Lustre

Escola Supercomputador SDUMONT

Introdução E/S Paralela no SDUMONT
Lustre

André Ramos Carneiro
Bruno Alves Fagundes

Lustre

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica

Roteiro:

- O Sistema de Arquivos Lustre
- Arquitetura
- I/O + Stripe
- Lustre no SDumont
- Utilitário lfs
  - Configuração de Stripe
  - Comandos Básicos
- Melhores Práticas e Recomendações

O Sistema de Arquivos Lustre

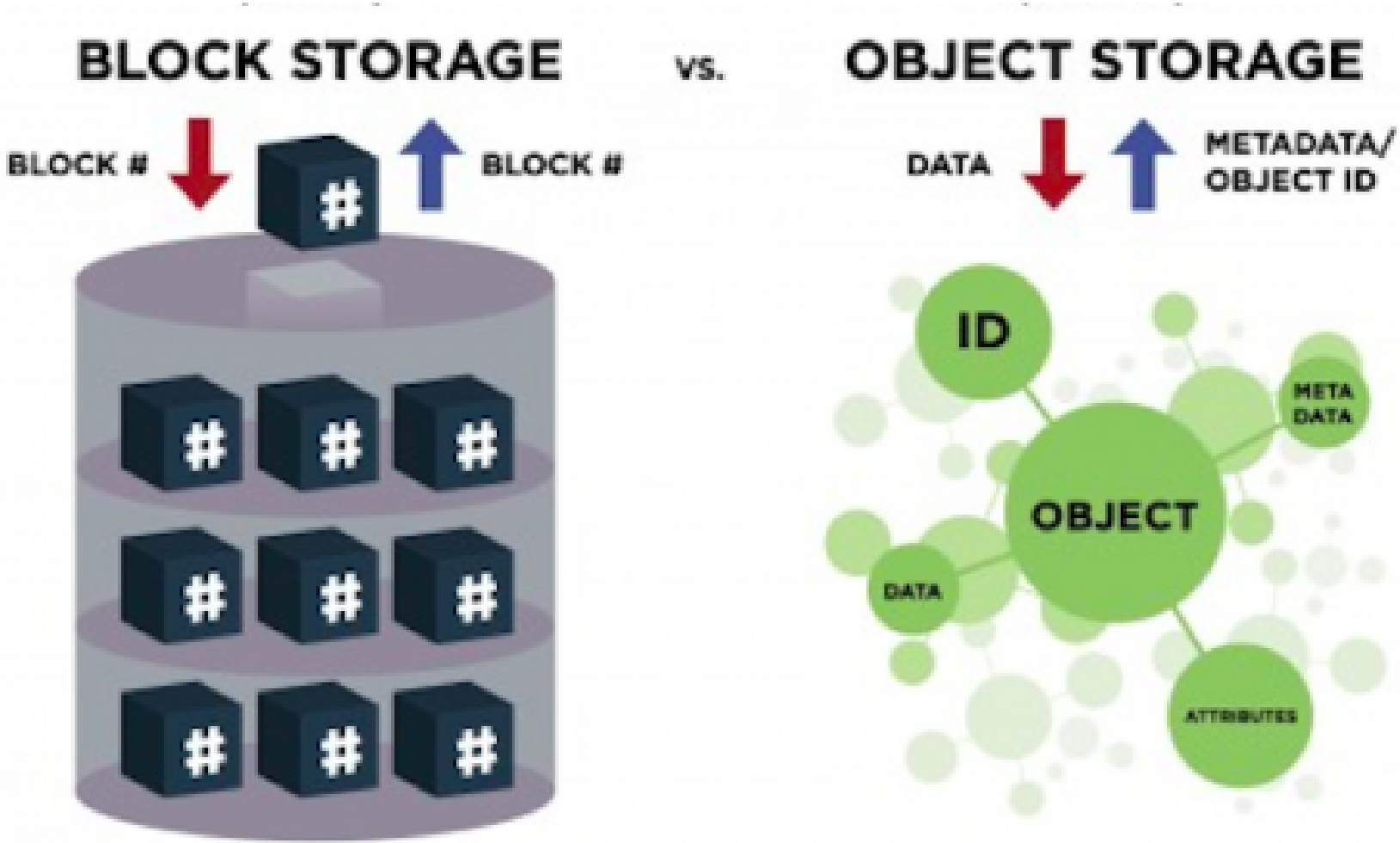
- Filesystem OpenSource
- Projetado para ser escalável
- Capaz de lidar com um grande volume de dados e um enorme número de arquivos
- Alta disponibilidade
- Coerência de dados e metadados
- Armazenamento Baseado em Objetos

O Sistema de Arquivos Lustre

- Filesystem ext4 (modificado) ou zfs:
- Conformidade com o Padrão POSIX:
- Rede heterogênea de alto desempenho:
  - InfiniBand, TCP (10GE e IpoIB), Myricom, Cray Seastar, Omni-Path
- Locks de arquivo e metadados, na granularidade de byte (LDLM)
- Striping
- MPI I/O – ADIO (*Abstract-Device Interface for I/O*)
- Escalável
- NFS e CIFS

Object Based Storage
Armazenamento Baseado em Objeto

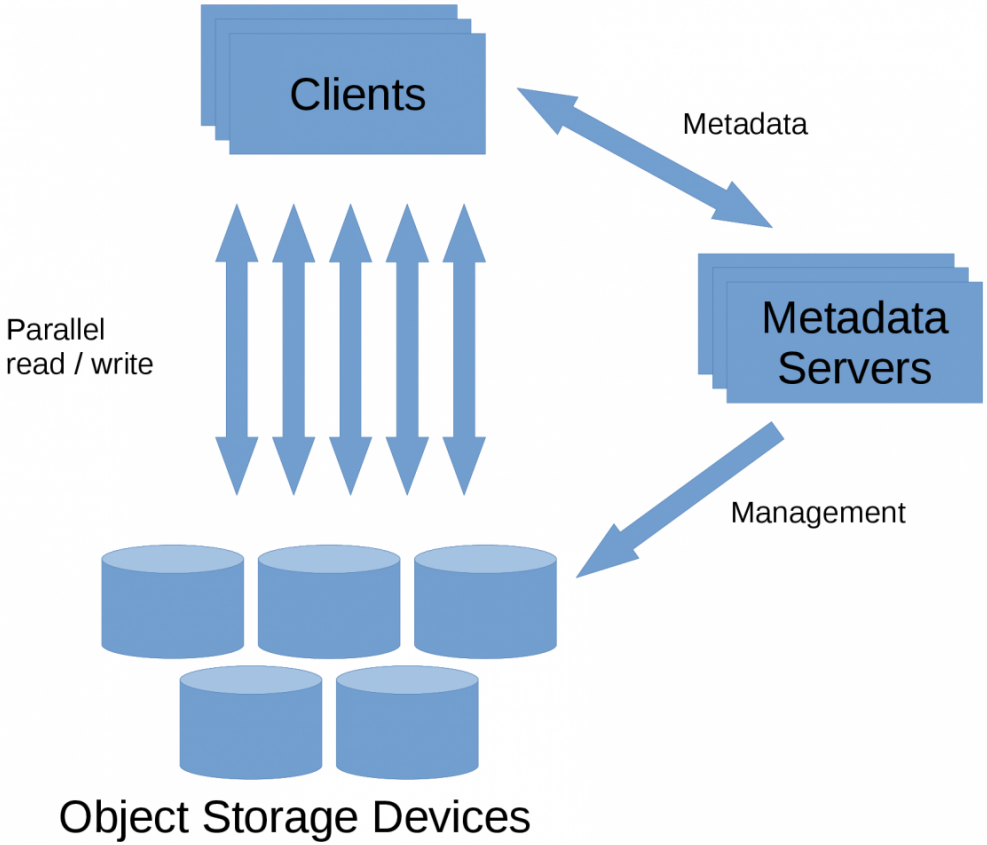
LNCC
www.lncc.br
Laboratório Nacional de Computação Científica



Object Based Storage
Armazenamento Baseado em Objeto

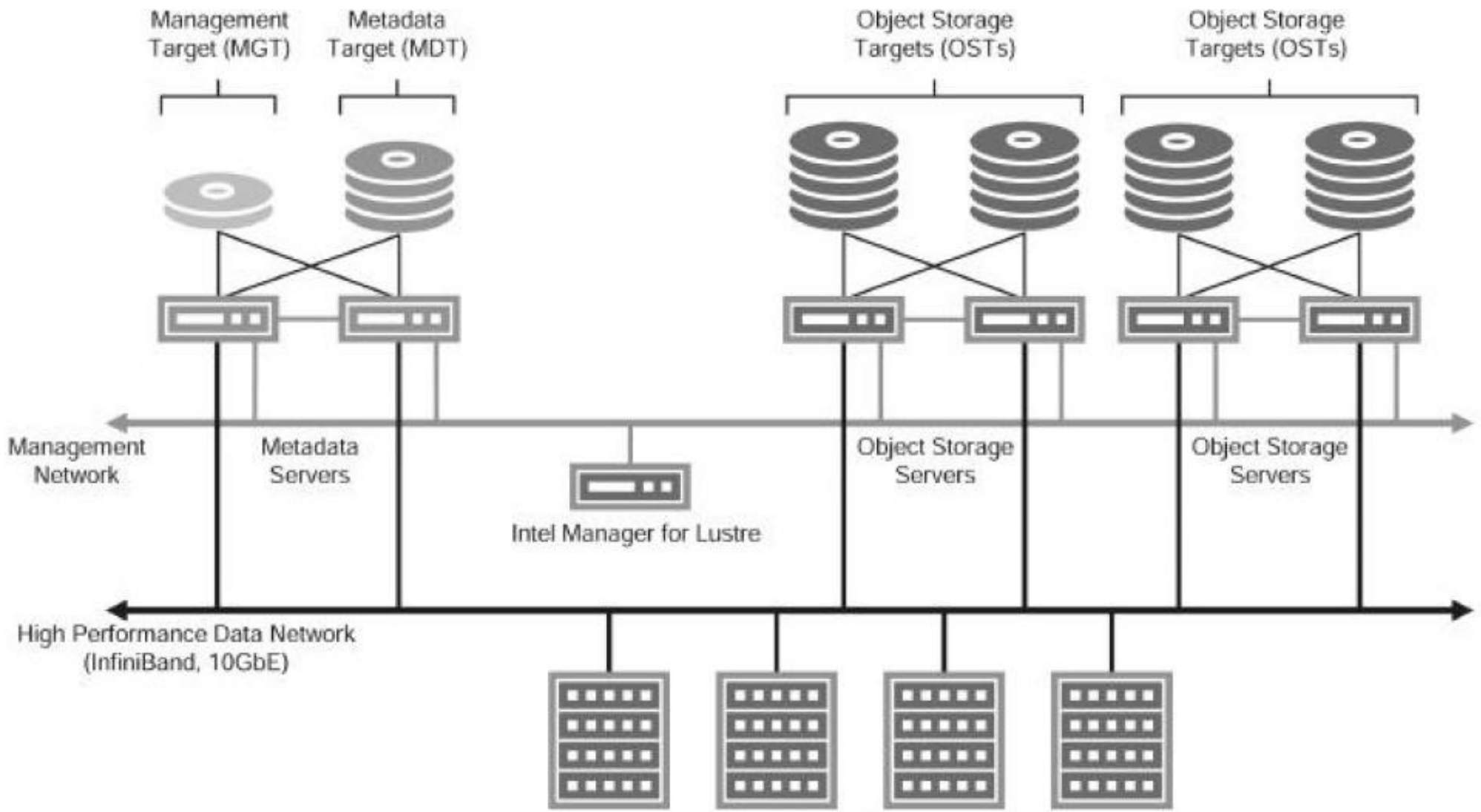
LNCC
www.lncc.br
Laboratório Nacional de Computação Científica

Object Based



Arquitetura

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica



Arquitetura

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica

- MGS – Management Server – Servidor de Gerenciamento
- MDS – Metadata Server – Servidor de Metadados
- MDT – Metadata Target – Onde os objetos de metadados são armazenados
- OSS – Object Storage Server – Servidor de Armazenamento de Objetos
- OST – Object Storage Target – Onde os objetos de dados são armazenados
- Clientes Lustre
- LNET (ptlrpc)

Arquitetura

- Objetos Lustre:
  - (1) objetos de dados: arrays de bytes para armazenar os dados dos arquivos
  - (2) objetos de índice (metadados): contêineres que armazenam pares de chave-valor – implementando a estrutura de diretórios POSIX
- Objetos implementados pelo Lustre OSD (object storage device)
  - ldiskfs (ext4) ou zfs
- Dispositivo de block (disco, lun) → 1 OSD → storage target → MDT ou OST

Arquitetura

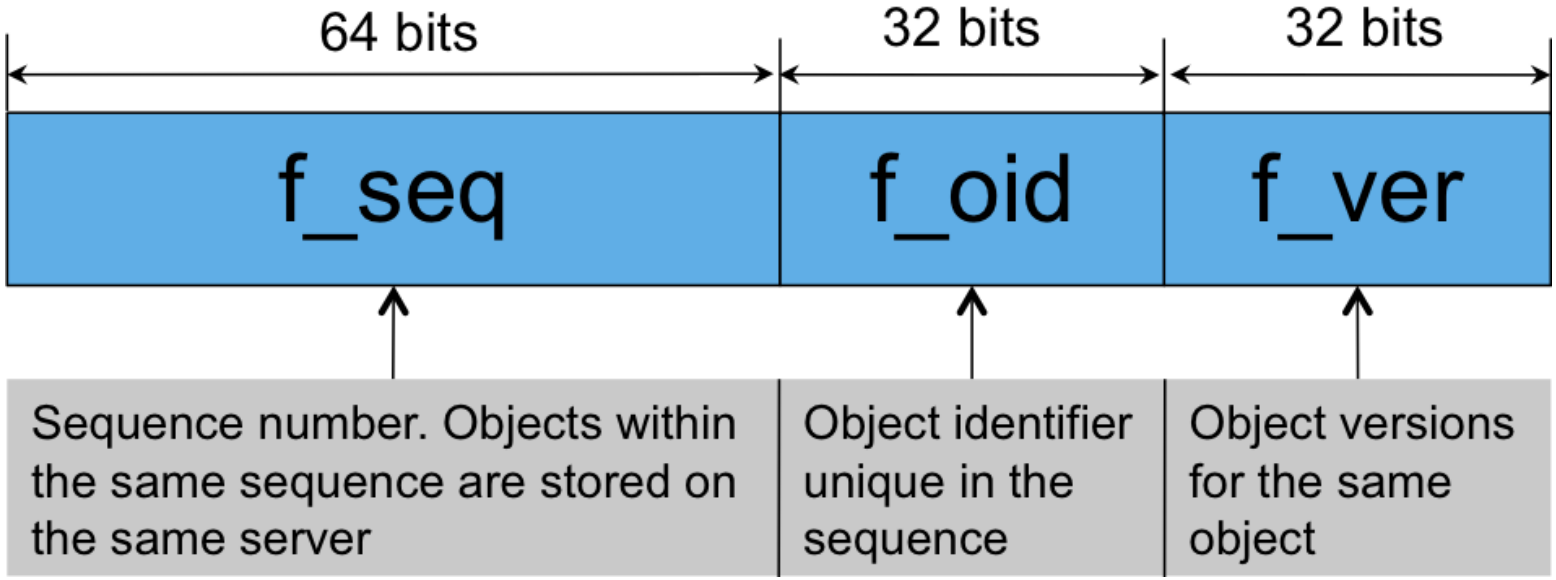
- storage target
  - MDTs – Metadata targets: Armazenam os metadados (como nomes de arquivos, diretórios, permissões de layout de arquivo)
  - OSTs – Object Targets: Armazenam os dados dos arquivos
- MDSs – Servidores de Metadados: Exportam os MDTs. Operações no namespace (como busca de arquivos, criação de arquivo, e manipulação de atributos de arquivo e diretório)
- OSSs – Servidores de Objetos: Exportam os OSTs. Fornecem serviço de I/O de arquivo, além de tratamento de requisições de rede para um ou mais OSTs locais.

Arquitetura

- LNET
  - Abstração das redes físicas (IB / TCP/IP)
  - Troca de mensagens de RDMA
- Lustre RPC (ptlrpc): Comunicação entre clientes e servidores

- LDLM – Lustre Distributed Lock Manager: serviço de lock fornecido pelos storage targets (MDT ou OST)
  - Serializa operações conflitantes
  - Garante coerência do cache distribuído

Armazenamento e Operações de E/S



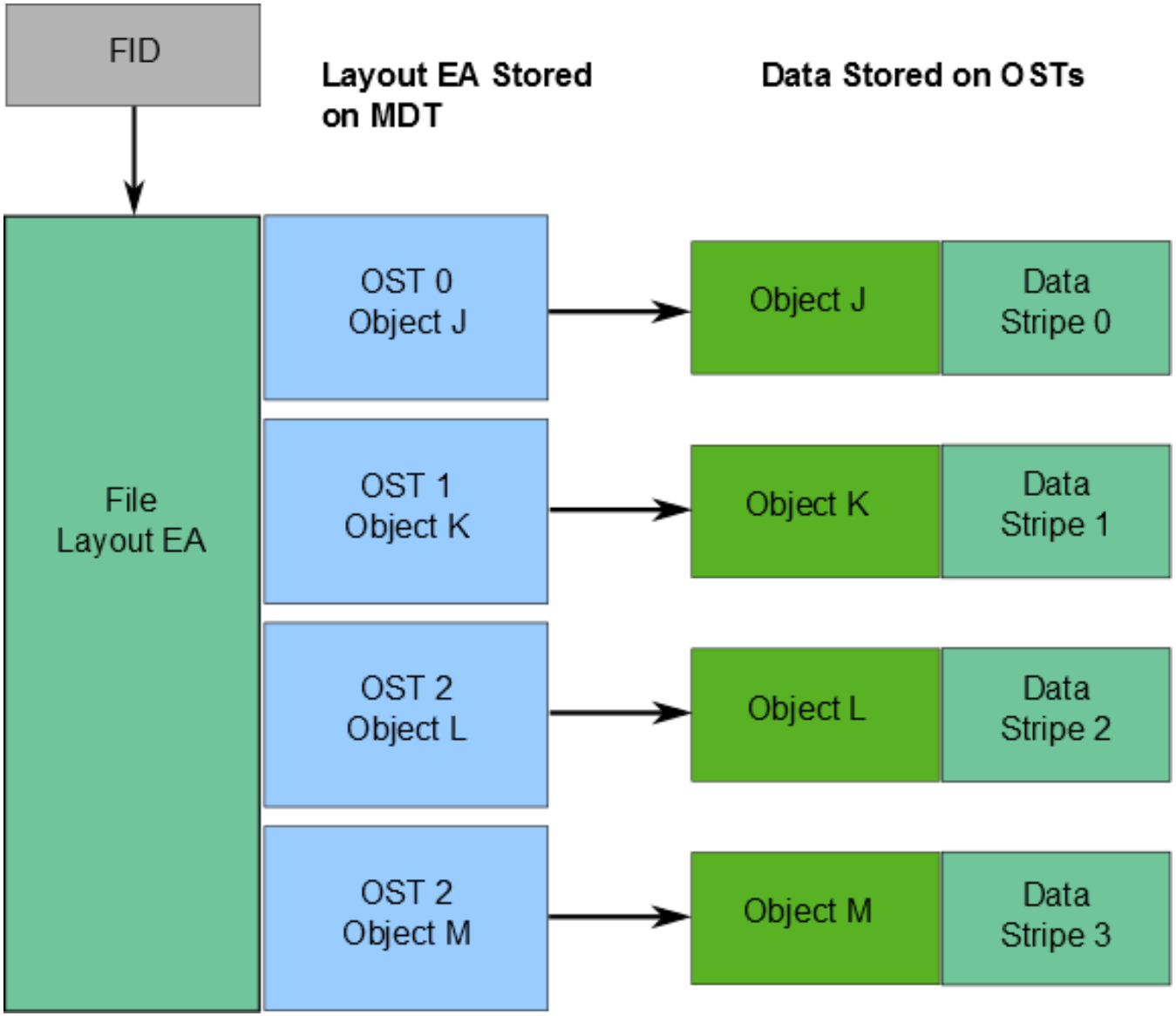
- FID – File Identifiers: Introduzido na versão 2.0. Substitui os números de inodes para identificar arquivos ou objetos de forma única

Armazenamento e I/O no Lustre FS

- ***Layout EA***: Atributo estendido que armazena informação sobre onde os dados do arquivo estão localizados no(s) OST(s)
  - Armazenado no Objeto MDT, identificado pelo FID do arquivo.
- Se o arquivo é um arquivo regular, o objeto MDT possui um relacionamento que aponta de 1-para-N objeto(s) OST
- Se o objeto MDT apontar para mais do que um objeto → arquivo foi dividido em stripes, sendo cada objeto (stripe) armazenado em um OST diferente

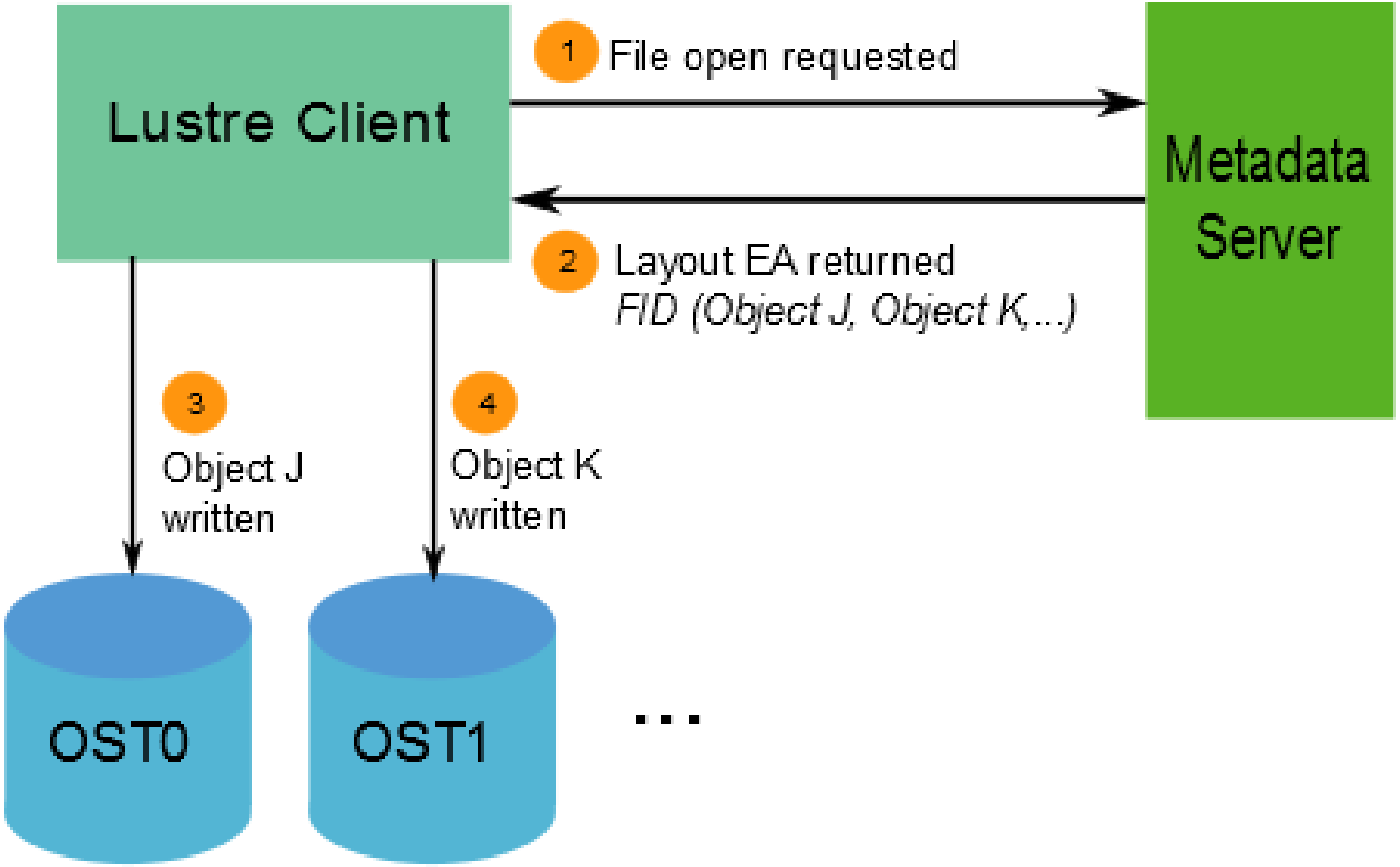
Armazenamento e I/O no Lustre FS

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica



Armazenamento e I/O no Lustre FS

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica

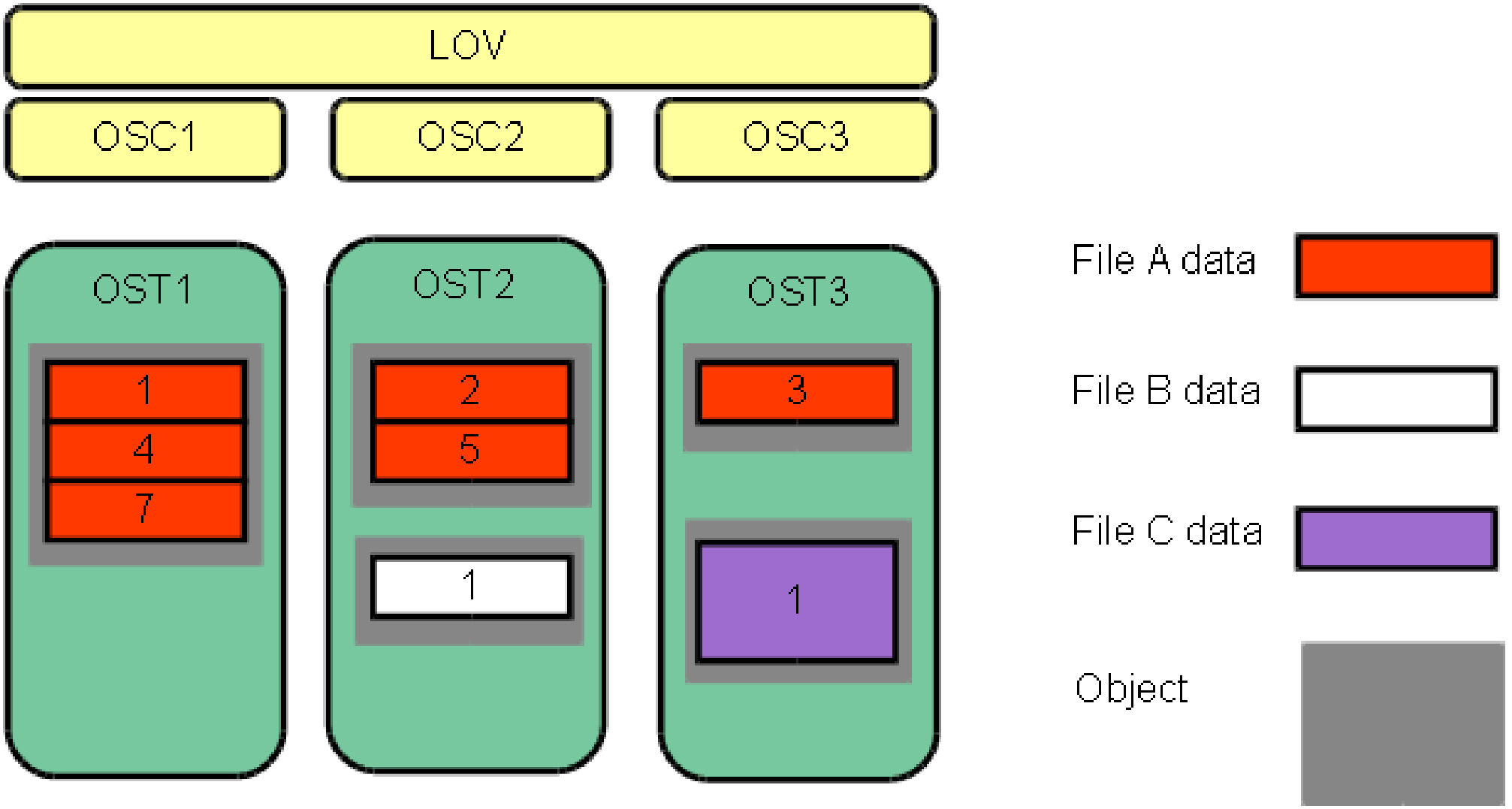


Armazenamento e Striping

- Dividir arquivos em stripes (“faixas”) em um esquema RAID 0, sendo cada uma armazenada em objetos/OSTs diferentes
- Utilizado para melhorar o desempenho
  - Largura de banda agregada para um único arquivo excede a largura de banda de um único OST/OSS
  - OST não possui espaço suficiente para armazenar o arquivo inteiro
- stripe_count: determina em quantos objetos de dados um arquivo será dividido
- stripe_size: determina a quantidade de dados que será armazenada em cada pedaço objeto

Armazenamento e I/O no Lustre FS
Striping

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica



Armazenamento e I/O no Lustre FS Striping

- O tamanho máximo do arquivo não é limitado pelo tamanho de um único target
- O valor **padrão** para o stripe_count é **1** (limitado pelo número de OST / Máx 2000)
- O valor **padrão** para o stripe_size é **1MB**

- A largura de banda de I/O para acessar um único arquivo é a largura de banda agregada de I/O para os objetos do arquivo

Lustre do SDUMONT

LNCC
www.lncc.br
Laboratório Nacional de Computação Científica

- Solução ClusterStor da Xyratex (Seagate)
- 2 nós de Administração
- 2 nós MGS / MDS (Ativo/Ativo Alternados)
- 10 nós OSS (5 pares Ativo/Ativo)
- Utilizando interconexão Infiniband
  - `172.20.250.103@o2ib,172.20.250.104@o2ib:/cstor → /scratch`
- Tamanho total do volume – 1.7 PB
- Versão Servidor v2.1 e Clientes v2.4.3

lfs – Configurar o Striping

- **lfs** *setstripe* -s stripe_size -c stripe_count \ filename|dirname
- Exemplo – configura o layout para um diretório. Todos os arquivos criados dentro do diretório possuirão o mesmo layout:

```
cd $SCRATCH

mkdir full_stripe

lfs setstripe -s 4m -c -1 full_stripe

##

mkdir single_stripe

lfs setstripe -s 1m -c 1 single_stripe
```

lfs – Configurar o Striping

- **lfs** *getstripe* filename|dirname
- Exemplo – lista informações do diretório (único stripe):

```
cd $SCRATCH

lfs getstripe single_stripe

single_stripe

stripe_count: 1 stripe_size:  1048576 stripe_offset:  -1
```

lfs – Configurar o Striping

- **lfs** *getstripe* filename|dirname
- Exemplo – cria um arquivo e lista suas informações (único stripe):

```
dd if=/dev/zero of=single_stripe/teste_single_stripe bs=1M count=1k

lfs getstripe single_stripe/teste_single_stripe

single_stripe/teste_single_stripe

lmm_stripe_count:   1

lmm_stripe_size:    1048576

lmm_layout_gen:     0

lmm_stripe_offset:  3

obdidx   objid             objid          group

 3         1167913         0x11d229       0
```

lfs – Configurar o Striping

- **lfs** *getstripe* filename|dirname
- Exemplo – lista informações do diretório (full stripe):

```
lfs getstripe -d full_stripe

full_stripe

stripe_count: -1 stripe_size:  4194304 stripe_offset:  -1
```

lfs – Configurar o Striping

Exemplo – cria um arquivo e lista suas informações (full stripe):

```
dd if=/dev/zero of=full_stripe/teste_full_stripe bs=4M count=256
lfs getstripe full_stripe/teste_full_stripe
full_stripe/teste_full_stripe
lmm_stripe_count:   10
lmm_stripe_size:    4194304
lmm_layout_gen:     0
lmm_stripe_offset:  8
obdidx   objid   objid   group
     8         1167882      0x11d20a              0
     2         1167720      0x11d168              0
     9         1166249      0x11cba9              0
     1         1167274      0x11cfaa              0
     7         1165803      0x11c9eb              0
     5         1152459      0x1195cb              0
     3         1167915      0x11d22b              0
     6         1167210      0x11cf6a              0
     4         1166409      0x11cc49              0
     0         1167722      0x11d16a              0
```

lfs – Comandos Básicos

- `lfs df`: exibe a utilização do espaço no Lustre

- `lfs find`: realiza busca no Lustre
  - `lfs find . -atime -1 -name '*.f90'`
  - Procura por códigos fonte em Fortran acessados no último dia

- `lfs cp`: realiza cópias no Lustre

lfs – Comandos Básicos

- `lfs ls`: Listar diretórios e arquivos
- `lfs quota -u|-g <filesystem>`: Obtém a quota → Utilização do espaço no Lustre
  - `lfs quota -g PROJETO /scratch/PROJETO`
- Para uma lista completa dos comandos:
  - `lfs help`
- Para informações específicas de um comando:
  - `lfs help "comando"`

Boas Práticas - striping

- Algumas razões para utilizar o striping incluem:
  - Fornece alta largura de banda para o acesso → quanto mais OSTs = mais vias para acessar o arquivo
  - Fornecendo espaço para arquivos muito grandes → quanto mais OSTs = mais espaço agregado para armazenamento
- Cuidados ao utilizar o striping:
  - Aumento do overhead → Aumenta o número de locks e pode gerar “Disputas por I/O”
  - Aumento do risco → Em caso de falha de um OSS/OST, parte do arquivo se perde. Agravado em ambientes com muitos OSTs

Boas Práticas - striping

- O *stripe_size* não possui efeito em um arquivo com *stripe_count* igual a um
- O tamanho do stripe deve ser um múltiplo do tamanho da página (64KB)

```
lfs setstripe -s 614400 full_stripe

error: bad stripe_size 614400, must be an even multiple of 65536 bytes

error: setstripe: create stripe file 'full_stripe' failed
```

- O menor tamanho recomendado de stripe é 512KB
- Um bom tamanho de stripe para I/O sequencial utilizando redes de alta velocidade é entre 1MB e 4MB
- O tamanho máximo para o stripe é de 3.999GB

```
lfs setstripe -s 4096M full_stripe

warning: stripe size 4G or larger is not currently supported and would wrap

error: setstripe: create stripe file 'full_stripe' failed
```

- Escolher um padrão de stripe que leve em consideração os padrões de escrita da aplicação

Boas Práticas

- Evite utilizar o comando “ls -l” (especialmente em diretórios)
  - “`lfs ls`” se deseja verificar que um arquivo existe.
  - “`lfs ls -l nome-no-arquivo`” se deseja obter maiores informações de um arquivo específico.
- Evite ter um grande número de arquivos em um único diretório
- Evite acessar pequenos arquivos no Lustre (< 10MB)
- Utilize um “Stripe count” igual a 1 para diretórios com muitos arquivos pequenos

Boas Práticas

- Mantenha o código fonte no HOMEDIR
  - Compilar dentro do HOME → Mover para o SCRATCH
- Experimente diferentes valores para o stripe count/size para as operações de escritas coletivas do MPI (*mais adiante*)
- Alinhar as operações de E/S com o stripe para minimizar a contenção/disputa

Dúvidas?