Центр компетенций и обучения

XV Международная конференция
Научный сервис в сети Интернет: все грани параллелизма
23-28 сентября 2013, Абрау-Дюрсо

ПРЕДПРИЯТИЕ ГОСКОРПОРАЦИИ «РОСАТОМ»

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. РЕАЛИЗАЦИЯ НЕБЛОКИРУЮЩЕГО ВВОДА-ВЫВОДА.

ООО ЦКО Нагорный Д.В., НИЯУ «МИФИ» Леонова Н.М.

S-MPI

- Отечественная библиотека MPI, удовлетворяющая современным требованиям и превосходящая зарубежные аналоги по основным характеристикам, таким как производительность, масштабируемость и отказоустойчивость.
- Объединение в одном комплексе программные средства для разработки, исполнения, анализа и оптимизации параллельных приложений (сбор трасс, анализ на предмет эффективности, нахождения узких мест в производительности, проверка корректности кода и т.п.)
- Эффективная поддержка многоуровневых вычислительных сред.
- Смешивание разных моделей распараллеливания – на общей и распределенной памяти.
- Работоспособность на смешанных по архитектуре системах.
- Адаптация к возможным отечественным компонентам аппаратного

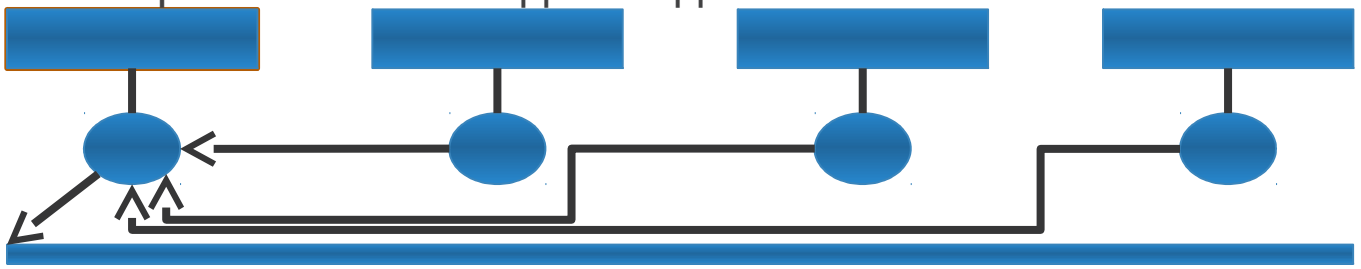
Обеспечение высокопараллельных расчетов в разных отраслях экономики (авиация, космонавтика, судо- и автомобилестроение, прогноз погоды, биоинформатика и биоинженерия, нефте- и газодобыча и разведка и т.п)

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. РЕАЛИЗАЦИЯ НЕБЛОКИРУЮЩЕГО ВВОДА-ВЫВОДА.

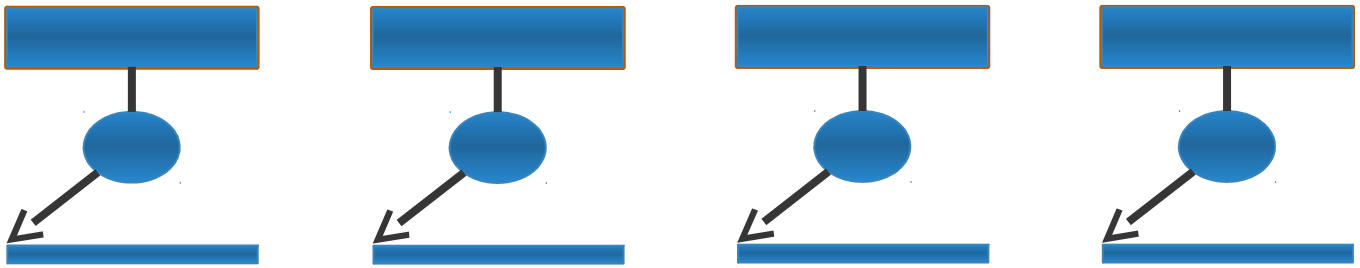
- Ввод-вывод - критическое место параллельных приложений.
- MPI-2 актуальный промышленный стандарт параллельного программирования
- Унификация средств ввода-вывода в рамках стандарта MPI-2

ВВОД-ВЫВОД В ПАРАЛЛЕЛЬНОМ ПРИЛОЖЕНИИ

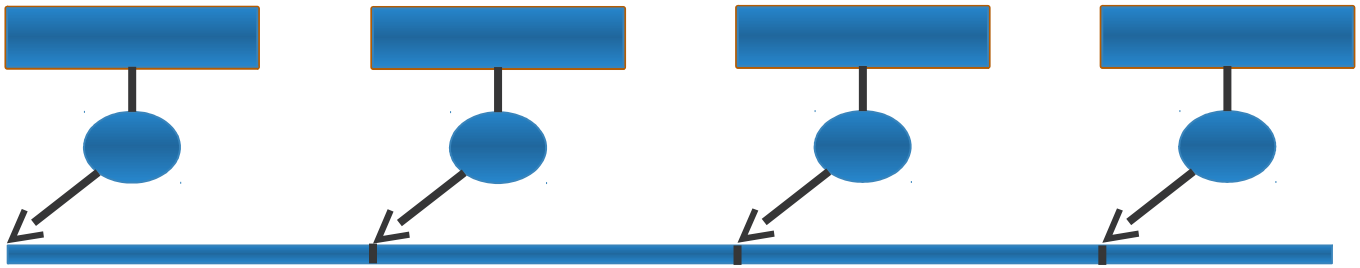
Не параллельный ввод-вывод



ввод-вывод в отдельные файлы



Параллельный ввод-вывод



ВВОД-ВЫВОД В ПАРАЛЛЕЛЬНОМ ПРИЛОЖЕНИИ
требования стандарта MPI-2

1. Управление файлами
2. Представление данных
3. Доступ к данным:
   - **Позиционирование:** явное, индивидуальные файловые указатели, общие файловые указатели
   - **Координирование:** коллективные, неколлективные
   - **Синхронность:** блокирующие, **неблокирующие**

```
#define N 100
MPI_Datatype arraytype;
MPI_Offset disp;
disp = rank*sizeof(int)*N; etype = MPI_INT;
MPI_Type_contiguous(N, MPI_INT, &arraytype);
MPI_Type_commit(&arraytype);
MPI_File_open( MPI_COMM_WORLD,
"/pfs/datafile",
              MPI_MODE_CREATE |
MPI_MODE_RDWR,
              MPI_INFO_NULL, &fh);
MPI_File_set_view(fh, disp, etype,
arraytype,
                  "native",
MPI_INFO_NULL);
MPI_File_write(fh, buf, N, etype,
MPI_STATUS_IGNORE);
```

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
Проблемы с неблокирующим вводом-выводом в исходной версии (OpenMPI 1.5.4, gcc 4.4.6, IMB 3.2.4)

```
#---------------------------------------------------------------------------
# Benchmarking S_IWrite_Indv
# #processes = 1
#---------------------------------------------------------------------------
#    MODE: AGGREGATE
#
       #bytes #repetitions t_ovrl[usec] t_pure[usec]  t_CPU[usec]   overlap[%]
            0           50     10083.50         0.08      9964.94         0.00
            1           50     10707.46       507.66      9964.94         0.00
            2           50     10645.58       520.34      9964.94         0.00
            4           50     10690.36       522.56      9964.94         0.00
            8           50     10688.12       490.38      9964.94         0.00
           16           50     10751.28       508.64      9964.94         0.00
           32           50     10648.54       474.66      9964.94         0.00
           64           50     10797.96       475.82      9964.94         0.00
          128           50     10725.22       491.52      9964.94         0.00
          256           50     10736.64       541.24      9964.94         0.00
          512           50     10736.58       524.42      9964.94         0.00
         1024           50     10706.30       543.90      9964.94         0.00
         2048           50     10720.32       562.80      9964.94         0.00
         4096           50     10738.32       588.38      9964.94         0.00
         8192           50     10784.58       605.30      9964.94         0.00
        16384           50     10969.08      3149.40      9964.94        63.12
        32768           50     11195.66      1198.42      9964.94         0.00
        65536           50     11693.20      1558.38      9964.94         0.00
       131072           50     12230.70      4262.38      9964.94        45.84
       262144           50     13672.50      3333.54      9964.94         0.00
       524288           32     16240.41      6024.41      9964.94         0.00
      1048576           16     21075.87     11203.32      9964.94         0.93
      2097152            8     31714.89     21515.64      9964.94         0.00
      4194304            4     51676.99     41915.54      9964.94         2.04
      8388608            2    360165.00     82630.99      9964.94         0.00
     16777216            1    175130.13    164592.03      9964.94         0.00
```

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
А что с неблокирующим вводом-выводом в коммерческих реализациях? (Intel MPI 4.1.0, icc 13.1.2, IMB 3.2.4)

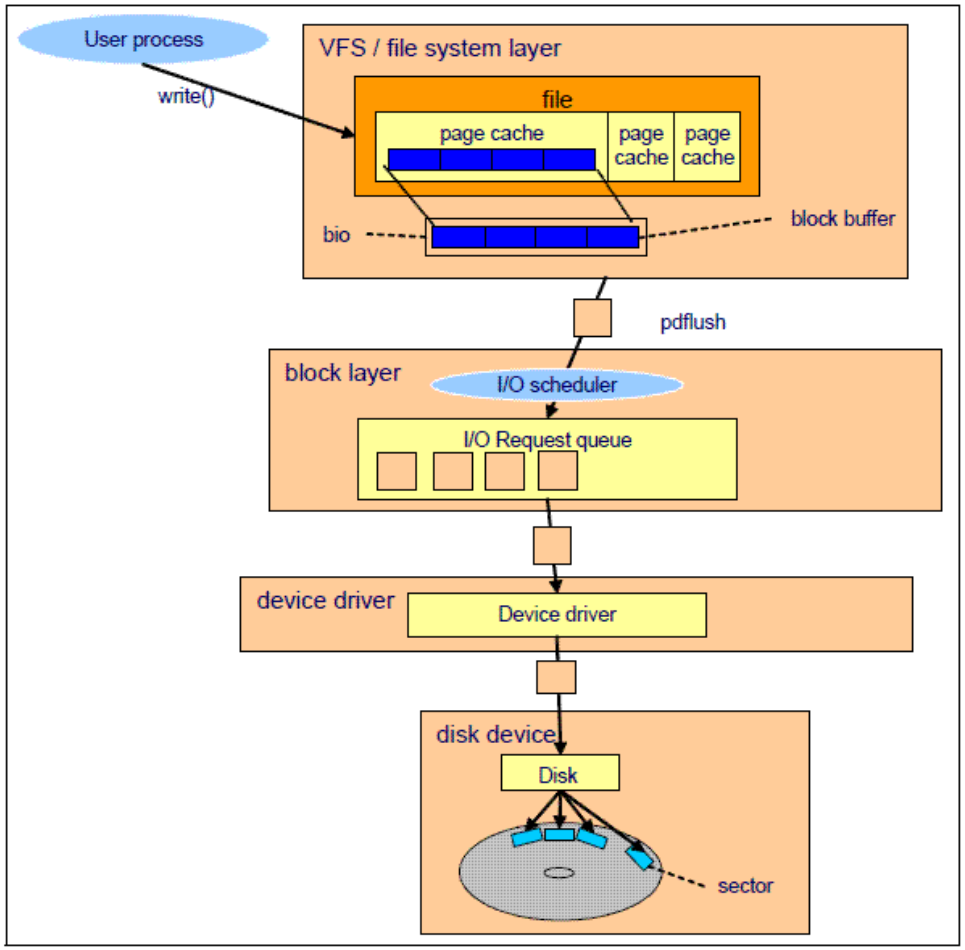
```
#---------------------------------------------------------------------------
# Benchmarking S_IWrite_Indv
# #processes = 1
# ( 11 additional processes waiting in MPI_Barrier)
#---------------------------------------------------------------------------
#    MODE: AGGREGATE
#
       #bytes #repetitions t_ovrl[usec] t_pure[usec]  t_CPU[usec]    overlap[%]
            0           50      9738.88         0.10      9748.94        100.00
            1           50     12430.68       396.02      9748.94             0.00
            2           50     10173.90       391.96      9748.94             0.00
            4           50     10148.18       387.00      9748.94             0.00
            8           50     10159.14       391.42      9748.94             0.00
           16           50     10137.48       386.76      9748.94             0.00
           32           50     10163.40       425.46      9748.94             2.58
           64           50     10640.22       390.00      9748.94             0.00
          128           50     10180.00       417.76      9748.94             0.00
          256           50     10195.06       431.76      9748.94             0.00
          512           50     10164.82       457.16      9748.94             0.03
         1024           50     10149.74       459.60      9748.94            12.79
         2048           50     10195.48       476.88      9748.94             5.36
         4096           50     10226.82       503.48      9748.94             5.08
         8192           50     10260.04       516.80      9748.94             1.10
        16384           50     10377.18       673.36      9748.94             5.70
        32768           50     10624.00       989.92      9748.94            11.60
        65536           50     11081.72      1365.18      9748.94             2.37
       131072           50     11856.80      2079.76      9748.94             0.00
       262144           50     12830.42      3290.72      9748.94             5.36
       524288           32     15544.09      5819.37      9748.94             0.42
      1048576           16     20655.44     11029.20      9748.94             1.26
      2097152            8     33746.00     21640.75      9748.94             0.00
      4194304            4     54378.45     42198.72      9748.94             0.00
      8388608            2     92724.08     83836.91      9748.94             3.84
     16777216            1    172389.03    163548.95      9748.94             9.32
```

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
OS Linux. Проблемы с асинхронным вводом-выводом.

“But the point is that AIO is needed just to cover up the fundamental idiocy in the interface. If the interface had been properly designed, it would have been useful _without_ AIO.”

Linus

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
OS Linux. Архитектура подсистемы ввода-вывода.



ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
OS Linux. Проблемы с прямым вводом-выводом.

“The thing that has always disturbed me about O_DIRECT is that the whole interface is just stupid, and was probably designed by a deranged monkey on some serious mind-controlling substances [*].”

“The right way to do it is to just not use O_DIRECT. The whole notion of "direct IO" is totally braindamaged. Just say no.

This is your brain: O
This is your brain on O_DIRECT: .

Any questions?.”

Linus

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
OS Linux. Прямой асинхронный ввод-вывод.

“For O_DIRECT to be a win, you need to make it asynchronous.”

Linus

Asynchronous I/O Support Linux 2.5, Ottawa Linux Symposium 2003

- io_submit
- io_setup
- io_getevents
- io_cancel
- io_destroy

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. Библиотека ROMIO.

- ROMIO - реализация ввода-вывода MPI. Используется в MPICH, MPICH2, MVAPICH, MVAPICH2, OpenMPI, IntelMPI...

- ADIO - набор низкоуровневых реализаций для различных файловых систем

  - ad_bgl
  - ad_bglockless
  - ad_gridftp
  - ad_lustre
  - ad_nfs
  - ad_ntfs
  - ad_panfs
  - ad_pfs
  - ad_pvfs
  - ad_pvfs2
  - ad_sfs
  - ad_testfs
  - ad_ufs
  - ad_xfs
  - ad_zoidfs
  - **common**

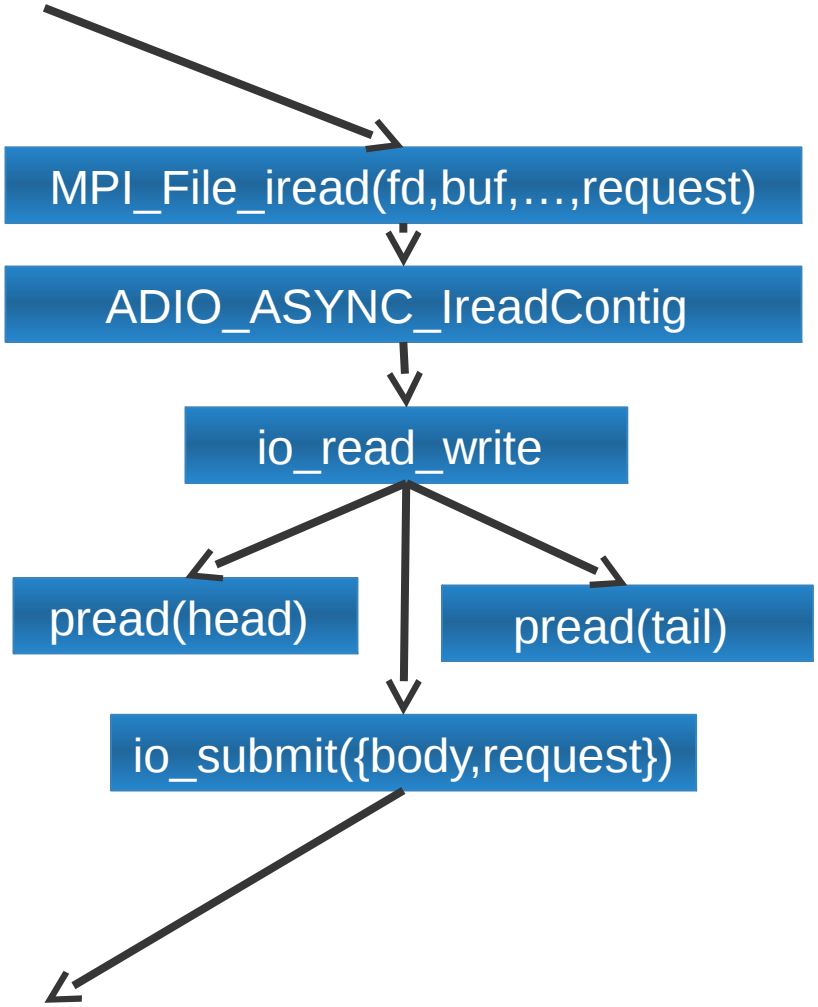
ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. Библиотека ROMIO.

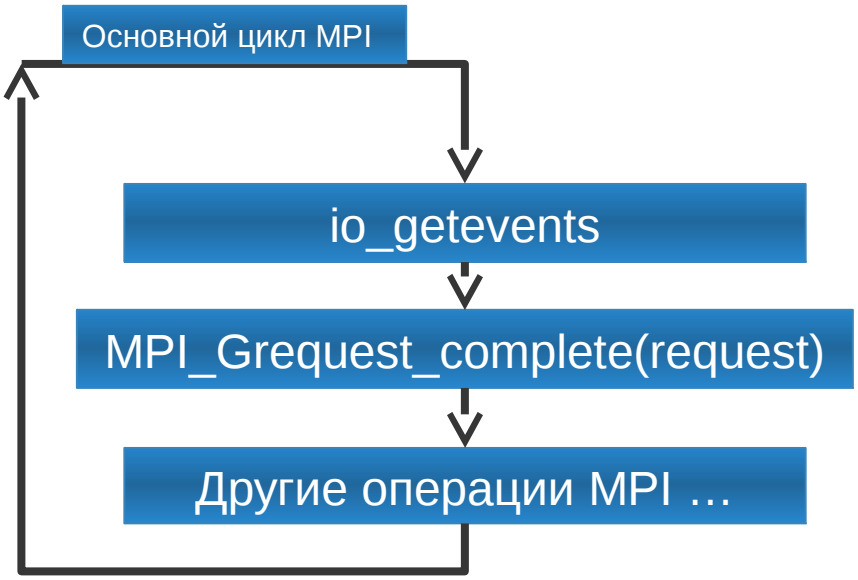
- ROMIO - реализация ввода-вывода MPI. Используется в MPICH, MPICH2, MVAPICH, MVAPICH2, OpenMPI, IntelMPI...

- ADIO - набор низкоуровневых реализаций для различных файловых систем

  - **ad_async**
  - ad_bgl
  - ad_bglockless
  - ad_gridftp
  - ad_lustre
  - ad_nfs
  - ad_ntfs
  - ad_panfs
  - ad_pfs
  - ad_pvfs
  - ad_pvfs2
  - ad_sfs
  - ad_testfs
  - ad_ufs
  - ad_xfs
  - ad_zoidfs
  - **common**

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI.
Библиотека ROMIO. Модуль ad_async. Принцип работы.





ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. РЕАЛИЗАЦИЯ НЕБЛОКИРУЮЩЕГО ВВОДА-ВЫВОДА. (S-MPI 0.3, gcc 4.4.6, IMB 3.2.4)

```
#---------------------------------------------------------------------------
# Benchmarking S_IWrite_Indv
# #processes = 1
#---------------------------------------------------------------------------
#    MODE: AGGREGATE
#
       #bytes #repetitions t_ovrl[usec] t_pure[usec]  t_CPU[usec]   overlap[%]
            0           50      9820.66         0.08      9835.00        100.00
            1           50      9627.30        15.74      9835.00        100.00
            2           50      9652.98        13.58      9835.00        100.00
            4           50      9676.26        16.04      9835.00        100.00
            8           50      9631.00        15.18      9835.00        100.00
           16           50      9757.94        12.44      9835.00        100.00
           32           50      9628.08        15.16      9835.00        100.00
           64           50      9750.88       264.56      9835.00        100.00
          128           50      9887.00        14.82      9835.00          0.00
          256           50      9632.78        17.72      9835.00        100.00
          512           50      9625.70        25.38      9835.00        100.00
         1024           50      9625.16        28.72      9835.00        100.00
         2048           50      9635.78        47.82      9835.00        100.00
         4096           50      9781.78        60.14      9835.00        100.00
         8192           50      9802.34       104.78      9835.00        100.00
        16384           50      9855.96       181.38      9835.00         83.45
        32768           50      9816.96       335.92      9835.00        100.00
        65536           50      9825.28       652.42      9835.00        100.00
       131072           50      9671.36      1287.52      9835.00        100.00
       262144           50     10466.78     12959.12      9835.00        100.00
       524288           32      9686.38      5079.60      9835.00        100.00
      1048576           16      9739.07     10171.82      9835.00        100.00
      2097152            8     17798.48     21010.99      9835.00        100.00
      4194304            4     35558.52     40842.77      9835.00        100.00
      8388608            2     76880.46     82368.49      9835.00        100.00
     16777216            1    155488.97    163975.00      9835.00        100.00
```

ПРОГРАММНЫЙ КОМПЛЕКС S-MPI. РЕАЛИЗАЦИЯ НЕБЛОКИРУЮЩЕГО ВВОДА-ВЫВОДА.