

Escola Supercomputador SDUMONT

Introdução E/S Paralela no SDUMONT
MPIIO - Tuning

André Ramos Carneiro (andrerc@lncc.br)
Bruno Alves Fagundes (brunoaf@lncc.br)

MPIIO - Tuning

Roteiro:

- MPI-IO
- ROMIO
- OMPIO

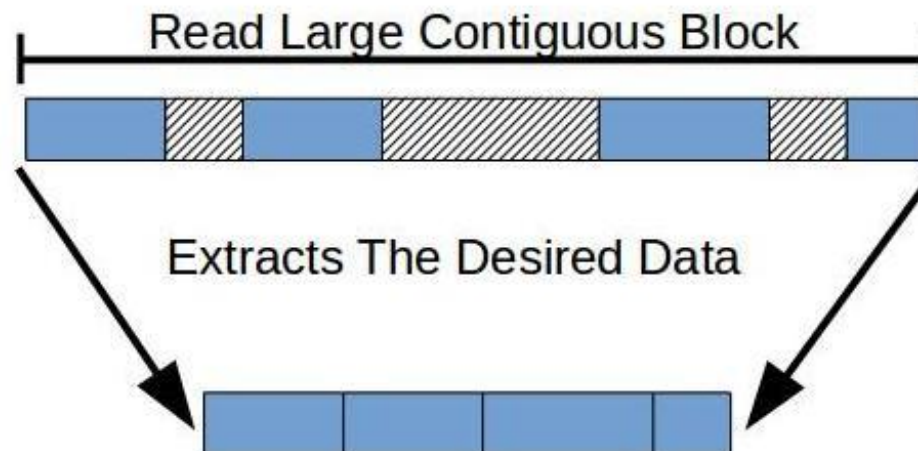
MPIIO

- Desenvolvido em 1994 no Laboratório Watson da IBM
- Fornecer suporte de Operações de E/S paralelas ao MPI
- 1996: Adotado pela NASA e incorporado pelo Forum do MPI no MPI-2.
- 1997: Publicação do MPI 2 com o MPI-IO já definido dentro.
- Chamadas de função do MPI-IO -> chamadas MPI.
- Escrever arquivos MPI -> enviar mensagens MPI.
- Ler arquivos MPI -> receber mensagens MPI.
- Versatilidade e flexibilidade dos tipos de dados MPI
- Define as visões de arquivo MPI (MPI file views).

- Implementação do MPI-IO portátil
- Utilizada por qualquer implementação MPI
- Já faz parte do código da maioria das implementações
- Duas técnicas de otimização de desempenho

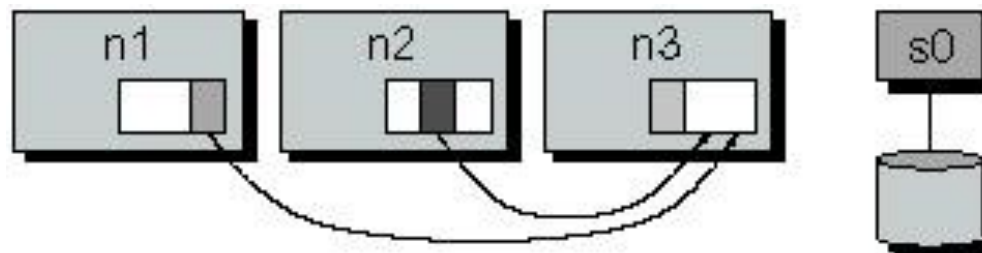
1) “Data Sieving”

- Focado em operações independentes com dados não contíguos
- Lê grandes blocos contíguos de dados e depois extrai as áreas de interesse

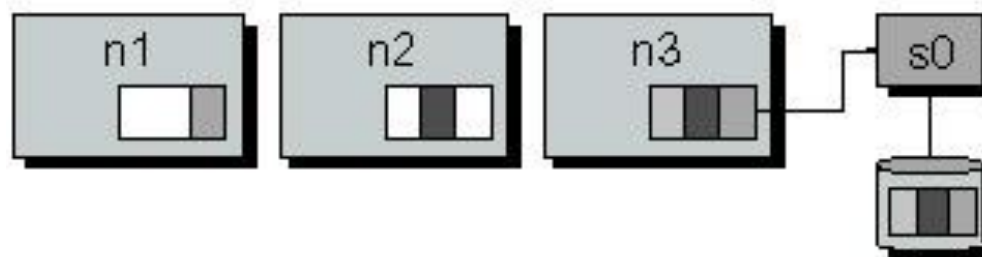


2) “Collective buffering” ou “Two phase I/O”

- ❑ Focado em operações coletivas
- ❑ Conjunto de processos (agregadores)
- ❑ Leem os dados do disco e distribuem para os demais processos
- ❑ Coletam os dados de todos os processos e os escrevem no disco



In step one, data is transferred to aggregators.



In step two, aggregators write data to storage.

- Data Sieving

- ind_rd_buffer_size (BYTES)

- Tamanho do buffer intermediário ao realizar operação de leitura

- ind_wr_buffer_size (BYTES)

- Tamanho do buffer intermediário ao realizar operação de escrita

- romio_ds_read [enable, disable, **automatic**]

- Determina realizar o data sieving para operações de leitura

- romio_ds_write [enable, disable, **automatic**]

- Determina realizar o data sieving para operações de escrita

- Collective Buffering

- `cb_buffer_size`: (BYTES)
 - ◆ Tamanho do buffer intermediário
- `cb_nodes`: (Num nós/processos)
 - ◆ Número máximo de agregadores (Hosts únicos)
- `romio_cb_read`: [enable, disable, **automatic**]
 - ◆ Controla quando o buffering coletivo é aplicado às operações de leitura coletiva
- `romio_cb_write`: [enable, disable, **automatic**]
 - ◆ Controla quando o buffering coletivo é aplicado às operações de escrita coletiva
- `cb_config_list`: Pode ser utilizado para um controle mais refinado
 - ◆ `*:2` → Cada host utilizará 2 processos para realizar E/S
 - ◆ `sdumontXXXX:24,*:0` → O nó especificado utilizará 24 processos para serem agregadores. Os demais nós não serão utilizados.

- Lustre

- romio_lustre_co_ratio: [Int] *C*liente/*O*ST

- Número máximo de cliente de E/S por OST. É configurado por padrão como CO = 1

- romio_lustre_coll_threshold: (BYTES) – limite para realizar E/S Col.

- Tamanho limite da requisição para realizar as operações coletivas. Acima do valor definido, não são realizadas. Configurado com o valor 0 significa sempre realizar as operações coletivas.

- romio_lustre_cb_ds_threshold: (BYTES) - limite para realizar Sieving.

- Otimiza as operações de E/S coletivas com uma versão do “data sieving”. Se a requisição de E/S for menor do que o valor desse hint, o “data sieving” não será utilizado.

- romio_lustre_ds_in_coll: [enable, disable]

- Habilitar/Desabilitar o Data Sieving nas chamadas coletivas

- * Para que essas Hints estejam disponíveis, é necessário que o ROMIO tenha sido compilado com suporte ao Filesystem Lustre

- Configurar

- Um arquivo texto contendo:

- hint_1 valor
 - hint_2 valor
 - etc....

- Configurar a variável de ambiente ROMIO_HINTS, apontando para o arquivo:

- `export ROMIO_HINTS=/scratch/PROJETO/USUARIO/arquivo_de_hints`

- Executar a aplicação através do mpiexec/mpirun/srun

- Exibir as Hints utilizadas pelo ROMIO

- `export ROMIO_PRINT_HINTS=1`

MPIIO – ROMIO - Ativar

- Intel MPI - PSXE 2016, 2017 e 2018
 - `export I_MPI_EXTRA_FILESYSTEM=on`
 - `export I_MPI_EXTRA_FILESYSTEM_LIST=lustre`
- Intel MPI - PSXE 2019
 - `export I_MPI_EXTRA_FILE_SYSTEM=on`
- OpenMPI - Ativado através do MCA = Modular Component Architecture
 - `2.X | 3.X: export OMPI_MCA_io=romio314`
 - `4.X: export OMPI_MCA_io=romio321`

MPIO – OMPIO

- Implementação desenvolvida pelo Open-MPI.org
- Introduzida a partir da versão 1.7 (default a partir da versão 2.0)
- Suporte ao Lustre a partir da versão 2.0
- Objetivos
 - Aumenta a modularidade da biblioteca de E/S Paralela
 - Permite aos frameworks utilizar diferentes algoritmos de decisão
 - Melhora a integração das funções de E/S paralelas
- Sub-frameworks:
 - **fs framework**: Gerenciamento de todas as operações com arquivos
 - **fbtl framework**: Operações individuais de E/S blocking e non-blocking
 - **fcoll framework**: Operações coletivas de E/S blocking e non-blocking
 - **sharedfp framework**: Operações de arquivos compartilhados

- **io_ompio_cycle_buffer_size:** (BYTES)
 - Tamanho dos dados utilizados por chamadas de I/O individuais.
- **io_ompio_bytes_per_agg:** (BYTES)
 - Tamanho do buffer temporário para operações coletivas de I/O nos processos agregadores. **(32MB)**
- **io_ompio_num_aggregators:** (INT)
 - Número de processos agregadores utilizados nas operações coletivas de I/O.
- **io_ompio_grouping_option:** [1-7]
 - Algoritmo utilizado para decidir automaticamente o número de agregadores utilizados. 1: Data volume based | 2: maximizing group size uniformity | 3: maximimze data contiguity | 4: hybrid optimization | 5: simple (default) | 6: skip refinement step | 7: simple + grouping based on default file view
- **fs_lustre_stripe_size:** (BYTES)
 - Define o tamanho do stripe size de um arquivo no filesystem lustre
- **fs_lustre_stripe_width:** (INT)
 - Define o stripe count de um arquivo no filesystem lustre

- Lista completa de todos os parâmetros:
 - ❑ `mpi_info --level 9 --param io ompio`
 - ❑ `mpi_info --level 9 --param fcoll all`
 - ❑ `mpi_info --level 9 --param fs all`
 - ❑ `mpi_info --level 9 --param fbt1 all`
 - ❑ `mpi_info --level 9 --param sharedfp all`

MPIIO – OMPIO – Parâmetros

- Para utilizar:

- ☐ export OMPI_MCA_io=ompio
- ☐ export OMPI_MCA_io_ompio_bytes_per_agg=536870912
- ☐ export OMPI_MCA_io_ompio_num_aggregators=5
- ☐ export OMPI_MCA_fs_lustre_stripe_width=5
- ☐ export OMPI_MCA_fs_lustre_stripe_size=5242880

FIM

Dúvidas?