Putting Xroot and Ceph at the Heart of CASTOR

Sébastien Ponce
sebastien.ponce@cern.ch

CERN

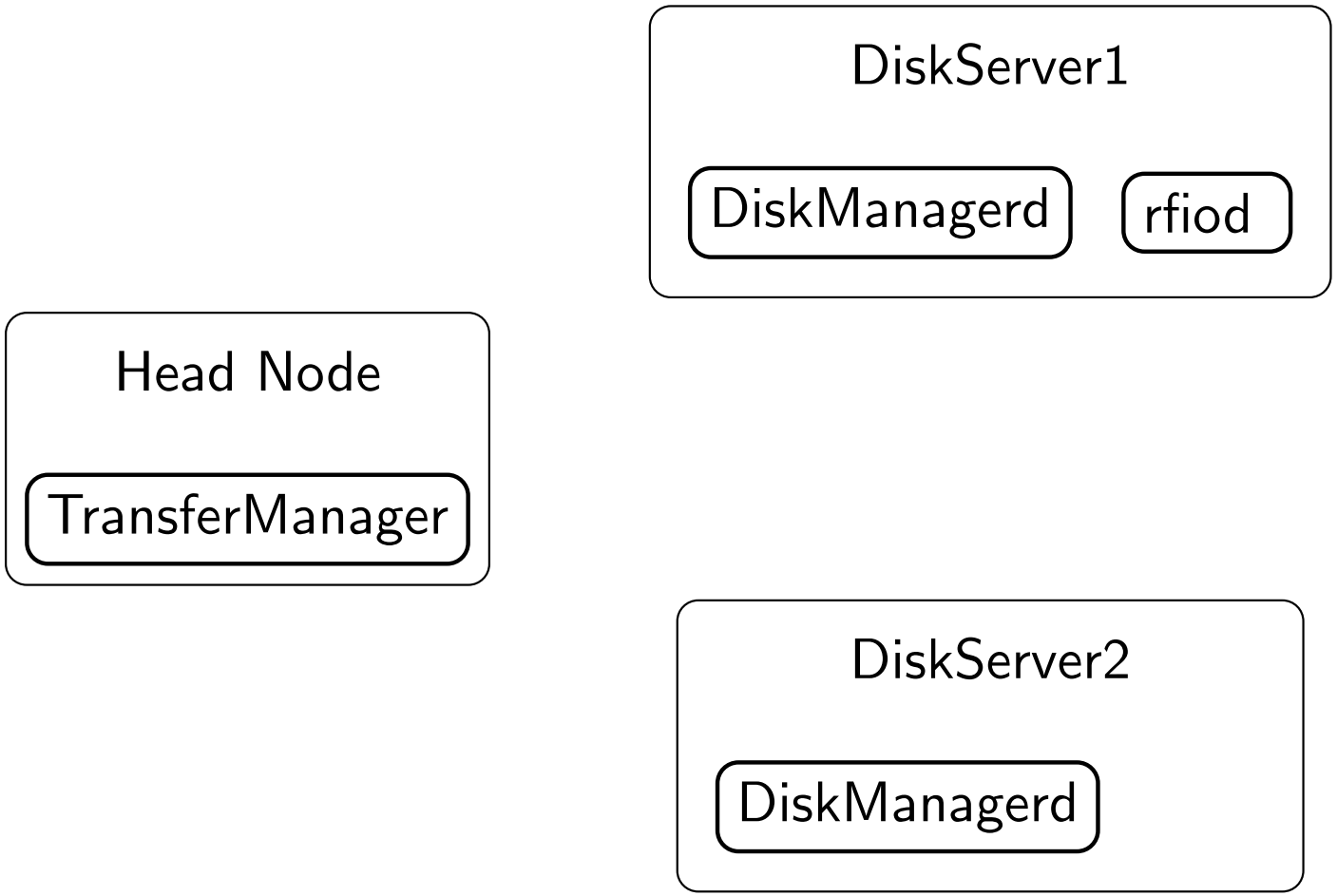
May 19th 2014

Outline

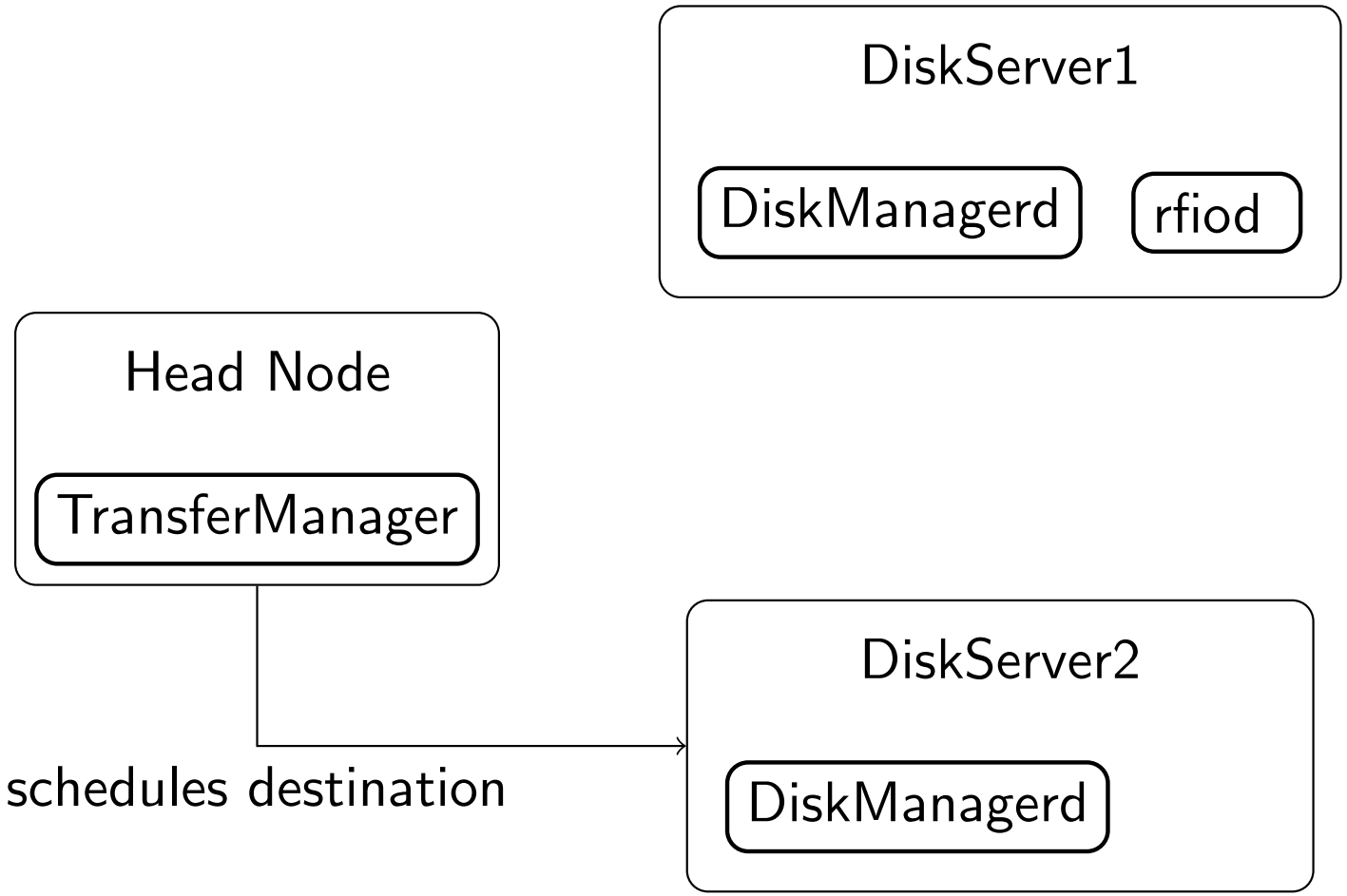
1. Moving disk to disk copies to Xroot
2. Moving tape transfers to Xroot
3. Introducing Ceph
4. Current status and future deployment

Moving disk to disk copies to Xroot

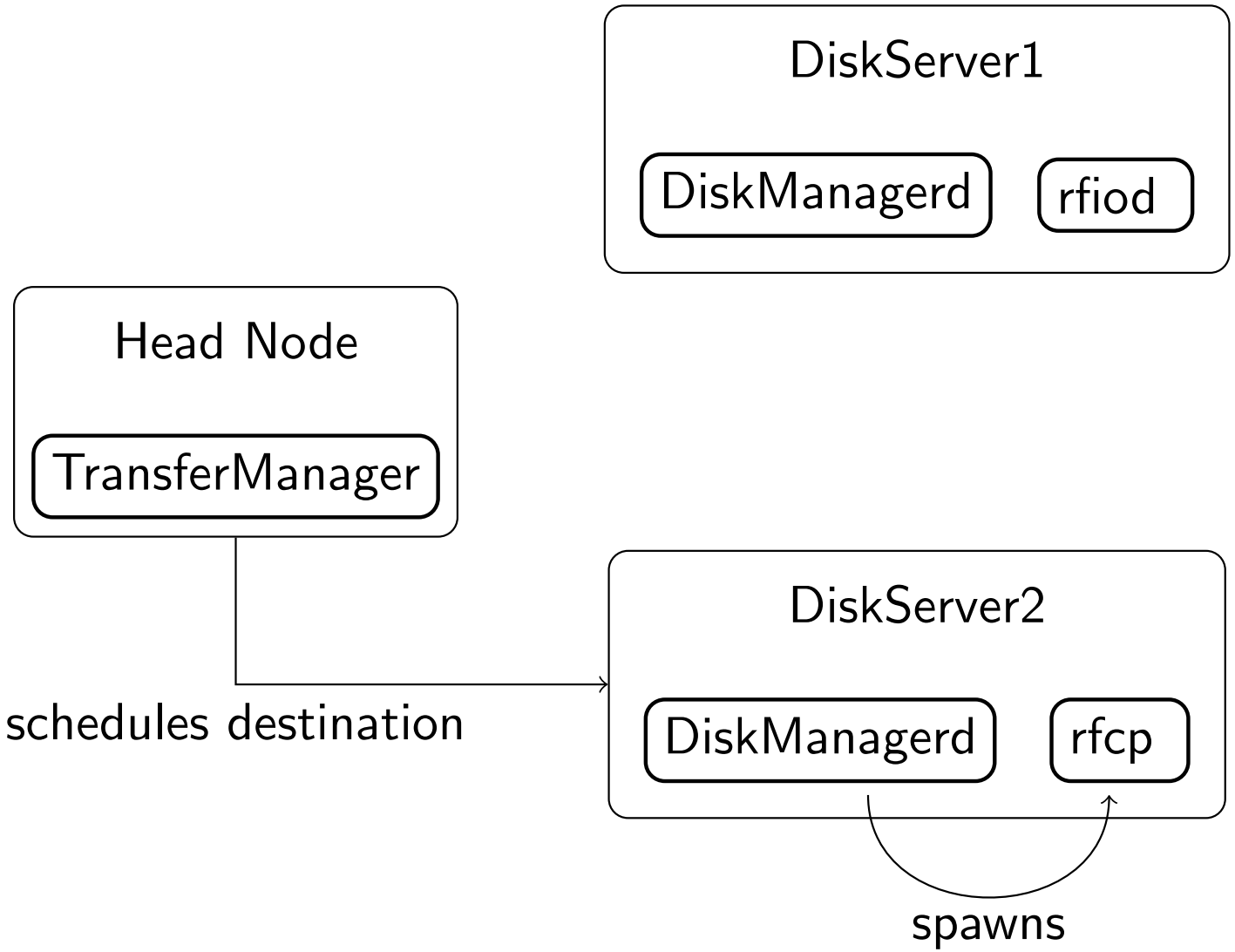
Sketch of a disk to disk transfer



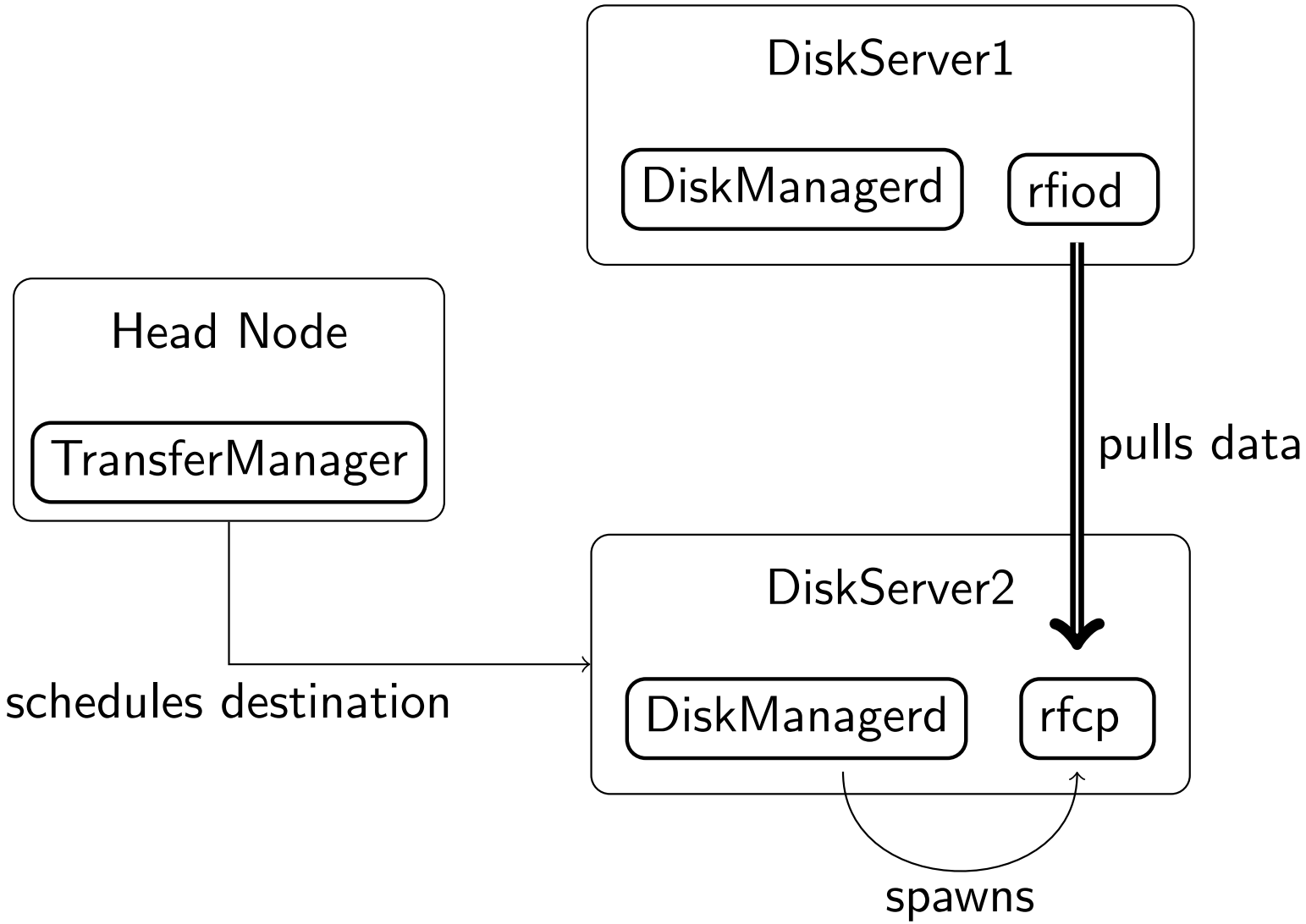
Sketch of a disk to disk transfer



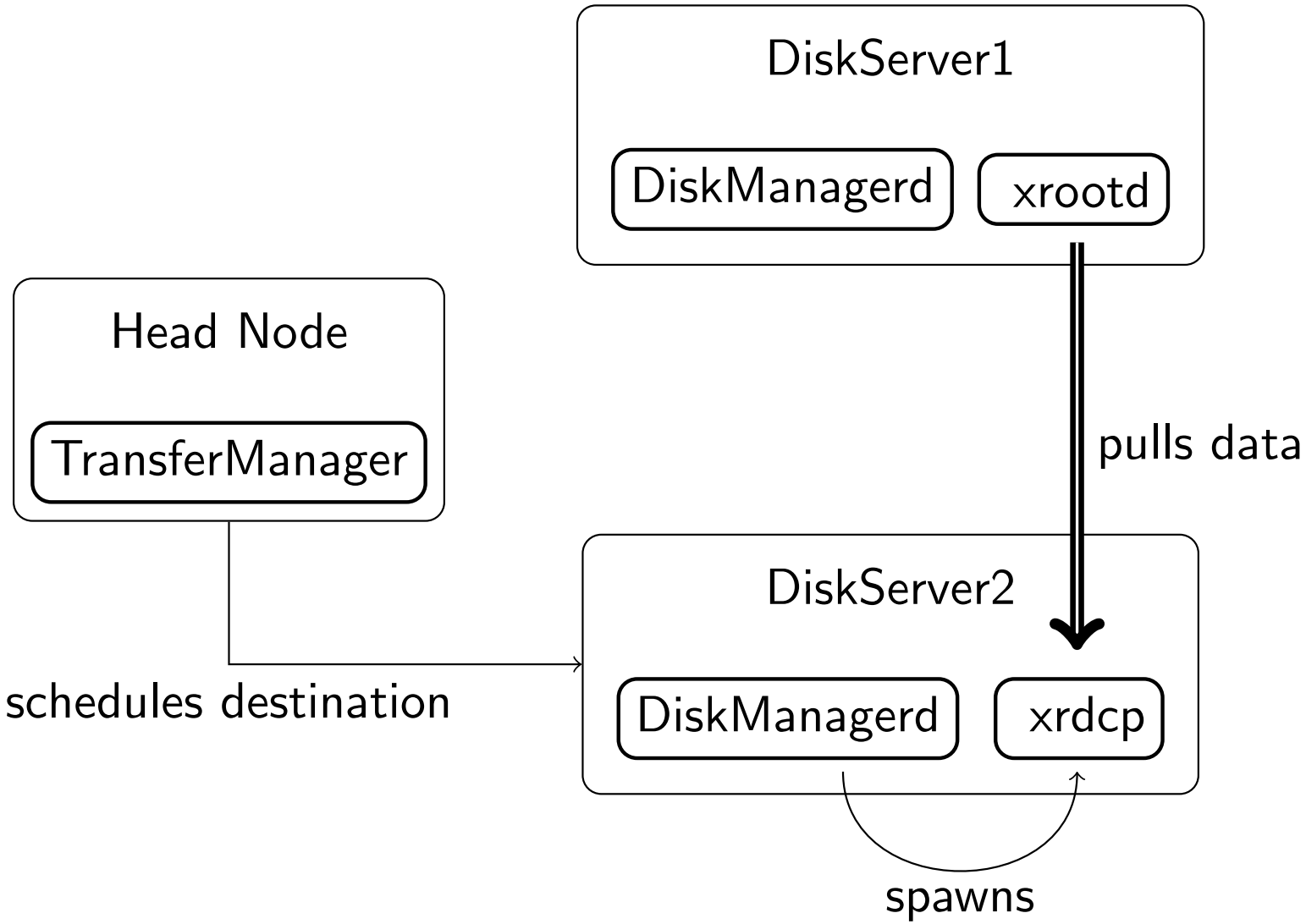
Sketch of a disk to disk transfer



Sketch of a disk to disk transfer



Moving to xroot for d2d transfers



code change

```
diff --git a/castor/scheduler/diskmanager/activitycontrol.p
index aa8146b..0bed2b7 100644
@@ -30,6 +30,9 @@ import connectionpool, dlf
+def buildXrootURL(diskserver, path):
+  return 'root://'+diskserver+':1095//dummy?castor2fs.pfn1
+
@@ -68,7 +71,8 @@ class ActivityControlThread(threading.Thr
-    cmdLine = ['rfcp', srcDcPath, destDcPath]
+    srcDS, srcPath = srcDcPath.split(':',1)
+    cmdLine = ['xrdcp',
+               buildXrootURL(srcDS, srcPath),
+               buildXrootURL('localhost', destDcPath)]
```

code change

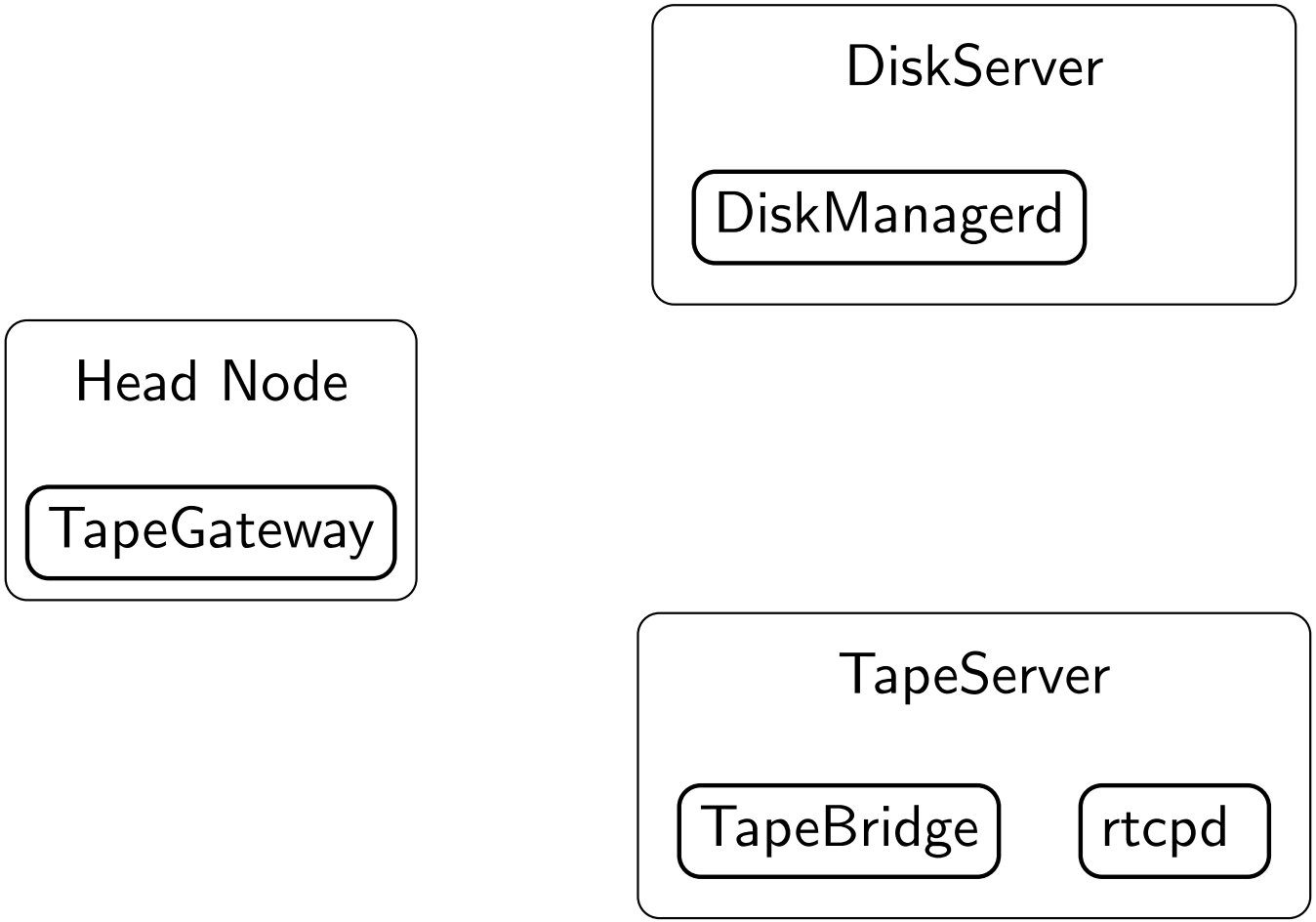
```
diff --git a/castor/scheduler/diskmanager/activitycontrol.p
index aa8146b..0bed2b7 100644
@@ -30,6 +30,9 @@ import connectionpool, dlf
+def buildXrootURL(diskserver, path):
+  return 'root://'+diskserver+':1095//dummy?castor2fs.pfn1
+
@@ -68,7 +71,8 @@ class ActivityControlThread(threading.Thr
-    cmdLine = ['rfcp', srcDcPath, destDcPath]
+    srcDS, srcPath = srcDcPath.split(':',1)
+    cmdLine = ['xrdcp',
+               buildXrootURL(srcDS, srcPath),
+               buildXrootURL('localhost', destDcPath)]
```

Security

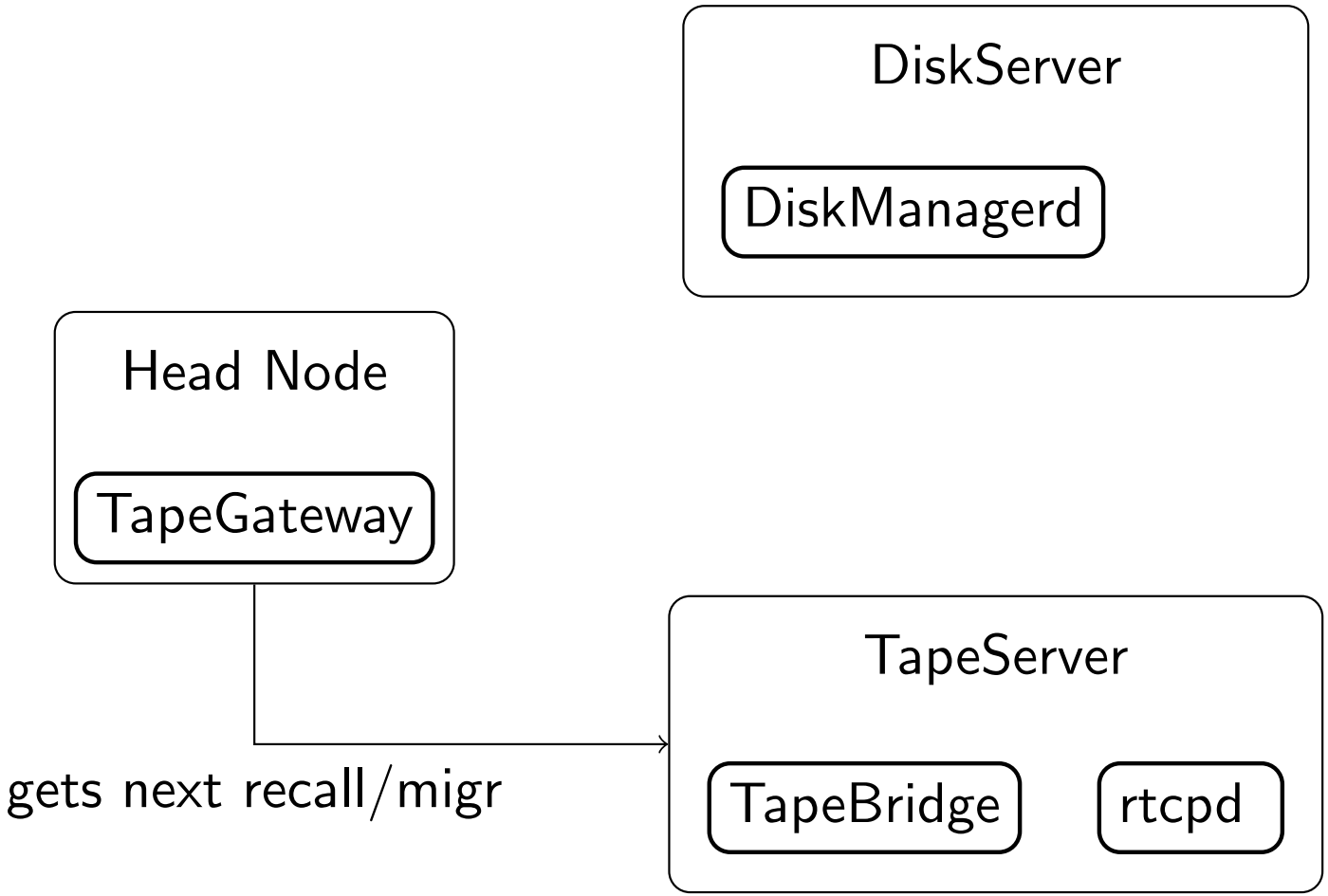
- url needs to be encrypted
- 3 mores lines using openssl module

Moving tape transfers to Xroot

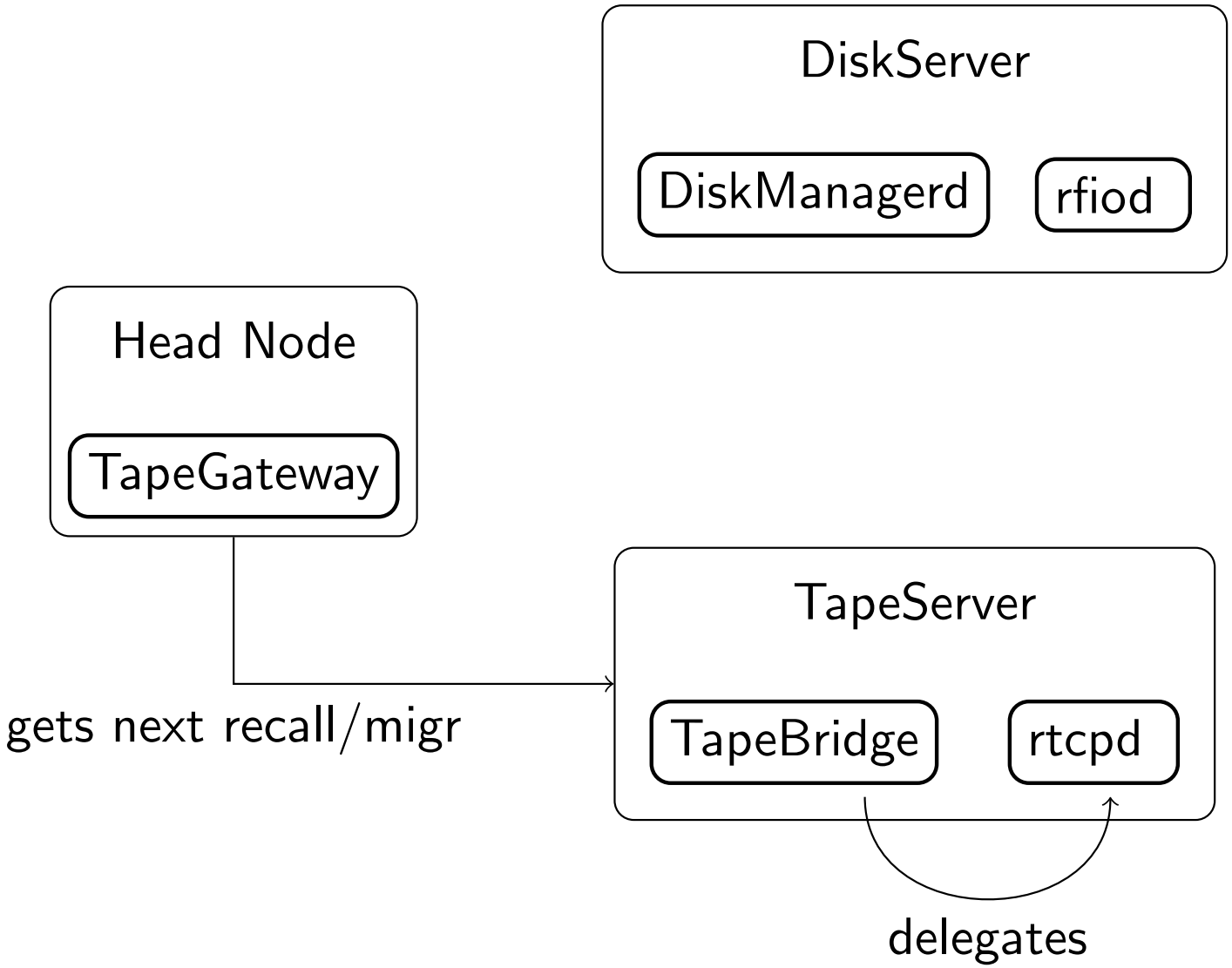
Sketch of a tape transfer



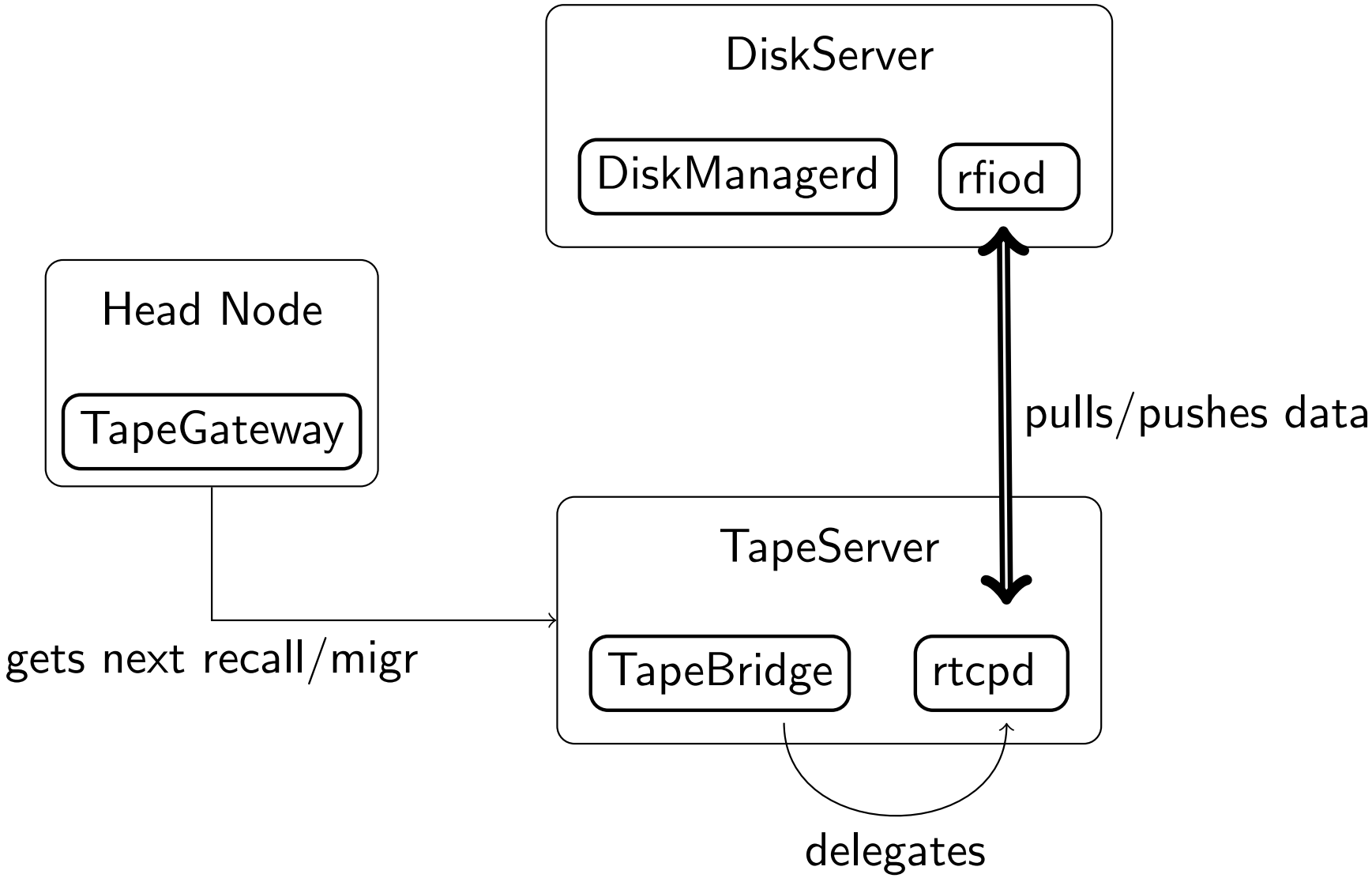
Sketch of a tape transfer



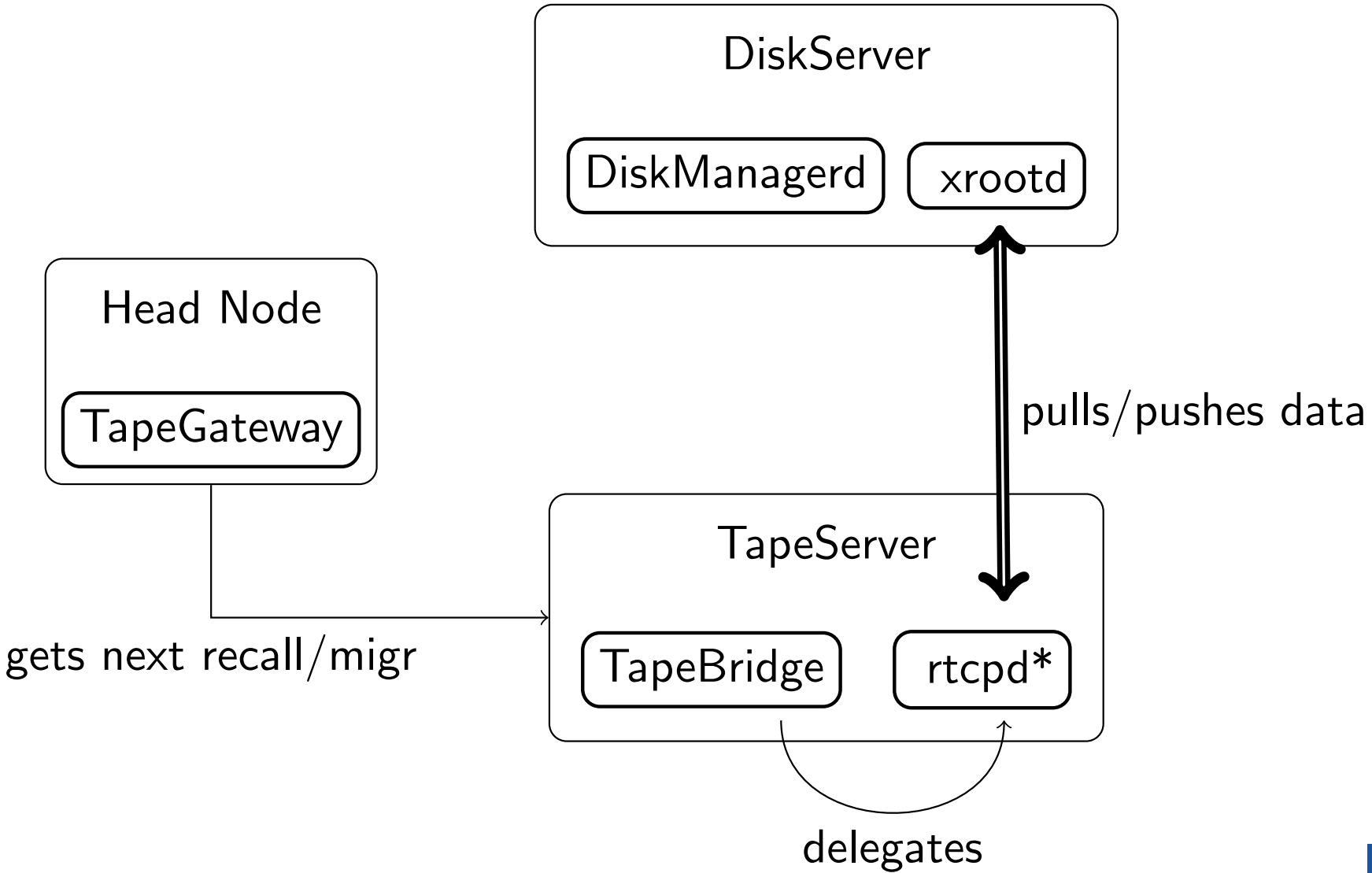
Sketch of a tape transfer



Sketch of a tape transfer



Moving to xroot for tape transfers



code change

```
>git diff --stat master xrootTape
 cmake/Findxrootd.cmake |   7 +-
 h/rtcp_xroot.h         |  22 +++++
 rtcopy/CMakeLists.txt  |   8 +-
 rtcopy/rtcp_CheckReq.c |  94 ++++++++++++++++++------
 rtcopy/rtcp_xroot.c    |  49 ++++++++++++
 rtcopy/rtcpapi.c       |   2 -
 rtcopy/rtcpd_Disk.c    | 278 ++++++++++++++++++++++++++++++++++++++++++++++++++
 7 files changed, 275 insertions(+), 185 deletions(-)
```

code change

```
>git diff --stat master xrootTape
 cmake/Findxrootd.cmake |   7 +-
 h/rtcp_xroot.h         |  22 +++++
 rtcopy/CMakeLists.txt  |   8 +-
 rtcopy/rtcp_CheckReq.c |  94 ++++++++++++++++++------
 rtcopy/rtcp_xroot.c    |  49 ++++++++++++
 rtcopy/rtcpapi.c       |   2 -
 rtcopy/rtcpd_Disk.c    | 278 ++++++++++++++++++++++++++++++++++++++++++++++++++
 7 files changed, 275 insertions(+), 185 deletions(-)
```

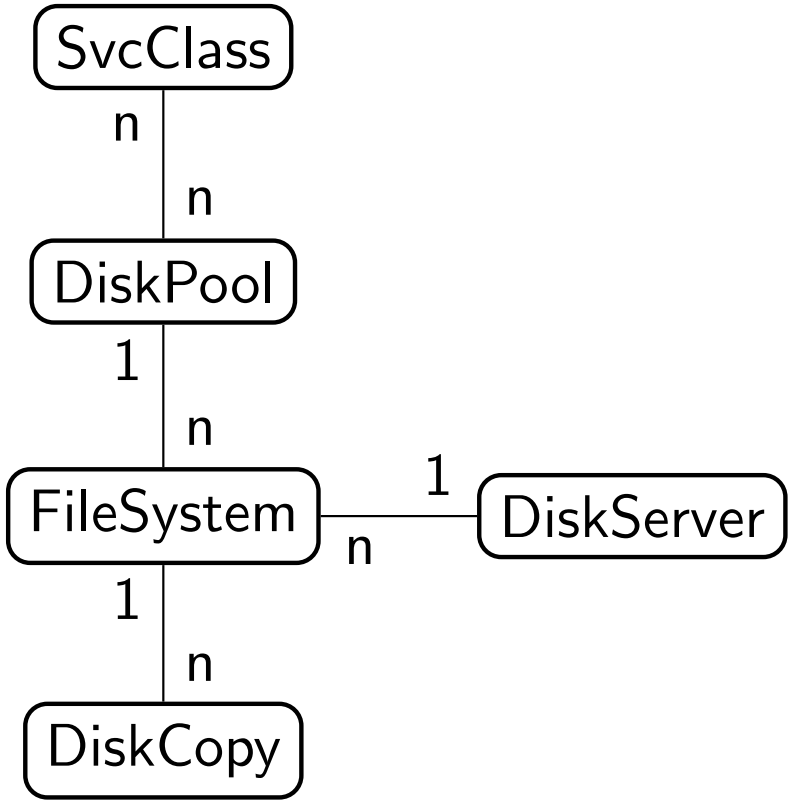
Security

- url needs to be encrypted
- mores lines using openssl module

Introducing Ceph

Evolution of the stager DB schema

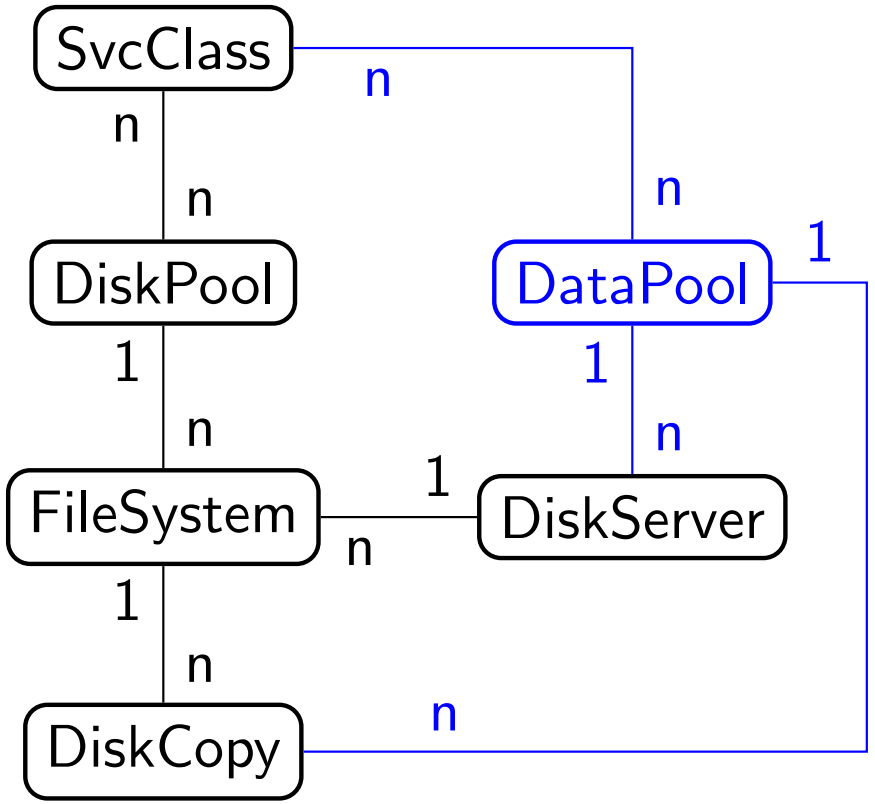
- DiskPools are a set of FileSystems
- Each FileSystem belong to a DiskPool
- DiskCopies reside in FileSystems



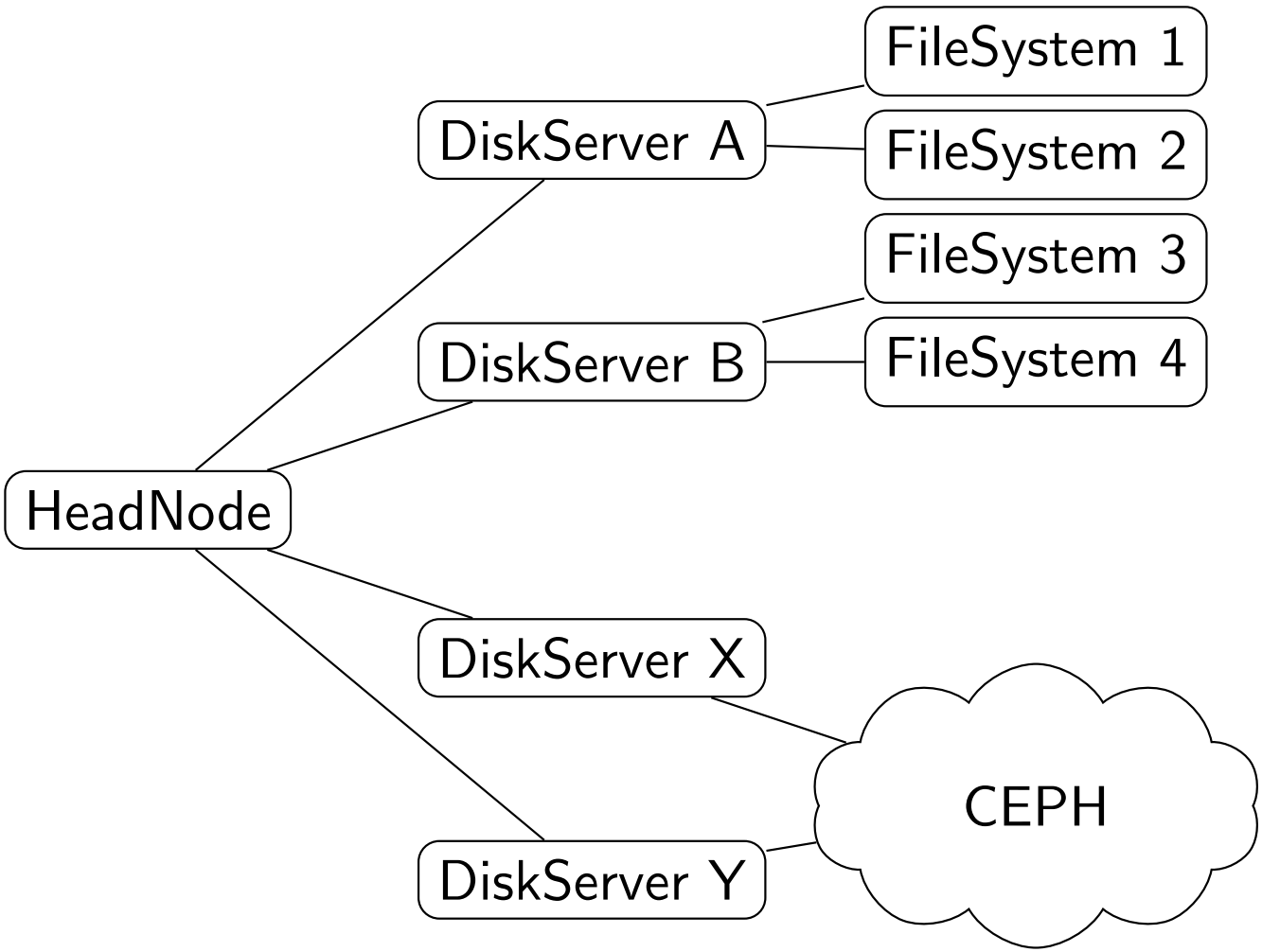
Evolution of the stager DB schema

- DiskPools are a set of FileSystems
- Each FileSystem belong to a DiskPool
- DiskCopies reside in FileSystems

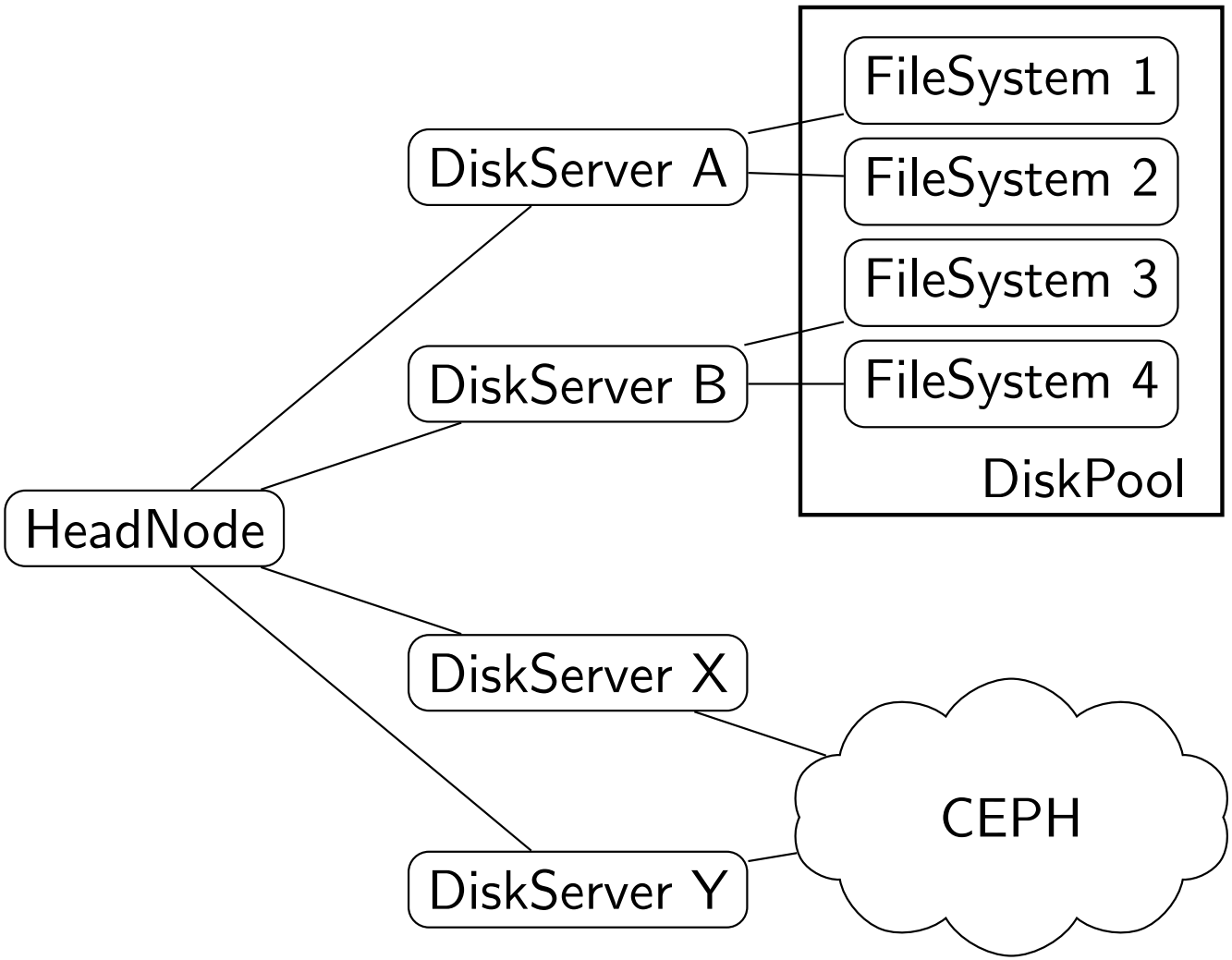
- DataPools are independent entities
- Each DiskServer serves a given DataPool
- DiskCopies reside in the DataPool



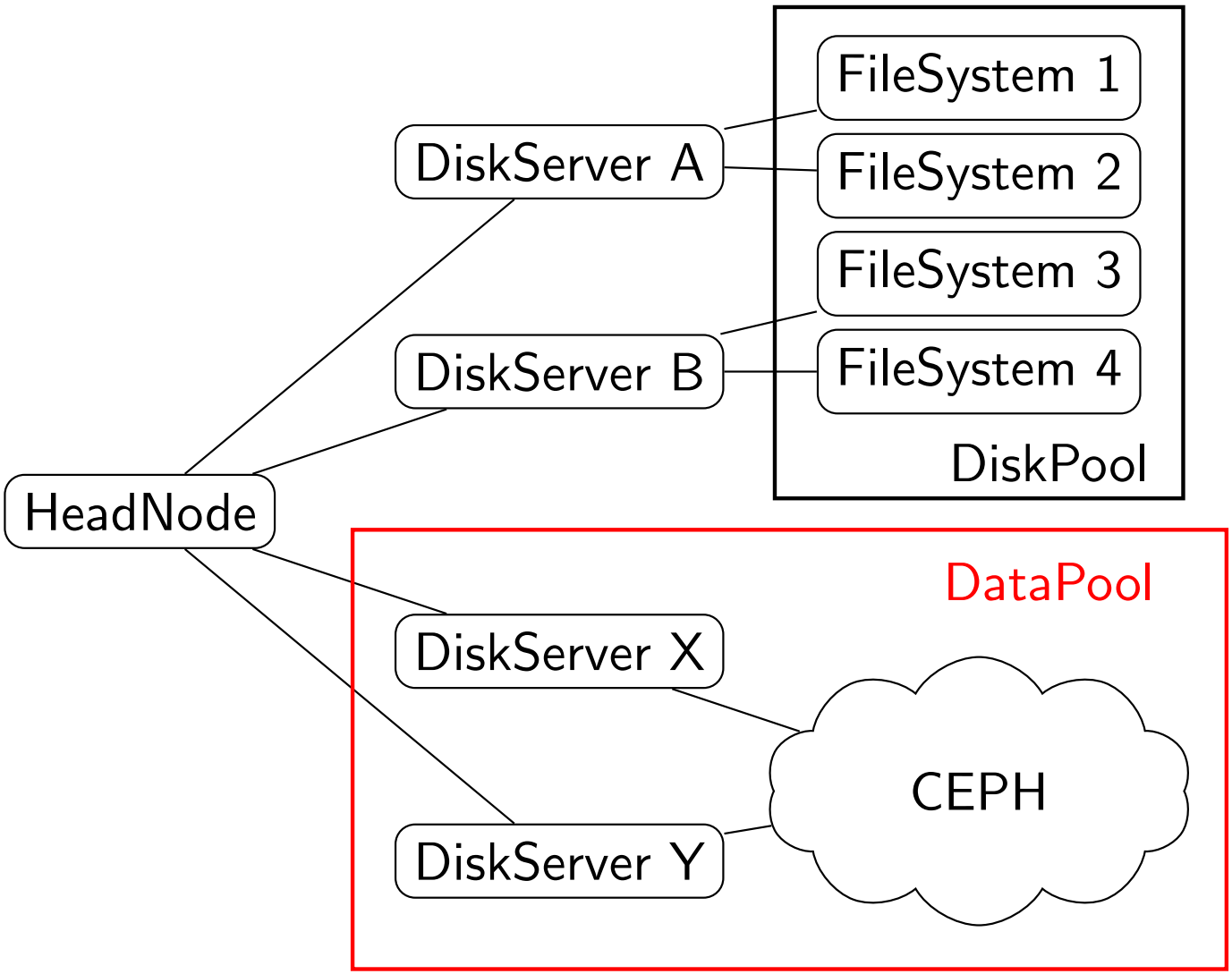
Practically



Practically



Practically



Practical Implementation

Code to be changed

- DB code has to be adapted
- transfermanagerd does not change at all
- protocols need to be CEPH enabled
- GC and synchronization need to be adapted

Practical Implementation

Code to be changed

- DB code has to be adapted
- transfermanagerd does not change at all
- protocols need to be CEPH enabled
- GC and synchronization need to be adapted

Protocol details

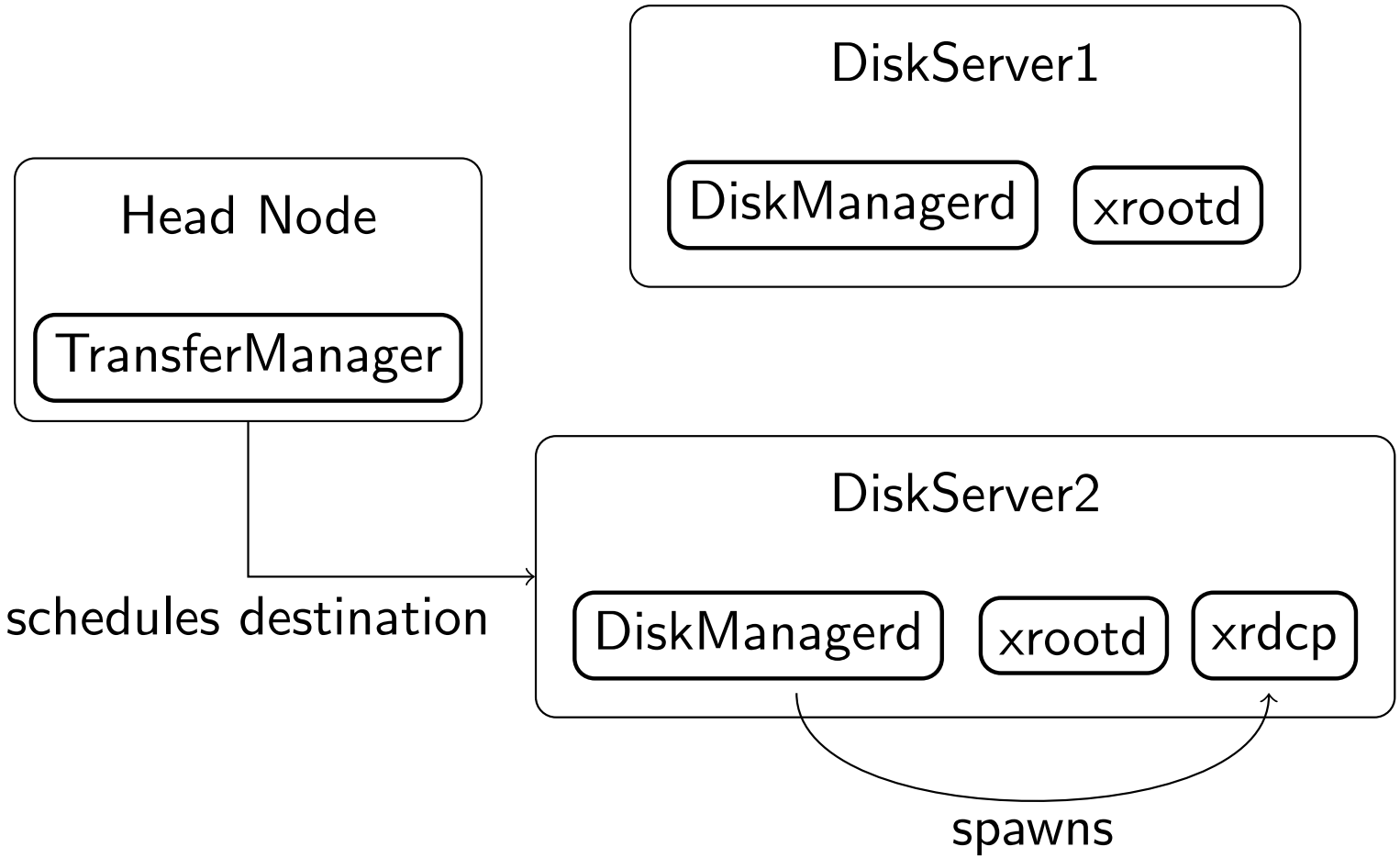
- root is discontinued
- rfiod needs to be adapted
- gridftp needs to be adapted
- xrootd needs an OSS plugin

Disk2 disk copies with ceph

- in principle rfcp/xrdcp should be adapted too and talk ceph
- but xroot can used its local daemon and its OSS plugin

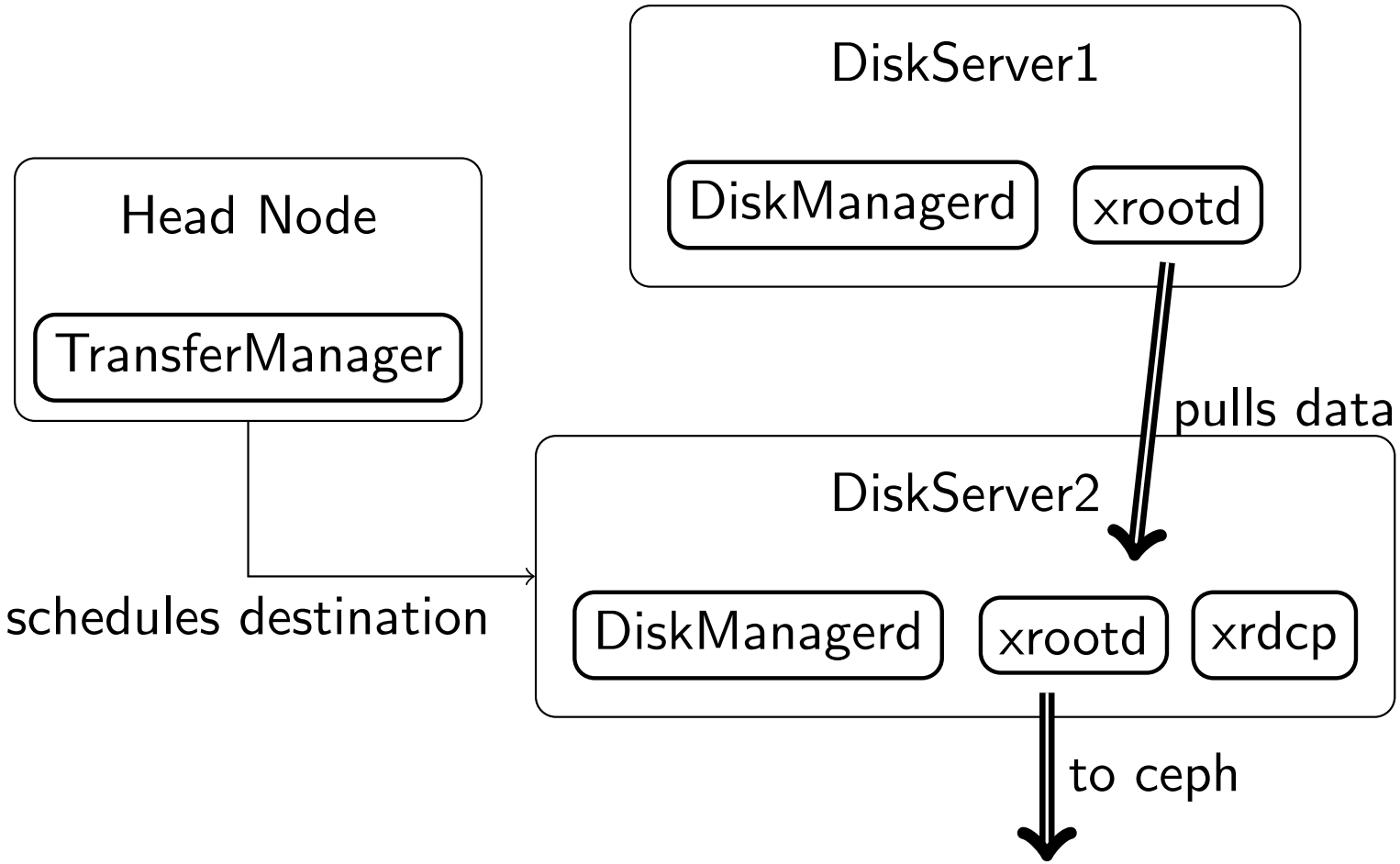
Disk2 disk copies with ceph

- in principle rfcp/xrdcp should be adapted too and talk ceph
- but xroot can used its local daemon and its OSS plugin



Disk2 disk copies with ceph

- in principle rfcp/xrdcp should be adapted too and talk ceph
- but xroot can used its local daemon and its OSS plugin

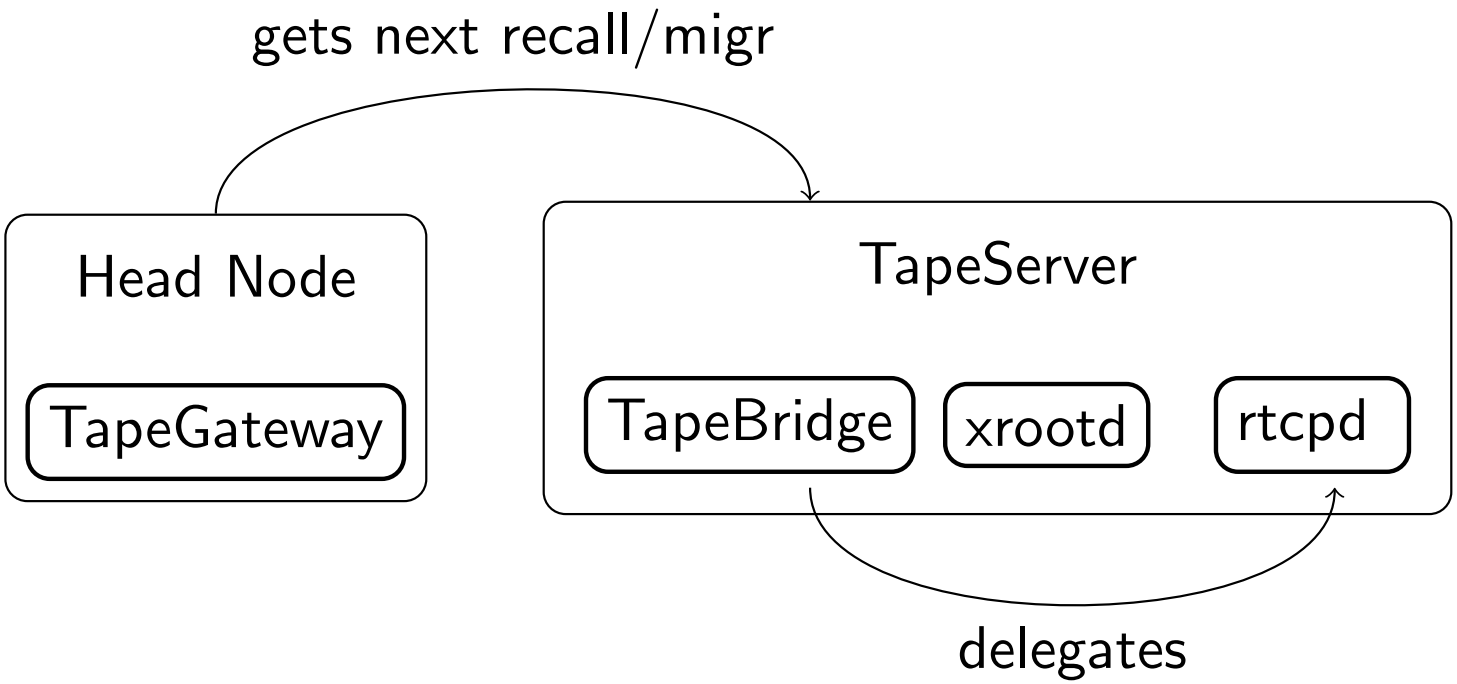


Tape transfers with ceph

- Same situation as for disk to disk copies
- and we run an xroot daemon on the tapeservers

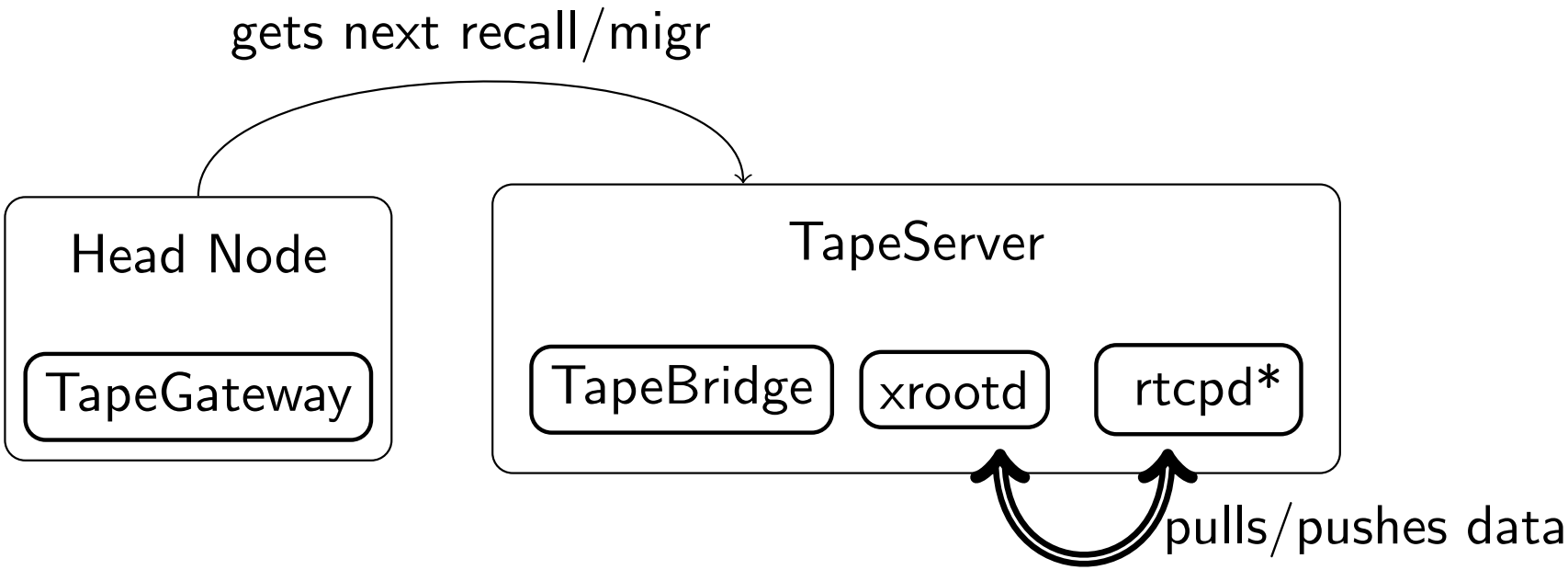
Tape transfers with ceph

- Same situation as for disk to disk copies
- and we run an xroot daemon on the tapeservers



Tape transfers with ceph

- Same situation as for disk to disk copies
- and we run an xroot daemon on the tapeservers



Current status and future deployment

Status right now

- ✓ D2d copies through xroot
  - × security has to be added
- ✓ Tape transfers through xroot
  - × security has to be added
- ✓ striping in ceph
  - although will only be in giant
- ✓ DB schema and PL/SQL changes
- ✓ adaptation of all tools (e.g. printpool)
- × adaptation of all protocols
  - ✓ rfio under test
  - × gridFTP to be done
  - × xroot OSS plugin to be written
- × GC and synchronization to be done

Deployment

Certitudes

- 2.1.15 will have xroot for internal transfers and ceph enabled
- But LHC production should start without DataPools/ceph
- Major repack campaign should finish without DataPools/ceph
- External institutes should not play with it and real data

Deployment

Certitudes

- 2.1.15 will have xroot for internal transfers and ceph enabled
- But LHC production should start without DataPools/ceph
- Major repack campaign should finish without DataPools/ceph
- External institutes should not play with it and real data

Proposals

- Heavy testing to be done on ITDC/pps for the stability/deployment of CEPH itself
- Then we could test it on non critical repacks
- Before a test production pool can be created (2015)
  - "use at your own risk SLA"