Управление вычислительными ресурсами Сибирского Суперкомпьютерного Центра

Б.М. Глинский, И.Г. Черных, Н.В. Кучин, С.В. Ломакин, И.Н. Макаров

Институт Вычислительной Математики и Математической Геофизики СО РАН

Разбираются вопросы эксплуатации суперкомпьютера НКС-30Т Сибирского Суперкомпьютерного Центра Коллективного Пользования с системой управления очередью заданий PBS Pro.

1. Введение

Центр коллективного пользования «Сибирский Суперкомпьютерный Центр» (ССКЦ) организован как структурное подразделение ИВМиМГ СО РАН в соответствии с постановлением Президиума СО РАН от 06.03.2001 № 100 «О создании Сибирского суперкомпьютерного центра коллективного пользования СО РАН». В ССКЦ были установлены и эксплуатировались разнообразные системы пакетной обработки (batch system):

- Distributed Queuing System (DQS) на комплексе RM600 E30;
- Система управления прохождением параллельных заданий (СУППЗ) на кластерах МВС-1000/32 и МВС-1000/128;
- Sun Grid Engine на кластере НКС-160;
- PBS Pro 11.1 на суперкомпьютере НКС-30Т.

2. Вычислительные ресурсы и программное обеспечение ЦКП ССКЦ

2.1 Кластерный суперкомпьютер НКС-30Т

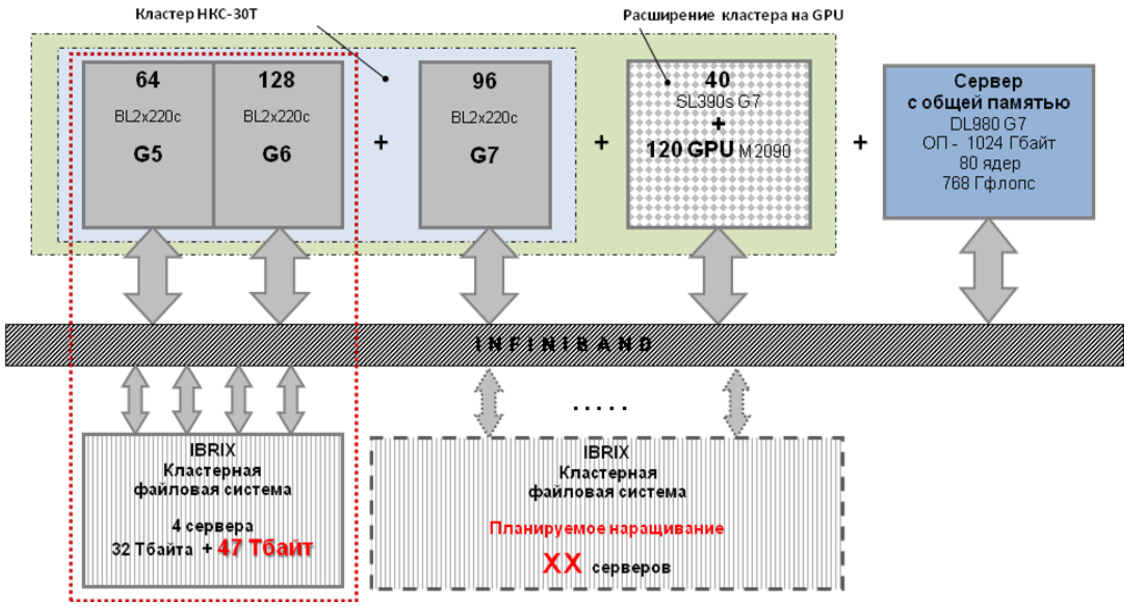

Рис. 1. Логическая схема гетерогенного кластера НКС-30Т

Гетерогенный высокопроизводительный кластер НКС-30Т с пиковой производительностью 115 ТФлопс это основной вычислительный ресурс ССКЦ [1-3]. Гибридное расширение на GPU NVIDIA Tesla M2090 занимает 30 место, а базовый кластер на процессорах Intel Xeon - пятидесятое место в 22 списке Top 50 от 31.03.2015.

Коммуникационная сеть кластера QDR Infiniband, транспортная и сервисная сети – Gigabit .Ethernet. Кластерная файловая система Ibrix была модернизирована в 2014 году.

Сервер HP ProLiant DL980 G7 с одним терабайтом оперативной памяти включен в состав НКС-30Т как нестандартный вычислительный узел.

Программно-аппаратная среда кластера позволяет использовать облачные технологии для создания специализированных вычислительных сред. Гетерогенность ресурсов кластера позволяет гибко варьировать параметры выделяемых в облако виртуальных ресурсов.

В настоящее время в ССКЦ СО РАН функционирует две, основанные на KVM виртуализированные вычислительные среды, являющихся частями более крупных виртуальных кластеров. Первая - ИЯФ-НГУ-ССКЦ [4]. Используется для обработки данных физических экспериментов в физике высоких энергий, осуществляемых в ИЯФ СО РАН. Вторая - Академпарк-ССКЦ. Разработана совместно с Академпарком Технопарка Новосибирского Академгородка и предназначена для решения задач BigData (геофизика, обработка медицинских данных и другие). Оба виртуальных кластера для обмена данными с ССКЦ используют Суперкомпьютерную сеть Новосибирского Научного Центра (10 Гбит/с).

2.2 Программное обеспечение

Системное программное обеспечение включает в себя Red Hat Enterprise Linux 5.4, HP Cluster Management Utility (CMU) 7.0 и систему управления пакетной обработкой PBS Pro 11.1.

Средства разработки включают Intel MPI 4.1, Intel TraceAnalyzer/Collector, компиляторы Intel C++ и Intel Fortran из состава Composer XE 2013 SP1, включающие библиотеки Intel MKL, Intel IPP и Intel TBB. Дополнительно установлены компиляторы и библиотеки из Intel Parallel Studio XE 2015. Такой подход позволяет использовать дорогостоящие лицензии на старые версии программного обеспечения Intel.

Из коммерческих пакетов установлены Gaussian g09 Rev D.01 w/LINDA, ANSYS CFD версии 14.5.7. с лицензиями HPC, обеспечивающими параллельное выполнение программ Fluent, а также ANSYS CFD 16.1 (без лицензий HPC)..

Архитектура ССКЦ поддерживает две современных парадигмы параллельных вычислений – **MPI** для систем с распределенной памятью (MPP-кластеров)и **OpenMP** для систем с общей памятью. Смешанная схема вычислений(**MPI**+**OpenMP**) позволяет запуск на каждый вычислительный узел кластера по одному MPI-процессу, который запускает внутри каждого вычислительного модуля несколько потоков с помощью OpenMP.

Для гибридной архитектуры: суперкомпьютер состоит из набора соединенных между собой узлов, для обмена данными используется MPI; каждый узел состоит из двух CPU и трёх GPU; на каждом узле запускается 1 процесс MPI, управляющий вычислениями (процесс выполняется на CPU); из MPI процесса запускаются потоки (threads) OpenMP, каждый из которых управляет работой одного GPU. Другой вариант: запускаются три MPI процесса на узел, каждый управляет закрепленным за ним GPU.

3. Использование вычислительных ресурсов

Таблица 1. Использование машинного времени по годам

Статистика по кластерам (НКС-30Т + НКС-160)	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
∑ производительность (тфлопс)	7,1	17,5	31	116	115	115
∑ CPU (дни)	80 179,52	121 201,46	209 997,57	533 324, 55	713 960,42	529 708,63
∑ количество заданий	38 914	39 750	35 952	83 797	103 840	89059
TOP50 (места) С 2012г. два кластера	26	34	32	16, 30	21, 36	28, 45

Таблица 2. Отчет по работе пользователей за 2014 год

Всего пользователей – **155** Всего организаций – **29** Академических организаций – **24** Университетов – **3** (СФУ, Красноярск), НГУ, НГТУ) Другие организации – **2** (СибНИА им. Чаплыгина, СибНИГМИ)	Всего грантов, программ, проектов, тем — **176** Из них Российских — **171** Международных — **5** Грантов РФФИ – **67** Программ РАН – **22** Проектов СО РАН – **30** Программ Минобрнауки – **18** Другие – **34**	Всего публикаций – **158** Российских – **88** Зарубежных – **70** Доктор. диссерт. – **1**, Кандидат. диссерт. – **5**, Дипломы – **7**, Патенты – **2.**

Приведенная статистика показывает, что ЦКП ССКЦ действительно центр коллективного пользования, крупнейший за Уралом. Его пользователи в основном сотрудники СО РАН. С увеличением вычислительных ресурсов число заданий и загрузка вычислительной техники – возрастает. Спад в 2014 году связан с уменьшением потока заданий пользователей и потерями машинного времени на модернизацию программного обеспечения кластерной файловой системы IBRIX и наращивание дискового пространства в августе - сентябре. Основные затраты пришлись на сохранение по частям 27 Терабайт пользовательских данных и последующее их восстановление.

4. Обеспечение безопасности

Установлен сетевой экран (firewall), разрешающий доступ по протоколу ssh/scp/sftp только из сетей, в которых есть зарегистрированные пользователи ССКЦ. На исходящий трафик ограничений не налагается. Более точные ограничения накладываются средствами Linux (hosts.allow , AllowUsers).

Пользователи авторизуются по ключу ssh, который создаётся каждым пользователем индивидуально. Секретная часть ключа ssh хранится у пользователя, а открытая посылается администратору кластера.

5. Использование возможностей PBS Pro

PBS Professional [5] это коммерческая система для планирования заданий и управления загрузкой высокопроизводительных вычислительных кластеров.

5.1 Очереди заданий

На кластере организовано несколько очередей, каждая очередь поддерживает работу с однотипными серверами одного поколения. Написан скрипт pbsinfo, выдающий число свободных вычислительных узлов и отдельных ядер в каждой очереди.

```
[kuchin@nks-g6 ~]$ pbsinfo

QUEUE          NODES FREE  CORES FREE  NODES TOTAL GPUS FREE   GPUS TOTAL
-----------    ----------  ----------  ----------  ----------  ----------
workq(G5)          0           0*8          64          0           0
g6_q(G6)          36          36*8+37      128          0           0
G7_q(G7)           1           1*12         96          0           0
SMP_G7_q(G7)       0           0*80+67       1          0           0
SL_q(GPU)          0           0*12         40          0          120
```

Рис. 2. Пример выдачи pbsinfo

Для специальных целей можно создать очередь, поддерживающую работу с неоднородными серверами. Например, сервер с большой оперативной памятью и несколько серверов с графическими ускорителями Nvidia Tesla.

5.2 Ограничение числа выполняемых заданий для конкретных пользователей

Пользователь, поставивший в очередь большое число заданий, которые выйдут в счет и захватят все свободные ресурсы, фактически будет использовать кластер монопольно и блокировать счет заданий других пользователей. Для таких пользователей вводится лимит на число решаемых задач и лимит на число выделяемых процессорных ядер.

```
 [root@nks-g6 ~]# qmgr -c "print queue G7_q "
#
# Create queues and set their attributes.
#
#
# Create and define queue G7_q
#
create queue G7_q
set queue G7_q queue_type = Execution
set queue G7_q max_run = [u:user1=10]
set queue G7_q max_run += [u:user2=10]
set queue G7_q max_run += [u:user3=10]
set queue G7_q max_run += [u:user4=10]
set queue G7_q max_run += [u:user5=10]
. . . . . .
set queue G7_q max_run_res.ncpus = [u:user1=120]
set queue G7_q max_run_res.ncpus += [u:user4=120]
set queue G7_q max_run_res.ncpus += [u:user5=300]
set queue G7_q enabled = True
set queue G7_q started = True
[root@nks-g6 ~]#
```

Рис. 3. Предотвращение монопольного использования очереди заданиями одного пользователя.

5.3 Учет использования заданиями пользователей машинного времени

За выполненное задачей пользователя время принимается время решения задачи умноженное на число выделенных задаче процессорных ядер. За основу подсистемы учета заданий пользователей взята [6] версии pbsacct-1.4.7.tar.gz. Пользователи каждого института СО РАН объединяются в отдельную группу. Информация по загрузке вычислительных ресурсов выдается по каждой очереди и интегрально по всем очередям кластера. Дополнительно выдается информация по использованию заданиями графических ускорителей. Отметим что в приведенном примере информация о пользователях (login, Full name) не приведена и урезана до первых пяти строк.

```
GPU May 2015

Portable Batch System accounting statistics

------------------------------------------
Processing a total of 31 accounting files... done.
The first job record is dated Fri 01 May 2015 02:38:12 AM NOVT.
The last  job record is dated Sun 31 May 2015 12:23:12 AM NOVT.

QUEUE(S): all (GPU only)
```

Group	#jobs	Wallclock days	Percent	Average #cores	Average #GPUs	GPUs days	Average q-days
TOTAL	611	7113.76	100.00	40.36	10.20	1797.94	0.41
itam	95	5592.59	78.62	109.38	27.34	1398.15	2.18
uiggm	28	695.73	9.78	101.43	12.00	82.31	0.15
niboch	14	593.17	8.34	12.00	3.00	148.29	0.41
icmmg	340	184.87	2.60	8.36	1.32	29.21	0.07
altstu	28	46.64	0.66	1.00	3.00	139.91	0.04
nsu	106	0.77	0.01	12.16	1.12	0.07	0.06

Рис. 4. Пример выдачи статистики за май: организации использующие GPU.

QUEUE(S): all (GPU only)

Username	Group	#jobs	Wallclock days	Percent	Average #cores	Average #GPUs	GPUs days	Average q-days	Full name
TOTAL	-	611	7113.76	100.00	40.36	10.20	1797.94	0.41	
User1	itam	85	5558.26	78.13	110.77	27.69	1389.57	2.43	
User2	uiggm	28	695.73	9.78	101.43	12.00	82.31	0.15	
User3	niboch	14	593.17	8.34	12.00	3.00	148.29	0.41	
User4	icmmg	12	147.07	2.07	7.81	1.00	18.83	0.00	
User5	altstu	28	46.64	0.66	1.00	3.00	139.91	0.04	

Рис. 5. Пользователи, заказывающие GPU.

6. Виртуальный кластер ИЯФ-НГУ-ССКЦ

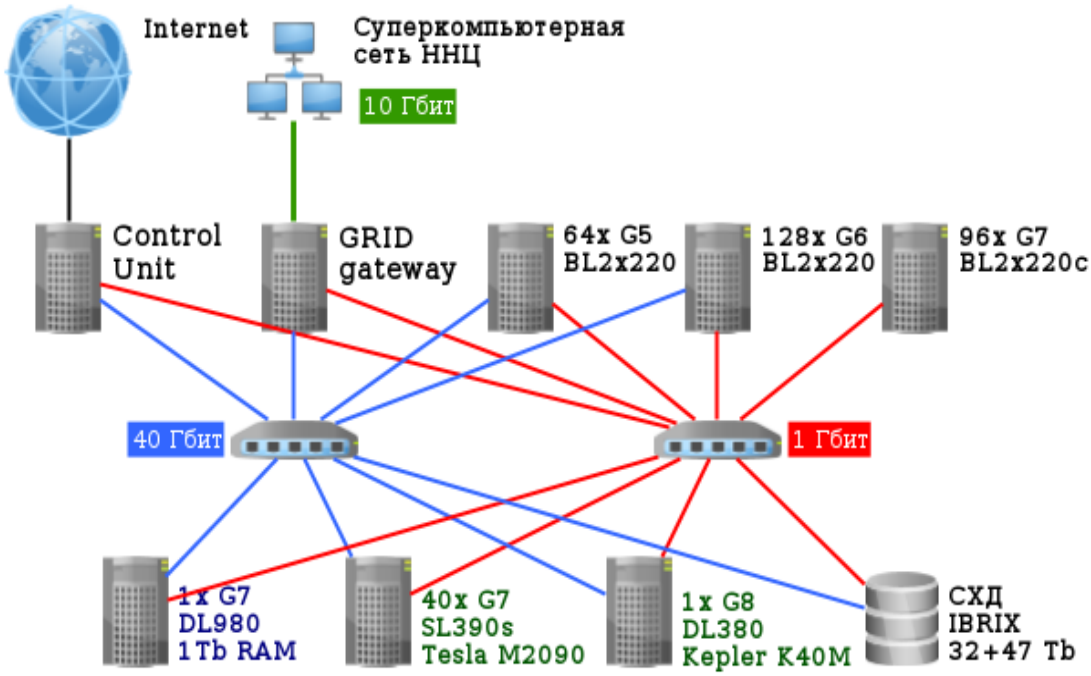

Рис. 6. Общая схема ресурсов ССКЦ

Виртуальный кластер ИЯФ-НГУ-ССКЦ используется для обработки данных физических экспериментов в физике высоких энергий, осуществляемых в ИЯФ СО РАН [4]. Задачи характеризуются использованием однопоточных программ и хорошей параллелизацией на уровне данных. Для создания виртуальных ресурсов на локальном кластере используются штатные возможности PBS Pro. PBS Pro запускает требуемую виртуальную машину, описанную в скрипте (prologue) и останавливает ее по завершению задания (epilogue), после чего вычислительный узел может использоваться в обычном режиме другими заданиями. В качестве среды виртуализации используется KVM. Обмен данными между ИЯФ СО РАН и ССКЦ осуществляется через суперкомпьютерную сеть Новосибирского Научного Центра (10 Гбит/с).
Основные задачи, решаемые на виртуальном кластере ИЯФ-НГУ-ССКЦ:
* Эксперимент **КЕДР**
Работа проводится на электрон-позитронном коллайдере ВЭПП-4М с детектором КЕДР. Эксперименты в области рождения ψ-резонансов (J/ψ, ψ(2S), ψ(3770)) и τ-лептона. Использует Scientific Linux CERN 3.
* Эксперимент **ATLAS**
Работа проводится на Большом адронном коллайдере (БАК) (ЦЕРН, Швейцария). Анализ данных эксперимента ATLAS в рамках ATLAS Exotics Working Group.
* Эксперимент **СНД**
Работа проводится на коллайдере ВЭПП-2000 со Сферическим нейтральным детектором (СНД). Изучение процессов электрон-позитронной аннигиляции в области энергии до 2 ГэВ в системе центра масс.

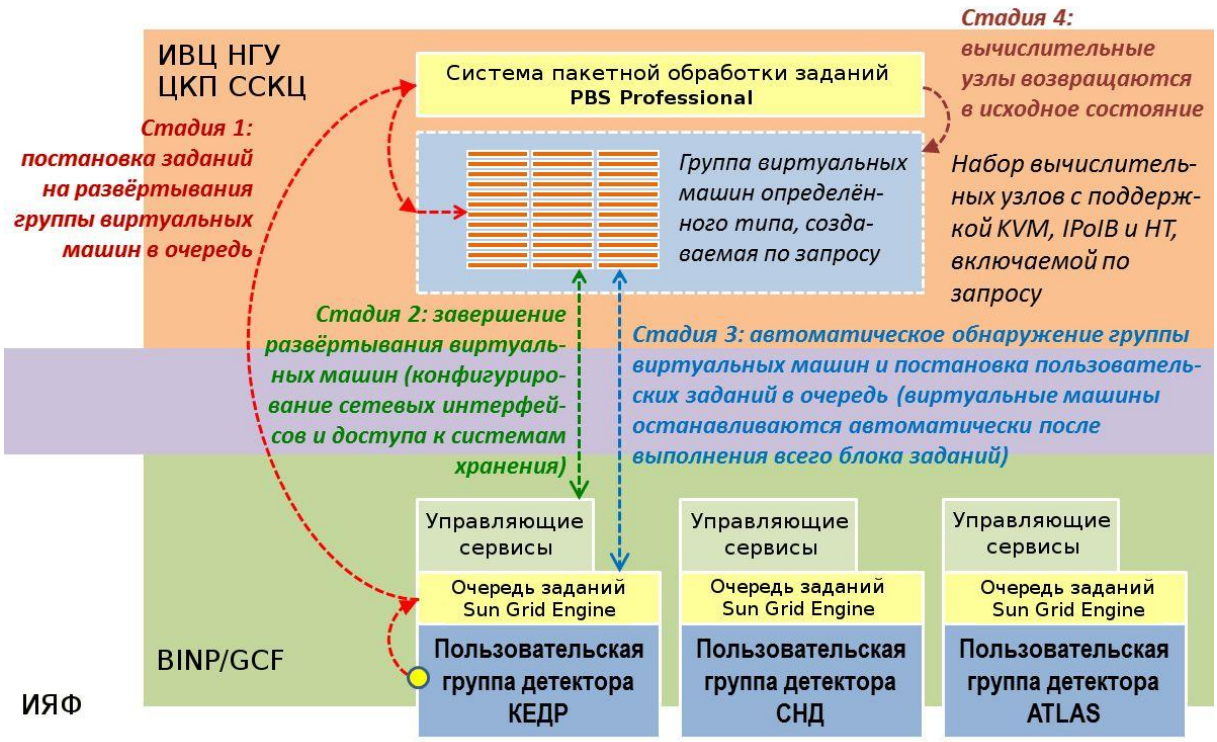

Рис. 7. Схема виртуального кластера ИЯФ-НГУ-ССКЦ

7. Виртуальный кластер Академпарк-ССКЦ

Виртуальный кластер Академпарк-ССКЦ создан совместно с департаментом системной интеграции Компании ТехноСити для Академпарка Технопарка Новосибирского Академгородка. Предназначена для решения задач обработки BigData. В первую очередь это геофизика, обработка медицинских данных, обработка трафика. Функционирует вне очереди заданий на MPP-G6 части кластера ССКЦ.

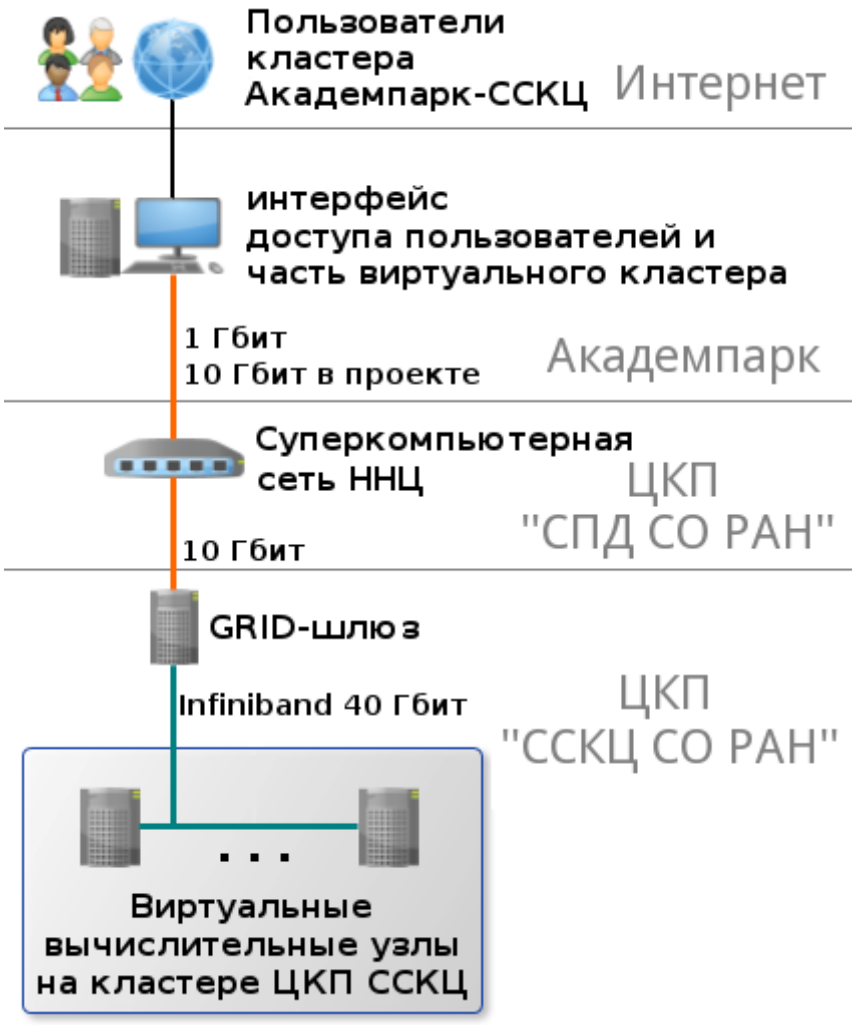

Рис. 8. Схема виртуального кластера Академпарк-ССКЦ

8. Работа с пользователями

11 -14 марта 2014 г.совместно с NVIDIA и Учебным центром по технологии CUDA (CUDA Teaching Center) НГУ была проведена Школа-тренинг по углубленному изучению технологий программирования графических процессоров [7]. На школе были рассмотрены вопросы профилирования, отладки, оптимизации кода на CUDA, применения технологии OpenACC. Практическая часть занятий школы проходила на гибридном расширении кластера НКС-30Т с GPU NVIDIA TESLA M2090.

15 -16 мая 2014 г.на ресурсах ССКЦ сотрудниками Intel проведен **ТРЕНИНГ** Intel® Software Development Excellence Program Применение Intel® Parallel Studio XE и Intel® Cluster Studio XE в решении исследовательских задач [8].

Проводится регулярный семинар «Архитектура, системное и прикладное программное обеспечение кластерных суперЭВМ» на базе ССКЦ, кафедры Вычислительных систем НГУ и Центра Компетенции по высокопроизводительным вычислениям СО РАН – Intel [9].

9. Проблемы и выводы

1. Моральное устаревание вычислительной техники.
2. Завершение гарантийного обслуживания, т.е. все ремонты за счет института.
3. Недостаточная пропускная способность Ibrix, соответственно, в конечном итоге скорость обменов определяется быстродействием дисковой полки.

4. Неприспособленность Ibrix к работе с большими файлами размером в сотни Гигабайт.

5. Новый кластер должен быть с жидкостным охлаждением..

6. Часть вычислительных узлов должна быть с сопроцессорами Intel Xeon Phi, часть с GPU Nvidia и приблизительно половина без сопроцессоров и ускорителей только с процессорами Intel Xeon.

7. Несколько вычислительных узлов должны быть серверами с большой оперативной памятью в 2 – 4 Терабайта и локальным дисковым массивом не менее 10 Терабайт.

8. В качестве коммуникационной среды должен быть Mellanox Infiniband и/.или Intel Omni-Path.

9. Обязательно должна быть параллельная файловая система Lustre или Panasas, а также файловое хранилище большой емкости для хранения и архивирования неиспользуемых данных.

Выявленные во время эксплуатации проблемы будут учтены при наращивании вычислительных ресурсов и закупке следующего кластера.

Литература

1. Глинский Б.М., Кучин Н.В., Ломакин С.В., Черных И.Г. Сибирский суперкомпьютерный центр СО РАН. Материалы международной конф. «Методы создания, исследования и идентификации математических моделей», 2013, С. 28-29
2. Б.М. Глинский, Д.А. Караваев, И.М. Куликов, Н.В. Кучин, Н.В. Снытников. Масштабируемые вычисления с применением гибридного кластера// Материалы международной конференции «Mathematical and Informational Technologies, MIT-2013», с.89.
3. Гибридный кластер НКС-30Т
   URL: http://www2.sscc.ru/HKC-30T/HKC-30T.htm
4. С.Д. Белов, А.С. Зайцев, В.И. Каплин, А.А. Король, К.Ю. Сковпень, А.М. Сухарев, А.С. Адакин, В.С. Никульцев, Д.Л.. Чубаров, Н.В. Кучин, С.В. Ломакин, В.А. Калюжный //Использование виртуализованной суперкомпьютерной инфраструктуры Новосибирского научного центра для обработки данных экспериментов физики высоких энергий. Журнал «Вычислительные технологии», 2012 г., том 17, №6, стр.36-46.
5. PBS Professional.URL: http://www.pbsworks.com/Product.aspx?id=1
6. pbsacct - Accounting Report Tool.
   URL:http://www.mcs.anl.gov/research/projects/openpbs/patches/pbsacct/README.txt
7. Школа – тренинг по программированию на GPU
   URL: http://www2.sscc.ru/Seminars/Nvidia%20Cuda-2014.htm
8. Применение Intel® Parallel Studio XE и Intel® Cluster Studio XE
   URL: http://www2.sscc.ru/SORAN-INTEL/documents_2014.htm
9. «Архитектура, системное и прикладное программное обеспечение кластерных суперЭВМ»
   URL: http://www2.sscc.ru/Seminars/NEW/Seminars.htm

Control and managing the HPC cluster in Siberian Supercomputer Center

Nikolay Kuchin, Boris Glinsky, Igor Chernykh, Sergey Lomakin and Igor Makarov

Keywords: HPC, Bach system, queue, PBS

The paper presents the experience of exploitation high-performance computing cluster installed in the Siberian Supercomputer Center (SSCC ICMMG SB RAS). SSCC has more than 150 users from more than 20 academic institutions. One important example of virtual computing environment is the integration KVM with batch system PBS Pro.