Вычислительные ресурсы УрО РАН. Состояние и перспективы*

М.Л. Гольдштейн[1], А.В. Созыкин[1,2], Г.Ф. Масич[3], А.Г. Масич[3]

Институт математики и механики УрО РАН[1], Уральский федеральный университет[2], Институт механики сплошных сред УрО РАН[3]

Рассматриваются вопросы создания киберинфраструктуры УрО РАН - распределенных вычислительных ресурсов, информационных систем и устройств хранения данных, объединенных высокоскоростными каналами связи. Киберинфраструктура служит основой для проведения научных исследований на мировом уровне. Описываются вычислительные ресурсы суперкомпьютерного центра, подходы к построению высокоскоростной магистральной сети УрО РАН на основе DWDW технологии и особенности построения вычислительной облачной платформы УрО РАН.

1. Введение

Решение фундаментальных и прикладных ключевых задач, обусловливающих научно-технический прогресс в экономике и бизнесе и ведущих к получению новых знаний и информации, требует новых всё более производительных вычислительных систем, скоростных сетей передачи данных и распределенных инфраструктур коллективного пользования. Обеспечение научно-образовательного сообщества Уральского отделения РАН вычислительным инструментарием мирового уровня, отвечающим актуальным потребностям ученых и соответствующим приоритетным направлениям фундаментальных и прикладных исследований – одна из целей стратегии развития Уральского отделения РАН до 2025г [1].

В качестве ключевых этапов развития вычислительной инфраструктуры УрО РАН на сегодняшний день можно выделить:

- развитие суперкомпьютерного вычислительного центра (СКЦ) ИММ УрО РАН до мирового уровня (уровень Т1, вхождение в списки TOP500, Graph500, Green500) [2],
- построение высокоскоростной магистрали, объединяющей вычислительные ресурсы и системы хранения данных Научных центров УрО РАН в городах Архангельск, Сыктывкар, Ижевск, Пермь и Екатеринбург [3],
- интеграция грид-среды УрО РАН с российской грид-средой (проект Министерства связи и массовых коммуникаций России),
- создание распределенной среды высокопроизводительных вычислений [2],
- развитие облачной вычислительной платформы и создание пакета сервисов для пользователей СКЦ [4].

Работы по данным направлениям деятельности в УрО РАН ведутся в Институте математики и механики (ИММ) УрО РАН и Институте механики сплошных сред (ИМСС) УрО РАН.

2. Вычислительные ресурсы

Главный вычислительный ресурс СКЦ - суперкомпьютер (СК) “УРАН” (5 место в рейтинге TOP50, 17-ая редакция от 18.09.2012г.) имеет гибридную архитектуру:

- графические процессоры GPU: Nvidia Tesla M2090 (80 шт.) и M2050 (160 шт.)
- центральные процессоры: Intel Xeon E5450 3 ГГц (416 шт.), Intel Xeon X5675 3,06 ГГц (60 шт.)
- сети: Infiniband DDR (коммуникационная), GigabitEthernet (ввода-вывода и управления)

* Работа поддержана грантами УрО РАН № № 12-П-1-2012, 12-П-1-1029, 12-С-1-1012 и РФФИ № 11-07-96001-р_урал_а

- системное ПО: Linux, MPI (OpenMPI и MPICH), компиляторы gcc и Intel, система запуска задач SLURM

Прикладное ПО: Matlab, Ansys.

Инженерная инфраструктура:

- кондиционеры Liebert CR-V CR035RA – 4 шт.,
- кондиционер Liebert 50 -1шт.
- ИБП APC Smart-UPS VT 40kVA - 4 шт.
- ИБП APC Symmetra 16kVA – 7 шт.

Структурная схема СК приведена на рис. 1.

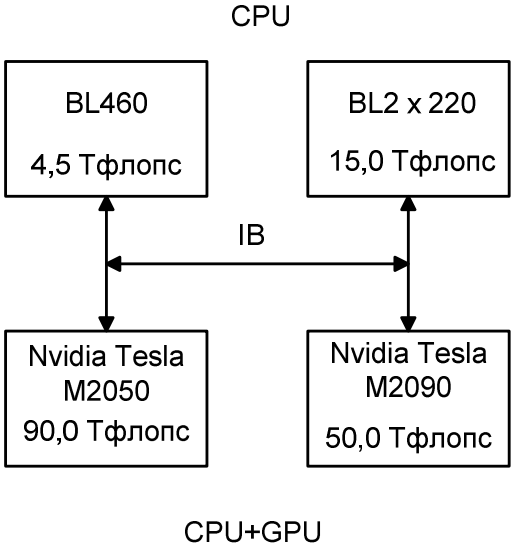

Рис. 1. Структурная схема суперкомпьютера “УРАН” (2012г.)

Развитие СК “УРАН” предполагает дальнейшее наращивание его производительности до 240 Тфлопс (пиковая) и объема дисковой памяти до 190 Тбайт (полная емкость). В качестве элементной базы расширения СК выбраны узлы на основе CPU Intel Xeon X5675 3,06 ГГц + GPU Nvidia Tesla M2090 и дисковые массивы на основе 2-х серверов хранения Supermicro в конфигурации: процессор CPU Intel Socket 1366 Xeon E5607 2.26Ghz tray и 48 жестких дисков HDD Seagate SATA3 3Tb. Структурная схема очередной модификации СК “УРАН” представлена на рис. 2.

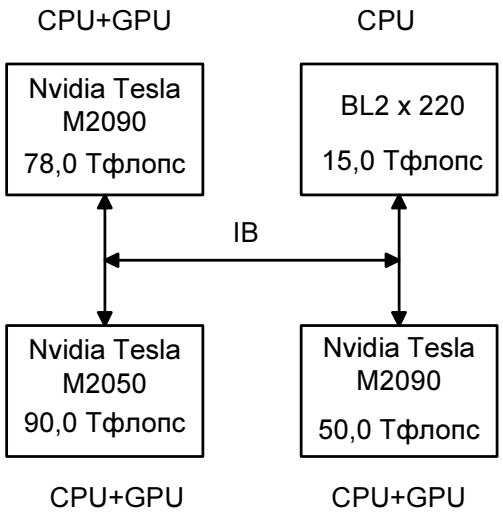

Рис. 2. Структурная схема очередной модификации СК “УРАН” (реализация январь-февраль 2013г.)

Динамика роста ключевых параметров СК “УРАН” приведена на рис. 3.

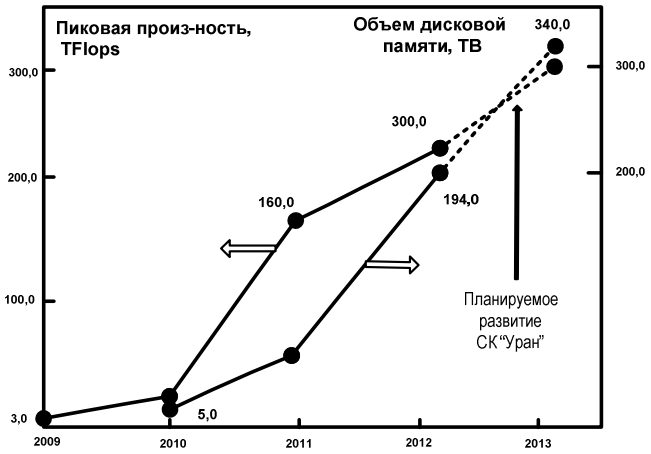

Рис. 3. Динамика роста ключевых параметров СК "УРАН" (2013г.)

Развитие грид-инфраструктуры УрО РАН предполагает создание на первом этапе ВЦ уровня Т2 в ИМСС. Для реализации этого проекта вычислительное поле производительностью 4,5 Тфлопс на модулях BL-460 (см. рис. 1) передаются в ИМСС вместе с дисковой системой хранения на основе сервера хранения Supermicro в конфигурации: процессор CPU Intel Socket 1366 Xeon E5607 2.26Ghz tray и 12 жестких дисков HDD Seagate SATA3 3Tb. Выбор ИМСС обусловлен наличием высокоскоростного (2 х 10Гбит/с + 8 х 1Гбит/с) канала связи между гг. Екатеринбург-Пермь с одной стороны и наличием экспериментальных установок (например, PIV), генерирующих большие потоки данных.

Структурная схема грид УрО РАН (первый этап) представлена на **рис.4.**

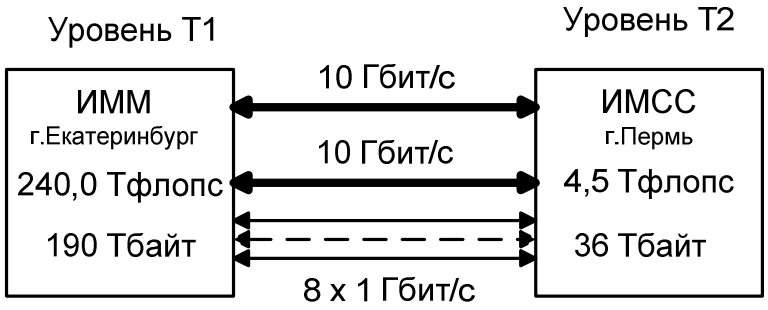

Рис.4. Структурная схема грид УрО РАН (первый этап - 2013г.)

Дальнейшая перспектива развития СКЦ заложена в ряде долгосрочных планов УрО РАН и вкратце выглядит таким образом (**табл. 1**).

Таблица 1. Планы развития вычислительных ресурсов УрО РАН на период 2013-2020гг.

Год	2013	2014	2015	2016	2017	2018	2019	2020
R(peak), в Pflops	0,35- 0,5	0,6-0,8	1,0-1,3	1,5-1,8	2,0-2,3	2,5-3,0	3,5	4,0
Объем дисковой памяти, в PB	0,35- 0,5	0,6-0,8	1,0-1,3	1,5-1,8	2,0-2,3	2,5-3,0	3,5	4,0

3. Распределенная система хранения

Современные экспериментальные установки, которые вместе с СК и магистральными каналами связи формируют киберинфопространсто, в результате проведения научных исследо-

ваний создают большие объемы данных. Эти данные нужно хранить и быстро передавать на СК и вычислительные кластеры для обработки. Для этих целей в УрО РАН создается распределенная система хранения (РСХ). В качестве основы для создания такой системы выбрано программное обеспечение dCache [5], которое активно используется для создания РСХ в крупных международных проектах, таких как Worldwide LHC Computing Grid, European Grid Infrastructure, NorduGrid и др. Отличительной особенностью dCache является поддержка не только специфичных для грид протоколов доступов к данным (SRM, dCap, xRoot), но и открытых протоколов (NFS, HTTP). В результате РСХ на основе dChache можно использовать как в составе грид, так и подключать к обычным вычислительным кластерам и персональным компьютерам.

РСХ УрО РАН строится из отдельных серверов хранения, которые представляют собой серверы стандартной архитектуры под управлением Linux с большим количеством дисков (12-24 шт.). Объединение серверов хранения в единую систему выполняется с помощью dCache. На первом этапе создания РСХ серверы хранения устанавливаются в ИММ УрО РАН и ИМСС УрО РАН (рис. 5.), в дальнейшем планируется установка серверов хранения в других регионах присутствия УрО РАН.

Рис. 5. Схема распределенной системы хранения данных УрО РАН (первый этап)

На первом этапе создания к РСХ подключаются суперкомпьютер «УРАН» и вычислительный кластер в ИМСС УрО РАН по протоколу NFS. Также апробируется подключение к РСХ экспериментальных установок в ИМСС УрО РАН. В дальнейшем, при подключении СКЦ ИММ УрО РАН к грид-проектам, планируется использование грид-протоколов РСХ.

4. Магистральные каналы связи

Ключевой вопрос построения скоростных научно-образовательных сетей – создание собственных или аренда существующих каналов связи. Выполненные в ИМСС УрО РАН исследования позволили найти экономически эффективные решения построения собственной скоростной оптической магистрали «Архангельск–Сыктывкар–Ижевск–Пермь–Екатеринбург», названной «Инициатива GIGA UrB RAS» [3]. Отличительная особенность этих решений - преодоление сложившейся в России негативной практики аренды дорогостоящих каналов связи путем использования темных волокон (dark fiber) в магистралях национальных операторов связи и использовании собственного оборудования спектрального уплотнения каналов. Цель – достижение стратегической обеспеченности научно-образовательного сообщества территории в коммуникационном сервисе.

Рис. 6. Инициатива GIGA UrB RAS

4.1. Этапы становления проекта

2008 год – в ИМСС УрО РАН разработан эскизный проект «Инициатива GIGA UrB RAS» (рис. 5) [3].

2009 год – в работе [6] сформулирована LambdaGrid парадигма распределенных вычислений, а в УрО РАН сформирован проект ГИГА, включающий помимо создания скоростной оптической сети, развитие суперкомпьютерных ресурсов, хранилищ данных и систем визуализации.

2010 год – на ИМСС УрО РАН возложено исполнение проекта ГИГА в части создания скоростной научно-образовательной оптической магистрали «Архангельск–Екатеринбург» [7]. В этом же году введен в эксплуатацию участок магистрали «Пермь–Екатеринбург» на скорости 1 Гбит/с по технологии Ethernet на двух «темных» оптических волокнах.

2012 год – выполнена структурная оптимизация DWDM тракта в рамках выделенного бюджета и проведена пуско-наладка DWDM системы 20 Гбит/с (2 λ канала по 10 Гбит/с) на участке «Пермь-Екатеринбург», с возможностью масштабирования до 16 лямбда каналов по 10-40Гбит/с) [8].

4.2. DWDM тракт Пермь-Екатеринбург

Используется DWDM оборудование компании ECI-Telecom [9]: платформы XDM-2000 на оконечных узлах и XDM-40 на промежуточных узлах. DWDM мультиплексоры обеспечивают возможность наращивания до 16 λ-каналов любой емкости каждый (10/40 Гбит/с). Комплектация установленного оборудования обеспечивает передачу двух λ-каналов по 10 Гбит/с в каждом со следующими характеристиками:

- Первый λ-канал 10 Гбит/с соединяет по схеме точка-точка коммутаторы CESR AS9215 с 24-мя портами GE и 4-мя портами 10GE с соответствующими «цветными» транспондерами для сопряжения с DWDM оборудованием.
- Структура второго λ -канала 10 Гбит/с: гарантированная и независимая передача 8x1GE каналов с терминацией в клиентские порты с применением соответствующих SFP модулей.

Спецификация промежуточных узлов обеспечивает только оптическое усиление сигнала, однако закладываемое DWDM оборудование предполагает возможность установки ROADM мультиплексоров для выделения и маршрутизации λ -каналов не только на оконечных, но и на промежуточных узлах. Система мониторинга и управления магистральной ВОЛС (Sun Fire V245 и программное обеспечение LightSoft) располагается в основном центре управления сетью в ИМСС УрО РАН (Пермь).

Настраиваемый коммуникационный сервис для нужд конкретных проектов и инициатив предоставляет следующие классы услуг передачи данных:

- λ услуга – предоставление в пользование фиксированного диапазона длин волн ("лямбда") на участке DWDM магистрали;
- FrameNet услуга – представление двухточечного или многоточечного Ethernet канала связи между Ethernet портами DWDM магистрали;
- PacketNet услуга – присоединение к Public Интернет;

λ услуга предусматривает, что оконечные точки канала находятся в двух различных узлах DWDM магистрали. Оконечными точками канала являются интерфейсные платы узлов DWDM магистрали. Эта услуга может предполагать установку лямбда каналообразующего оборудования в DWDM мультиплексоре. λ услуга обеспечивает точка-точка соединение на физическом уровне эталонной модели взаимодействия открытых систем (L1 OSI RM).

FrameNet услуга предполагает использование доступных посредством гигабит Ethernet GE (LX или на ZX) интерфейсов, таких как 1-10 GE портов оконечного оборудования магистрали. Эта услуга обеспечивает точка-точка и многоточечную доставку на Ethernet (канальном) уровне (L2 OSI RM). Доступны два базовых типа услуг FrameNet: (1) специализированная Ethernet услуга, гарантирующая заказанную пропускную способность; (2) неспециализированная Ethernet услуга, не гарантирующая постоянную пропускную способность.

PacketNet услуга предоставляет базовую услугу IP маршрутизации для доступа к public Интернет, обеспечивая соединения на сетевом уровне (L3 OSI RM). Прием анонсов IP-сетей, как и анонсирование IP-сетей, выполняется согласно политикам маршрутизации, отраженным в RIPE NCC.

Предоставляемый магистралью λ сервис позволяет создавать принципиально новый инструментарий для решения задач, основанный на зарождающихся LambdaGrid парадигмах распределенных вычислений и использующих параллелизм гарантированных и независимых каналов связи.

5. Вычислительная облачная платформа УрО РАН

Вычислительная облачная платформа (ВОП) УрО РАН создается для обеспечения гибкости при проведении вычислений. Пользователям иногда требуются оборудование или программное обеспечение в конфигурации, отличной от той, что предоставляется суперкомпьютерами. Например, пользователям может быть нужна операционная система Windows (а суперкомпьютеры работают под Linux), или набор выделенных серверов (на суперкомпьютере ресурсы предоставляются системой запуска задач во временное пользование).

Важным направлением использования ВОП является установка пакетов прикладных программ, интегрированных с суперкомпьютерами. Пакеты прикладных программ устанавливаются на виртуальные машины ВОП и настраиваются таким образом, чтобы задачи можно было запускать на счет на суперкомпьютерах прямо из графического интерфейса пакетов. Схема ВОП показана на рис. 7.

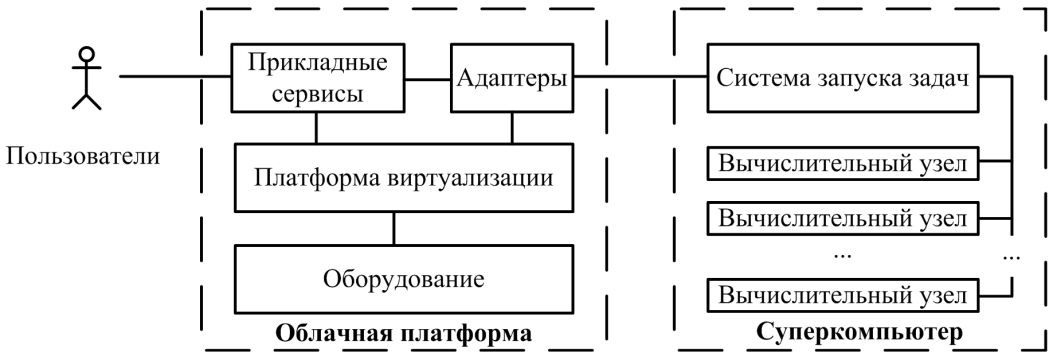

Рис. 7. Структура вычислительной облачной платформы УрО РАН

В настоящее время ВОП используется для следующих проектов:

- Виртуальная Web-лаборатория «Параллельное программирование в Matlab». На ВОП установлен Matlab, интегрированный с кластером. Создается Web-интерфейс, который будет служить единой точкой доступа к Matlab в облачной конфигурации, учебному курсу по параллельному программированию в Matlab и практическим руководством по применению Matlab на СК «УРАН».
- Обработка больших объемов изображений на параллельных вычислительных системах [10]. Для проекта в облачной платформе выделено четыре сервера, на которые установлен Hadoop. С помощью Hadoop выполняется автоматизация распараллеливания обработки изображений, а распределенная файловая система HDFS (входит в состав Hadoop) позволяет реализовать подход перемещения вычислений к данным.
- Разработка морфологического анализатора русского языка [11]. Для проекта в ВОП создан набор виртуальных машин, работающих в режиме распределенных вычислений без скоростного интерконнекта. Морфологический анализатор реализован в виде Web-сервиса.

Первая очередь реализации ВОП построена на базе четырех серверов Fujitsu-Siemens RX330 в каждом по 2 процессора AMD Opteron 2220, 8 ГБ памяти, жесткий диск 250 ГБ, сеть Gigabit Ethernet. В качестве системы хранения используется EMC Celerra NS-480. Платформа виртуализации создана на основе бесплатно распространяемого программного обеспечения oVirt (Open Source Vitrualization Manager [12]).

Важным аспектом реализации ВОП является то, что для ее создания используются серверы, ранее входившие в состав вычислительных кластеров, но производительность которых, по современным меркам, делает их дальнейшее использование для проведения высокопроизводительных вычислений нерациональным. Применение этих серверов для создания ВОП позволило продлить срок эксплуатации оборудования.

6. Заключение

Заложены основы построения киберинфраструктуры Уральского отделения РАН, объединяющей по скоростной оптической магистрали вычислительные ресурсы и системы хранения данных научных центров отделения на территории от Архангельска до Екатеринбурга. На основе суперкомпьютерного центра ИММ УрО РАН сформулирована трехуровневая иерархия распределенных по научным центрам вычислительных ресурсов. Описана облачная вычислительная платформа УрО РАН, в рамках которой пользователям предоставляются пакеты прикладных программ, интегрированные с суперкомпьютерами.

Литература

1. Стратегия развития Уральского отделения Российской академии наук до 2025г. Екатеринбург: УрО РАН, 2010. 236 с.

2. Гольдштейн М.Л., Созыкин А.В. Концептуальная модель и прототип ГРИД УрО РАН // Системы управления и информационные технологии, №3.1(49), 2012. – С. 132-136.

3. Масич А.Г., Масич Г.Ф. Инициатива GIGA UrB RAS // Вычислительные технологии. 2008. Т. 13. Вестник КазНУ им. Аль-Фараби. Серия математика, механика, информатика 2008. №3(58). Совместный выпуск. - Ч.II. - С.413-418.

4. Созыкин А.В., Гольдштейн М.Л. Модель взаимодействия прикладных облачных сервисов и суперкомпьютеров // Вестн. Пермского университета. Сер. Математика. Механика. Информатика. 2012. Вып. 3 (11). С. 65 - 69.

5. Millar P. dCache, agile adoption of storage technology // International Conference on Computing in High Energy and Nuclear Physics. New York. 2012.

6. Масич А.Г., Масич Г.Ф. От «Инициативы GIGA UrB RAS к Киберинфраструктуре УрО РАН. // Вестник Пермского научного центра. – Пермь: Изд-во ПНЦ УрО РАН, 2009. С. 41-56.

7. Матвеенко В.П., Масич А.Г., Масич Г.Ф. Региональная научно-образовательная оптическая сеть УрО РАН – фундамент киберинфраструктуры // Материалы X Международной конференции «Высокопроизводительные параллельные вычисления на кластерных системах (HPC2010)». – Пермь: Изд-во ПГТУ, 2010. Т.1. С. 11-21.

8. Масич А.Г., Масич Г.Ф., Матвеенко В.П., Тирон Г.Г. Инициатива GIGA UrB RAS: методология построения и архитектура научно-образовательной оптической магистрали Уральского отделения РАН // Межд. конф. Математические и информационные технологии, MIT-2011. Белград, 2012. С. 257-265.

9. Масич А.Г., Масич Г.Ф. Аспекты реализации научно-образовательной оптической магистрали УрО РАН // Труды XIX Всероссийской научно-методической конференции «Телематика’2012». - С-Пб.: изд-во ИТМО, 2012. – С.247-250.

10. Созыкин А.В., Гольдштейн М.Л. Система обработки изображений с автоматическим распараллеливанием на основе MapReduce // Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование», № 27, 2012. С. 109-118.

11. Усталов Д.А., Гольдштейн М.Л. Распределенная инструментальная среда морфологического анализа для обработки русского языка // Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование», № 27, 2012. С. 119-127.

12. oVirt Project. URL: http://www.ovirt.org/ (дата обращения 28.11.2012).