

# CASTOR2 deployment

Olof Barring, IT-FIO/DS

30 May 2005

# Outline

- What is CASTOR2?
- Constraints & observations
- Deployment order
- Managing co-existing old & new stagers
- Deployment architectures
- The steps
- Timeline
- Summary

# What is CASTOR2? (1)

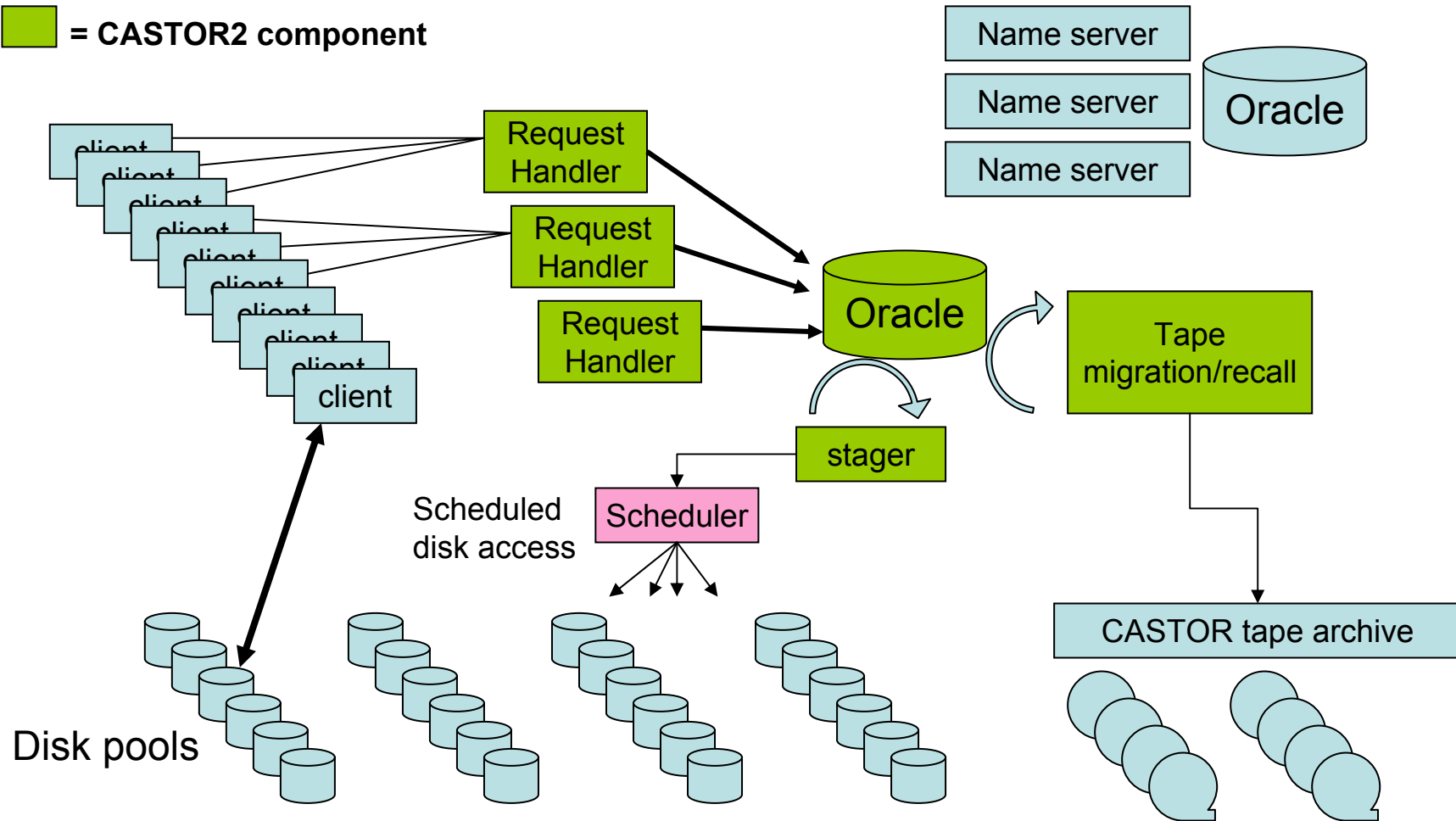
- CASTOR2 is the last major steps in enabling CASTOR for LHC data
  - It is a complete replacement of the stager (disk cache mgmt) component
  - Central services (name server, VMGR, ...) remain untouched
  - Minor changes to tape archive (RTCOPY), most are already deployed in production since one year
- CASTOR2 also comes with strong authentication but this is not enabled in the first deployment

# What is CASTOR2? (2)

- Request scheduling
  - Throttle under high load (CASTOR never said “no” until it was too late)
  - Supports Maui and LSF
  - Flexible resource administration
  - (Fair-) Sharing and guaranteed resources
- DB centric
  - Use of standardized and proven DB technology
  - Stateless daemons
  - Requests processing shielded from the request registration
- Scalable
  - Database for disk file residence
  - LSF scales to >100k queued jobs
- Overcome tape queue limitations
  - Requests for same tape are bundled together and new requests are appended
- Already tested with ALICE MDC6, StorageTank and CMS

# What is CASTOR2? (3)

 = CASTOR2 component



# Constraints & observations (1)

- SC3 service phase involves all LHC experiments' production groups
  - Must be migrated to CASTOR2 in time otherwise they won't be migrated until after SC3, i.e. beginning of 2006
  - SC4 starts April 2006 → service must be ready by end-February(?): not enough time
- Conclusion: missing SC3 → missing SC4 → CASTOR2 not deployed for LHC in time for stable operation (Sept'06)

# Constraints & Observations (2)

- SC4 involves physics analysis
  - Throughput phase in April'06
  - General LHC user must be migrated to CASTOR2 well in time: >2 months before throughput phase

# Constraints & observations (3)

- LHC Experiments' production groups are already suffering from limitations with the current stager
  - For the experiments' sharing production and normal users, the latter are also affected
    - This is in particular a problem for ATLAS these days
- All other user groups are OK
  - There is no particular urgency to migrate the non-LHC groups (including data-taking fixed target experiments) off the current stager



# Constraints & observations (4)

- CASTOR1 - CASTOR2 compatibility issues
  - RFIO API and command line backward compatible
  - Stager API and command line *not* backward compatible by design
    - CASTOR1 commands stagein, stageqry, ...
    - CASTOR2 commands stager\_get, stager\_qry, ...
    - The new and old stager command sets are non-overlapping
  - Old and new client API and commands can and do co-exist on the same machine
  - Resource hungry queries will be limited to administrators
  - Managed storage – internal information (e.g. disk path) not exposed to end-users
    - No backdoors - all access is scheduled
    - Avoid permissions conflicts between disk and CASTOR files

# Constraints & observations (5)

## Today's production stagers ...

<b>afs83</b>	<b>lxfsrk4506</b>	<b>stagealicedc04</b>	<b>stagecms</b>	<b>stagelhcb</b>
<b>stagedelphi</b>	<b>lxfsrk4507</b>	<b>stagealicedc04a</b>	<b>stagecmsprod</b>	<b>stagen45</b>
<b>lxfs6142</b>	<b>lxfsrk4508</b>	<b>stagealicedc05</b>	<b>stagecompass</b>	<b>stagen48</b>
<b>lxfsrk4501</b>	<b>stageopal</b>	<b>stageams</b>	<b>stagedirac</b>	<b>stagen49</b>
<b>lxfsrk4502</b>	<b>pubcdr005d</b>	<b>stageatlas</b>	<b>stagegridsc</b>	<b>stagenomad</b>
<b>lxfsrk4503</b>	<b>pubcdr006d</b>	<b>stageatlasdc2</b>	<b>stageharp</b>	<b>stagentof</b>
<b>lxfsrk4504</b>	<b>stagealeph</b>	<b>stagecast</b>	<b>stageisolde</b>	
<b>lxfsrk4505</b>	<b>stagealice</b>	<b>stagechorus</b>	<b>stageI3</b>	

**... must all be migrated**

# Deployment order

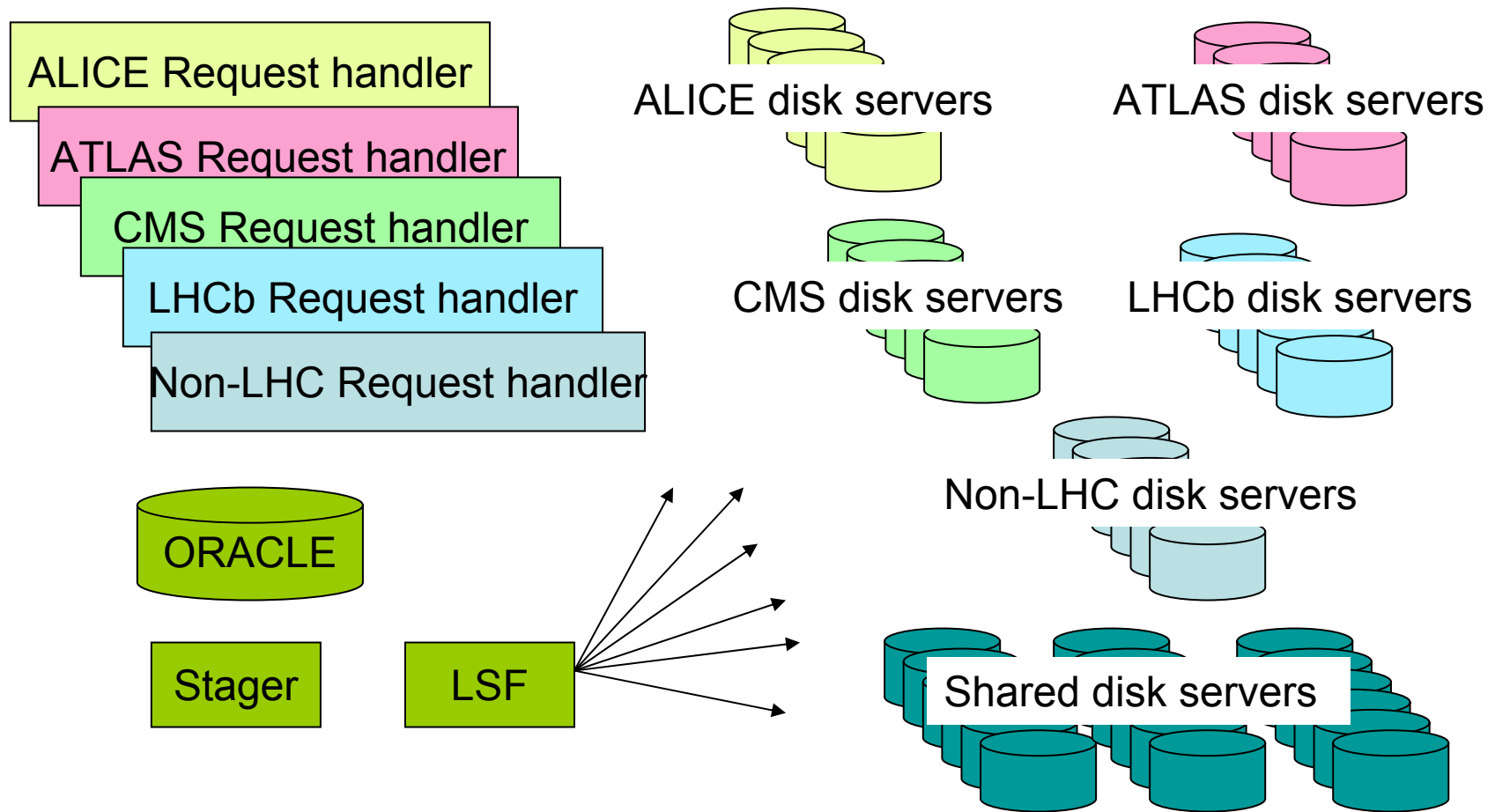
- Priority 1: deploy CASTOR2 for LHC experiments' production groups
  - In order they have declared their participation for the SC3 service phase:
    - CMS, ALICE
    - ATLAS, LHCb
- Priority 2: deploy CASTOR2 for general LHC users (in any order)
- Priority 3: deploy CASTOR2 for all other users
  - Begin with active groups (e.g. COMPASS, NA48)

# Co-existing old & new CASTOR

- The new and old stagers have to co-exist for a considerable time
  - Share the central services and tape archive
  - Clients (Ixplus, Ixbatch)
    - Both old and new client installed
      - RFIO compatible with both; switch is manual (environment variable or configuration)
      - Stager API and commands are different
  - Clients (grid)
    - Gridftp uses RFIO and works with both new and old stagers
      - Stager mapping used for switching
    - SRM uses both new and old stager API
      - Stager mapping used for the switching

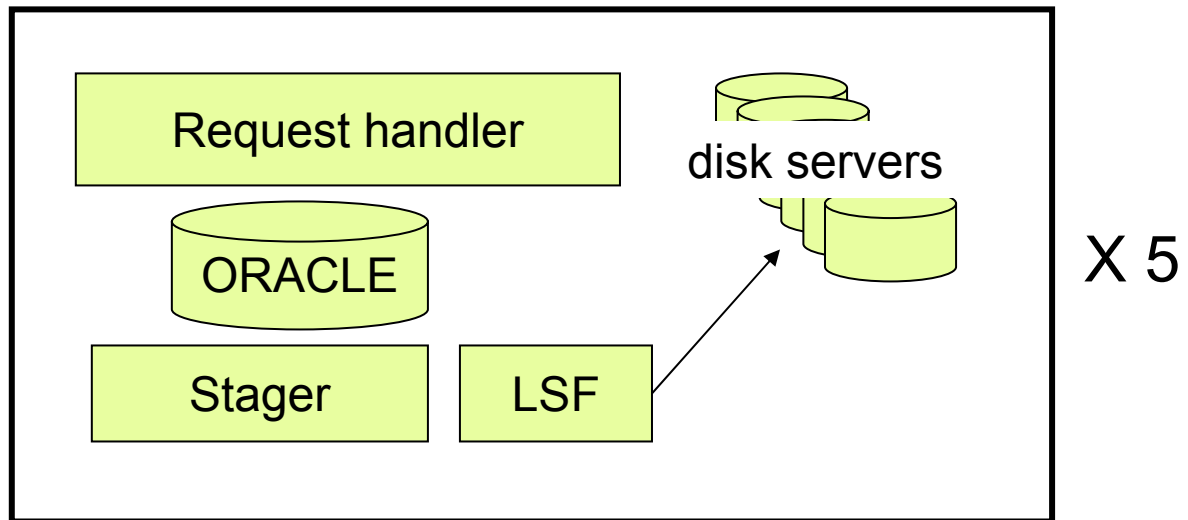
# Deployment architectures (1)

## Shared database and scheduling



# Deployment architectures (2)

Separate everything for each group



# Deployment architectures (3)

- Start with shared instance of Oracle and LSF
  - Similar architecture as for the CASTOR name server
  - Separate request handlers removes the interference between experiments
- Experience during the SC3 service phase will tell if we need to deploy independent instances
- In whatever scenario:
  - CASTOR2 instance is proposed in parallel with the experiments' production stagers
  - Old stager resources are kept until the experiments' have successfully migrated

# Oracle

- The CASTOR2 databases
  - Name server
    - All CASTOR files (30M today, billions tomorrow?)
    - Simple schema
    - High query rate, low insert/update
    - Shared among all groups
    - High availability required
  - VMGR, Cupv
    - Small and modest query/update/insert
    - Simple schema
    - High availability
  - Stager
    - Complex schema
    - ~10M rows
    - High query/insert/update
    - Extensive DB tuning during ALICE MDC6
    - High availability
  - DLF (logging)
    - High insert, low query, no updates
    - Simple schema
    - Billions of rows → use Oracle partitioning
- Use of “dataguard” in order to reduce maintenance periods
- Start with normal disk servers, one for each of:
  - Stager DB
  - Dataguard for stager DB
  - DLF
- No change to name server, VMGR and Cupv Oracle instance (old but stable disk server)

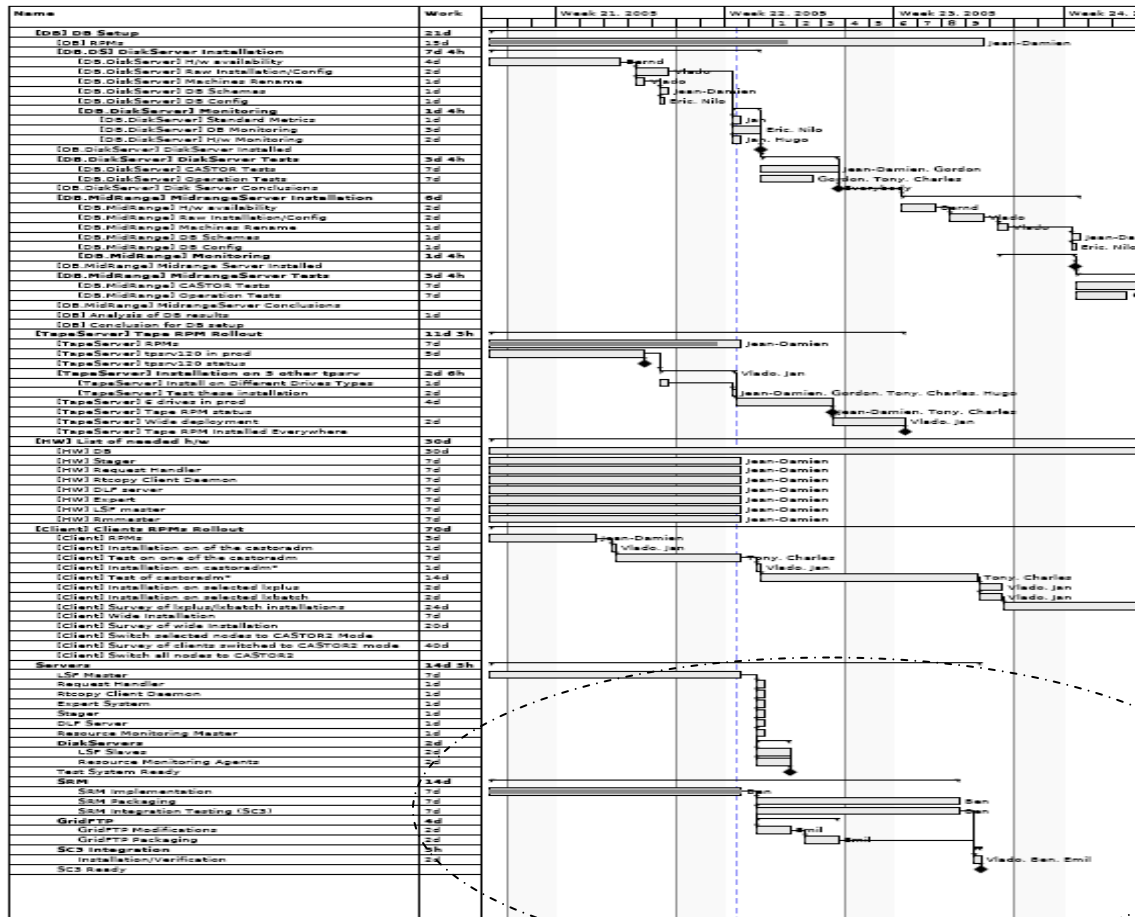


# The steps

- Install a pre-production instance
  - ~10TB disk pool to begin with
  - LSF, Oracle
- Install lxplus and lxbatch cells with new and old client
  - Verify that old client is not broken
- Stress-testing
  - Oracle tuning and hardware constraints
- Upgrade and add SC3 disk pool and upgrade SRM
- Talk with 4 LHC experiments to identify the target applications
  - Priority 1 is to cover applications used in SC3 service phase
- Widely install client on lxplus/batch
  - Enable new stager for selected (production) applications
- If necessary, split-up in several instances
- Migrate the LHC general user groups
- Migrate the non-LHC user groups
  - Probably involving tricky dependencies on the old stager commands.
- Deploy strong authentication

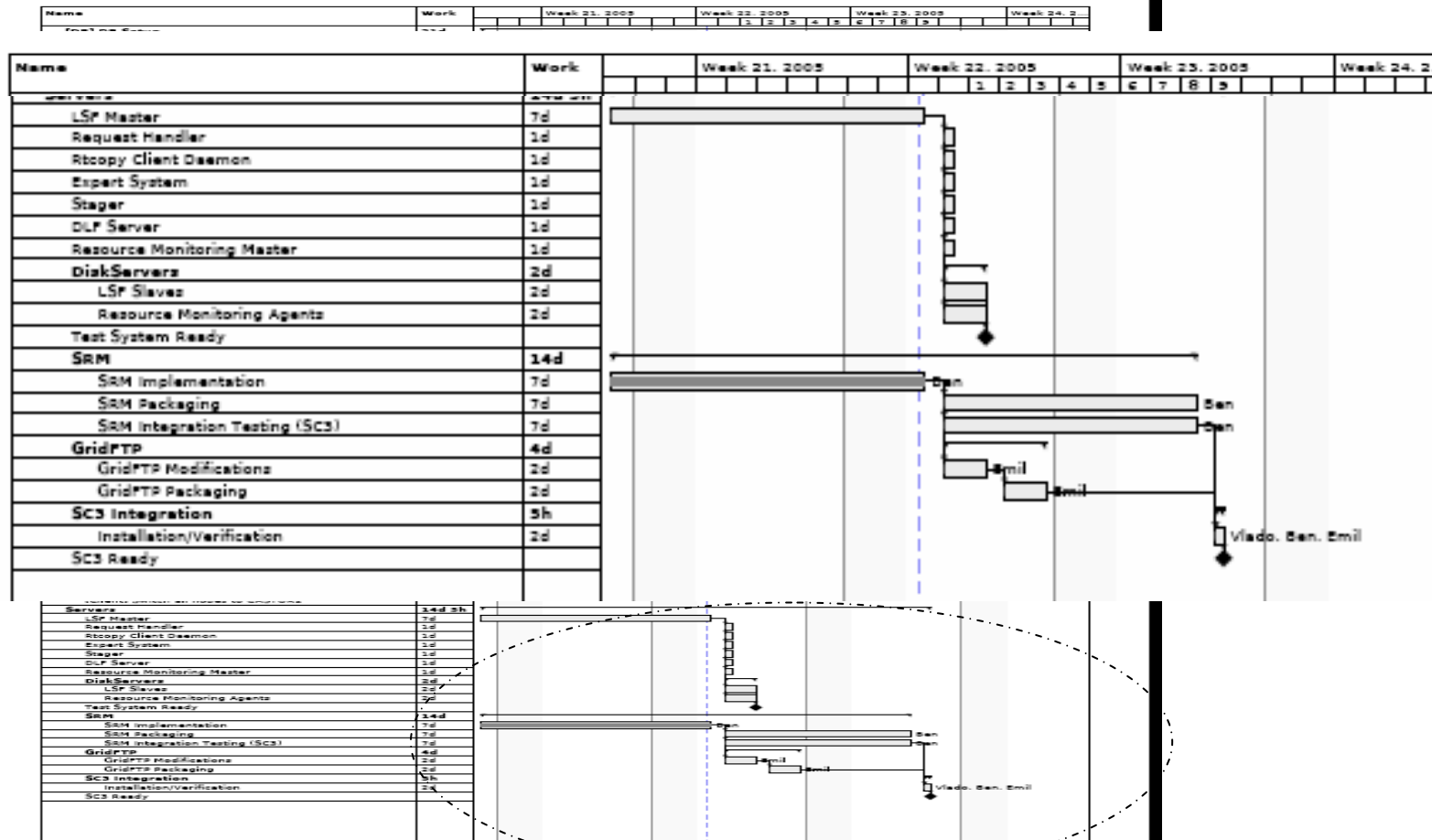
# Pre-production plan

<https://savannah.cern.ch/projects/castordeployment>

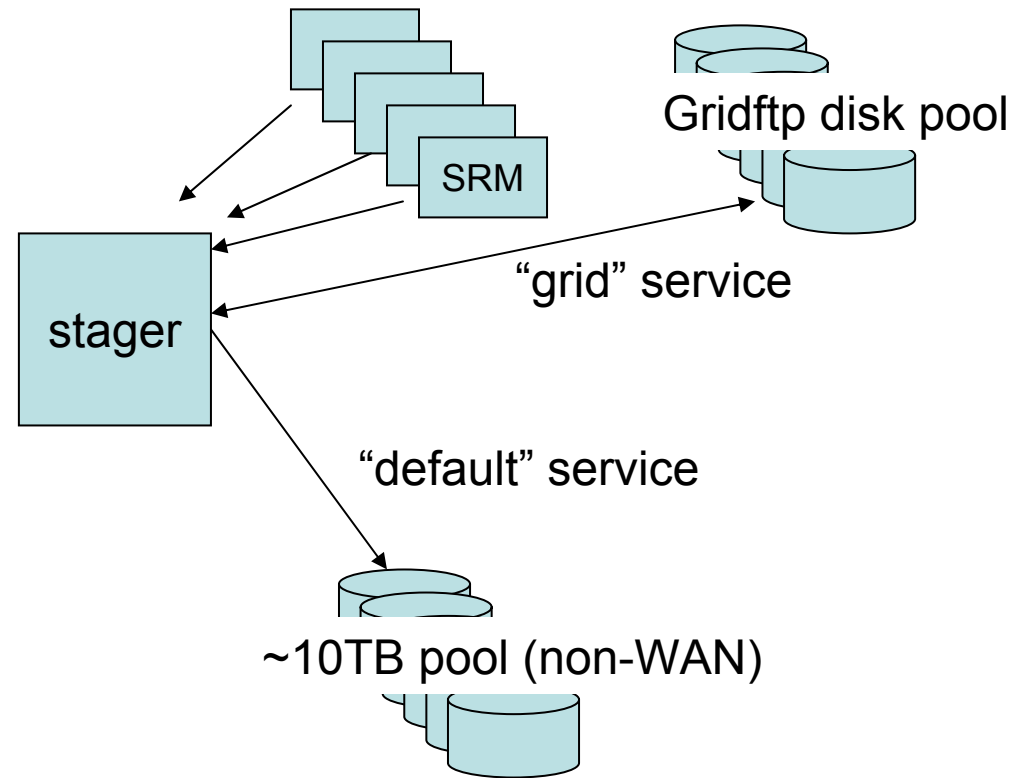


# Pre-production plan

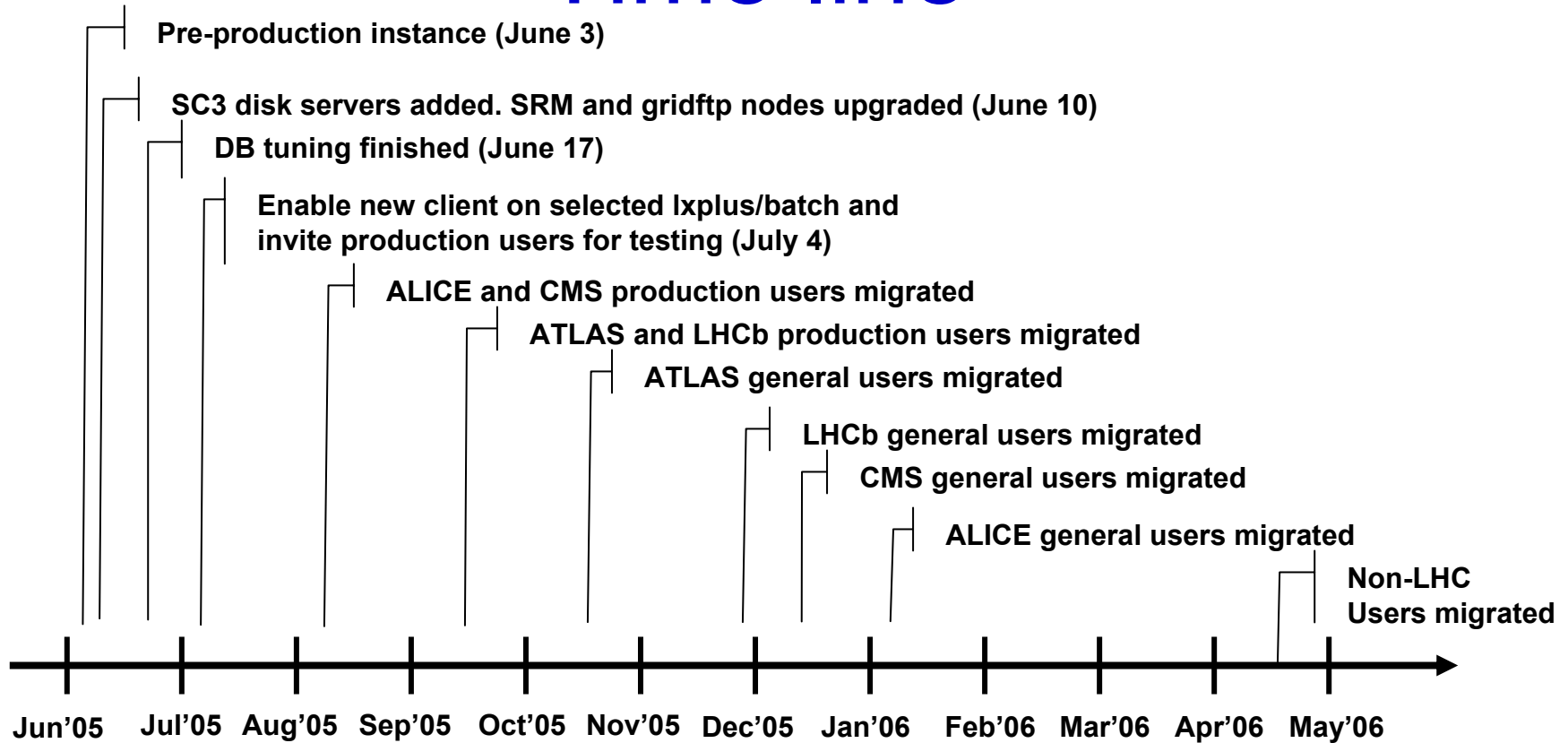
<https://savannah.cern.ch/projects/castordeployment>



# Preproduction instance



# Time line



SC3  
 Throughput  
 phase

SC3  
 Service phase

SC4

# Risks

- Hardware delivery
  - 0.5PB disk arriving beginning of September are required if we want to guarantee parallel setups while the experiments are migrating to CASTOR2. Any delay would put this at a stake
  - Mid-range PCs also for September. Any delay is a risk if a split deployment turns out to be required
- Bottlenecks (LSF, Oracle)
  - Database may require SMP
  - LSF scales to few 100,000s jobs
- Availability of production users in time for migration
  - We rely on the relevant SC3 users to be around in July (for ALICE and CMS) and September (for ATLAS and LHCb)
- Time to open a file
  - Scheduler introduces a file-open latency of a few seconds
  - May be a stopper for some client applications (e.g. interactive analysis?)

# Summary

- Mid-June: A pre-production instance, with ~10TB of disk + SC3 disk servers
- LHC experiments' production groups are the first to be migrated
  - Top priority to applications used in the SC3 service
- LHC general user groups migrated by beginning of 2006
- All other users groups are migrated by 2Q06
- Strong authentication deployed afterwards (precise plan to be worked out in early 2006)