СИСТЕМА ХРАНЕНИЯ И ЗАПИСИ ДАННЫХ СУПЕРКОМПЬЮТЕРА «УРАН»

А.В. Созыкин, А.Ю. Берсенёв, Р.А. Степанов, М.Л. Гольдштейн, А.С. Игумнов

Введение

Параметры системы хранения и записи данных (СХЗД) суперкомпьютера (СК) существенным образом влияют на его производительность в целом. К сожалению, при проектировании и создании СК зачастую не уделяется достаточного внимания сбалансированности параметров СХЗД структуре вычислительного поля. В частности, СХЗД СК «УРАН» [1], решенная традиционным путем (в рамках хост-машины), оказалась недостаточно емкой и производительной при увеличении количества и сложности решаемых на СК «УРАН» задач.

В связи с активным развитием Суперкомпьютерного центра (СКЦ) ИММ УрО РАН - одного из четырех главных ресурсных центров УрФО - возникла необходимость развития СХЗД с целью устранения существующих недостатков.

Актуальность данной работы заключается в том, что создание новой сбалансированной специализированной СХЗД для СК «УРАН», позволит увеличить производительность решения прикладных задач и повысить надежность работы.

В процессе годовой практики активного счета на СК «УРАН» было установлено, что исходная емкость СХЗД СК «УРАН» в 1,8 ТБ явно недостаточна для решения крупных научно-технических задач. С появлением СК «УРАН» в СКЦ пришли пользователи с задачами, требующими более 2 ТБ дискового пространства, что превышает весь имеющийся объем системы хранения. Низкая пропускная способность СХЗД СК «УРАН» привела к снижению общей производительности на задачах, активно использующих дисковую подсистему. Более того, при большой нагрузке на СХЗД происходили сбои, приводящие к полной остановке СК и отказу в обслуживании пользователей.

Целями модернизации СК «УРАН» являются:

- Повышение производительности и увеличение дисковой емкости СХЗД СК «УРАН».
- Организация эффективного и простого в использовании обмена данными между измерительной аппаратурой экспериментальных установок и СК «УРАН».

Для достижения данных целей поставлены следующие задачи:

- Модернизация архитектуры СХЗД СК «УРАН».
- Практическая реализация модернизированной архитектуры.
- Подключение экспериментальных установок к системе хранения для прямой передачи данных на СК «УРАН».
- Тестирование производительности модернизированной СХЗД [2].
- Определение параметров СК «УРАН», позволяющих оптимизировать характеристики системы под специфические требования прикладных задач.

Исходная архитектура СХЗД

Проблемы недостаточной производительности и емкости исходной СХЗД СК «УРАН» вызваны недостатками ее архитектуры (рис.1).

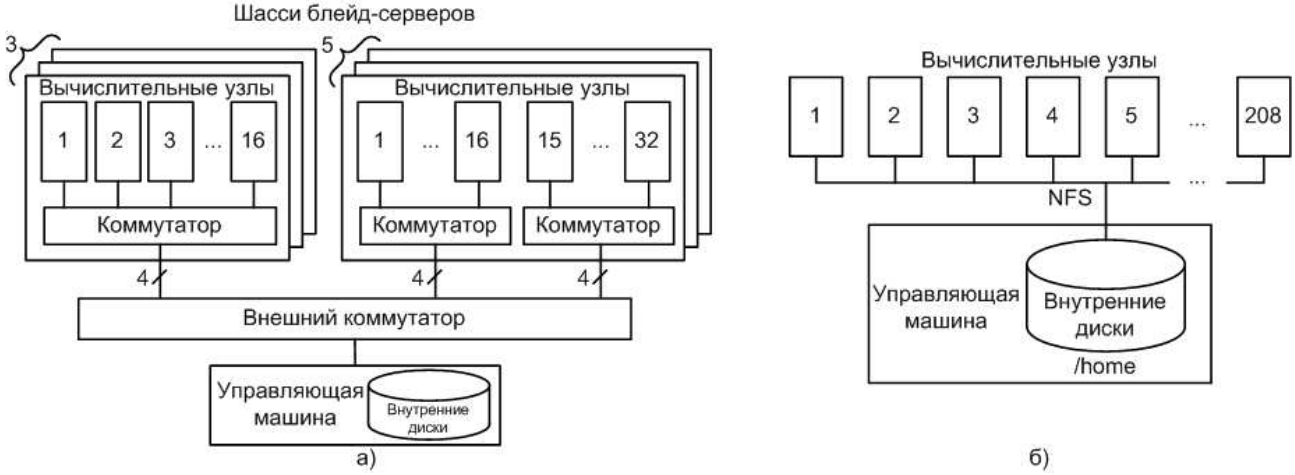

Рис.1. Исходная архитектура СХЗД СК «УРАН»: а) физическая архитектура, б) логическая архитектура

В качестве устройства хранения данных использовались внутренние диски управляющей машины СК

«УРАН», которой является сервер HP Proliant DL180G5 под управлением ОС Linux. Количество внутренних дисков 5 шт., тип SATA, объем 500 ГБ. С помощью внутреннего RAID-контроллера диски объединяются в логический том RAID5 полезной емкостью 1,8 ТБ.

Доступ к дискам вычислительные узлы получают через коммуникационную сеть ввода-вывода, построенную по технологии Gigabit Ethernet и имеющей двухуровневую организацию:

- Первый уровень – коммутаторы Gigabit Ethernet в шасси блейд, к которым подключаются все вычислительные узлы в данном шасси.
- Второй уровень – внешний коммутатор Gigabit Ethernet, объединяющий коммутаторы первого уровня.

Коммутаторы первого и второго уровней соединяются четырьмя каналами GigabitEthernet, объединенными в логический канал для повышения пропускной способности и отказоустойчивости. Управляющая машина подключена к коммутатору второго уровня по одному каналу Gigabit Ethernet.

Логически подключение дисков организовано по протоколу NFS. На внутренних дисках управляющей машины создана файловая система, подключенная к каталогу /home. Эта файловая система экспортируется по NFS и все вычислительные узлы кластера подключают ее также в каталог /home.

Недостатком исходной архитектуры СХЗД СК «УРАН» является совмещение сервером HP Proliant DL180G5 функций устройства хранения и управляющей машины, приводящее к следующим проблемам:

- Отсутствие масштабируемости дисковой емкости, вследствие того, что в сервере HP Proliant DL180G5 нет места для дополнительных дисков.
- Низкая производительность дисковой подсистемы, так как она построена на основе малопроизводительного RAID-контроллера, встроенного в сервер HP Proliant DL180G5, и использует небольшое количество дисков.
- Уменьшенная надежность работы СК «УРАН», связанная с тем, что большая нагрузка на подсистему ввода-вывода приводит к исчерпанию ресурсов управляющей машины. Несмотря на то, что вычислительные узлы продолжают работу, сбой в работе управляющей машины приводит к отказу в обслуживании пользователей, так как запуск задач на выполнение выполняется только через управляющую машину.

Модернизация архитектуры СХЗД

Для устранения недостатков СХЗД СК «УРАН» предложено разделить функции управляющей машины суперкомпьютера и устройства хранения данных. Модернизированная архитектура системы хранения СК «УРАН» показана на рис. 2 и рис. 3.

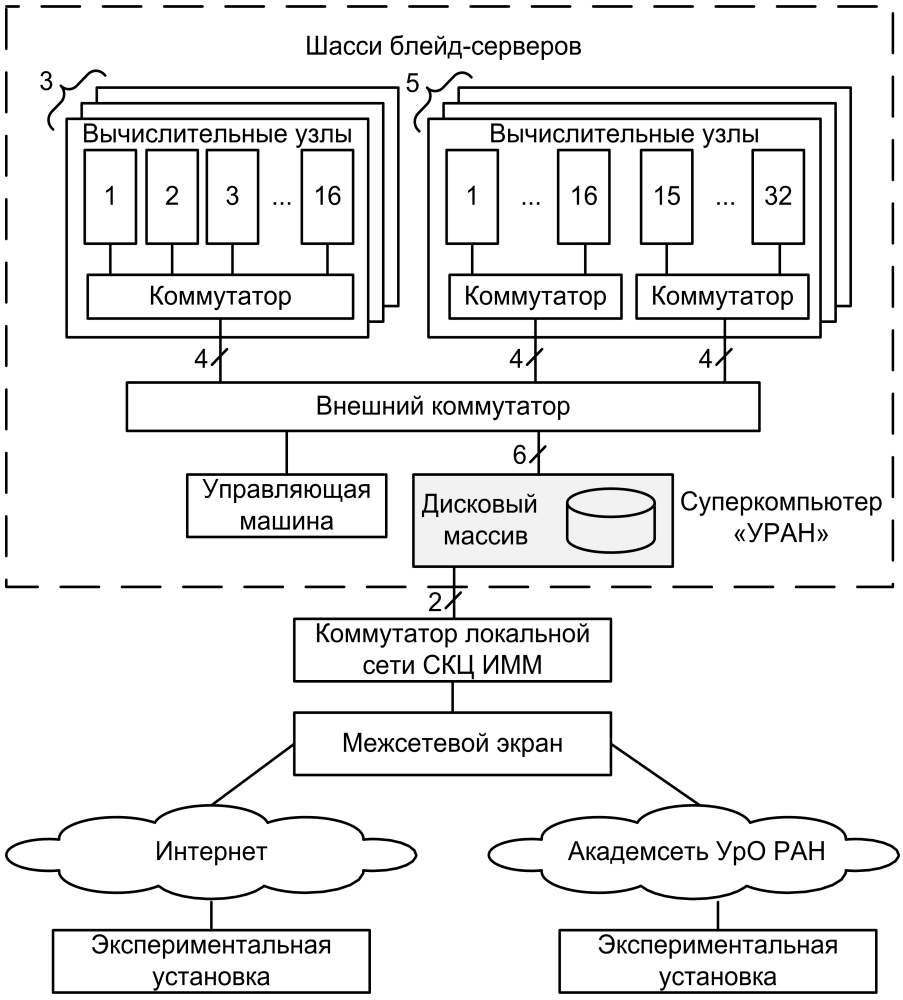

Рис.2. Модернизированная физическая архитектура системы хранения СК «УРАН»

В состав архитектуры введено специализированное устройство хранения данных – дисковый массив. Управляющая машина СК «УРАН» для хранения данных больше не используется.

Дисковый массив включает в себя высокопроизводительные RAID-контроллеры, диски для хранения данных и серверы ввода-вывода. Массив подключается к сети Gigabit Ethernet 8 каналами. При этом 6 каналов используется для подключения к коммуникационной сети ввода-вывода СК «УРАН», и 2 канала для подключения к локальной сети СКЦ ИММ УрО РАН. С помощью этих каналов к дисковому массиву можно подключать удаленные экспериментальные установки институтов УрО РАН через Академсеть УрО РАН и интернет.

Логически дисковый массив представляет собой несколько томов, часть из которых выделена для использования СК «УРАН», а другая часть применяется для передачи на суперкомпьютер данных от экспериментальных установок.

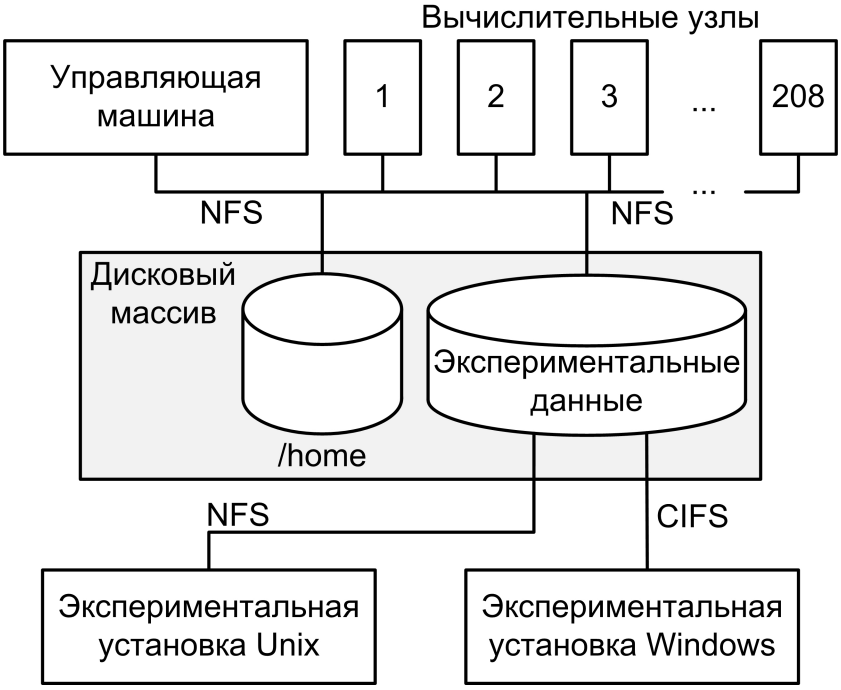

Рис.3. Модернизированная логическая архитектура системы хранения СК «УРАН»

Файловая система /home, ранее размещавшаяся на управляющей машине, перенесена на дисковый массив и экспортируется по протоколу NFS. Вычислительные узлы и управляющая машина по-прежнему подключают эту файловую систему в каталог /home.

Логический том для экспериментальных данных экспортируется одновременно по разным сетевым протоколам, включая NFS, CIFS и FTP. Через коммуникационную сеть ввода-вывода по протоколу NFS логический том подключается к вычислительным узлам СК «УРАН». В то же время данный логический том может быть подключен через интернет или Академсеть УрО РАН к компьютерам экспериментальных установок. При этом компьютеры под управлением ОС Unix могут использовать протокол NFS, а под управление ОС Windows – CIFS. За счет того, что логический том одновременно подключен как к вычислительным узлам СК «УРАН», так и к экспериментальным установкам, данные, получаемые в ходе эксперимента, после записи на дисковый массив сразу будут доступны для обработки на СК «УРАН».

Модернизированная архитектура обладает следующими преимуществами:

- Высокая масштабируемость путем установки дополнительных дисков в массив.
- Высокая производительность дисковой подсистемы за счет использования мощных RAID-контроллеров и большого количества дисков.
- Увеличенная надежность работы СК «УРАН», так как нагрузка на подсистему ввода-вывода не приводит к остановке управляющей машины и отказу в обслуживании пользователей, благодаря разделению функций управляющей машины и устройства хранения данных.
- Возможность прямого высокоскоростного (1 Гб/с) подключения экспериментальных установок к СК «УРАН», за счет большого количества портов в дисковом массиве.

Практическая реализация

Практическая реализация модернизированной архитектуры СК «УРАН» выполнена на основе дискового массива EMC Celerra NS-480, представляющего собой устройство типа Network Attach Storage (NAS), подключающегося к сети Gigabit Ethernet и поддерживающего сетевые протоколы NFS, FTP и CIFS.

Установленная конфигурация содержит 2 аппаратных RAID-контроллера с 8 ГБ кэша каждый, 15 дисков SATA 1 ТВ и 11 дисков FC 300 ГБ. Максимально в массив можно установить 480 дисков до 192 ТБ полезной емкости.

Тестирование производительности

Для оценки увеличения производительности, достигнутого за счет модернизации СХЗД СК «УРАН» было проведено тестирование двух типов: тестирование пропускной способности синтетическим тестом iozone [3] и тестирование решения реальной задачи механики сплошных сред.

Пропускная способность. Для оценки производительности СХЗД чаще всего используют две характеристики: пропускную способность и количество операций ввода/вывода в секунду (IOPS). Высокая пропускная способность важна при выполнении последовательных операций с файлами больших размеров, а количество IOPS – при частом выполнении операций небольшого размера. В СК «УРАН» преобладают последовательные операции с большими файлами, поэтому для оценки производительности СХЗД измерялась пропускная способность. Было протестировано несколько конфигураций дисковых томов, с разным типом диском (FC и SATA) и типом RAID (1 и 5).

Тестирование производили с помощью системы iozone. Количество клиентов – 16, выполняющих операции по чтению и записи файлов объемом 16 ГБ. Объем файлов выбран таким образом, чтобы файл не мог поместиться в кэш файловой системы в оперативной памяти сервера (16 ГБ). Результаты тестирования представлены в табл. 1.

Таблица 1. Результаты тестирования пропускной способности системы хранения СК «УРАН»

Система хранения	Тип дисков	Количество дисков	Тип RAID	Пропускная способность, МБ/с	
				Чтение	Запись
Управляющая машина	SATA, 500 ГБ	5	5	100	47
Дисковый массив	SATA, 1 ТБ	5	5	146	161
		4	1	110	120
	FC, 300 ГБ	5	5	178	160
		4	1	165	153

Тестирование показало, что пропускная способность СХЗД на основе дискового массива EMC Celerra NS-480 выше, чем у системы на внутренних дисках управляющей машины, но выигрыш для операций записи и чтения существенно отличается. Если пропускная способность на операциях записи, полученная на дисковом массиве с логическими томами типа RAID 5 (как с дисками FC, так и с SATA), превышает возможности управляющей машины более чем в три раза, то максимальная пропускная способность на операциях чтения, полученная на дисковом массиве (диски FC, RAID), всего в 1,78 раз выше, чем на управляющей машине. Более высокая пропускная способность записи, а не чтения, объясняется тем, что в массиве EMC Celerra NS-480 для операций записи производителем выделено 4 ГБ кэша, а для операций чтения всего 512МБ. Так как для прикладных задач скорость записи, как правило, важнее, чем скорость чтения, то данные настройки оставлены без изменения.

Другой интересной особенностью, выявленной в процессе тестирования, оказалось более высокая пропускная способность логических томов с RAID 5, чем логических томов с RAID 1. Это можно объяснить небольшим количеством дисков в системе, что не позволяет получить существенный выигрыш логическим томам RAID 1. В то же время производительность RAID 5 увеличена, благодаря аппаратным RAID-контроллерам, которые выполняют вычисление контрольных сумм RAID 5.

Прикладная задача. Для оценки влияния СХЗД на скорость решения задач на СК «УРАН» было выполнено тестирование производительности работы прикладной задачи на внутренних дисках управляющей машины и дисковом массиве. Задача выполняет расчет трехмерных магнитогидродинамических (МГД) течений с аппроксимацией пространственных производных до 7-го порядка точности [4]. Используется явное интегрирование методом Рунге-Кутта-Фельберга 5-го порядка точности с адаптивным шагом. Для тестирования использовалась сетка 128^3, в реальных расчетах применяют сетки до 4096^3. Распараллеливание реализовано путем деления расчетной области на подобласти и использованием MPI для организации межпроцессорного обмена. Время счета задачи, в расчете на количество процессорных ядер, включающее суммарные затраты на математические операции, межпроцессорный обмен и запись результатов, показано на рис. 4.

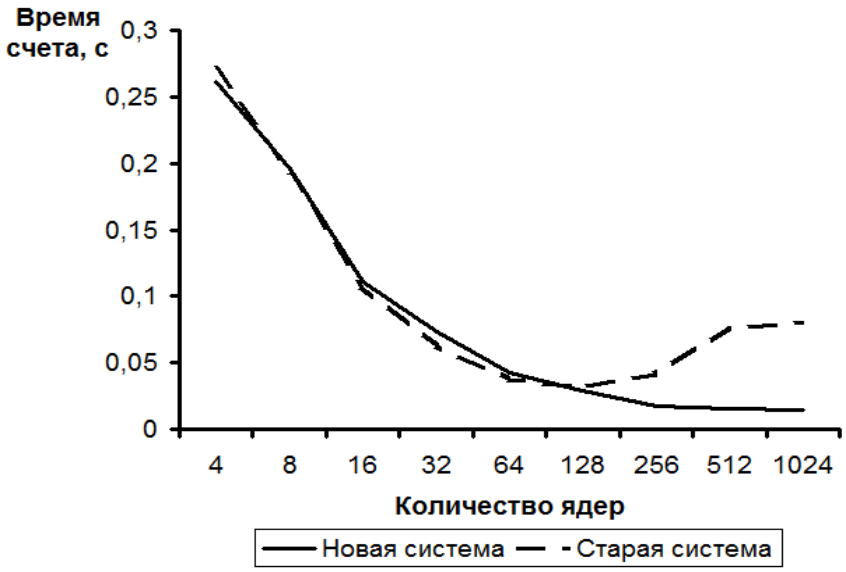

Рис. 4. Время расчета трехмерных МГД течений

Время счета при количестве ядер до 128 одинаково для обеих систем хранения. При использовании дискового массива, увеличение количества процессорных ядер больше 128 ведет к снижению времени счета, а при использовании внутренних дисков управляющей машины, наоборот, к увеличению. Пропускная способность системы хранения при многопоточном режиме работы становится критически важной для эффективности параллельных вычислений. Так при 1024 ядрах время счета с использованием дискового массива меньше в 5,75 раза.

Подключение экспериментальной установки

К модернизированной системе хранения СК «УРАН» была подключена экспериментальная установка измерения полей скорости жидкости или газа в выбранном сечении потока с помощью метода Particle Image Velocimetry (PIV) [5]. Установка находится в Институте механики сплошных сред (ИМСС УрО РАН), г.Пермь. Установка в максимальной конфигурации имеет две видеокамеры, генерирующие последовательность 4-х мегапиксельных изображений с частотой 30 кадров в секунду, что соответствует потоку данных около 500 Мб/с. Изображения, поступающие с установки, обрабатываются пакетом программ, который реализует алгоритм параллельного расчета мгновенных полей скорости. Скорость обработки одной пары изображений зависит от параметров расчета и производительности процессоров. Типичное время счета составляет 10 секунд, для синхронной обработки потребуется порядка 150 процессорных ядер. Вычислительной системы такой мощности в ИМСС УрО РАН нет, поэтому для обеспечения возможности высокоскоростной обработки экспериментальных данных было выполнено подключение установки к СК «УРАН» (рис. 5).

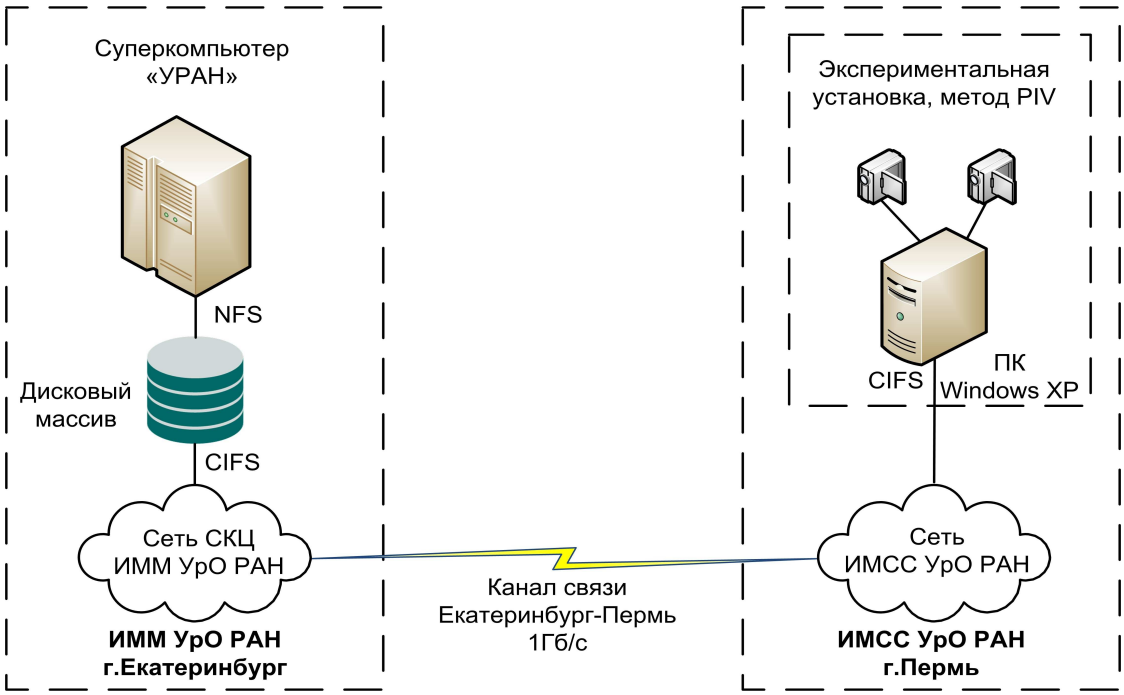

Рис. 5. Схема подключения экспериментальной установки к СК «УРАН»

Подключение выполнено с использованием дискового массива, в котором создан специальный логический том для экспериментальных данных. Логический том одновременно экспортируется в две сети:

- Коммуникационную сеть ввода-вывода СК «УРАН» для подключения к вычислительным узлам.
- Академсеть для подключения к экспериментальной установке через высокоскоростной канал связи 1Гб/с Екатеринбург-Пермь [6].

Экспериментальная установка управляется персональным компьютером с ОС Windows XP. Логический том для экспериментальных данных подключается к ней как сетевой диск по протоколу CIFS. Одновременно этот логический том подключается к вычислительным узлам СК «УРАН» под управление ОС Linux по протоколу NFS. Согласование протоколов CIFS и NFS при одновременном доступе обеспечивает дисковый массив. Таким образом, как только экспериментальная установка записывает данные на дисковый массив, они сразу же становятся доступны на СК «УРАН».

Достоинством предлагаемого подхода является простота в использовании: для подключения к СК «УРАН» в экспериментальной установке не требуется ничего менять, за исключением того, что данные записываются не на локальный диск, а на сетевой.

Заключение

Модернизация СХЗД суперкомпьютера «УРАН» обеспечила следующие преимущества:

- Увеличилась дисковая емкость СК «УРАН», появилась возможность масштабируемости.
- Повысилась производительность СК «УРАН» (как на тестах, так и на реальных задачах).
- Появилась возможность простого подключения экспериментальных установок к СК «УРАН»

Направлениями дальнейшей работы являются:

- Увеличение дисковой емкости СХЗД.
- Тестирование разных версий сетевых протоколов (NFS v3 и v4, CIFS v1 и v2).
- Исследование особенностей передачи данных с удаленных экспериментальных установок.

Работа выполнена при финансовой поддержке РФФИ (грант № 10-01-05006-б) и УрО РАН (грант № РЦП-11-П7).

ЛИТЕРАТУРА:

1. М.Л. Гольдштейн, А.В. Созыкин Инновационно-интеграционная политика суперкомпьютерного вычислительного центра ИММ УрО РАН // Труды международной суперкомпьютерной конференции “Научный сервис в сети Интернет: суперкомпьютерные центры и задачи”, г. Новороссийск, – М.: Изд-во МГУ, 2010. С.81-87.
2. Trayger, A., Zadok, E., Joukov, N., Wright, C. P. A nine year study of file system and storage benchmarking. ACM Transactions on Storage, 4(2), 2008.
3. IOzone Filesystem Benchmark. [Электронный ресурс]. URL: http://www.iozone.org/ (Дата обращения

15.05.2011).
4. Степанов Р.А., Чупин А.В., Фрик П.Г. Винтовое МГД динамо в торе // Вычислительная механика сплошных сред. 2008. Т. 1, № 1. С. 109 – 117.
5. Adrian R.J. Scattering particle characteristics and their effect on pulsed laser measurements of fluid flow: speckle velocimentry vs. particle image velocimentry // Appl. Opt. 1984. Vol. 23. Pp. 1690-1691.
6. А.Г. Масич , Г.Ф. Масич GIGA UrB RAS подход к LambdaGrid парадигмам вычислений // Труды международной суперкомпьютерной конференции “Научный сервис в сети Интернет: суперкомпьютерные центры и задачи”, г. Новороссийск, – М.: Изд-во МГУ, 2010. С.4-12.