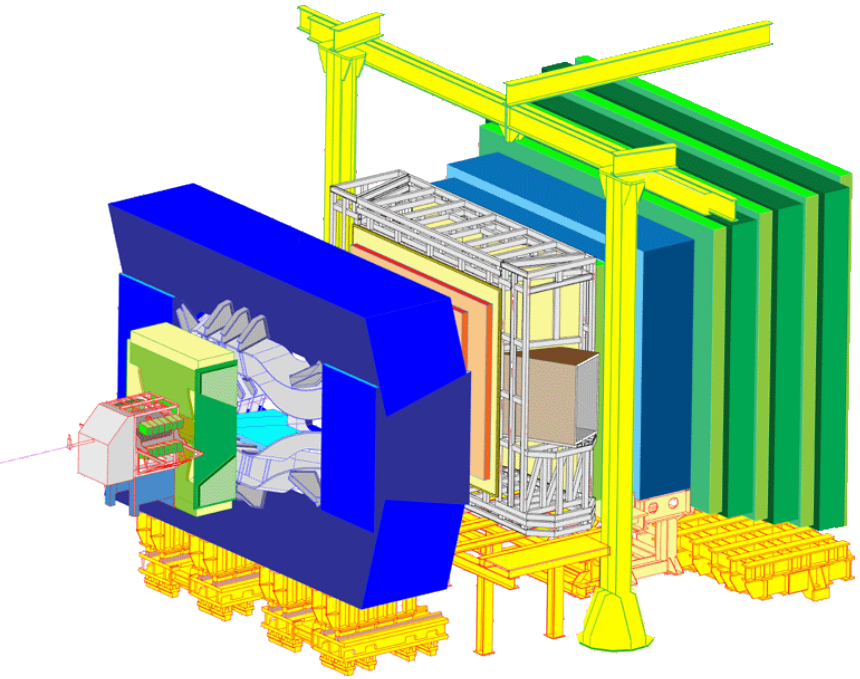
ECFS

A decentralized, distributed and fault-tolerant FUSE filesystem for the LHCb online farm

Session: Data Stores, Data Bases, and Storage Systems

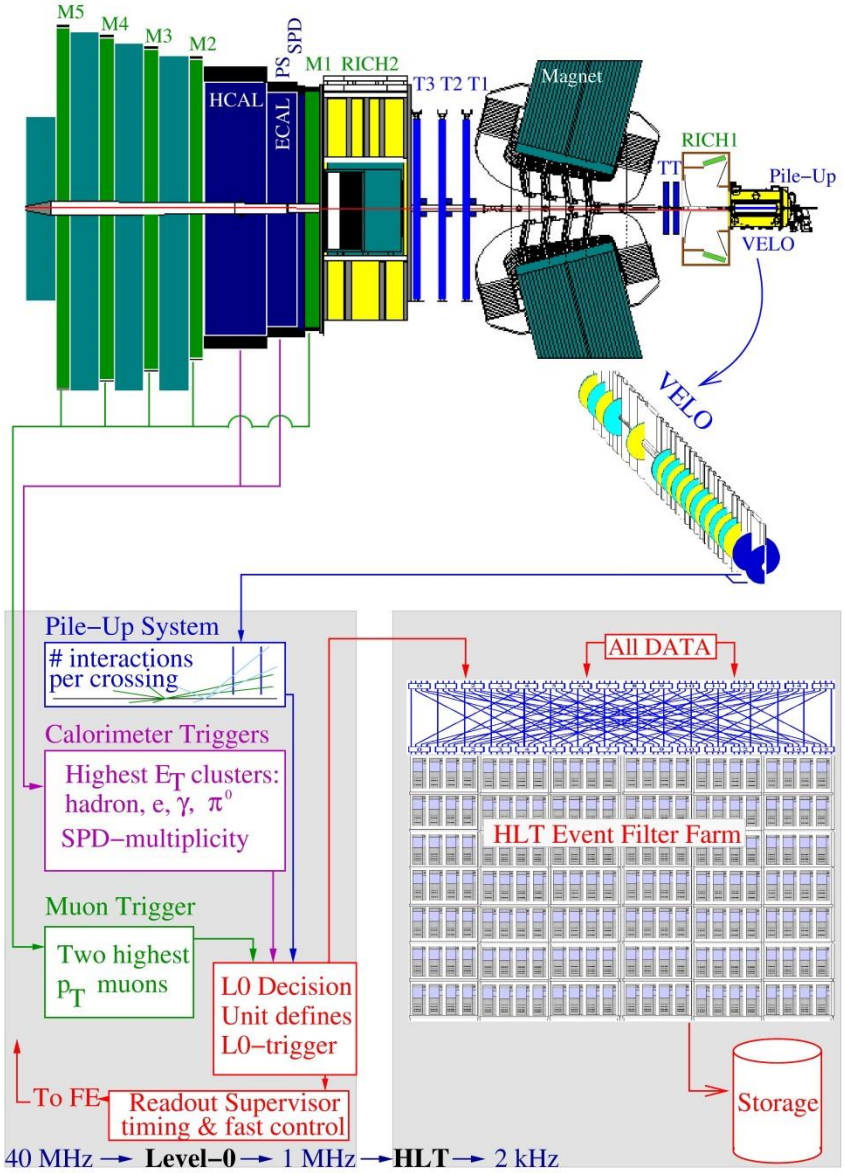
tomasz.rybczynski@cern.ch*
enrico.bonaccorsi@cern.ch
niko.neufeld@cern.ch

Introduction

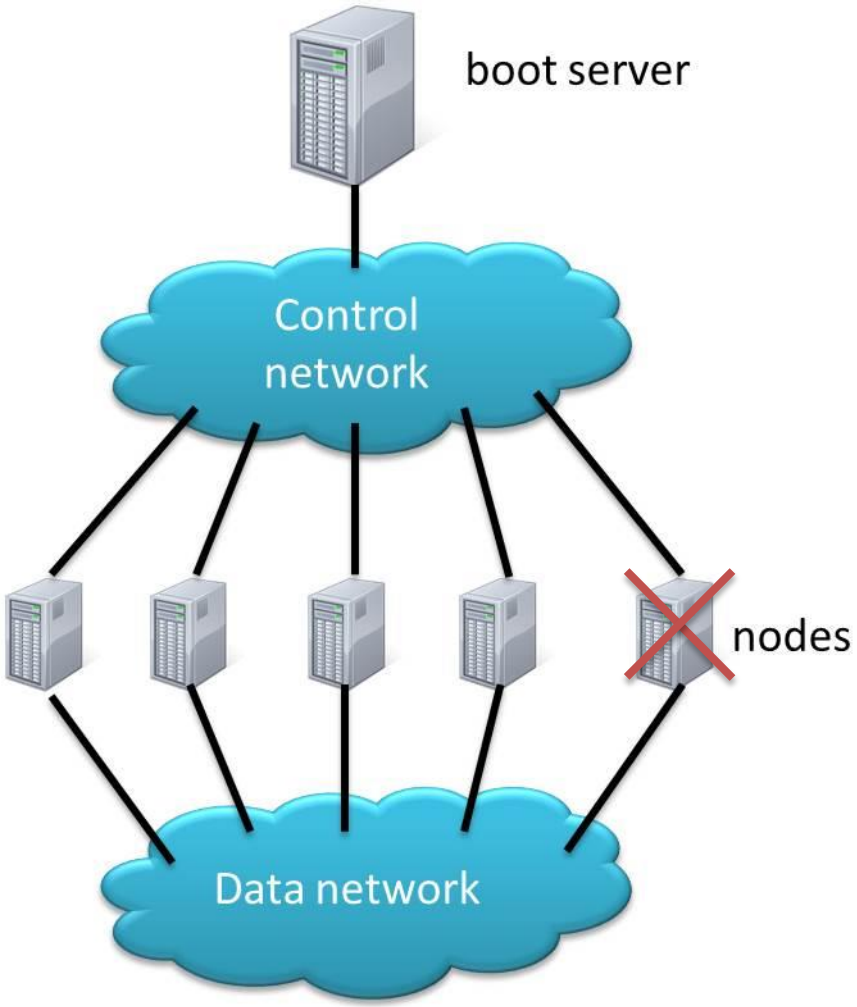


Event Filter Farm
~2000 nodes

High Level Trigger



Event Filter Farm



20-25% of the time

Requirements

- Fault tolerance – data redundancy
- Single namespace
- POSIX semantics
- Write once read many
- Sufficient performance

Searching for a solution

- Linux RAID over HA cluster
- GlusterFS
- Tahoe-LAFS tahoe-lafs.org
- …

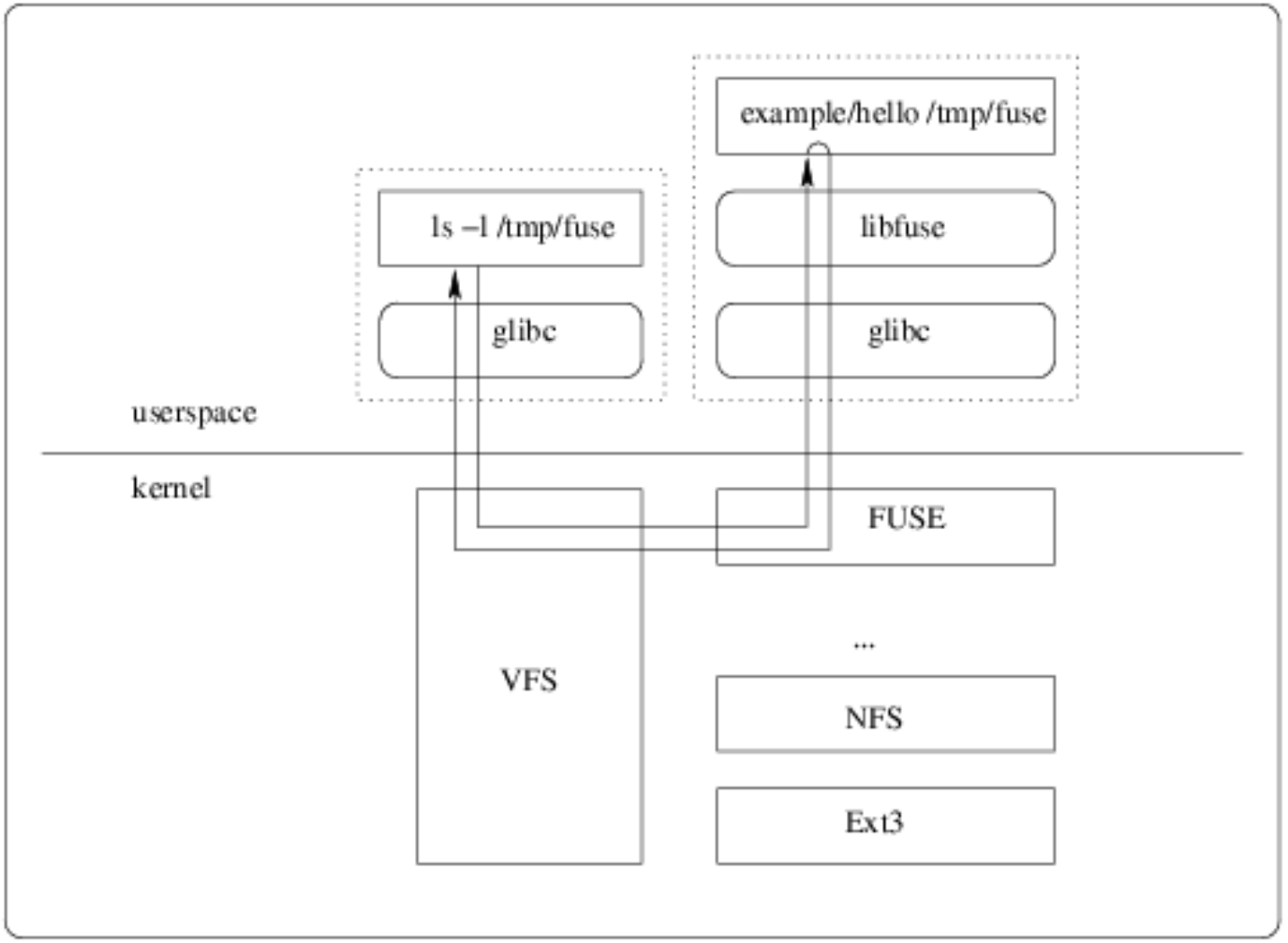
Erasure Codes File System

ECFS - distributed, fault-tolerant file system

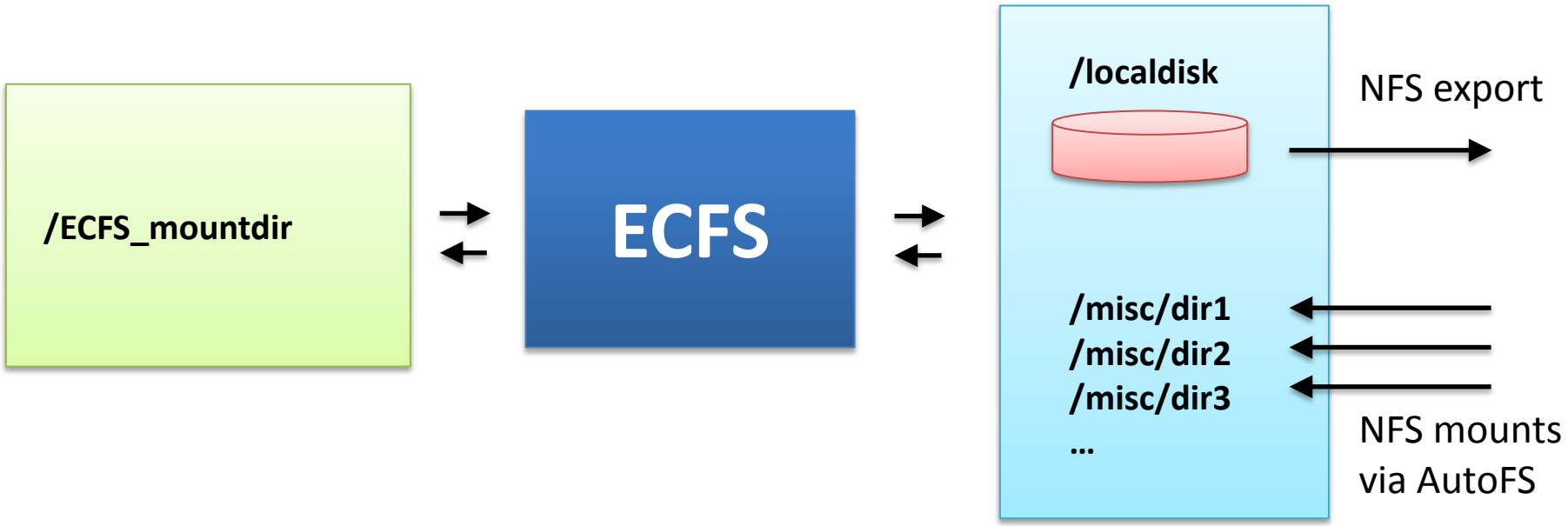
Project based on:

- FUSE
- NFS
- Erasure Codes

FUSE


example/hello /tmp/fuse
ls –l /tmp/fuse
libfuse
glibc
glibc
userspace
kernel
FUSE
VFS
...
NFS
Ext3

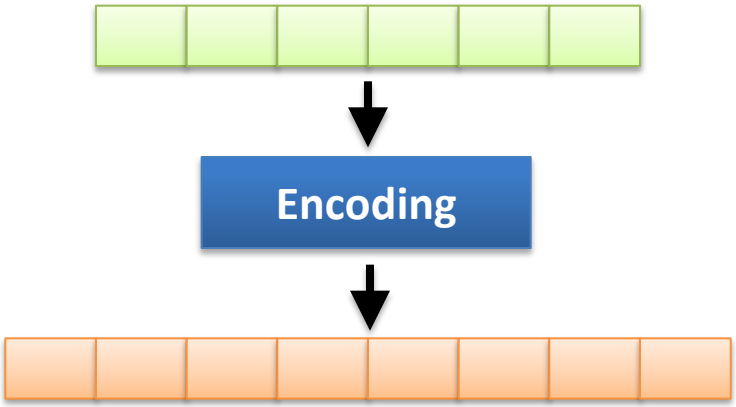
NFS


/ECFS_mountdir
ECFS
/localdisk
NFS export
/misc/dir1
/misc/dir2
/misc/dir3
…
NFS mounts
via AutoFS

Erasure Codes

Coding parameters:
$k = 6$
$n = 2$

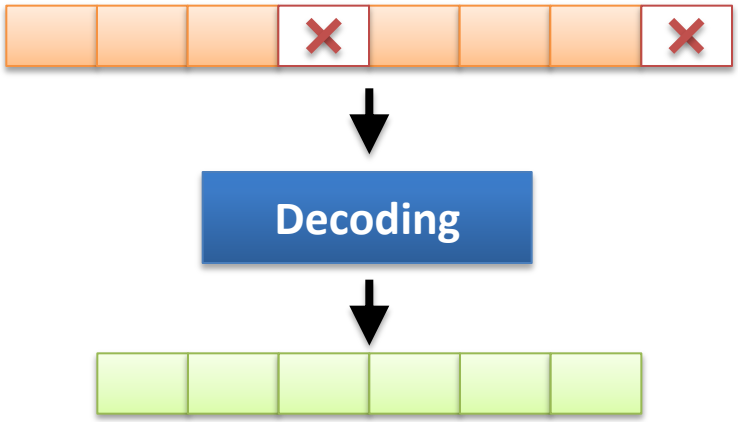
Original message of a size of k symbols



Code word of a size of $k + n$ symbols

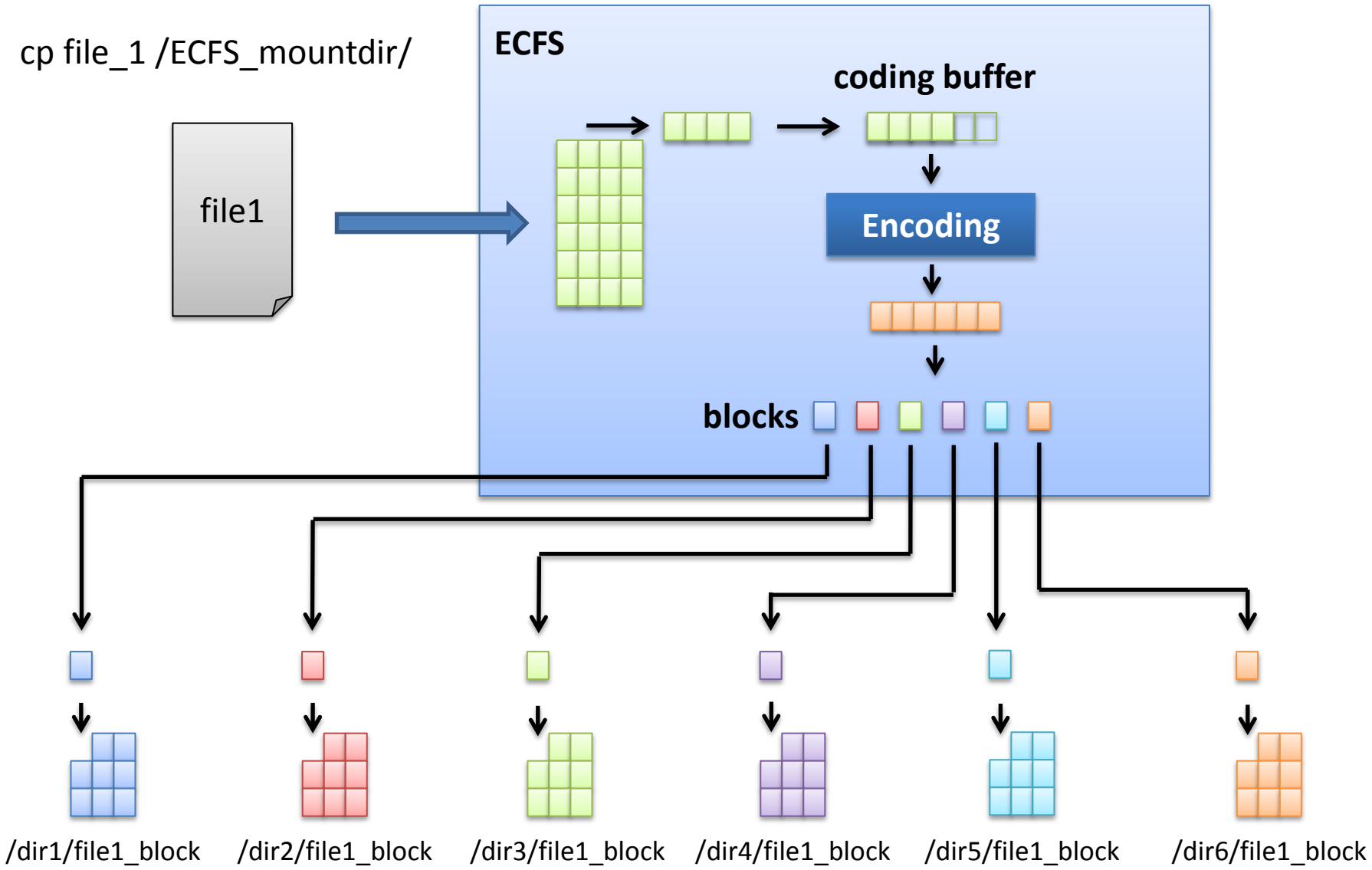
Successful decoding possible if the number of missing symbols is less or equal *n*

Code word with some symbols missing



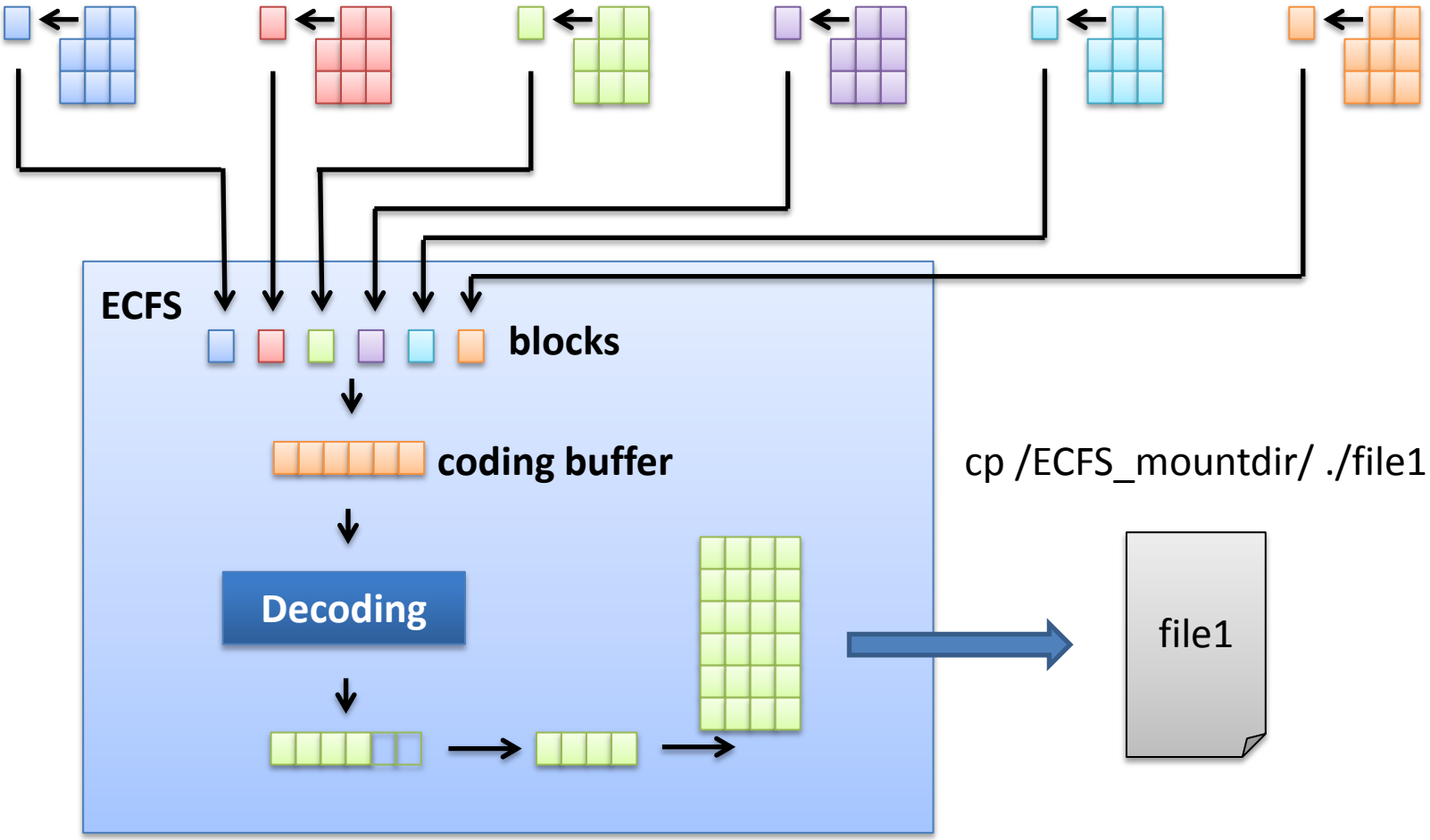
Original message

File encoding

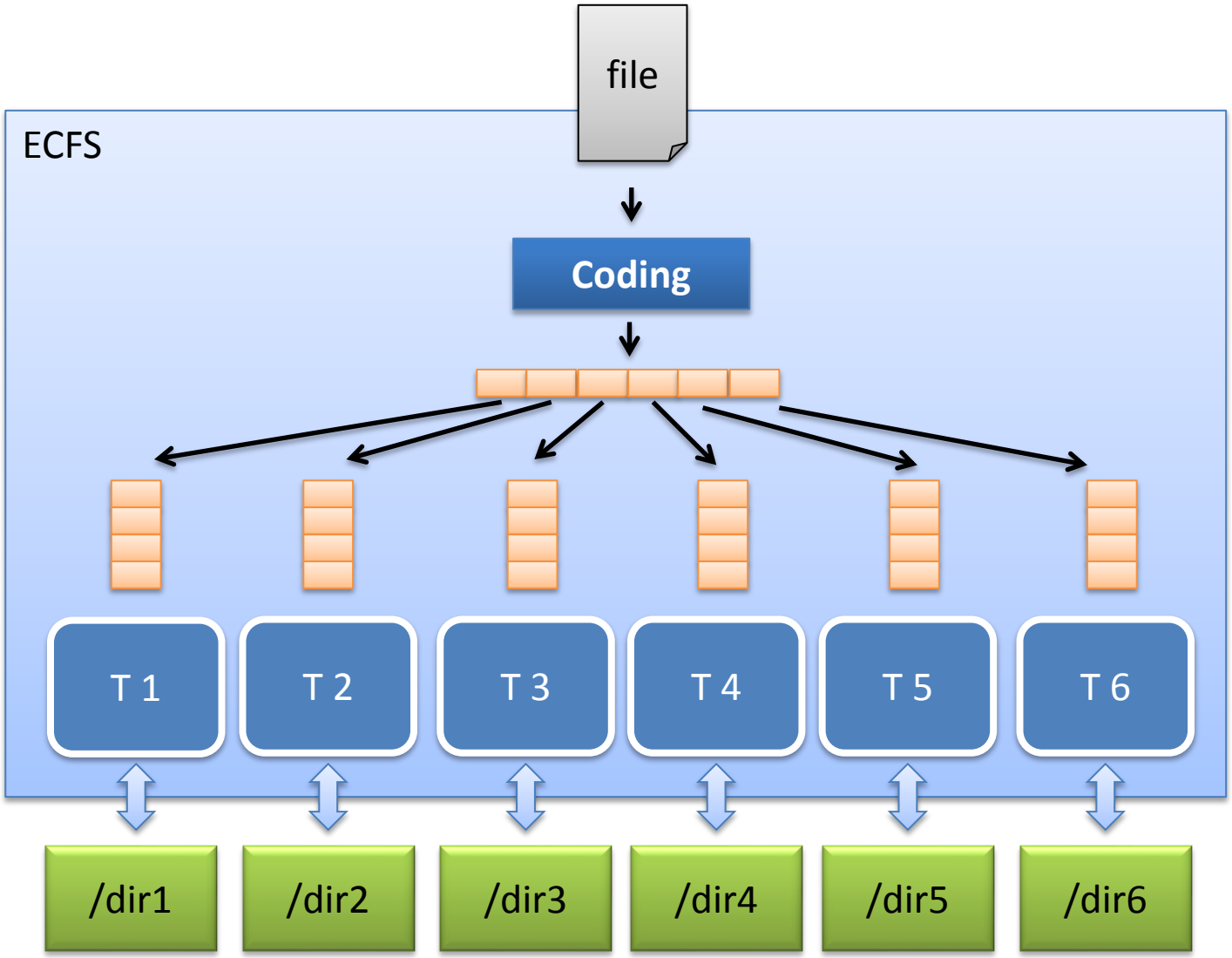


File decoding

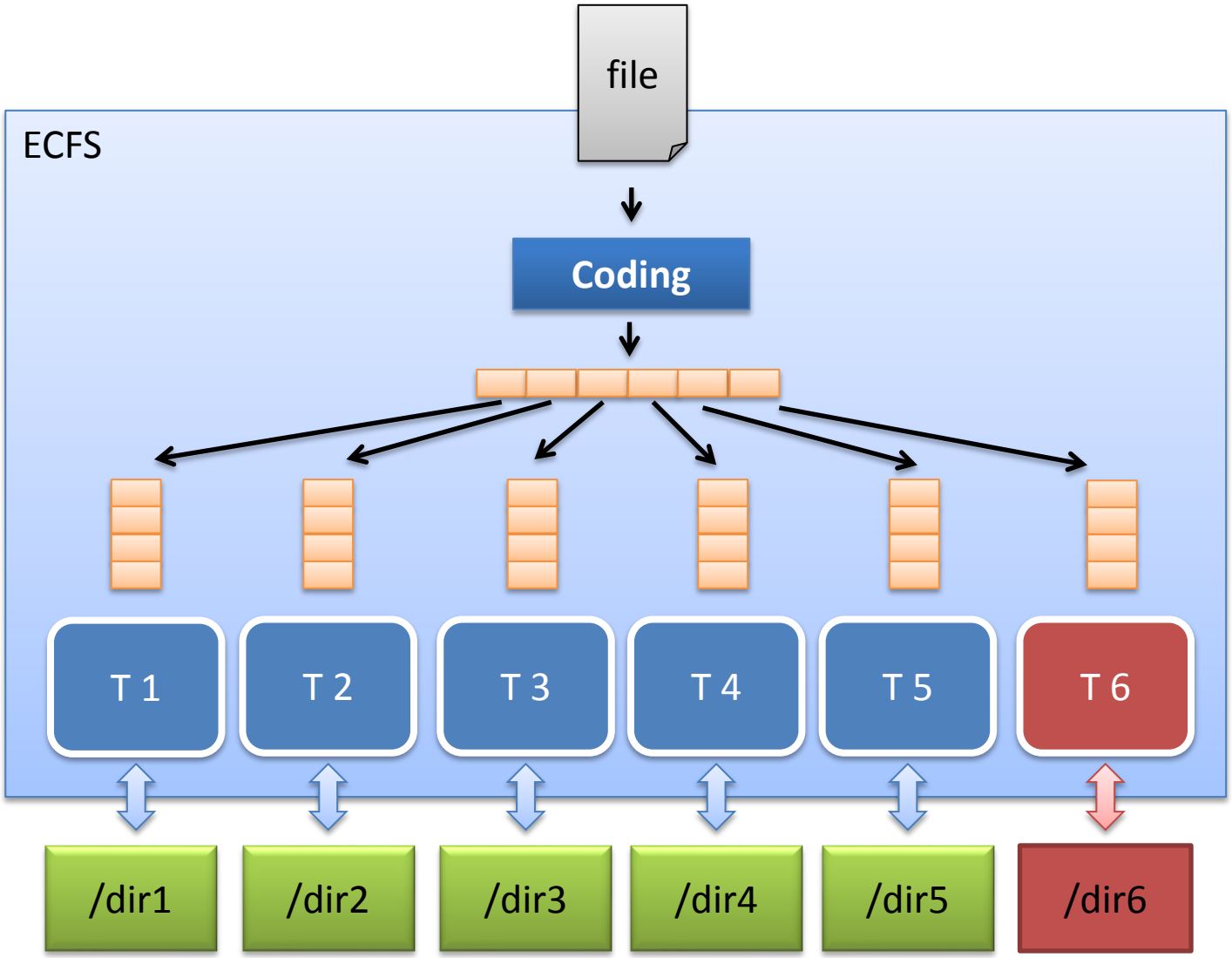
/dir1/file1_block /dir2/file1_block /dir3/file1_block /dir4/file1_block /dir5/file1_block /dir6/file1_block



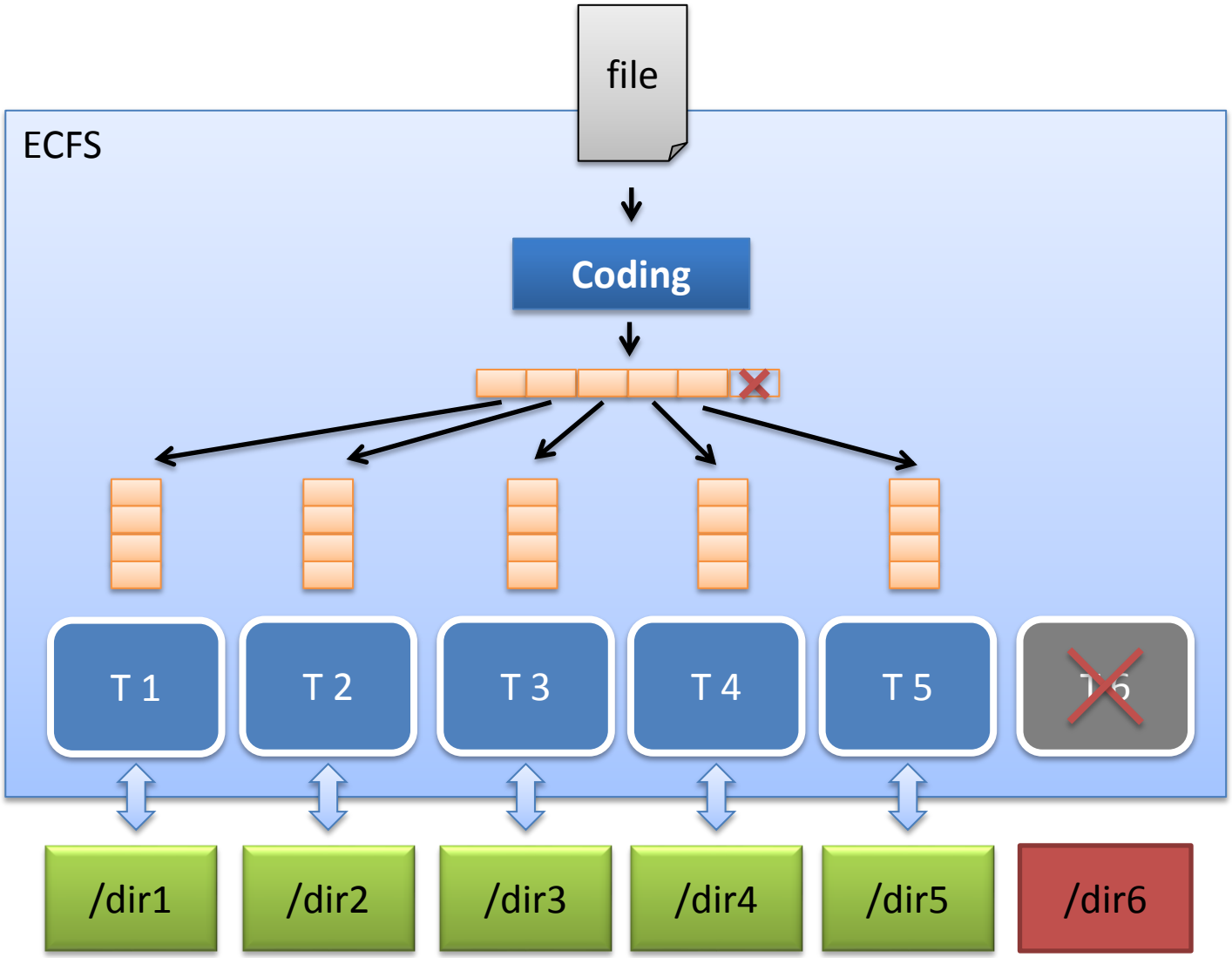
Parallel IO


file
ECFS
Coding
T 1
T 2
T 3
T 4
T 5
T 6
/dir1
/dir2
/dir3
/dir4
/dir5
/dir6

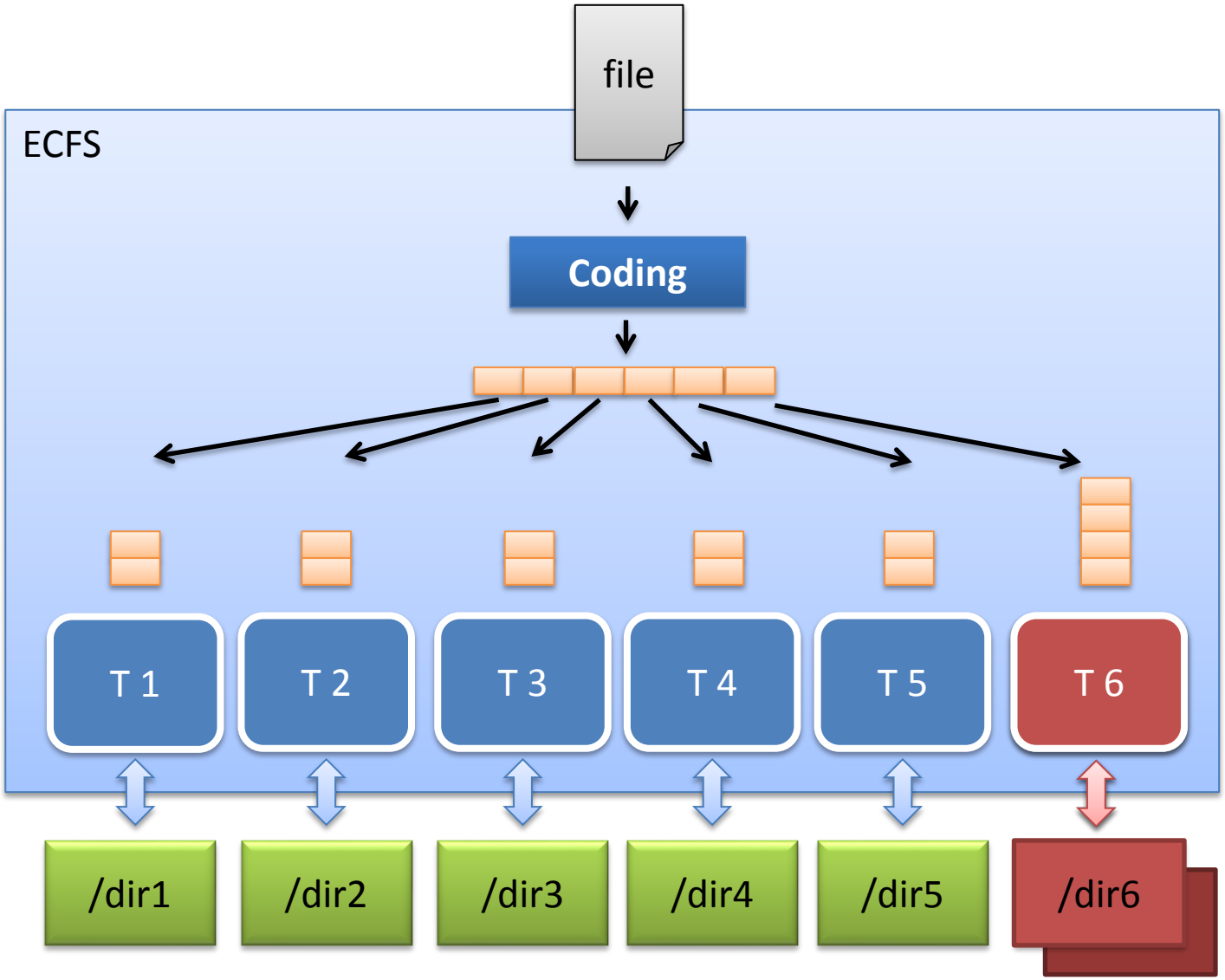
Fault detection



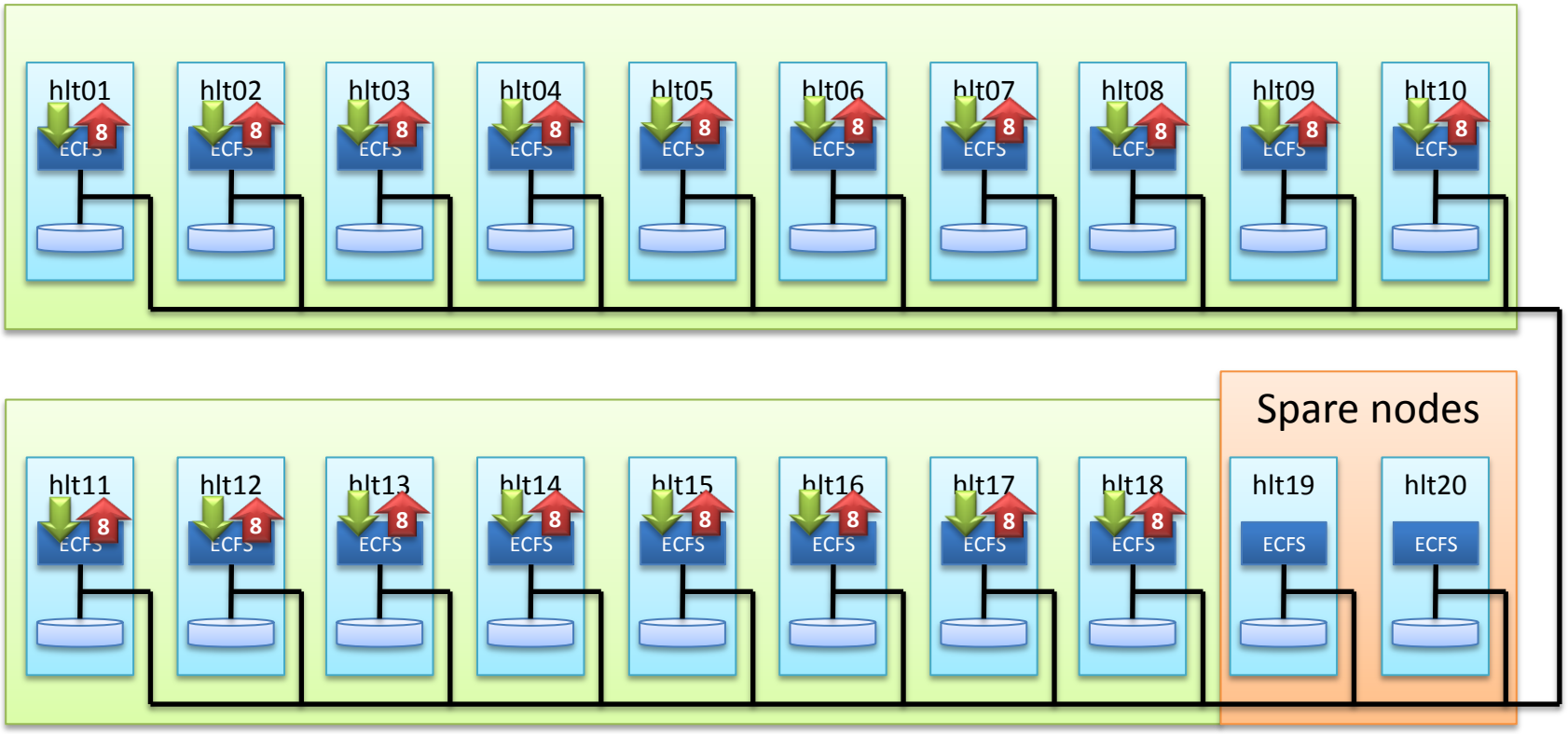
Degraded write



Write to a spare



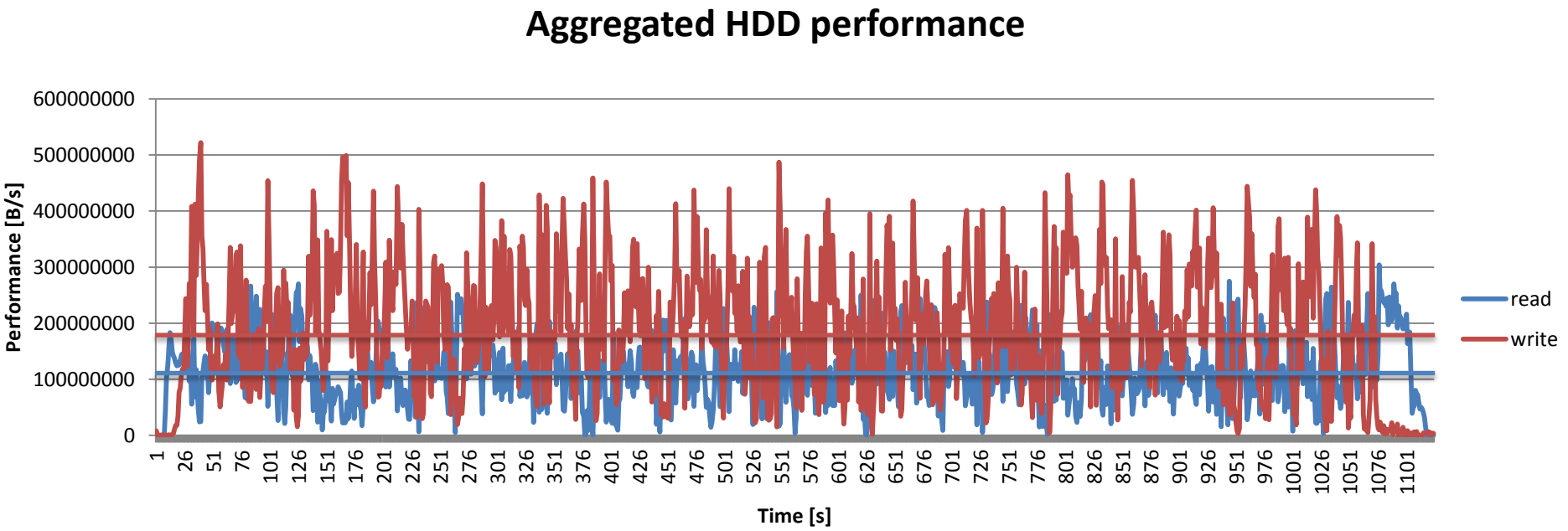
Test environment



Each of the nodes:

- Writes one file
- Reads eight files

Performance



Average write: 188 MB/s ~10 MB/s per node
Average read: 119 MB/s ~7 MB/s per node

Conclusion

- ECFS meets technical requirements of the project
- Performance lower than expected
- Much improvement has to be done