ОБРАБОТКА НА СУПЕРВЫЧИСЛИТЕЛЕ ПОТОКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Р.А. Степанов, А.Г. Масич, А.Н. Сухановский, В.А. Щапов, А.С. Игумнов , Г.Ф. Масич

Аннотация. Проведение современных газо- и гидродинамических экспериментальных исследований невозможно без качественных измерений полей скорости. Бесконтактные методы измерений (PIV, PTV, PLIF), основанные на обработке изображений, занимают ведущую позицию. Точность измерений зависит от характеристик видеокамер (разрешение и частота кадров) и от возможностей алгоритмов расчета. Ограниченность вычислительной производительности подключенных к экспериментальной установке компьютеров во многом сдерживает развитие математического аппарата и возможности проведения эксперимента. Перенос вычислений на многопроцессорные системы позволит использовать ресурсоемкие, но высокоточные алгоритмы, избегать хранения гигантских объемов избыточной информации, обрабатывать измерения «на лету» и проводить эксперименты с обратной связью. Это направление работ получило название «Распределенный PIV» [1], поддержано грантом РФФИ № 11-07-96001-р_урал_а и проектом Региональной целевой программой УрО РАН № РЦП-11-П10.

1. Постановка задачи экспериментального и численного исследования переноса мелкомасштабной спиральности и завихренности крупномасштабным адвективным потоком в плоском слое. Адвективный поток создается распределенным нагревом в центральной части дна модели. В данной задаче диаметр области нагрева существенно превышает толщину слоя. Генерация мелкомасштабных закрученных конвективных струй, характеризуемых ненулевым значением спиральности, будет производиться за счет точечных источников тепла. Эта задача имеет интересное приложение в связи с изучением начальной стадии формирования тропических циклонов. Вопрос о том, как из состояния вихревого возмущения происходит заметная интенсификация вихря, до сих пор остается открытым. Существенную роль могут играть так называемые «горячие башни» - интенсивные конвективные струи в области формирования циклона [2, 3]. Существует предположение, что эти конвективные струи могут приводить к концентрации завихренности, формированию мезомасштабных вихрей, которые, объединяясь, могут существенно усиливать крупномасштабное вихревое движение. Экспериментальные исследования в упрощенной, модельной постановке могут помочь в понимании данной задачи.

2. Экспериментальная установка. *Измерительная часть установка PIV* состоит из следующих компонент. Мощный импульсный лазер Quantel используется для освещения измерительной области лазерным ножом регулируемой толщины. Две цифровые камеры ВИДЕОСКАН-11002-2001 [4] на базе CCD матрицы разрешением 4004*2671 (пикселей) и разрядностью АЦП 12 бит/пиксель регистрируют мгновенные изображения с кадровой частотой от 1.2 до 5.8 Гц. Синхронизатор и персональный компьютер с программным обеспечением Actual Flow управляют экспериментом и сохраняют полученные с камер данные [5]. Камеры подключены к компьютеру через карту видеозахвата (PCI-контроллер VS2001). Компьютер работает под управлением операционной системы Windows XP, оснащен четырехядерным процессором (Phenom 2X4) с тактовой частотой 3.2 Ггц, имеет объем оперативной памяти 4 ГБ и дисковой памяти 1TB, подключен двумя портами 1GE к коммутирующему оборудованию ИМСС УрО РАН. На компьютере с ПО Actual Flow возможна также последующая обработка данных и визуализация результатов обработки, однако большой объем получаемых данных, а также ресурсоемкость расчетных программ не позволяют проводить обработку в реальном времени [5].

Экспериментальный фрагмент установки (рис. 1) представляет собой прямоугольную кювету (70x70 см2) с цилиндрической вставкой. Такая конфигурация позволяет обеспечить осевую симметрию и избежать оптических искажений в вертикальных сечениях. Конвективное течение создается набором распределенных нагревателей, что позволяет задавать структуру течения. Кювета располагается на вращающемся столе, который обеспечивает стабильное вращение в широком диапазоне угловых скоростей Ω. В кювету заливается силиконовое масло с кинематической вязкостью $5x10^{-6}$ $м^2/с$. Трассерами являются стандартные PIV частицы диаметром 20 мкм. Участники проекта имеют успешный опыт применения PIV систем при исследовании конвективных потоков в замкнутых объемах [6, 7].

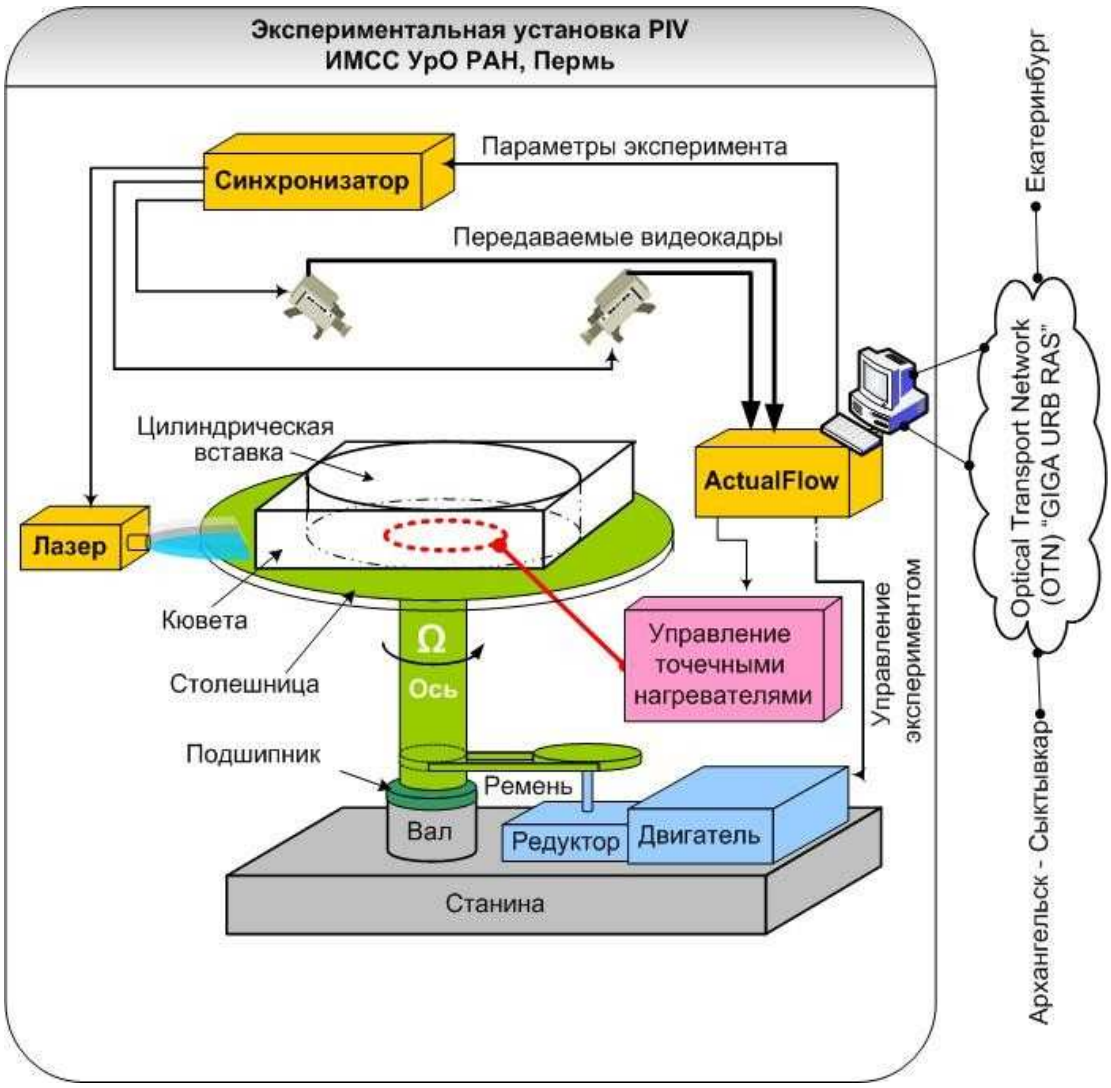

Рис. 1. Экспериментальная установка

3. Параллельный алгоритм обработки экспериментальных данных. Алгоритм параллельного расчета мгновенных полей скорости по изображениям, поступающих с измерительной установки, реализуется в двух вариантах. Первый вариант состоит в распределении непрерывного потока данных по процессорам. При этом каждый процессор выполняет последовательный алгоритм определения поля скорости по одной паре изображений во всей расчетной области. Второй вариант заключается в параллельном расчете на нескольких процессорах поля скорости для одной пары изображений. В этом случае целесообразна передача одной пары изображений от удаленной экспериментальной установки до внутренней сети кластера и последующее дублирование этой пары в оперативную память каждого процессора средствами межузловой сети кластера. Непосредственное распараллеливание расчетного алгоритма делается стандартным способом деления расчетной области на подобласти. Итерации циклов кросскорреляционного алгоритма являются информационно независимыми. Минимальный межпроцессорный обмен необходим лишь на стадии адаптационного уточнения.

Оба подхода параллельной обработки данных имеют свои преимущества и недостатки. Основное удобство использования первого подходя состоит в непосредственном применении любого последовательного пусть то стандартного, адаптивного или вейвлет кросскорреляционного алгоритма. Межпроцессорный обмен отсутствует. Недостатком является ограничение минимального времени обработки одной пары изображений временем работы последовательного алгоритма. Это может быть критично при использовании высокоточных и ресурсозатратных алгоритмов, а также при обработке данных с высоким разрешением. Так же отметим ограниченную масштабируемость. Максимальное число процессоров не может превышать отношение времени работы последовательного алгоритма и времени между двумя последовательными парами изображений. Второй способ требует доработки в плане организации межпроцессорного обмена, но обладает практически неограниченными возможностями по минимизации времени обработки одной пары изображений. Окончательный выбор метода имеет смысл делать при всех известных условия проведения эксперимента.

4. Идеальная модель обмена интенсивным потокам данных. Используется предложенная в [10]

идеальная модель (рис. 2.), суть которой — прямой доступ к оперативной памяти вычислительных узлов многопроцессорной системы. Задержка обмена данными между системами имеет две компоненты: задержки в оконечных системах и скорости передачи и распространения сигналов по протяженным оптическим линиям связи. Спектральное уплотнение сигналов в оптической магистрали (OTN) решает задачу скоростной передачи данных между взаимодействующими системами. Задержка приема/передачи в оконечных системах определяется скоростью порта и временем поступления данных от Ethernet порта сетевого адаптера (NIC) в буфер приложения и, наоборот, из буфера приложения в сеть (порт адаптера). Скорости портов непрерывно растут (1-10-100GE) и основная проблема заключается в передаче данных приложению. Эта задержка возникает как на передающей, так и на приемной стороне и определяется внутренней сущностью NIC. В качестве I/O портов связываемых по OTN оконечных систем целесообразно использование интеллектуальных NIC карт (Intelligent Ethernet adapter), которые аппаратно поддерживают стеки протоколов передачи данных (TOE NIC - TCP Offload Engine) и технологии удаленного прямого доступа к памяти (R-NIC) для разгрузки CPU узлов в связи с переходом на скорости 10-100GE.

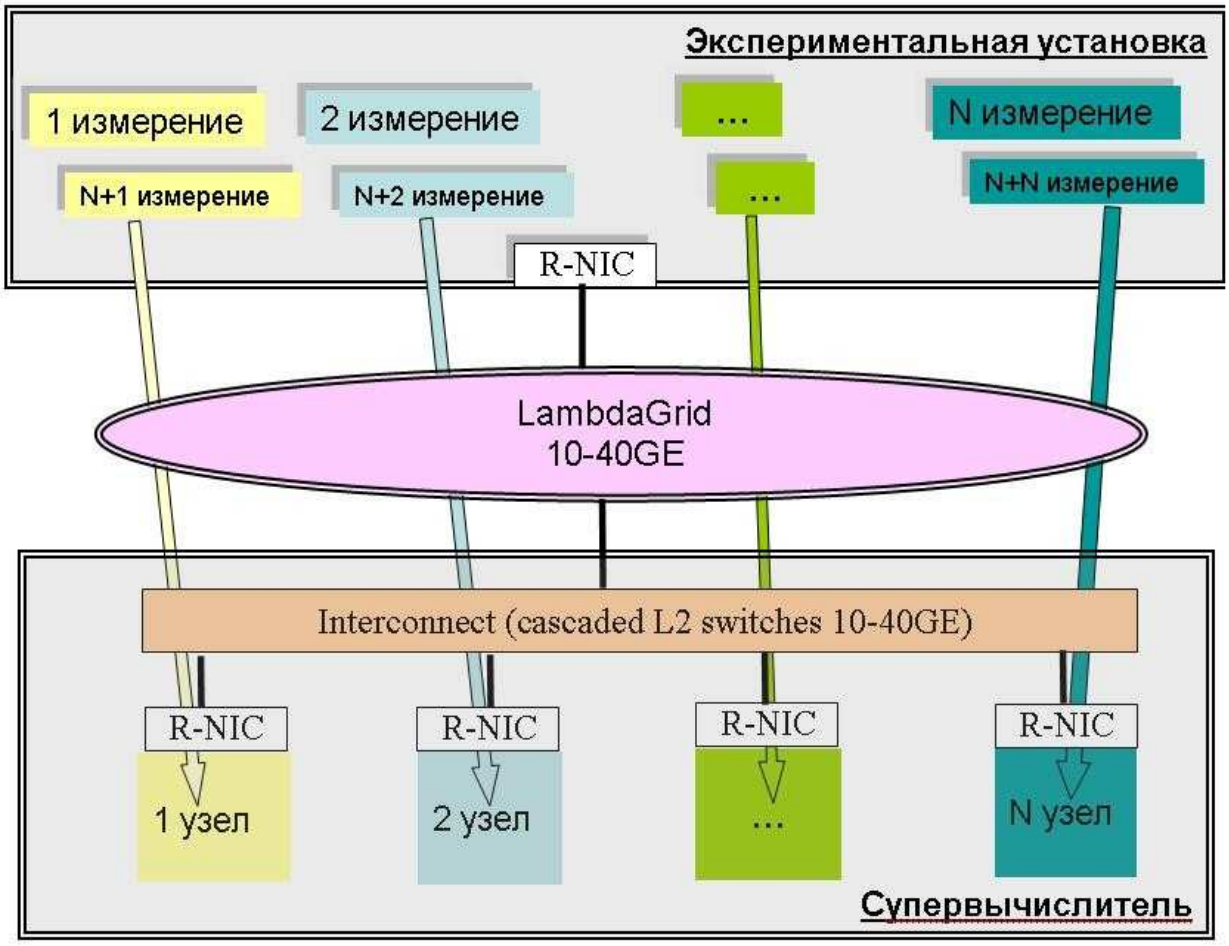

Рис. 2. Идеальная модель обмена интенсивным потоком данных

5. Коммуникационный фрагмент инфраструктуры использует оптическую транспортную сеть (OTN – optical transport network), создаваемую в рамках «Инициативы GIGA UrB RAS» [8, 9] для нужд научно-образовательного сообщества территории от Архангельска до Екатеринбурга. Реализованный участок Пермь-Екатеринбург построен по технологии Гигабит Ethernet (GE) на «темном» оптическом волокне. Обеспечение гибкости многоцелевого использования скоростных коммуникаций для нужд конкретных проектов достигается построением различных VLAN по транковым соединениям L2 коммутирующего оборудования. Как показано на рисунке 3 по VLAN A обеспечена L2 коннективность экспериментальной установки PIV в ИМСС УрО РАН с суперкомпьютером URAN в ИММ УрО РАН, а VLAN B используется для установления пиринговых отношений между маршрутизирующим оборудованием сопрягаемых IP-сетей передачи данных.

Таким образом, инфраструктура коммуникаций позволяет соединять оконечные системы как на уровне L2 OSI RM, в нашем случае через коммутируемый Ethernet, так и на уровне L3 OSI RM, т. е. через маршрутизирующее оборудование IP сетей передачи данных. Модели распределения IP-адресного пространства (Private и Public) и последующего их назначения портам оконечных систем позволяют увеличить число способов и технологии организации их взаимодействия. А вводимая в эксплуатацию на этом участке система плотного волнового мультиплексирования (DWDM) позволит отрабатывать перспективные лямбда грид парадигмы распределенных вычислений.

Локальное тестирование разрабатываемых протоколов передачи данных и алгоритмов обработки данных выполняется на МВС-1000/16П, расположенной в ИМСС УрО РАН.

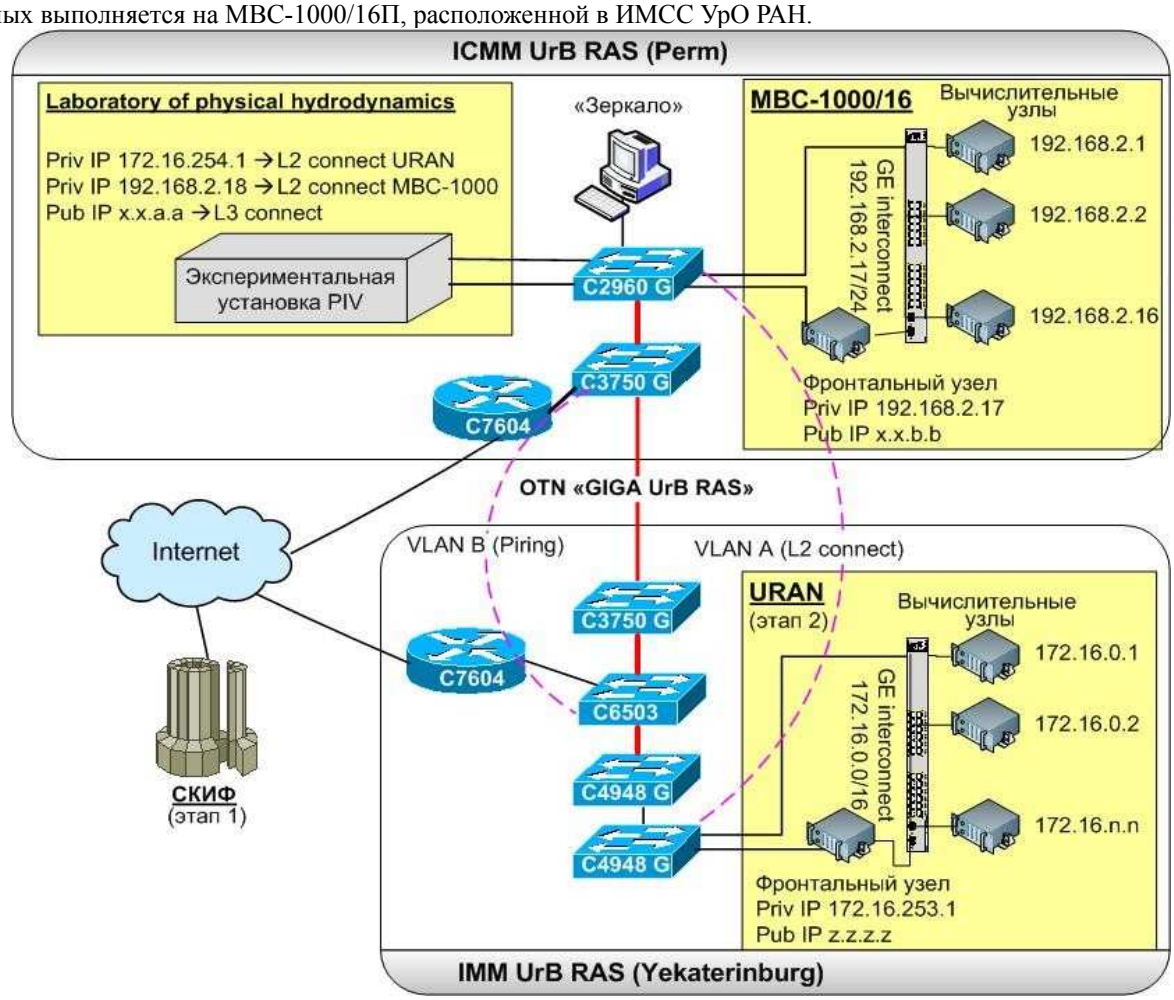

Рис. 3. Коммуникационный фрагмент инфраструктуры

6. Каналы ввода/вывода суперкомпьютера. Разрабатываемая инфраструктура согласно идеальной модели требует подключения к внутреннему interconnect вычислителя удаленной экспериментальной (или любой другой) установки. Наша попытка реализовать средствами MPI взаимодействие платформ UNIX и Windows была не успешна. Поэтому для поддержки многоплатформенности сопрягаемых устройств и минимальных доработок в оконечных системах используется описанный ниже «протокольный» подход. В процессе работы над проектом реализованы и апробированы две схемы подключения.

Согласно первой схеме ЭУ подключена к внутреннему interconnect вычислителя средствами VLAN в единый Ethernet сегмент. ЭУ выделяется свободный Private IP-адрес из единого блока IP-адресов внутреннего interconnect вычислителя. Это не классическая схема доступа пользователей к супервычислителю и, пожалуй, психологически трудно воспринимаемая администраторами этих систем.

Во второй схеме ЭУ подключается к фронтальной машине вычислителя по VLAN или через Public Интернет. Для этого на фронтальной машине выполняется PAT (Port Address Translation) трансляция Private IP адресов вычислительных узлов в один Public IP адрес с разными портами прикладных процессов, запускаемых на вычислителе. Этим достигается возможность доступа прикладных процессов (вычислительных узлов) к ЭУ через фронтальную машину. Такое решение ближе к классическим системам доступа пользователей к системам пакетной обработки и психологически более приемлемо. Однако узким местом становится фронтальная машина, транслирующая через себя интенсивный поток данных. Тем не менее, выполненные измерения скорости обмена между тестовой машиной и вычислителем URAN через фронтальную машину по каналу 1 Гбит/с с помощью запуска программы iperf на 24 узлах при 10 TCP соединениях с каждого узла вычислителя показали суммарную пропускную способность 850 Мбит/с, что практически не отличается от пропускной способности без использования NAT/PAT.

7. Протокол взаимодействия оконечных систем. Взаимодействие оконечных систем осуществляется, с использованием разработанного протокола передачи данных, получившего название «Протокол PIV». Разработанный протокол PIV базируется на идее отказа от однозначного отображения измерений на вычислительные узлы [11].

В случае наличия единого источника данных, который соответствует экспериментальной установке,

отказ от однозначного отображения измерений на вычислительные узлы позволил реализовать схему, в которой вычислительные узлы самостоятельно запрашивают данные для расчетов у экспериментальной установки. Это позволило упростить процедуру распределения экспериментальных данных на вычислительные узлы до процедуры, работающей по схеме: «отдай готовые для расчета данные первому обратившемуся узлу». Помимо этого отказ от однозначного отображения позволит:

- отказаться от синхронизаций вычислительных узлов между собой перед обработкой каждого измерения;
- изменять число задействованных узлов непосредственно во время эксперимента, добавляя в случае необходимости вычислительные мощности;
- минимизировать объем потерянной информации в случае выхода из строя одного или нескольких вычислительных узлов.

Протокол PIV является протоколом прикладного уровня, работает по схеме запрос-ответ и базируется на протоколе TCP. Текущая архитектура позволяет использовать вместо протокола TCP любой протокол, работающий в потоковом режиме и гарантирующий доставку данных. Протокол PIV позволяет обмениваться сообщениями (элементы пары запрос-ответ), состоящими из нескольких блоков бинарных данных. В одном пакете протокола можно передать от нуля до 65535 блоков, каждый из которых может иметь размер до 4 Гбайт. Поля заголовка пакета протокола кодируются в сетевом порядке байт. Формат пакета протокола PIV приведен на рисунке 4.

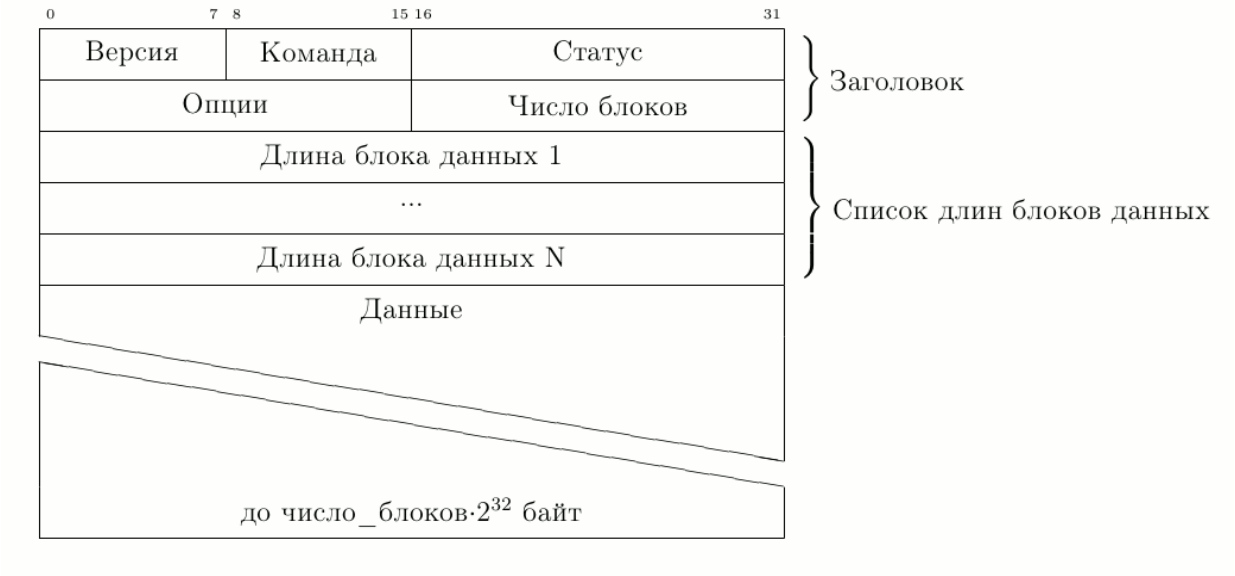

Рис. 4. Формат пакета протокола PIV

Текущая версия протокола поддерживает три типа пакетов (поле «Команда» протокола PIV):

- GET – предназначен для запроса данных вычислительными узлами у ЭУ;
- POST – предназначен для передачи данных с вычислительных узлов на ЭУ;
- RESPONSE – пакет ответа на запрос вычислительного узла.

Информация о подтипе RESPONSE-пакета кодируется в поле статуса. Через поле статуса вычислительным узлам сообщается успешно ли был обработан запрос или нет. В случае неуспеха конкретное значение поля статуса определяет произошедшую ошибку.

В текущей реализации определены 4 группы статусов:

- 1xx – информационные статусы. Используются для информирования клиента о том, что сервер активен или о том, что новых данных нет, но они будут.
- 2xx – статусы этой группы показывают, что запрошенное действие выполнено успешно.
- 4xx – ошибка клиента. Статусы этой группы показывают, что от клиента был получен некорректный запрос.
- 5xx – ошибка сервера. Статусы этой группы показывают, что при обработке запроса возникла внутренняя ошибка сервера.

Процесс получения данных вычислительным узлом состоит из следующих этапов:

1. Открытие соединения с сервером.
2. Отправка запроса GET.
3. Получение ответа от сервера. Если получено сообщение о прекращении работы – переход на шаг 8.
4. Обработка полученных от сервера данных.

5. Отправка на сервер результатов расчета в запросе POST.
6. Получение подтверждения об успешном приеме.
7. Переход на шаг 2.
8. Закрытие соединения с сервером.

8. Результаты измерений. Анализ эффективности разрабатываемых архитектурных решений по передаче интенсивного потока экспериментальных данных проводится путем обработки как log файлов, так и дампа трафика, снимаемого программным обеспечением wireshark. Для записи дампа трафика используется два подхода. В первом случае wireshark запускается непосредственно на компьютере экспериментальной установки. Недостатком этого способа являлся значительный рост нагрузки на процессор и дисковую подсистему ЭУ, что негативно отражалось на процессе передачи данных. Во втором случае трафик экспериментальной установки, средствами коммутатора Cisco C2960G (port mirroring) зеркалируется на отдельный сервер, который занимается только перехватом и записью дампа трафика. Использование зеркалирования трафика позволяет снять нагрузку от процесса записи трафика на компьютер экспериментальной установки, и перенести ее на специальный выделенный сервер.

Первичный анализ и визуализация трафика проводятся средствами Wireshark. Для анализа строились графики зависимости скорости передачи данных в каждой задействованной TCP-сессии от времени. Это позволяет наблюдать протокольные процедуры каждого соединения, степень загрузки канала связи и как следствие эффективность используемых протоколов. На рисунке 5 приведен график скоростей передачи данных между экспериментальной установкой и вычислительными узлами суперкомпьютера.

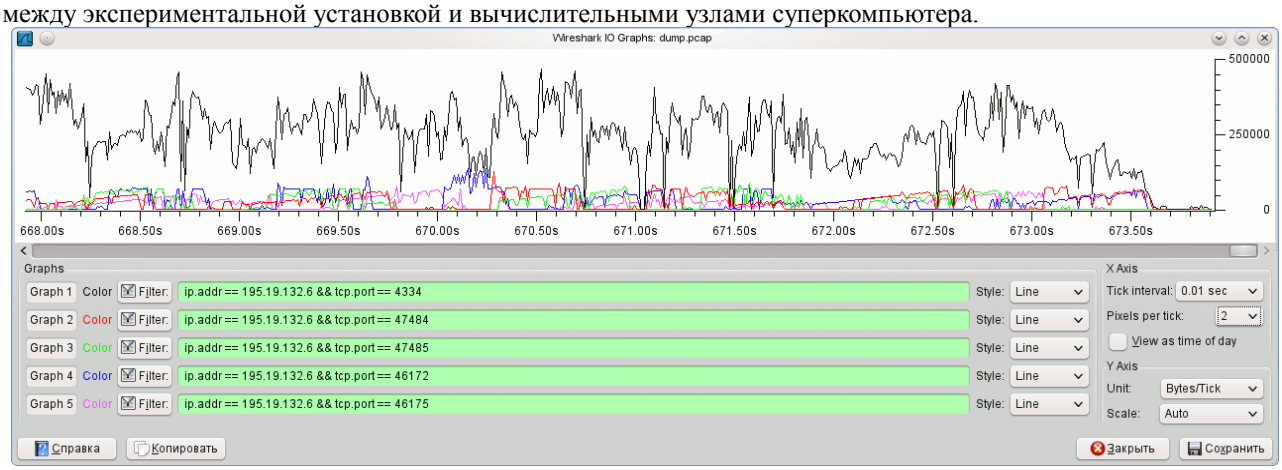

Рис. 5. Визуализация трафика четырех TCP-соединений средствами Wireshark

Черной линией на графике показана суммарная скорость передачи данных, а оставшиеся линии показывают скорости передачи данных в четырех из 10 используемых TCP-сессий.

9. Заключение. Созданная инфраструктура позволяет выполнять комплекс работ по передаче и обработке в реальном времени интенсивных потоков экспериментальных данных, является прототипом создаваемой в Уральском отделении РАН киберинфраструктуры региона, фундаментом которой являются высокоскоростные оптические сети. Успешность разрабатываемых решений определяется новыми парадигмами распределенных вычислений и как следствие новыми постановками научных задач (приложений). Это обстоятельство повлияло на состав участников проекта и данной публикации, в котором присутствуют экспериментаторы, архитекторы многопроцессорных систем и сетевых технологий, профессионалы по системному и параллельному программированию.

ЛИТЕРАТУРА:

1. Р.А. Степанов, А.Г. Масич, Г.Ф. Масич. Инициативный проект “Распределенный PIV" // Научный сервис в сети Интернет: масштабируемость, параллельность, эффективность: труды Всероссийской суперкомпьютерной конференции – М.: Изд-во МГУ, 2009. – С. 360-363. (ISBN 978-5-211-05697-8).
2. Eric A. Hendricks, Michael T. Montgomery and Christopher A. Davis. 2004: The Role of “Vortical” Hot Towers in the Formation of Tropical Cyclone Diana (1984). Journal of the Atmospheric Sciences: Vol. 61, No. 11, pp. 1209–1232.
3. M.T. Montgomery, M.E. Nicholls, T.A. Cram and A.B. Saunders. 2006: A Vortical Hot Tower Route to Tropical Cyclogenesis. Journal of the Atmospheric Sciences: Vol. 63, No. 1, pp. 355–386.
4. Видеоскан спецификация. http://videoscan.ru/page/802. 25.03.2011
5. Е.К. Ахметбеков, А.В. Бильский, Ю.А. Ложкин, Д.М. Маркович, М.П. Токарев, А.Н. Тюрюшкин. Система управления экспериментом и обработки данных, полученных методами цифровой трассерной визуализации (ActualFlow). Вычислительные методы и программирование, 2006. - Т.7, С. 79-85.
6. В.Г. Баталов, А.Н. Сухановский, П.Г. Фрик. Экспериментальное исследование спиральных валов в

адвективном потоке, натекающем на горячую горизонтальную поверхность. Известия РАН. Механика жидкости и газа. №4. 2007. С. 50-60.
7. Batalov V., Sukhanovsky A. and Frick P. Laboratory study of differential rotation in a convective rotating layer // J. Geophys.Astrophys.Fluid Dynam. 104: 4, pp. 349 — 368, 2010. – DOI: 10.1080/03091921003759876.
8. А.Г. Масич, Г.Ф. Масич. Инициатива GIGA UrB RAS // Совместный вып. журн. “Вычислительные технологии” и журн. “Вестник КазНУ им. Аль-Фараби”. Сер. “Математика, механика, информатика” №3 (58). По материалам Междунар. конф. “Вычислительные и информационные технологии в науке, технике и образовании”. - Казахстан, Алматы.-2008.-Т.13.- Ч. II. -С. 413-418 (ISSN 1560-7534).
9. А.Г. Масич, Г.Ф. Масич. От «Инициативы GIGA UrB RAS к Киберинфраструктуре УрО РАН. // Вестник Пермского научного центра (октябрь-декабрь 4/2009). – Пермь: Изд-во ПНЦ УрО РАН, 2009. С. 41-56 (ISSN 1998-2097).
10. А.Г. Масич, Г.Ф. Масич. GIGA UrB RAS подход к LambdaGrid парадигмам вычислений // Научный сервис в сети Интернет: суперкомпьютерные центры и задачи: Труды Международной суперкомпьютерной конференции. – М.: Изд-во МГУ, 2010. С. 4-11. ISBN 978-5-211-05916-0.
11. А.Г. Масич, Г.Ф. Масич, Р.А. Степанов, В.А. Щапов. Скоростной I/O-канал супервычислителя и протокол обмена интенсивным потоком экспериментальных данных // Сб. тез. докл. X международной конференции высокопроизводительные параллельные вычисления на кластерных системах “HPC-2010” – Пермь: Изд-во ПГТУ, 2010. - Т. 2. С. 119–128. (ISBN 978-5-398-00506-6).