланс свойств в сторону специализации для применения именно в суперкомпьютерах. Подробно характеристики решения рассмотрены в работе [1]. Установка «СКИФ-Аврора» в Южно-Уральском государственном университете является первым пилотным проектом по развертыванию системы такого класса.

1. Вычислительная подсистема

Проект системы, включая ресурсы систем охлаждения и бесперебойного электропитания, позволяет установить до восьми вычислительных шасси «СКИФ-Аврора». Каждое шасси включает 64 двухпроцессорных узла с четырехядерными процессорами Intel Nehalem X5570 с рабочей частотой 2,93 ГГц. Таким образом, в рамках одного монтажного шкафа удалось собрать 2048 процессорных ядер. Максимальная теоретическая производительность системы, состоящей из одного шкафа, составляет 24 Тфлопс.

Высокая плотность упаковки процессоров в вычислителе диктует необходимость использования жидкостной системы охлаждения. Вычислительные узлы выполнены в виде печатных плат с интегрированными на материнской плате коммуникационными, сервисными микросхемами, модулями памяти. Тестирование плат проводится на заводе-изготовителе, что уменьшает число отказов компонент при инсталляции и первичной настройке системы. Каждый узел-плата накрыт плотно прилегающей пластиной охлаждения. Пластины охлаждения оснащены быстроразъемными муфтами, что позволяет демонтировать отдельный вычислительный узел без демонтажа системы охлаждения корзины (шасси) в целом.

Каждый узел оснащен твердотельным накопителем объемом 80 Гбайт. Использование твердотельных накопителей также направлено на повышение надежности вычислителя - отказы шпиндельных дисковых накопителей составляют львиную долю причин отказов узлов в кластерных установках и вычислительных фермах.

Внешний вид монтажного шкафа вычислителя “СКИФ-Аврора”

2. Коммуникационные сети

Ключевым компонентом любого суперкомпьютера является его коммуникационная среда. Узлы «СКИФ-Аврора» обладают суммарным каналом пропускания до 100 Гбит/с, учитывая как системную, так и вспомогательную коммуникационные сети. Если во вспомогательной сети используются стандартные решения Infmiband QDR, то системная сеть является оригинальной разработкой.

Системная сеть имеет топологию трехмерного тора, маршрутизаторы сети реализованы на уровне адаптеров. Суммарная пропускная способность сети в пересчете на один узел составляет 60 Гбит/с. Сеть позволяет обойтись без использования дополнительного оборудования (маршрутизаторов) и задействовать при монтаже кабели одинаковой длины, вне зависимости от размера системы. Соединения на уровне половины шасси (корзины) выполнены на соединительной плате. Трехмерная организация сети позволяет легче распределить задачи между узлами кластера при моделировании объектов реального мира (трехмерных) и распараллеливании методом декомпозиции области. Для системной сети создана реализация MPI на основе MPICH2, удовлетворяющая спецификации версии MPI 2.O.

Вспомогательная сеть - сеть Infmiband QDR (40 Гбит/с) с полной бисекционной пропускной способностью. Адаптеры сети интегрированы на платах - узлах. Маршрутизаторы первого уровня интегрированы на уровне корзин (шасси) на так называемых «корневых платах». Соединения между узлами и маршрутизатором первого уровня выполнены на соединительной плате (backplane), что существенно уменьшает количество кабелей Infmiband, подключаемых вручную при установке системы. Поскольку маршрутизаторы первого уровня уже присутствуют в системе, на втором уровне сети можно использовать относительно недорогие 36-портовые маршрутизаторы - количество Infmiband кабелей и их длина от этого не меняется.

3. Подсистема мониторинга и управления

Подсистема мониторинга и управления обеспечивает надежное выполнение всех функций по удаленному обслуживанию установки, за исключением функций, требующих физических манипуляций. Подсистема использует как возможности стандартных IPMI средств мониторинга, так и оригинальную разработку - сеть Servnet. Компоненты Servnet присутствуют во всех основных модулях «СКИФ-Аврора»:

1) на уровне узлов интегрированы контроллер и датчики температуры и влажности;

2) на уровне «корневой платы» интегрированы датчики и контроллер управления;

3) на плате блока питания интегрированы датчики и контроллер управления питанием;

4) соединительная плата обеспечивает связь сети Servnet на уровне половины шасси.

Отличительной особенностью Servnet является возможность осуществления мониторинга даже в случае полного отключения электропитания всех основных систем - питание Servnet осуществляется независимо.

«Корневые платы» играют важную роль в системе управления установкой. Именно ПО, работающее на «корневой плате», позволяет отключать и включать электропитание отдельных узлов, осуществлять мониторинг характеристик системы во время работы. ПО «корневой платы» осуществляет вывод информации на сенсорные дисплеи, установленные в торцах шасси.

Программное обеспечение мониторинга интегрирует информацию из различных источников, включая подсистемы электропитания, охлаждения, хранения данных, отображает и хранит архив данных. Поскольку установка «СКИФ-Аврора» носит экспериментальный характер, под нужды управления и мониторинга выделен отдельный сервер.

4. Подсистема электропитания

Подсистема электропитания вычислителя «СКИФ-Аврора» осуществляется постоянным током с напряжением 48 В. За счет использования постоянного тока подсистема бесперебойного электроснабжения оказывается проще - содержит лишь выпрямитель и аккумуляторные батареи. Преобразователь постоянного тока в переменный, оказывается, не нужен.

Бесперебойное питание сервера мониторинга дополнительно резервировано для обеспечения автономной работы в течение полутора часов. Таким образом, система мониторинга вполне в состоянии выполнять роль «черного ящика» вычислительной системы.

5. Подсистема хранения данных

Подсистема хранения данных реализована на основе параллельной файловой системы Lustre. Общий объем подсистемы - более 50 Терабайт. Теоретически подсистема должна обеспечивать производительность более 4000 операций ввода-вывода (IOPS) и пропускную способность более 500 Мбайт/с. Узлы вычислителя имеют доступ к хранилищу данных по вспомогательной сети Infiniband QDR.

Заключение

Суперкомпьютеры, высокопроизводительные вычислительные системы, являются одним из приоритетов научно-технического развития. Платформа СКИФ ряда 4 предоставляет существенный задел для развития оригинальных высокопроизводительных

вычислительных установок не только на ближайшие годы, но и на более далекую перспективу. Установка опытного образца в Южно-Уральском государственном университете является первым проектом создания систем такого класса.

Список литературы

1. Абрамов, С.М. СуперЭВМ ряда 4 семейства СКИФ: штурм вершины суперкомпьютерных технологий / С.М. Абрамов [и др.] // Вестник Нижегородского университета им. Н.Н. Лобачевского. - 2009. - № 5. - С. 200-210.