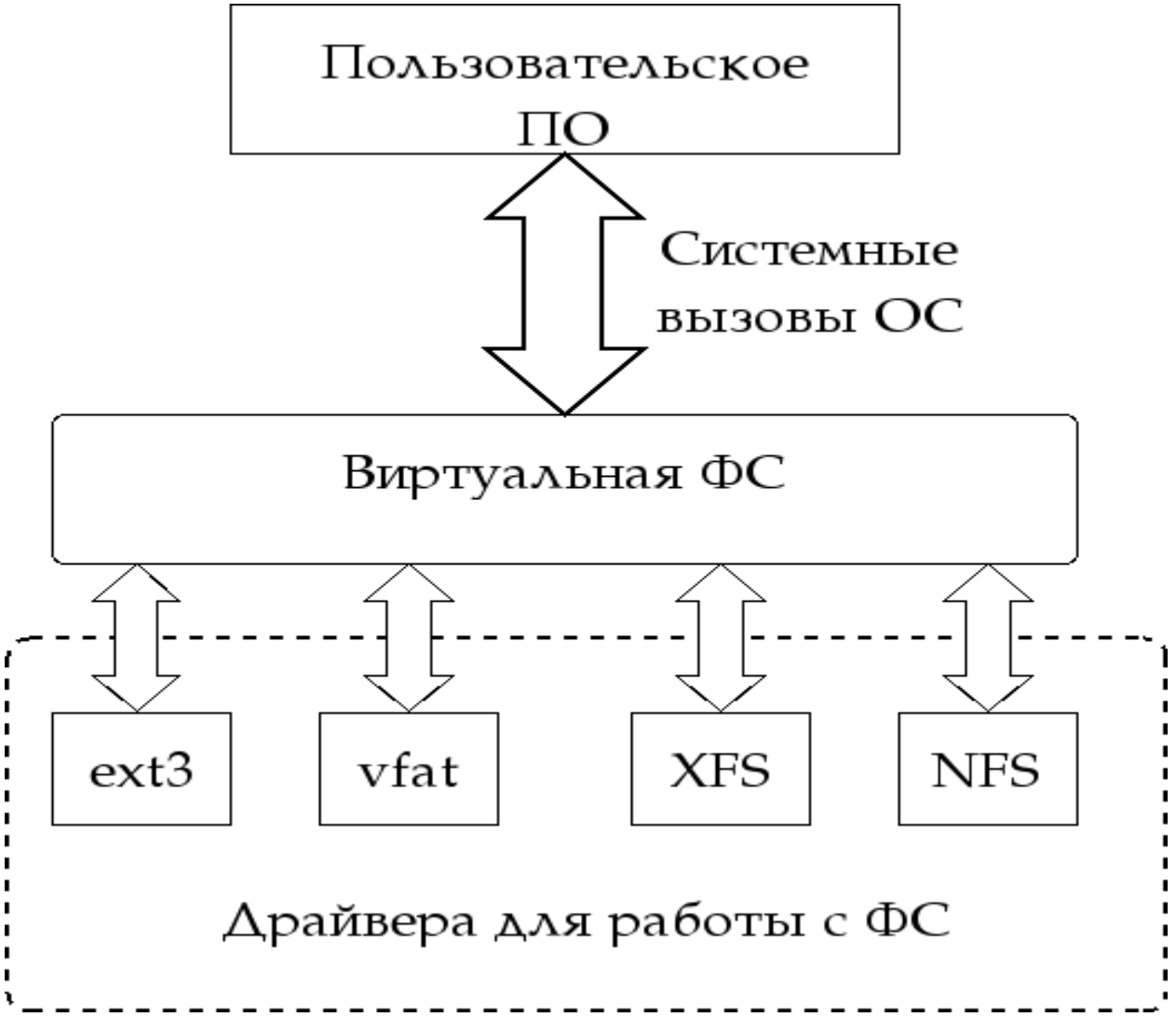
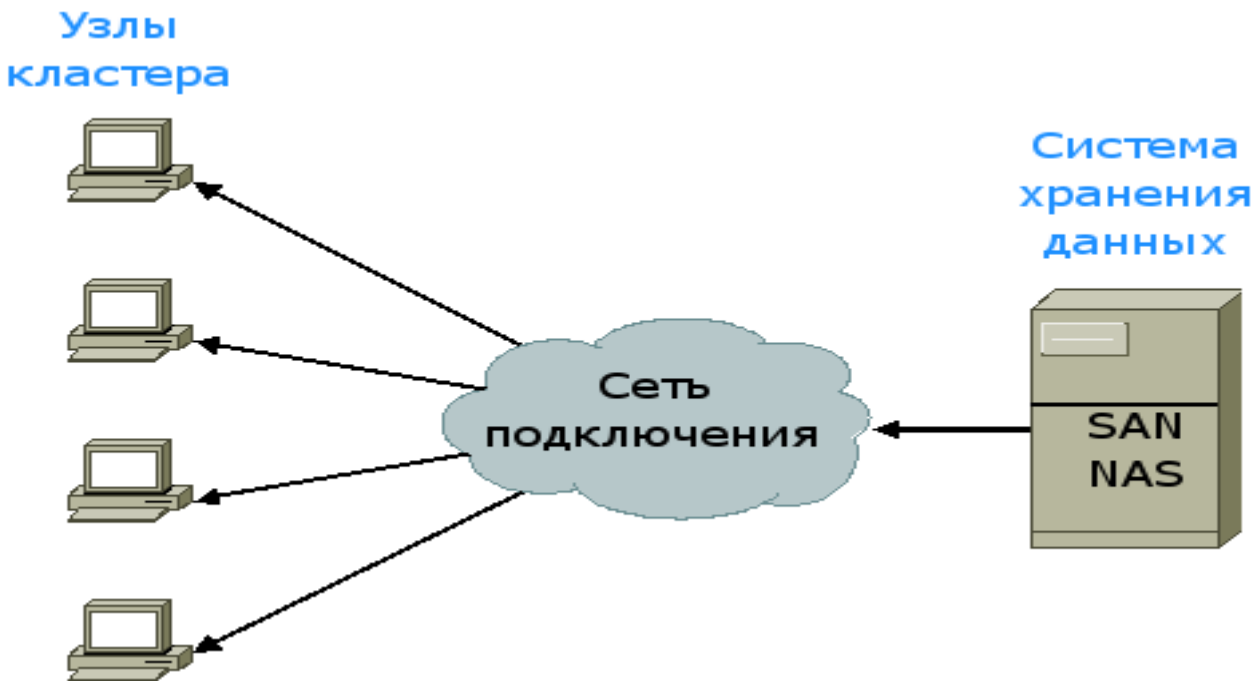
Кластерные файловые системы

Структура файловой подсистемы ОС

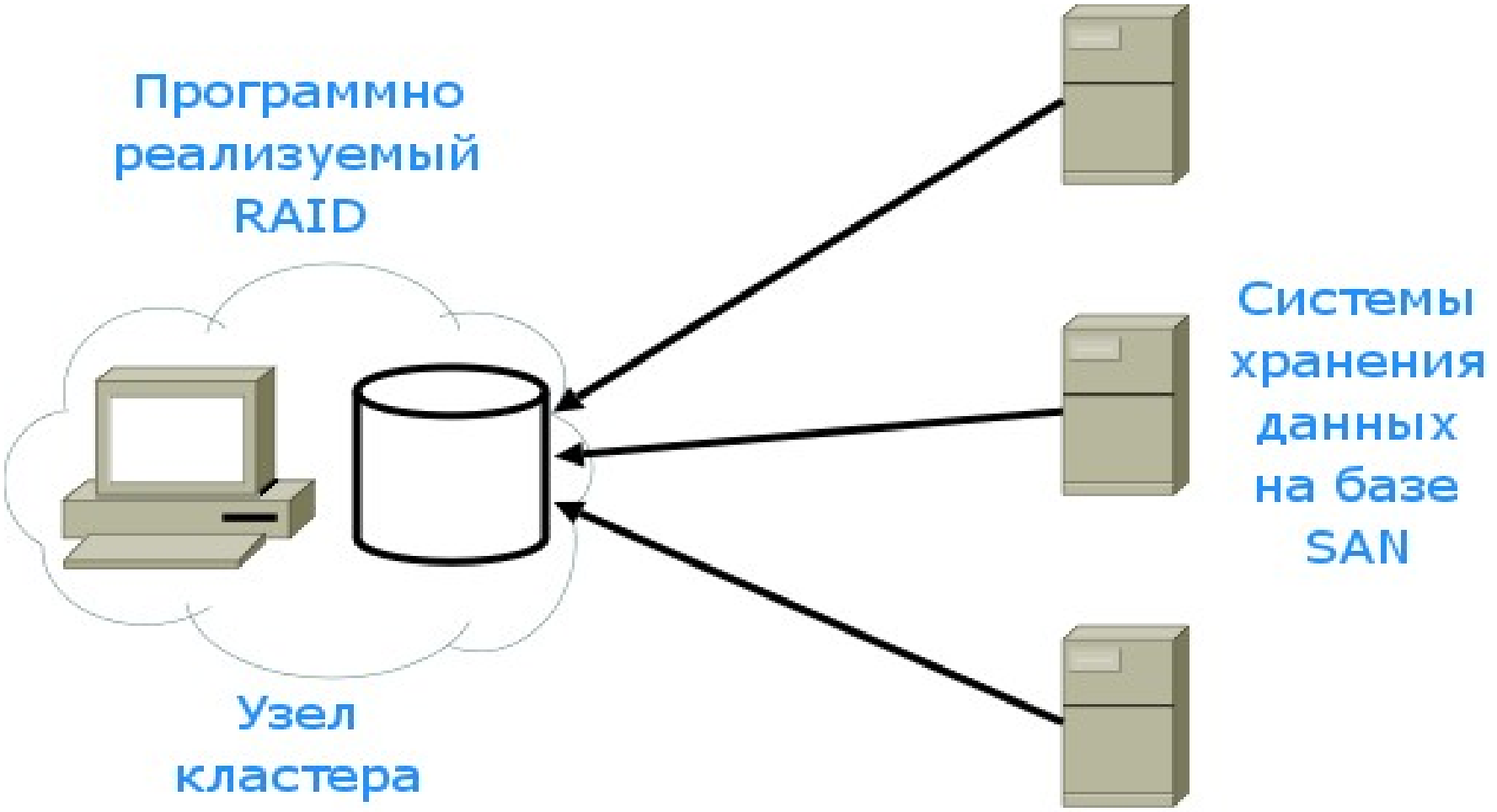


Классические сетевые хранилища данных

- NAS – Network Attached Storage
- SAN - Storage Attached Network



Распределенные хранилища на базе SAN

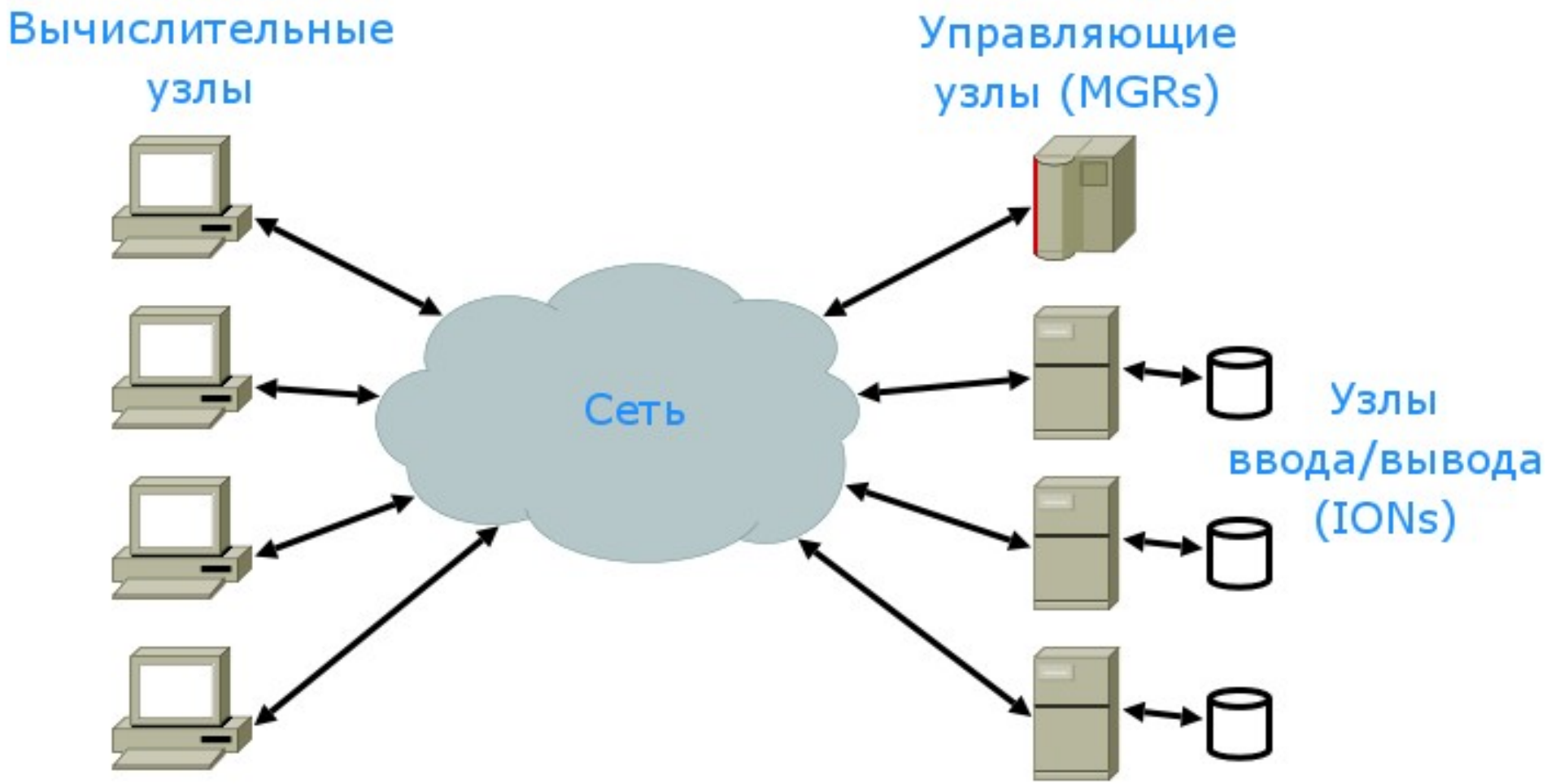


Распределенные файловые системы объединяют дисковые подсистемы узлов в единое пространство

Архитектура всех существующих кластерных файловых систем имеет схожую структуру:

- **CN** — вычислительные узлы, использующие ФС
- **ION** (Input-Output Nodes) — узлы ввода/вывода, к которым подключена дисковая подсистема (DAS, SAN, NAS)
- **MGR** — менеджер или серверы метаданных. Эти узлы отвечают за метаданные, распределение нагрузки по узлам ввода/вывода, и общее управление ФС.

Архитектура кластерных файловых систем



Кластерные файловые системы

GFS (Global File System) — Red Hat
Lustre — Sun (Oracle)
GPFS (General Parallel File System) — IBM
OCFS (Oracle Cluster File System) — Oracle
Cluster Shared Volumes — Microsoft (>server 2008)
PVFS (Parallel Virtual File System) — freeware
GlusterFS
HDFS – Hadoop Distributed File System

Пример: распределенная ФС PVFS

PVFS

PVFS — Parallel Virtual File System

Проект PVFS – это попытка создать высокопроизводительную и расширяемую параллельную файловую систему для кластеров.

Возможности PVFS

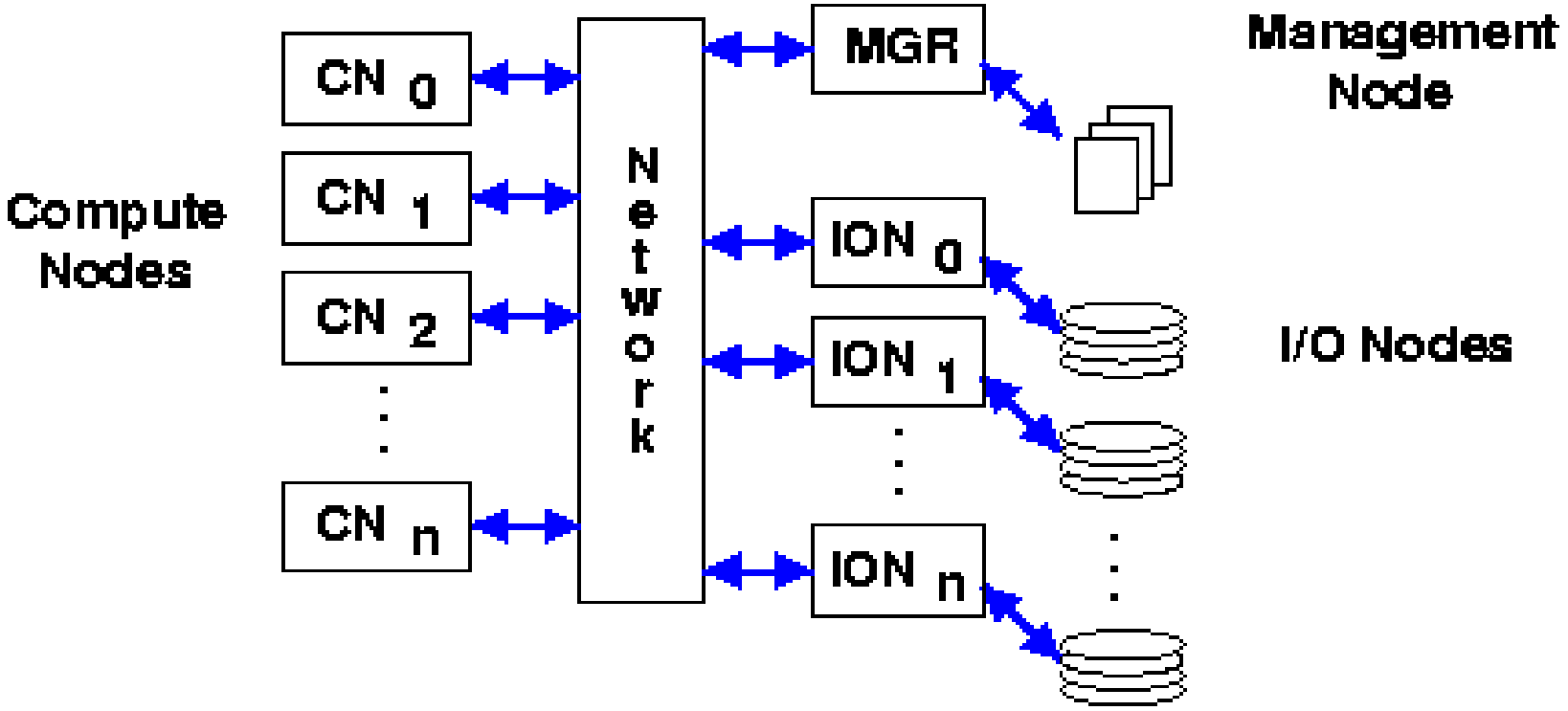
последовательное пространство файловых имен на отдельно взятой машине;

прозрачный доступ для существующих программ;

физическое распределение данных файла между несколькими дисками на множестве узлов кластера;

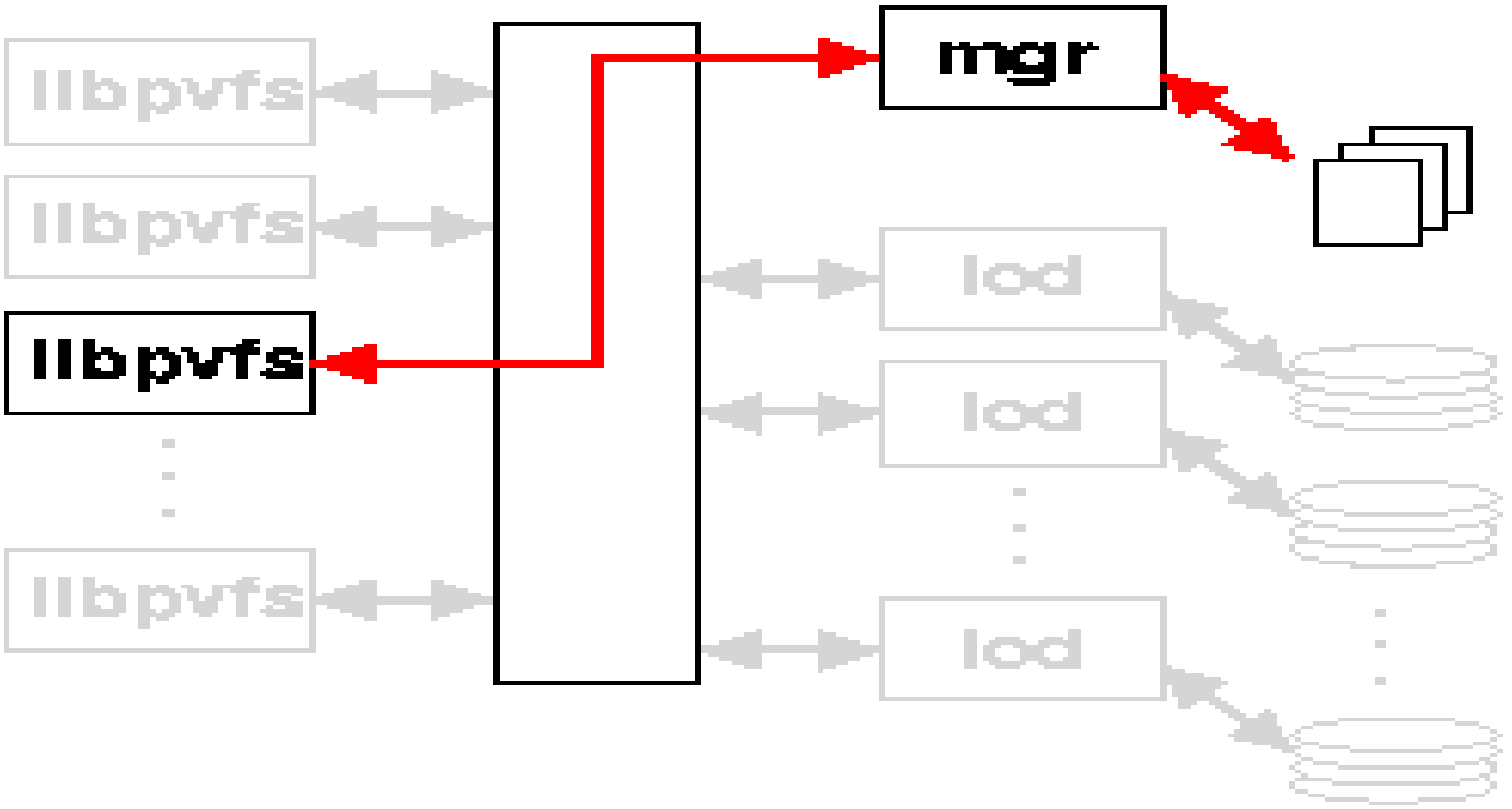
высокопроизводительный доступ для приложений выполняющихся в контексте пользователя.

Структура PVFS

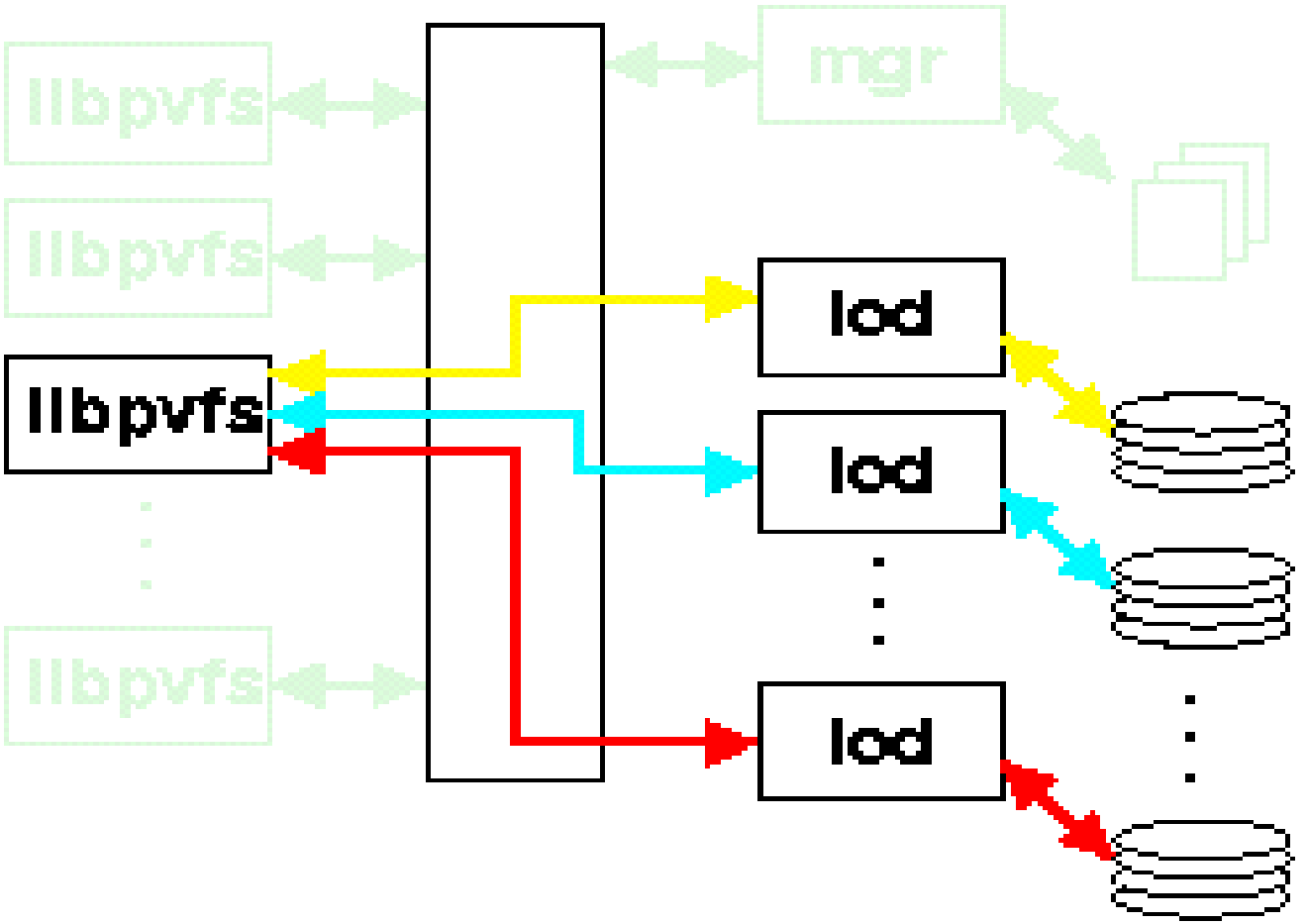


MGR — управляющие узлы
(содержат метаданные операций)
ION — узлы ввода/вывода

Доступ к метаданным



Доступ к данным



Интерфейсы

- оригинальное API PVFS
- интерфейс к ядру Linux
- интерфейс к ROMIO MPI-IO

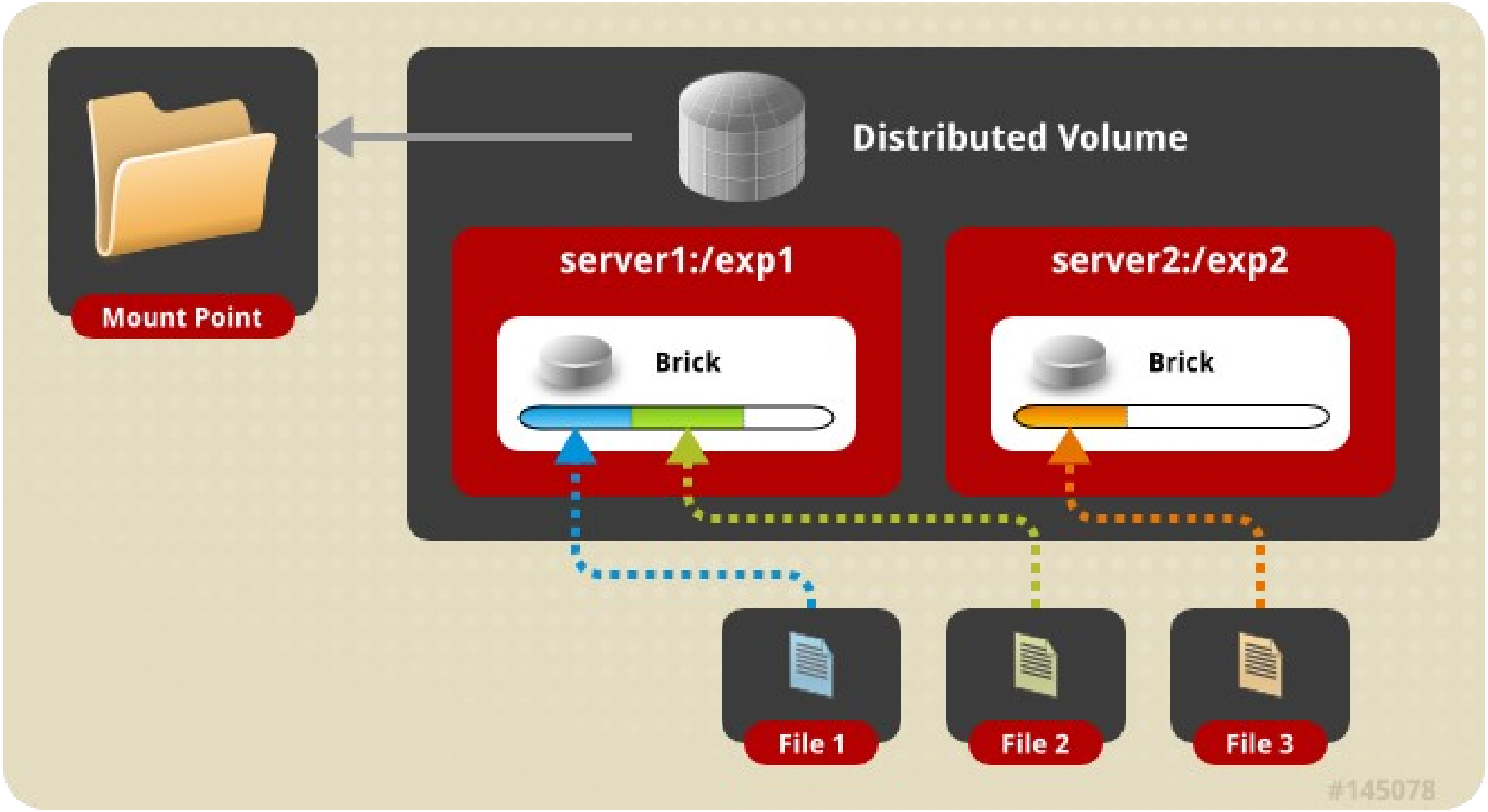
GlusterFS

- В отличие от других распределённых файловых систем, таких как Lustre и Ceph, для работы GlusterFS не требуется отдельный сервер для хранения метаданных.
- С помощью InfiniBand RDMA или TCP/IP GlusterFS может объединить хранилища данных, находящиеся на разных серверах, в одну параллельную сетевую файловую систему.
- GlusterFS работает в пользовательском пространстве при помощи технологии FUSE, поэтому не требует поддержки со стороны ядра операционной системы и работает поверх существующих файловых систем

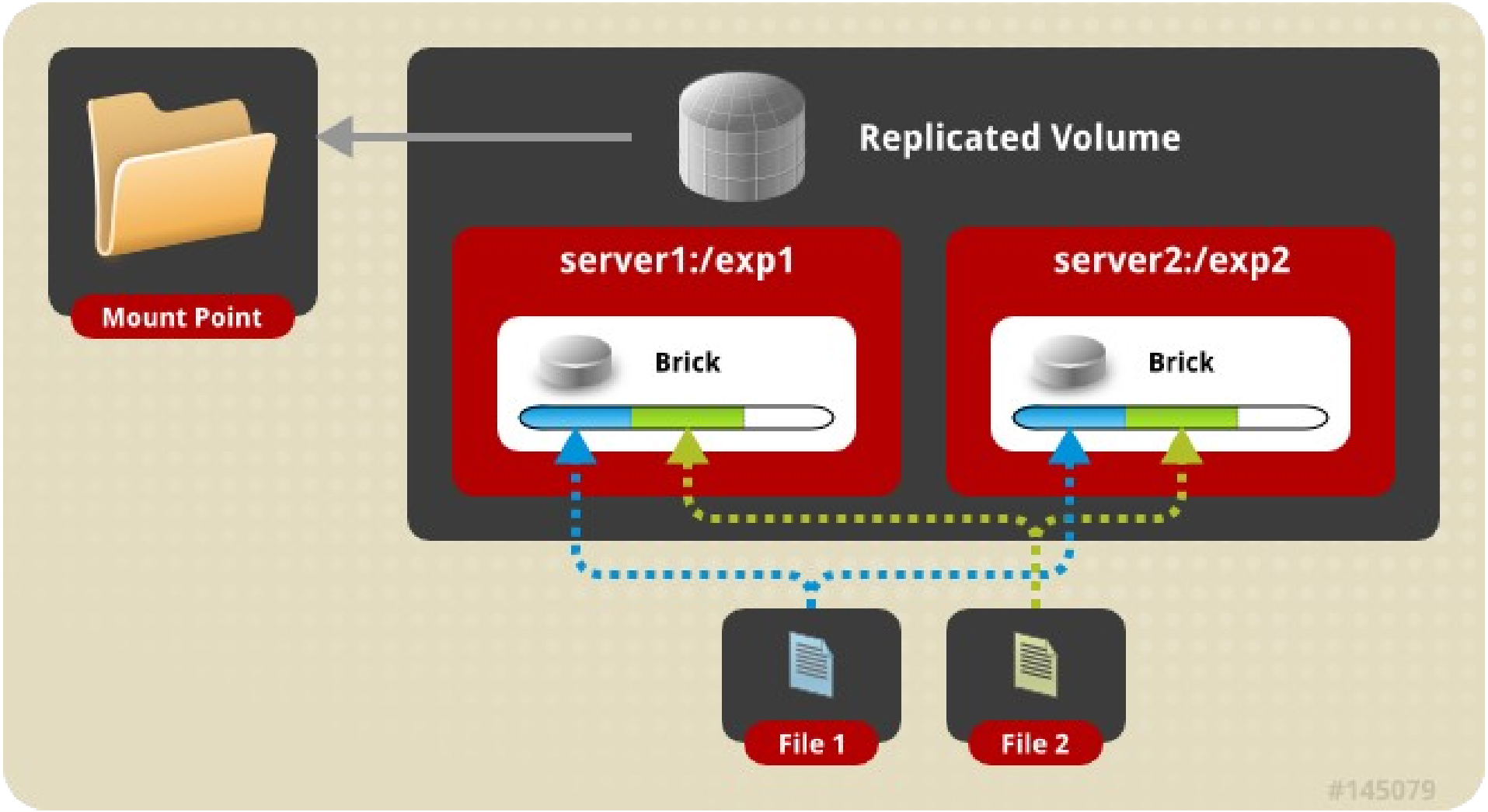
GlusterFS

- Синхронная репликация между серверами (нельзя расширить уже существующий том, добавив сервер для репликации)
- Чередование порций данных между серверами (Striping) Распределение файлов между серверами
- Балансировка нагрузки
- Восстановление после отказа узла (в ручном режиме с помощью опроса файлов (ls -lR или find на смонтированом томе))
- Опережающее чтение (read-ahead) и запаздывающая запись (write-behind) для увеличения быстродействия
- Дисковые квоты

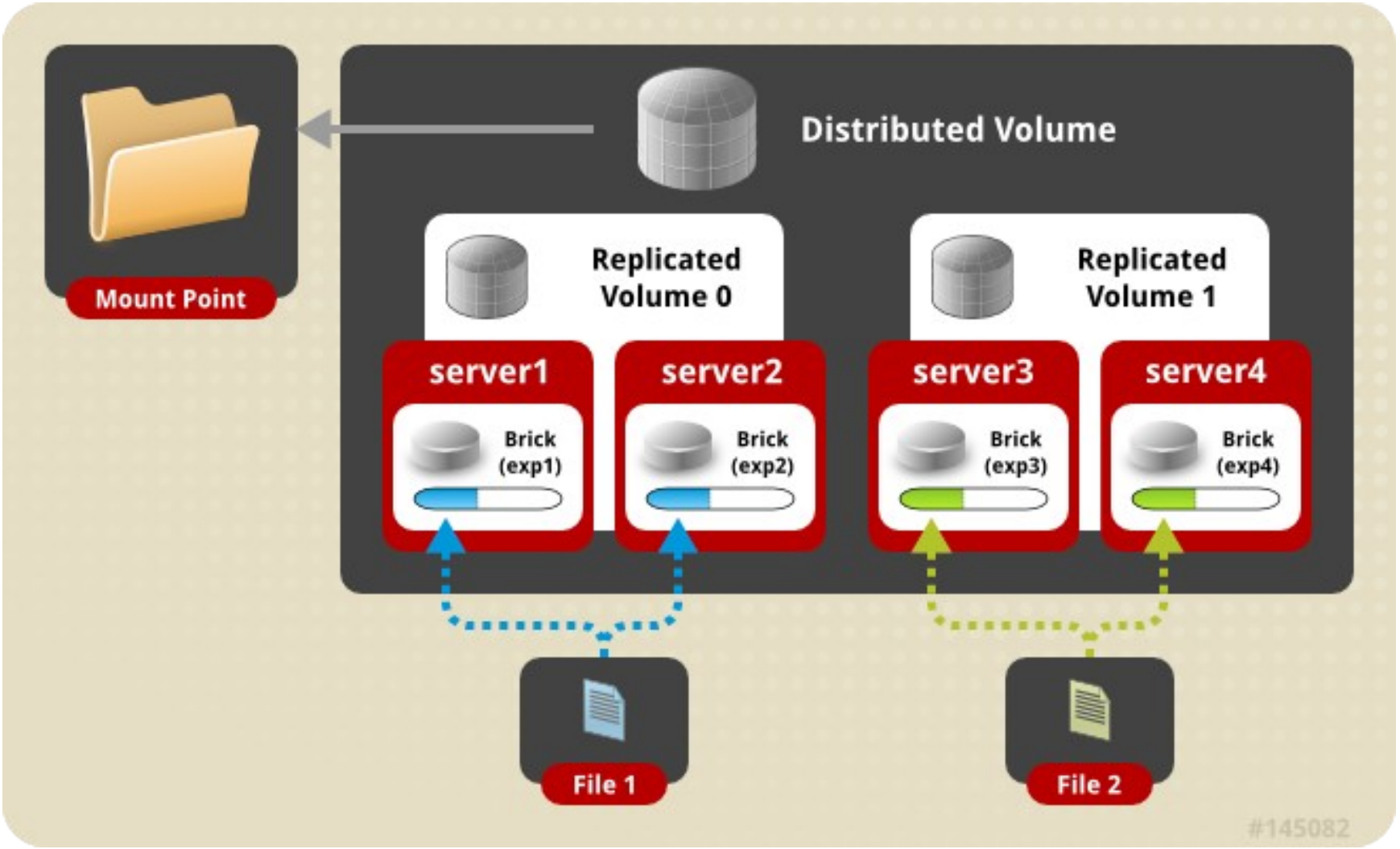
Distributed Glusterfs Volume



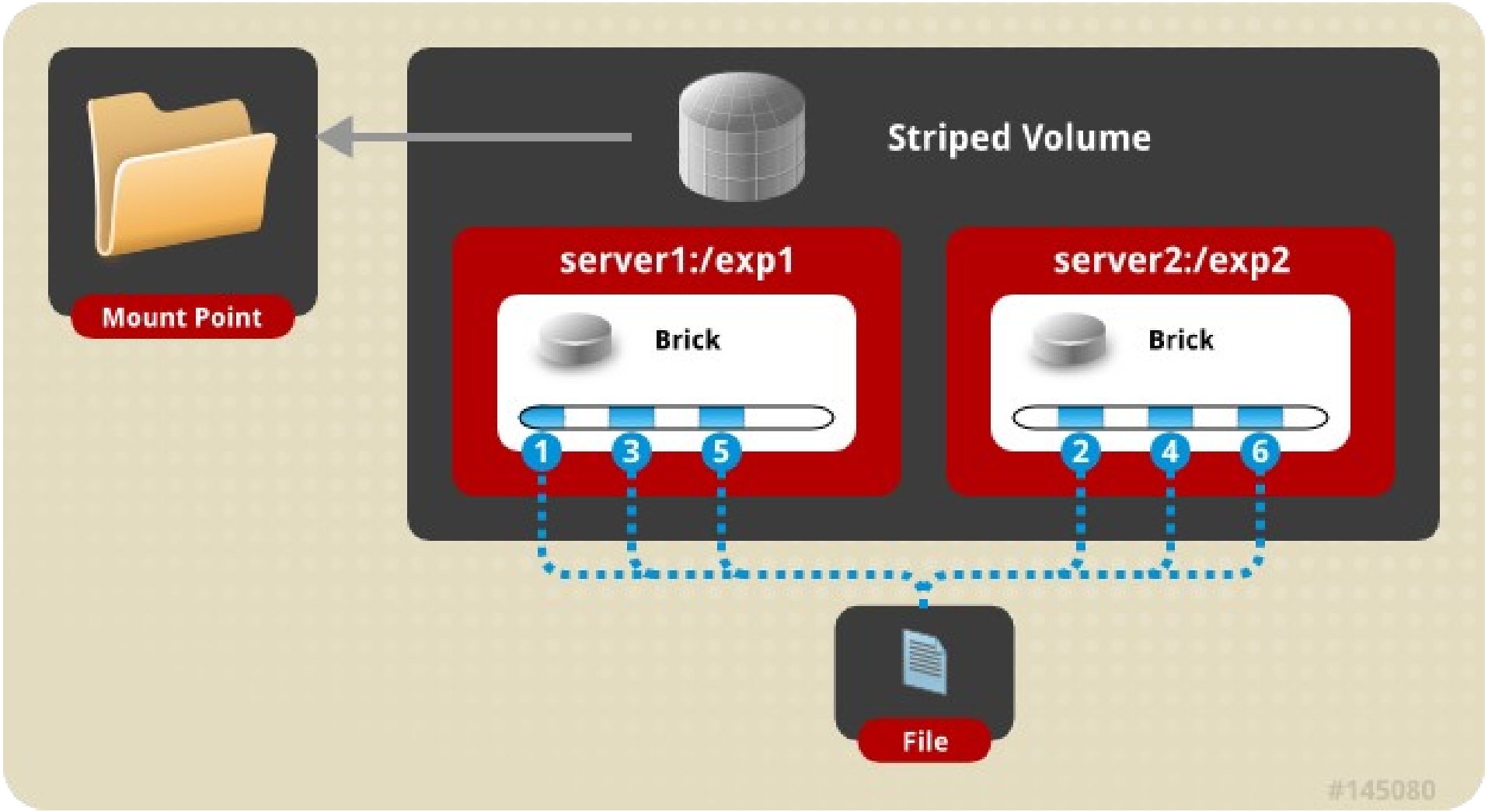
Replicated Glusterfs Volume



Distributed Replicated Glusterfs Volume



Striped Glusterfs Volume



Distributed Striped Glusterfs Volume