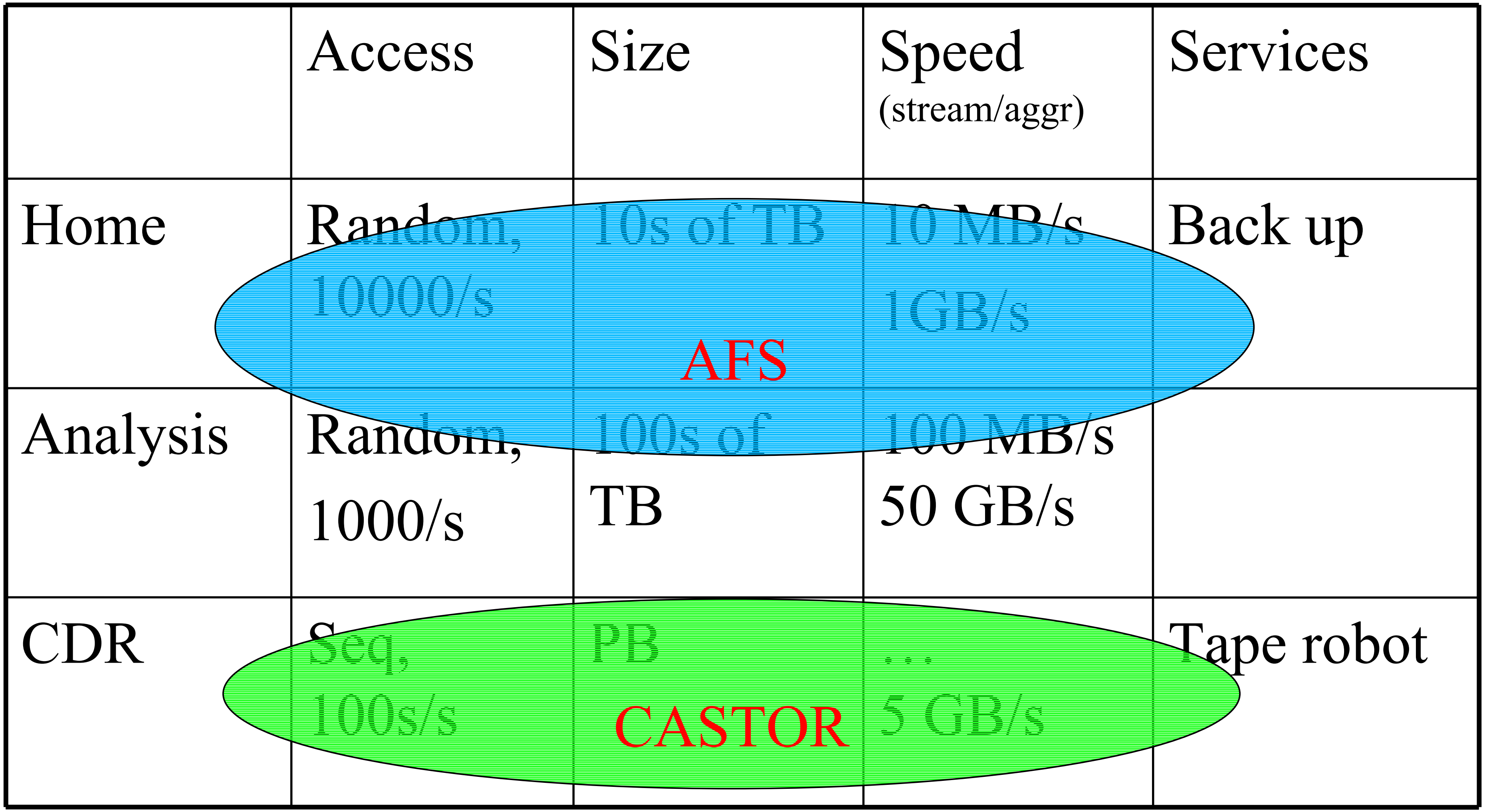
Storage and Storage Access

Rainer Többicke
CERN/IT

Introduction

- Data access
  - Raw data, analysis data, software repositories, calibration data
  - Small files, large files
  - Frequent access
  - Sequential access, random access
- Large variety

Usage



	Access	Size	Speed (stream/aggr)	Services
Home	Random, 10000/s	10s of TB	10 MB/s 1GB/s	Back up
Analysis	Random, 1000/s	100s of TB	100 MB/s 50 GB/s	
CDR	Seq, 100s/s	PB	… 5 GB/s	Tape robot

AFS

CASTOR

Plan A – CASTOR & AFS

- AFS for software distribution & home
- CASTOR for Central Data Recording, Processing
- Mass Storage
  - Disk layer, Tape layer
- Analysis
  - Combination of the AFS & CASTOR
  - Performance enhancements for AFS

AFS – Andrew file system

- In operation @ CERN since 1992
- ~7TB, 14000 users, ~20 servers
- Secure, wide area
- Good for high access rate, small files
- Open Source, community-supported
  - – Developed at CMU under IBM grant
  - – enhancements in Stability and Performance
- Service run by 2.2 staff at CERN

CASTOR

- HSM System being developed at CERN
- Tape server layer
  - Robotics (e.g. STK), tape devices (e.g. 9940)
  - 2 PB data, 13 million files
- Disk buffer layer
  - ~250 disk servers (~200 TB)
- Policy-based tape & disk management
- Development 4 staff, operation 5 staff

CASTOR & AFS Plans

- CASTOR 'Stager' rewrite
  - Design for performance and manageability
    - Demonstrated new concept October 2003
  - Security
  - Demonstrated pluggable scheduler
- AFS development
  - Performance enhancements
  - “Object” disk support

Plan B – Cluster File Systems

- Replacement for AFS
- Replacement for CASTOR disk server layer
- Replacement for CASTOR
- Basis for front-end to Storage-Area-Network-based storage

Shared File System Issues

- Optimization for a variety of access patterns
  - random/stream, tx rate, file sizes, data reuse
- Interface / Semantics / Platforms
- Security & trust model
- Scaling
- Operation
  - Policy-based management
  - Monitoring
  - Resilience
  - Reconfiguration

Storage
(Hardware aspects)

- File server with locally attached disks
  - PC-based [IDE] disk server
  - **N**etwork **A**ttached **S**torage appliance
- Fibre channel fabric
  - Confined to a **S**torage **A**rea **N**etwork
  - 'exporters' for off-SAN access
  - Robustness, manageability
- iSCSI – SCSI protocol encapsulated in IP

Storage Model
(Software aspects)

- NAS - data & “control” on same path
  - “control”: topology, access control, space mgmt
- SAN - data & “control” on separate paths
  - Performance
- Object Storage
  - Disks contain “objects”, not just blocks
  - Thin control layer, space mgmt
  - Thin authorization layer => Security!

Selection

- Dozens of experimental file systems
- Evaluations and tests
    - Data challenges at CERN
    - Hardware, Software technology, Benchmarks
    - Industry: Openlab collaboration with IBM
        - Storage Tank
    - Institutes: collaboration with CASPUR (Rome)
        - ADIC Storenext, DataDirect
- Search for “industrial strength” solutions

Candidates - I

- IBM SANFS (Storage Tank)
  - SAN based FCP & iSCSI support
  - Clustered Metadata servers
  - Policy-based lifecycle data management
  - Heterogeneous, native FS semantics
  - Under development
    - CERN 1st installation outside IBM, limited functionality in Rel.1, cluster-security model
  - Scaling?

Candidates - II

- Lustre
  - Object Storage
    - Implemented on Linux servers
    - “Portals” interface to IP, Infiniband, Myrinet, RDMA
  - Metadata cluster
  - Open Source, backing by HP

Candidates - III

- NFS
  - NAS model, Unix standard
  - Use case: access to exporter farm
- SAN file systems – basis for exporter farm
  - Storenext (ADIC)
  - GFS (Sistina)
  - DAFS – SNIA model
- Panassas – object storage based

Summary

- Ongoing development in improving of existing solution (CASTOR & AFS)
  - Limited AFS development
- Evaluation of new products has started
  - Expect conclusions by mid-2004