Государственный университет им. Н.И. Лобачевского

Национальный исследовательский университет

Суперкомпьютер «Лобачевский»

Работа с Linux-сегментом

Докладчик:

Алексей Сиднев

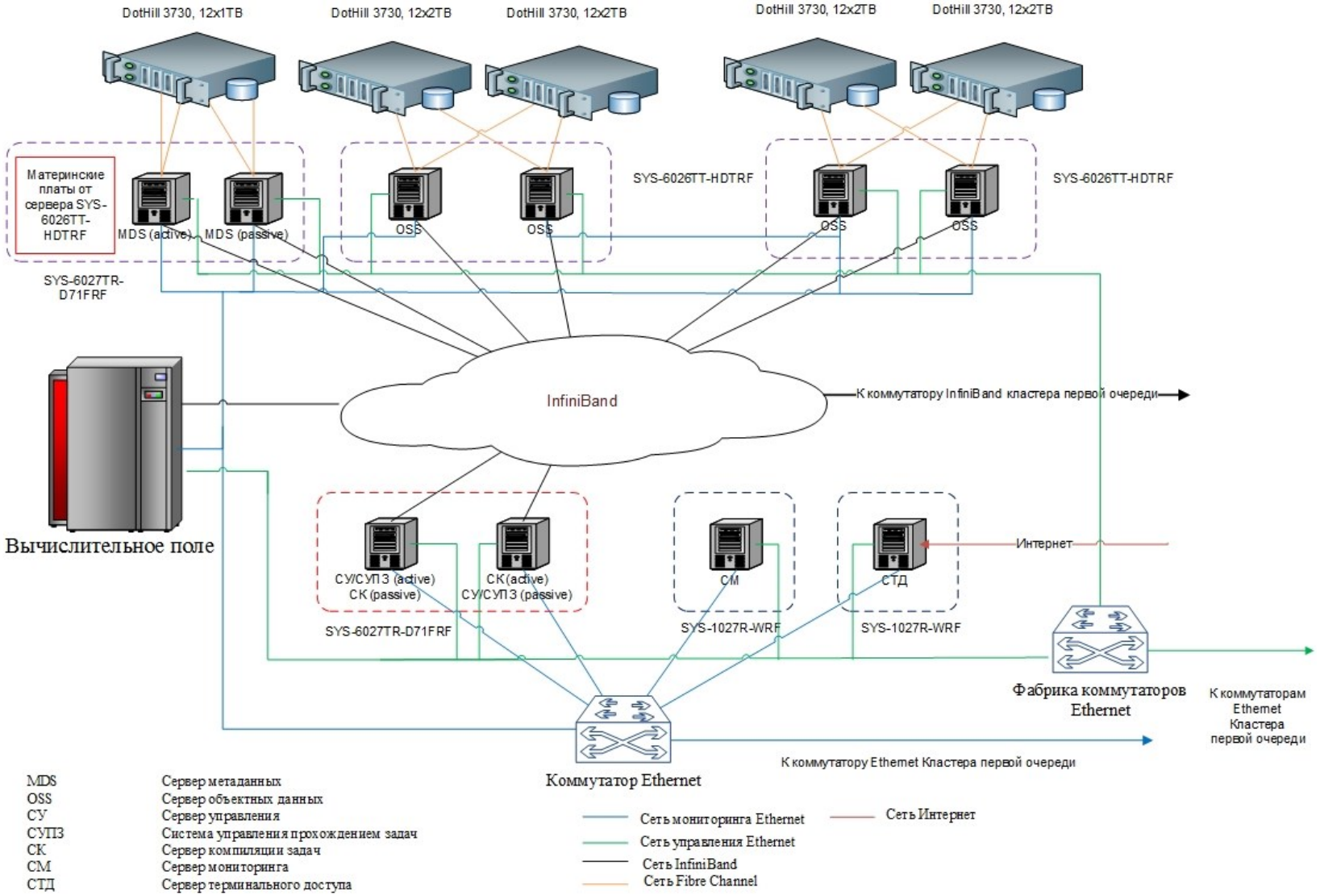
Содержание

- Архитектура
- Доступ
- Программное обеспечение
- Работа на кластере:
  - Информация об узлах
  - Запуск, компиляция и управление задачами
  - Рабочие директории

Архитектура (1)

- Состав кластера:
  - 120 блэйд-серверов:
    - Два универсальных процессора (16 ядер):
      - Intel Xeon CPU E5-2660 2.20GHz
    - 10 узлов с двумя сопроцессора Intel Xeon Phi
    - 100 узлов с тремя графическими ускорителями Nvidia Tesla Kepler
  - Вычислительная сеть:
    - Infiniband (Mellanox MSX6512-4R)
  - ОС:
    - CentOS 6.4

Архитектура (2)



Программное обеспечение (1)

- Windows-клиент:
  - PuTTY – ssh-клиент для доступа к узлам кластера
  - WinSCP – графический клиент SCP/FTP/SFTP

- SLURM (Simple Linux Utility for Resource Management) – система управления кластером
  - salloc – резервирование узлов кластера
  - sbatch – создание задания с использованием скрипта
  - scancel – отправка сигнала заданию
  - sinfo – просмотр информации об узлах кластера
  - squeue – просмотр информации о текущих заданиях
  - srun – запуск задания на выполнение

Программное обеспечение (2)

- Компиляторы:
  - GNU Compiler 4.4.7, 4.8.2
  - Intel Compiler 14.0.2
  - PGI 14.1
- Видеокарты:
  - Драйвер GPU NVidia 331.38
  - CUDA 5.5 Toolkit
- Intel Cluster Studio XE 2013
- MPI:
  - OpenMPI 1.6.5
  - MVAPICH2 2.0b
  - Intel MPI 4.1.3

Директории

- Домашняя:

  $HOME

- Общее хранилище (**предпочтительнее**):

  /common/$USER

Информация об узлах

`sinfo`

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
gpu          up 3-00:00:00     16  down* node[2,15,18,20,40-41,49-50,59,63,68,70-
   71,84,87,114]
gpu          up 3-00:00:00      7  drain node[10,42,53,67,82,90,118]
gpu          up 3-00:00:00     77   idle node[1,3-9,11-14,16-17,19,21-39,43-48,51-
   52,54-58,60-62,64-66,69,72-81,83,85-86,88-89,111-113,115-117,119-120]
cpu*         up 3-00:00:00     10  drain node[101-110]
phi          up 3-00:00:00      4  maint node[93-94,98,100]
phi          up 3-00:00:00      6   idle node[91-92,95-97,99]
all          up 3-00:00:00      4  maint node[93-94,98,100]
all          up 3-00:00:00     16  down* node[2,15,18,20,40-41,49-50,59,63,68,70-
   71,84,87,114]
all          up 3-00:00:00     17  drain node[10,42,53,67,82,90,101-110,118]
all          up 3-00:00:00     83   idle node[1,3-9,11-14,16-17,19,21-39,43-48,51-
   52,54-58,60-62,64-66,69,72-81,83,85-86,88-89,91-92,95-97,99,111-113,115-
   117,119-120]
```

Состояние узлов

- idle – узел свободен и готов для счета
- alloc – на узле считается задача
- comp – на узле выполняется epilog-скрипт
- maint – узел зарезервирован для определенных пользователей
- down, down*, drain, comp*, idle*,alloc* – узел недоступен/выведен из счета/неполадки на узле

Информация о задачах

```
squeue
```

```
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
5261       all     bash    kvant  R       3:29      1 node1
5263       gpu     bash    kvant  R    INVALID      3 node[60-62]
5262       phi     bash    kvant  R       1:07      2 node[91-92]
```

- Разделы:
  - **cpu** – 10 узлов node101 - node110 без ускорителей
  - **gpu** – 100 узлов node1 - node90, node111 - node120 с ускорителями Nvidia K20X
  - **phi** – 10 узлов node91 - node100 с ускорителями Intel Xeon Phi

Управление задачами

- Запуск задач
  - Интерактивный режим

```
srun [опции] [исполняемый файл]
```

  - Пакетный режим

```
sbatch –N <количество узлов> -p <название
  раздела> -t <лимит времени> <запускаемый
  скрипт>
```

- Отмена задач
  - `scancel <Номер задачи>`

Управление узлами

- Выделение узлов

```
salloc –N <количество узлов> -p <название
  раздела> -t <лимит времени>
```

- Пример

```
salloc -N 1 -p phi -t 120
ssh $SLURM_NODELIST
ssh mic0
export
  LD_LIBRARY_PATH=/common/intel/mkl/lib/mic:/co
  mmon/intel/lib/mic/
```

Компиляция

- Выбор MPI:
  - MVAPICH2

```
module load mva
```

  - OpenMPI

```
module load ompi
```

  - IntelMPI

```
module load impi
```

- Помимо штатной версии компиляторов GNU (4.4.7) доступна более новая (4.8.2 с библиотеками boost)

```
module load gcc-4.8.2
```

Запуск MPI-программ

- Интерактивный режим:

```
srun -t <time> -p <partition> --ntasks-per-node
  <ppn> -N <num_nodes> -n <num_procs> --
  cpu_bind=v,map_cpu:<core0>,<core1>...<coreM>
```

  - <time> – лимит времени на задачу
  - <partition> – раздел кластера
  - <ppn> – число процессов на узел
  - <num_nodes> – число узлов
  - <num_procs> – общее число процессов
  - <core0>,<core1>...<coreM> – привязка i-го процесса на узле к <corei> ядру

Запуск на Xeon Phi

- Количество используемых сопроцессоров на узле

```
MIC_NUM_PER_HOST
```

- Число потоков на узел

```
OMP_NUM_THREADS
```

- Число потоков на сопроцессоре

```
MIC_OMP_NUM_THREADS
```

- Передача параметров в mpiexec

```
MPIEXEC_FLAGS_HOST, MPIEXEC_FLAGS_MIC
```

```
sbatch -N num_nodes -p phi mpirun.mic -x
  ppn_host -c ./impi_native_hybrid -z ppn_mic -
  m ./impi_native_hybrid.mic
```