



dCache, managed storage

*Patrick Fuhrmann
for the dCache people*





Topics for today

The dCache team

The dCache spec's

dCache support and collaborations

Whats new in 1.6.6



On the horizon

d-Grid spin-off





Responsibility, dCache

Rob Kennedy

Patrick Fuhrmann

CORE

Jon Bakken

Mathias de Riese

Micheal Ernst

Alex Kulyavtsev

Birgit Lewendel

Dmitri Litvintsev

Tigran Mrktchyan

Martin Radecke

Ron Rechenmacher

Vladimir Podstavkov

Responsibility, SRM

Timur Perelmutov

External Support and Development

Nicolo Fioretti, BARI

Maarten Lithmaath, CERN

Zhenping (Jane) Liu, BNL

Jiri Mencak, RAL

Scott O'Hare, BNL

Stefan Piger, Hannover






Abhishek Singh Rana, SDSC

Owen Synge, RAL









Basic Specification

-  Single 'rooted' file system name space tree
-  File system names space view available through an nfs2/3 interface
-  Data is distributed among a huge amount of disk servers.
-  Supports multiple internal and external copies of a single file
-  Supports 'posix like' (authenticated) access as well as various FTP dialects and the Storage Resource Manager Protocol.








Scalability

-  Distributed Movers AND Access Points (Doors)
-  Automatic load balancing using cost metric and inter pool transfers.
-  Pool 2 Pool transfers on pool hot spot detection
-  Handles bunch requests by fast pool selection unit








Configuration

-  Fine grained configuration of *pool attraction scheme*.
(*write pools, subnet, directory tree, storage info*)
-  Pool to pool transfers on configuration of *forbidden transfers*
-  Fine grained tuning : Space vs. Mover cost preference




Tertiary Storage Manager connectivity

-  Automatic HSM migration and restore
-  HSM dCache interface by script (shell, perl ...)
-  Convenient HSM connectivity for
enstore, osm, tsm, preliminary for Hpss by BNL.








Administration

-  Using standard 'ssh' protocol for administration interface.
-  First version of graphical interface available for administration
-  Large (optional) command set per module
-> highly customizable.








Miscellaneous

-  CRC checksum calculation and comparison (partially implemented yet)
-  Pluggable door / mover pairs (gssdCap,gsiFtp,http ...)
(can be extended due to clear interface definition)
-  Data removed only if space is needed














Resilient dCache

-  Controls number of copies for each dCache dataset
-  Makes sure $n < \text{copies} < m$
-  Adjusts replica count on pool failures
-  Adjusts replica count on scheduled pool maintenance
-  Embedded in farm nodes (makes use of local farm disk space)
-  In production at various places
-  Is part of dCache 1.6.6









Local access library (dCap)

-  implements I/O and name space operations including 'readdir'
-  works on mounted *pnfs* and URL like syntax
-  available as standard shared object and preload library
 - `ls -l dcap://dcachedoor.desy.de/user/patrick`
-  positive tested for Linux, Solaris, Irix (partially for XP)
-  automatic reconnect on server door and pool failures
-  supports read ahead buffering and deferred write
-  supports ssl, kerberos and gsi security mechanisms
-  Thread safe
-  Interfaced by *ROOT* ®





LCG Storage Element Functionality

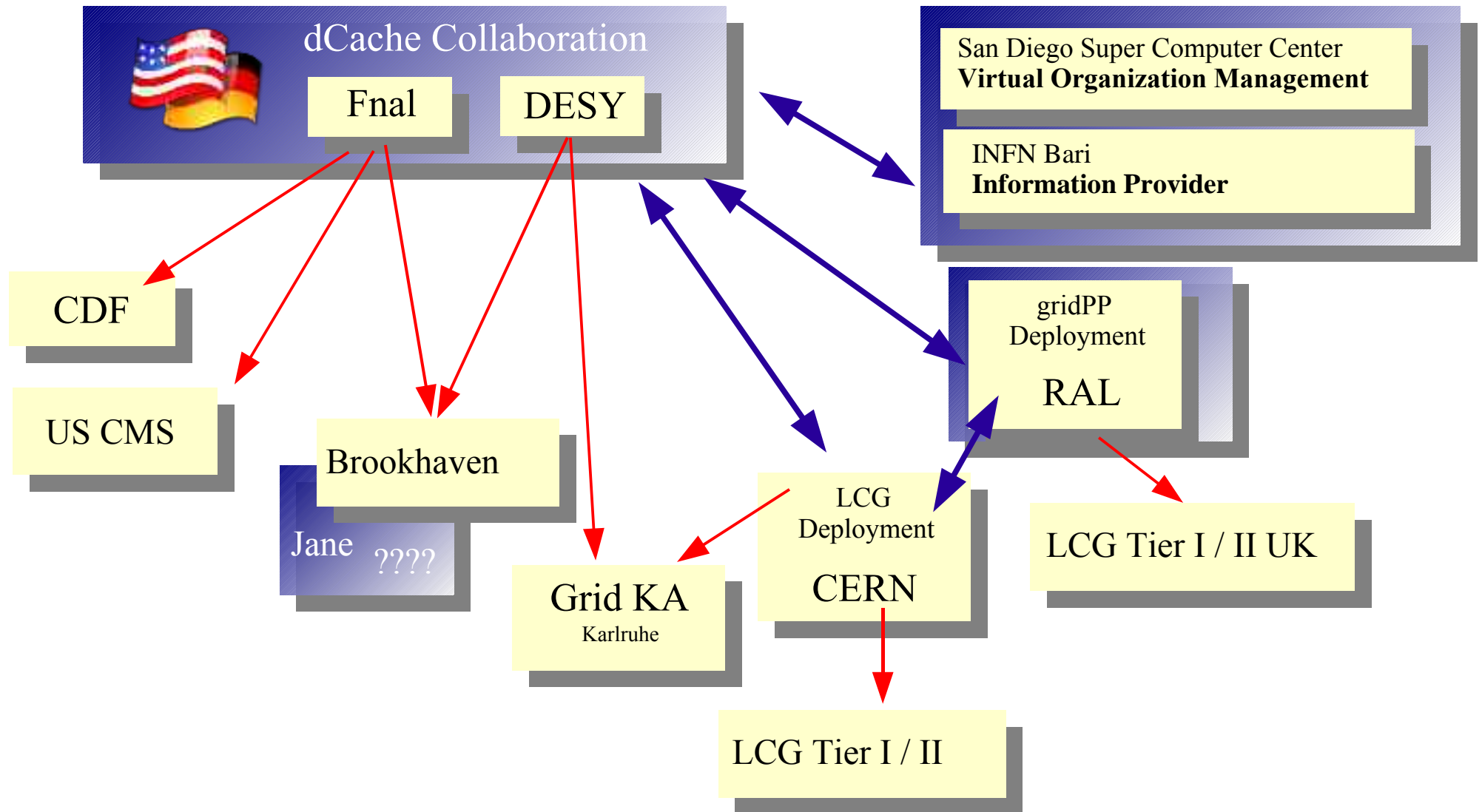
-  dCap, local posix like protocol
-  Wide area support : gsiFtp support
-  Control Protocol : SRM version ~ 1 (1.7) plus Space Reservation
-  Information Provider

Still 'poor mans' solution

BUT : real version already in 1.6.6 but not configured yet
planned to be done by RAL

Provides total and VO space plus protocol entry point







Central download and documentation area

www.dCache.ORG/manuals/Book

Special Ticket queues :

RAL

CERN

Generic Ticket queues :

support@dCache.org

Installation resp. setup support for huge installations

support@dCache.org and by (video/phone) conference

Mailing list for user communication

user-forum@dCache.org





HEP

France : Lyon

Germany : gridKa, RWTH, Aachen, DESY Hamburg, Zeuthen, GSI Darmstadt ...

Italy : INFN : Torino, Bari, CNAF Bologna

Netherlands : Amsterdam(Sara)

Spain : Madrid

UK : RAL, Manchester, Lancaster ...

Canada : Alberta, Toronto, Triumf

US : Fermi, Brookhaven

USCMS : UCSD, Florida, Nebraska, Purdue MIT soon.

International lattice data grid

DESY, Zeuthen

ZAM, Juelich (end of Nov)

ZIB, Berlin





Bug Fixes :

Slow SRM fixed

Pin Manager doesn't stuck any more

Space Management now works with LCG tools

Balanced use of pools for gridftp transfers (learned from SC3)



Improvements (significantly) learned from dCache workshop @ DESY

Documentation (dCache, the Book)

Installation





New features :

SRM, space reservation

resilient dCache

pnfs with postgres DB (improved backup and synchronization)

cache info in DB (companion)

grid ftp performance markers



In code, tested, but not configured

Multiple I/O queues per pool (slow versus fast transfers)

Improved information service (integrated into GIP)

In code, but not sufficiently tested

VOMS integration





White : observed numbers

Yellow : expected numbers in near future

Total amount of disk space

Number of "open" operations / second

2 PetaBytes

1000

50

300 – 500 Tbytes



300 TB/Day resp. 3 GB/sec

100 – 300 Pools

???

1000 pools

Sustained transfer rate

Number of pools attached





Identified shortcoming in 6 – 12 month future

File system operations (pnfs) too slow

Consequence

new project introduced @ chep 2004 by Tigran : Chimera

Properties

Totally db based file system implementation

Fast access of dCache into file system (no longer via nfs)

Should scale with price of database

Status

Fully functional prototype ready

Performance evaluation phase



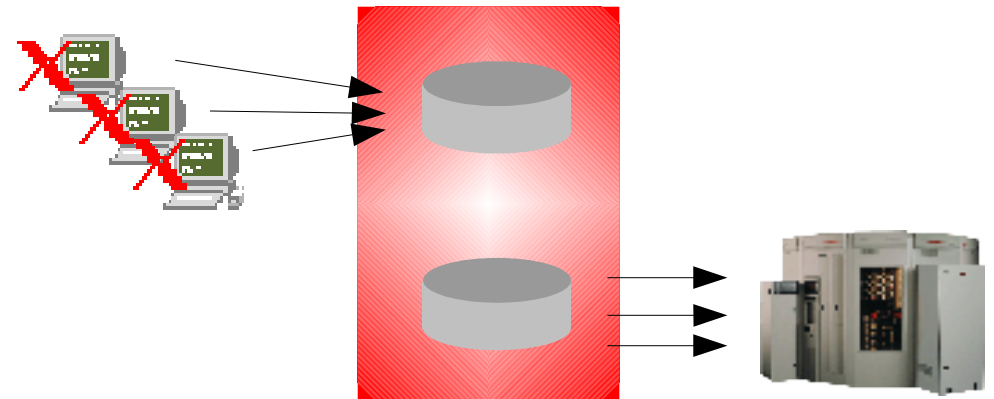


Improved interactions between dCache and HSM back end

Coordinated flush of all pools into HSM

Alternate disk write / HSM flush on different pools

never write while flushing
never flush while writing



Central HSM restore Manager

may collect restores (same as flush)

can give hints to HSM about next files to restore

Overcome problems with Tivoli Storage Manager (gridKa)





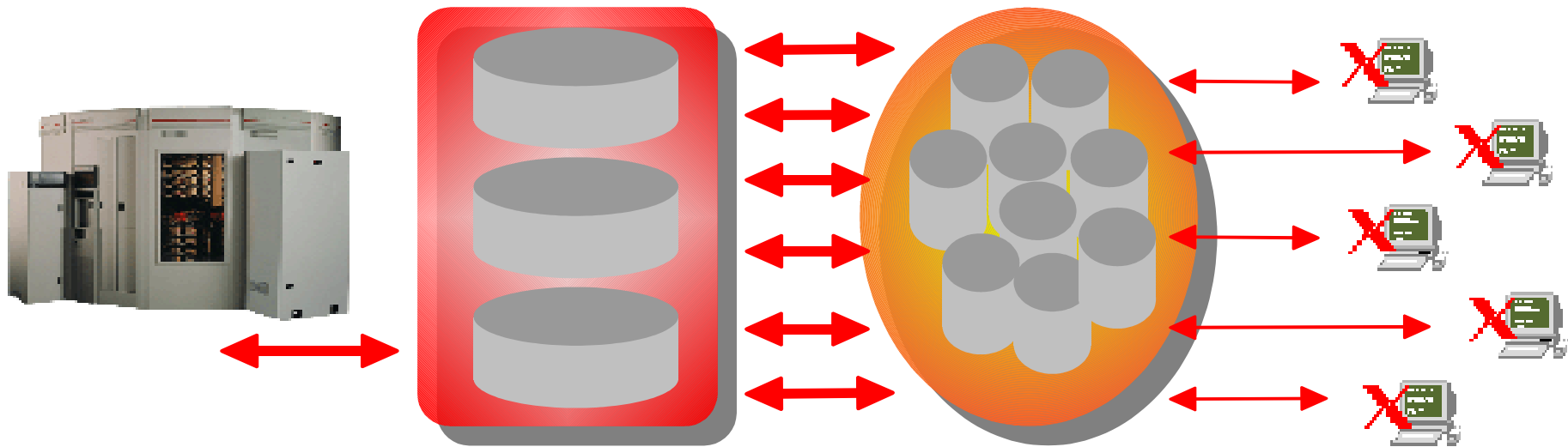
Treated like tape

inexpensive

high density

switch on/off

restore only on disk failures



Evaluated by Martin Gasthuber

easy for dCache Questions are :

does it make sense at all

which percentage of tape on disk > 90 % ??

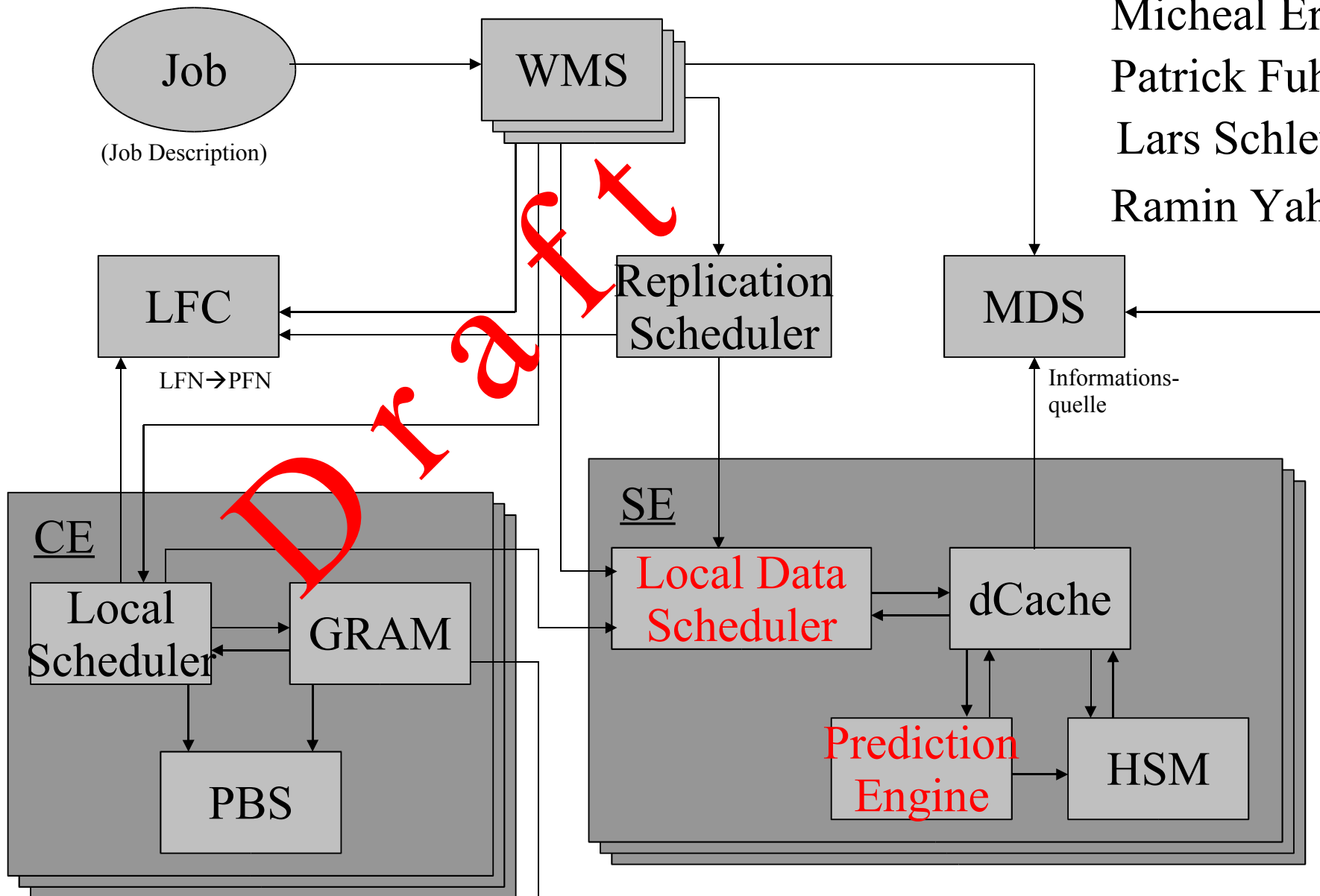
max price per GB (50 cent ??)

are there better solutions ?





Improved Scheduling



Micheal Ernst
Patrick Fuhrmann
Lars Schley
Ramin Yahyapour



