XtreemFS – a case for object-based storage in Grid data management

Jan Stender, Zuse Institute Berlin

an object-based file system for federated IT infrastructures.

XTREEMFS

In this talk...

Traditional Grid Data Management

Object-based file systems

XtreemFS

Grid use cases for XtreemFS

The XtreemOS Project

- **XtreemFS** is part of the XtreemOS project
- EU project - 18 partners from all over Europe, incl. NEC, SAP, Telefonica, Mandriva, Red Flag Linux
- Develops a distributed operating system around Kerrighed, a single system image Linux kernel
- The **XtreemFS** Team:
  - Zuse Institute Berlin
  - Barcelona Supercomputing Center
  - NEC High Performance Computing, Stuttgart
  - CNR, Rende, Italy
  - Universität Düsseldorf
  - SAP Research

In this talk...

Traditional Grid Data Management

Object-based file systems

XtreemFS

Grid use cases for XtreemFS

Traditional Grid Data Management

Access Daemon:

- uniform interface to heterogeneous storage resources
- conventional (network) file systems store data
  - geared towards local clusters, single data centers
  - lack of support for reliable organization-spanning WAN access

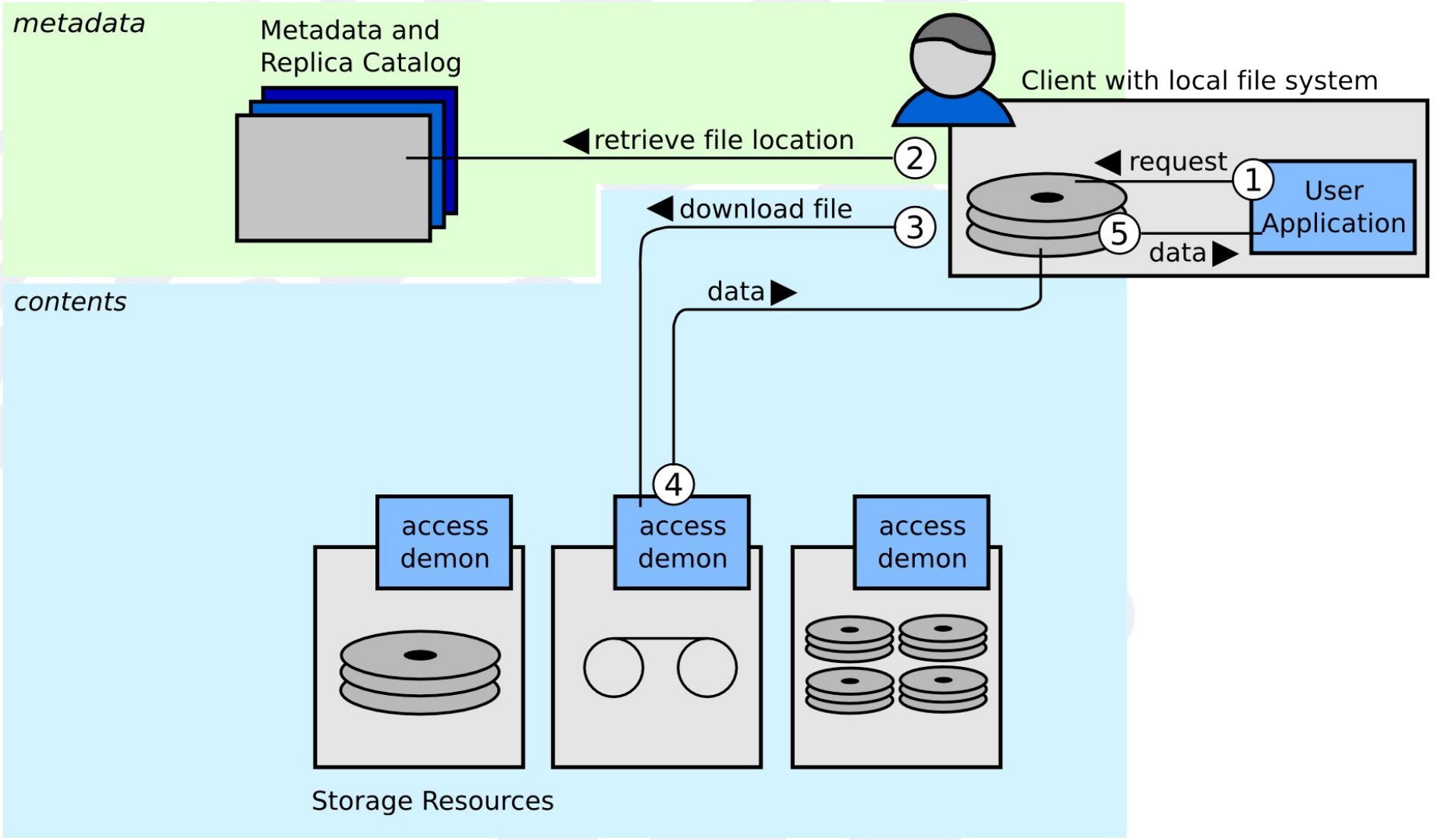
Metadata Catalog:

- hierarchical namespaces (Logical File Names)
- database-like queries

Replica Catalog:

- locations of file replicas (Physical File Names)

Traditional Grid Data Management



Traditional Grid Data Management

simple access to heterogeneous storage resources,

but ...

- in general, whole files have to be transferred and stored locally
  - high latency to first access
  - potential waste of network and storage resources
  - local access might be slower than network access
- no automatic replica consistency
  - usually restriction to write-once usage patterns: download of input files, upload of output files
- no access control on downloaded copies

In this talk...

Traditional Grid Data Management

Object-based file systems

XtreemFS

Grid use cases for XtreemFS

Object-based File Systems

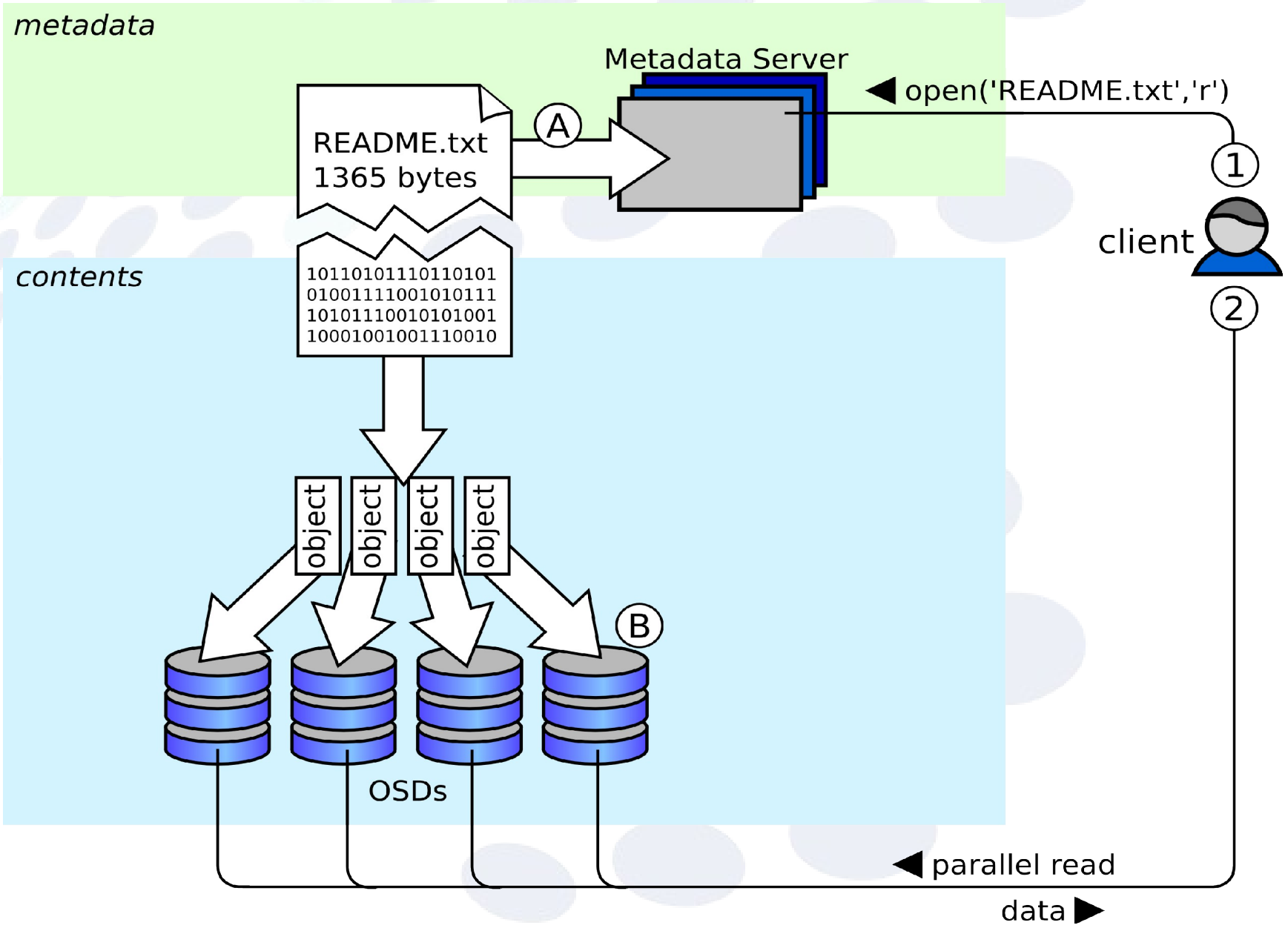
Block-based file systems:

- unit of distribution are disk blocks
- metadata and block management at central server
- file system addresses blocks over the network

Object-based file systems:

- storage devices can be more intelligent today
- split file in parts (objects) and distribute & address them
- only metadata at server, block management by storage devices

Object-Based File Systems



Object-based File Systems

architecture looks similar to Grid data management, but ...

- file content is accessed on OSDs
  - OSDs can exercise full control over any kind of access
- single files can be accessed in parallel
  - use of aggregate bandwidth to all storage devices

Object-based File Systems

several available...

- Lustre (Open-Source)
- Panasas ActiveStore (commercial)
- Ceph (Research, Open-Source)

common properties:

- parallel designs for high-performance LAN access
- centralized, one-datacenter, one-organization
- control over failures of hardware

In this talk...

Traditional Grid Data Management

Object-based file systems

XtreemFS

Grid use cases for XtreemFS

XtreemFS is an object-based file system designed for Grid environments

features:

- POSIX-compliant file system API ✓
- replication ✓ and partitioning of metadata
- extended metadata ✓ and queries
- parallel file access (striping) ✓
- replication of files ✓
- automatic, access pattern based replica creation
- client-side caching

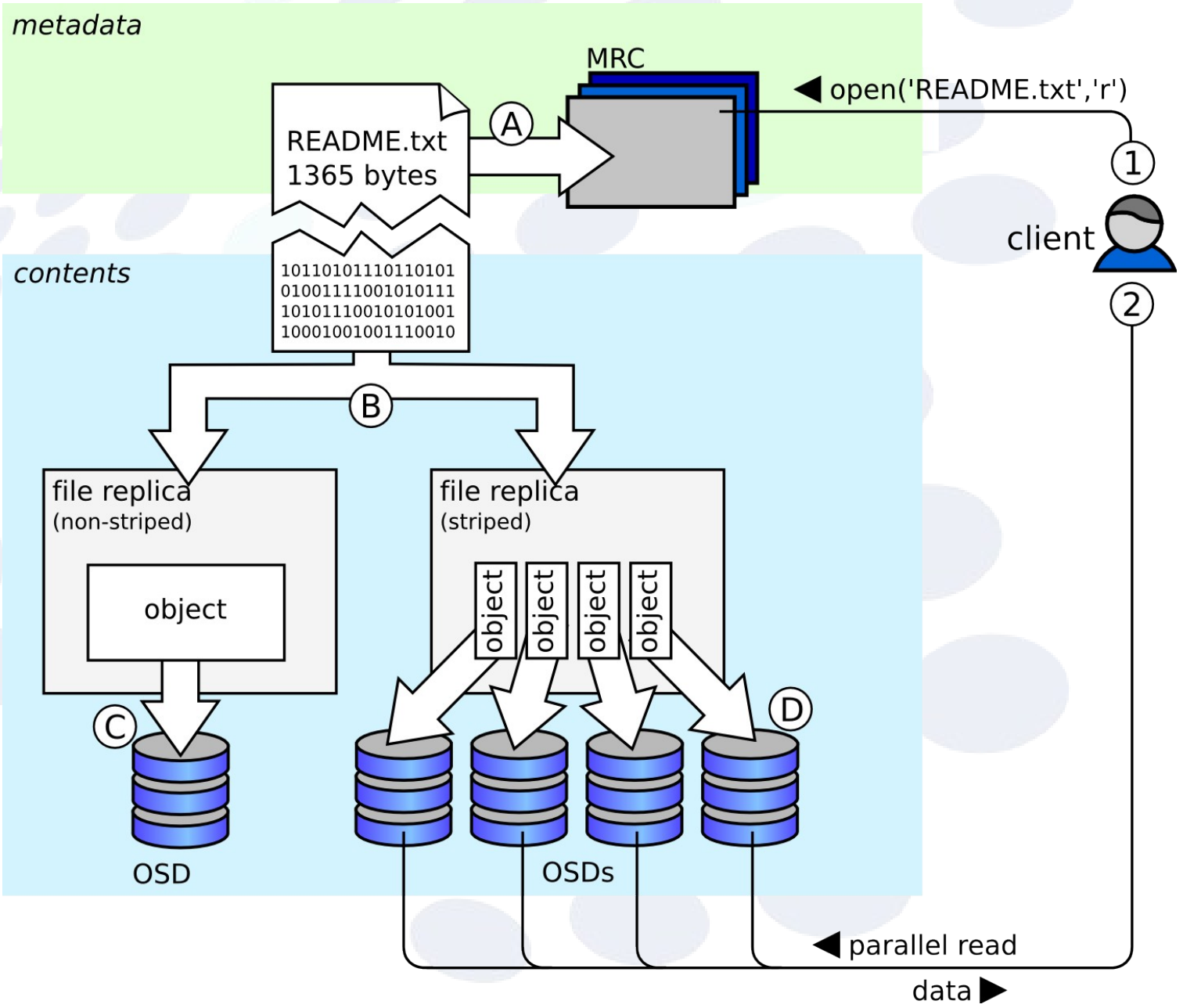
replication of files

- fully transparent to client
- guarantees POSIX consistency of data (ACID-like)
- can deal with failures

consistency coordination

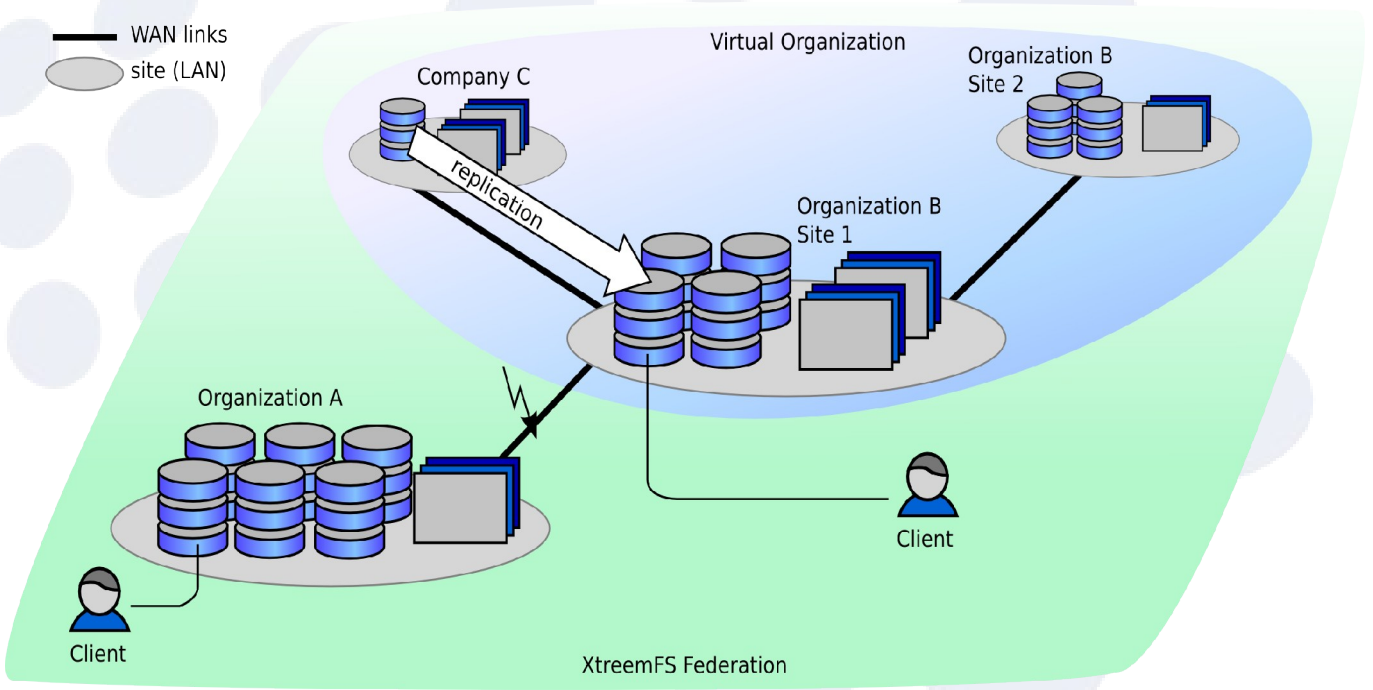
- currently at object level
- synchronous, asynchronous or on-demand

XtreemFS


metadata
MRC
README.txt
1365 bytes
A
open('README.txt','r')
1
client
2
contents
10110101110110101
01001111001010111
10101110010101001
10001001001110010
B
file replica
(non-striped)
object
C
OSD
file replica
(striped)
object
object
object
object
D
OSDs
parallel read
data

Replication of data and metadata at multiple sites:

- a site can continue working when network is down
- others can continue working if a site fails / leaves



In this talk...

Traditional Grid Data Management

Object-based file systems

XtreemFS

Grid use cases for XtreemFS

Grid use cases for XtreemFS

On-demand and asynchronous replication can significantly speed up Grid data processing jobs,

e.g. if ...

- some process stages or generates a huge file
- the file needs to be accessed by many clients
- each client only accesses a small portion of the file
- clients reside on different locations

Grid use cases for XtreemFS

XtreemFS creates an initially empty local replica for all remote clients

- clients can immediately work on their local replicas
- replicas are either updated in background, or when data is needed
- only such data is transferred which is actually needed

Summary

- Traditional Grid data management systems have inherent shortcomings
  - in terms of performance
  - in terms of resource usage
- Object-based storage can deal with these shortcomings
- XtreemFS is an object-based file system for wide area networks
  - it offers a POSIX-compliant interface
  - it provides sophisticated replication mechanisms

Thanks for your attention!