CASTOR status

Presentation to LCG PEB
09/11/2004
Olof Bärring, CERN-IT

Outline

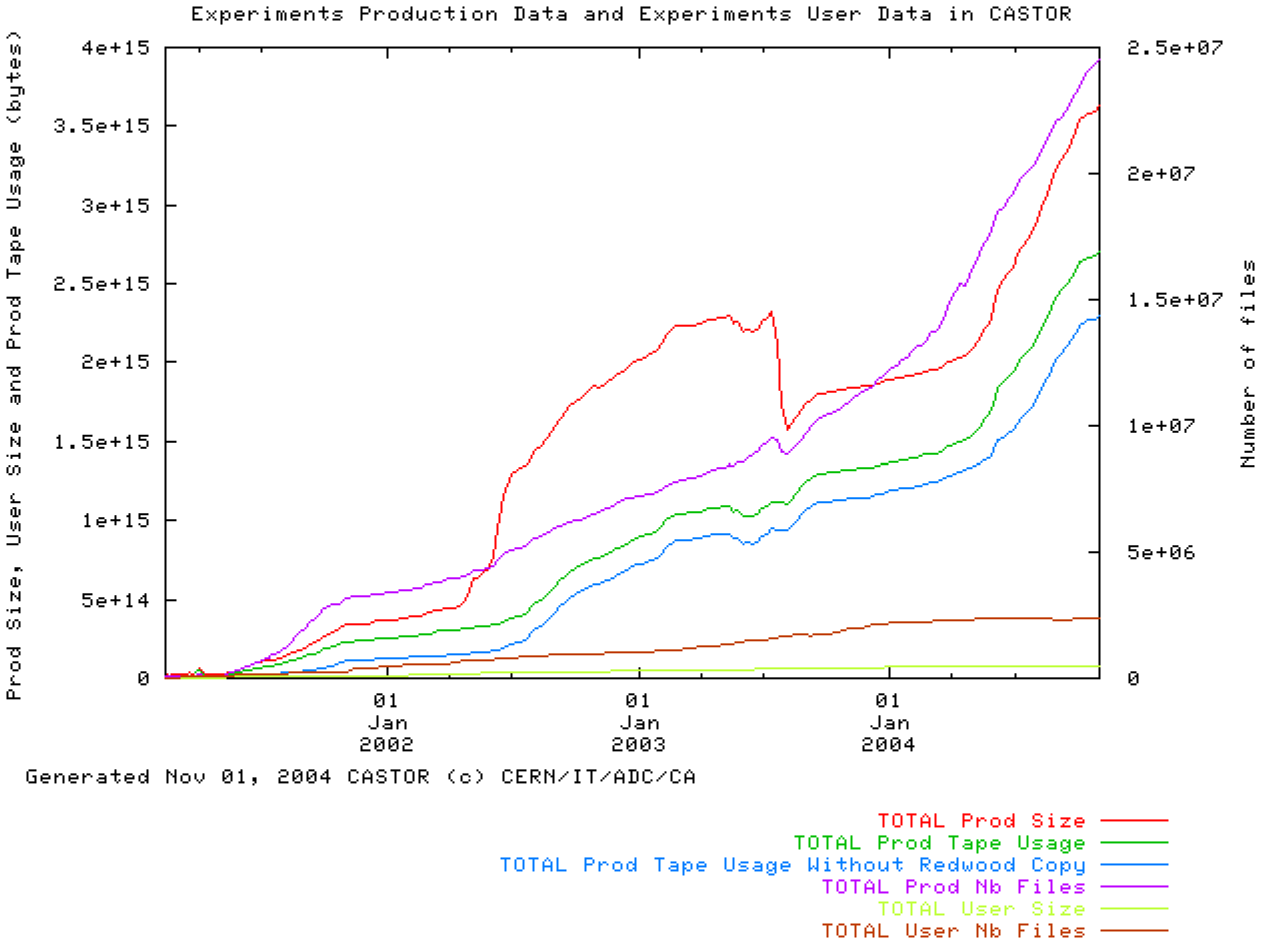
- CASTOR status
- New stager
  - Original plan
  - Delays
  - ALICE MDC-VI prototype
  - Current development status
- Conclusions

Status

CASTOR status

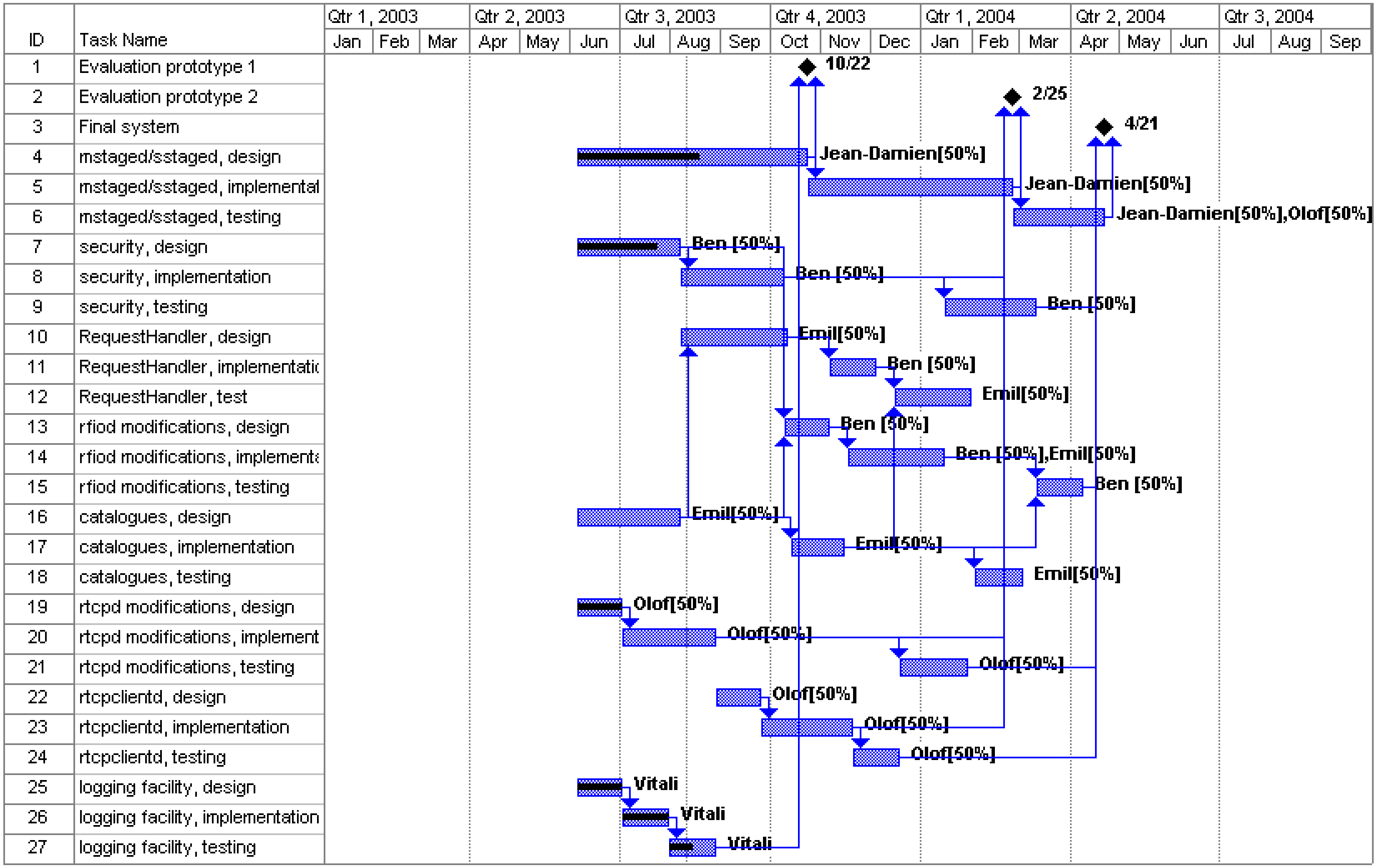
- Usage at CERN
  - ~3.4 PB data
  - ~26 million files
- Operation
  - Repack in production (since 2003): >1PB of data repacked
  - Tape segments checksum calculation and verification is in production since March 2004
  - Sysreq/TMS definitely gone in July
  - VDQM prioritize tape write over read → no drive dedication for CDR needed since September
  - During 2004 some experiments hit stager catalogue limitation (~200k files) beyond which the stager response can be very slow
- Support at CERN
  - 2nd and 3rd level separation works fine
  - Increasing support for SRM and gridftp users
- Other sites
  - PIC and IHEP contribute to CASTOR development at CERN → liberate efforts for better CASTOR operational support to other sites
  - CNAF may soon contribute(?)
  - RAL planning to evaluate CASTOR

CASTOR@CERN evolution



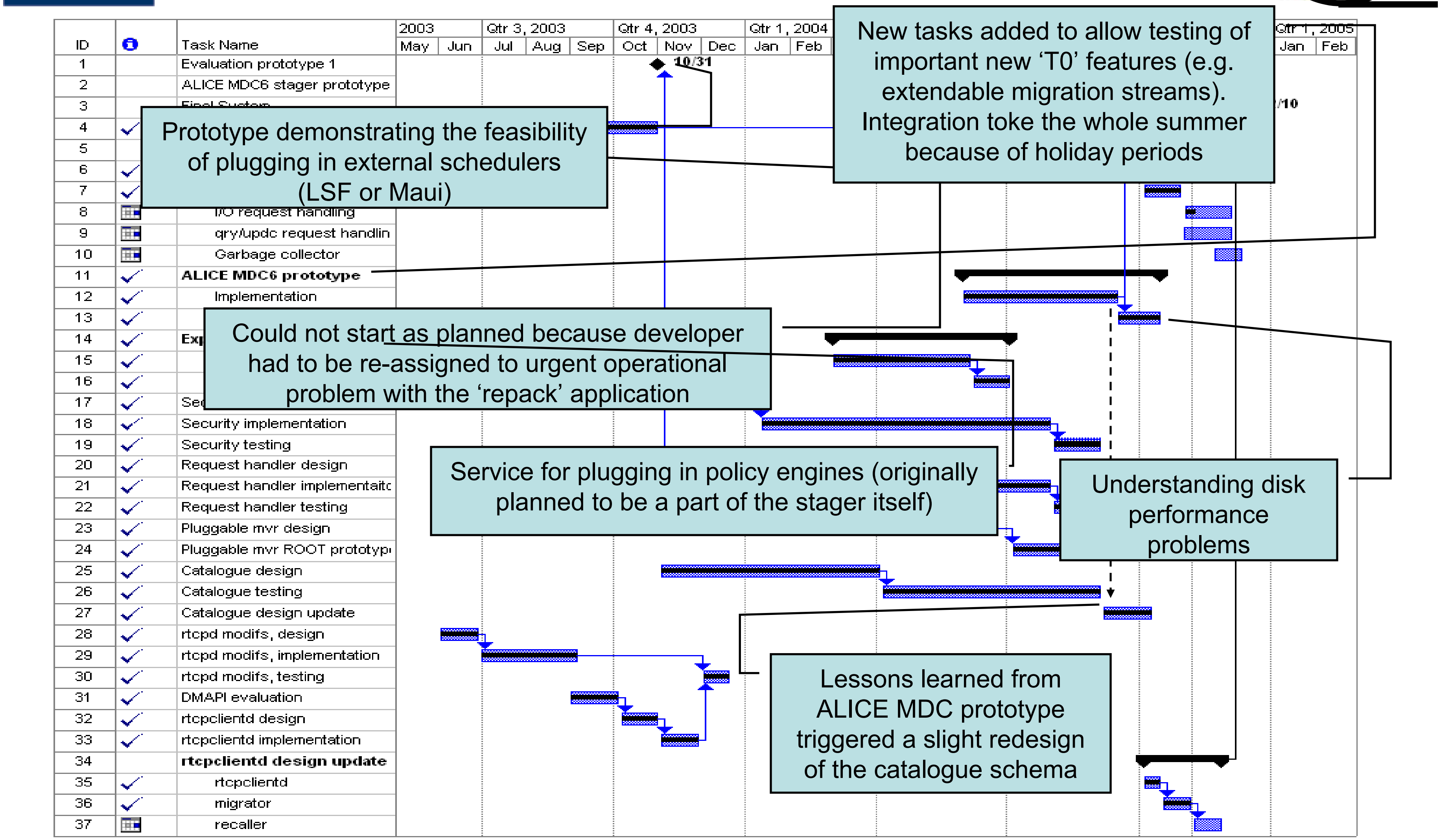
New stager, original plan

New stager developments
Original plan, PEB 12/8/2003



ID	Task Name
1	Evaluation prototype 1
2	Evaluation prototype 2
3	Final system
4	mstaged/sstaged, design
5	mstaged/sstaged, implementat
6	mstaged/sstaged, testing
7	security, design
8	security, implementation
9	security, testing
10	RequestHandler, design
11	RequestHandler, implementatic
12	RequestHandler, test
13	rfiod modifications, design
14	rfiod modifications, implementa
15	rfiod modifications, testing
16	catalogues, design
17	catalogues, implementation
18	catalogues, testing
19	rtcpd modifications, design
20	rtcpd modifications, implement
21	rtcpd modifications, testing
22	rtcpclientd, design
23	rtcpclientd, implementation
24	rtcpclientd, testing
25	logging facility, design
26	logging facility, implementation
27	logging facility, testing

New stager developments actual task workflows



ID	●	Task Name
1		Evaluation prototype 1
2		ALICE MDC6 stager prototype
3		Final System
4	✓	
5		
6	✓	
7	✓	
8	▦	I/O request handling
9	▦	qry/updc request handlin
10	▦	Garbage collector
11	✓	**ALICE MDC6 prototype**
12	✓	Implementation
13	✓	
14	✓	**Exp**
15	✓	
16	✓	
17	✓	Sec
18	✓	Security implementation
19	✓	Security testing
20	✓	Request handler design
21	✓	Request handler implementaitc
22	✓	Request handler testing
23	✓	Pluggable mvr design
24	✓	Pluggable mvr ROOT prototyp
25	✓	Catalogue design
26	✓	Catalogue testing
27	✓	Catalogue design update
28	✓	rtcpd modifs, design
29	✓	rtcpd modifs, implementation
30	✓	rtcpd modifs, testing
31	✓	DMAPI evaluation
32	✓	rtcpclientd design
33	✓	rtcpclientd implementation
34		**rtcpclientd design update**
35	✓	rtcpclientd
36	✓	migrator
37	▦	recaller

New tasks added to allow testing of important new ‘T0’ features (e.g. extendable migration streams). Integration toke the whole summer because of holiday periods

Prototype demonstrating the feasibility of plugging in external schedulers (LSF or Maui)

Could not start as planned because developer had to be re-assigned to urgent operational problem with the ‘repack’ application

Service for plugging in policy engines (originally planned to be a part of the stager itself)

Understanding disk performance problems

Lessons learned from ALICE MDC prototype triggered a slight redesign of the catalogue schema

New stager, delays

New stager developments delay
Main reason: The “repack problem”

- Repack: standard HSM utility to recover tape media:
  - ‘Holes’ created because of deleted files
  - Migration to higher capacity media
- A test version of the CASTOR repack utility was released in April 2003
  - Tested during summer for repacking CASTOR log files and other CASTOR operation files
  - Tests OK, started with some (mostly inactive) user files in September
- End November 2003: bug detected
  - Bug found in stager API during the certification of first production release of repack
  - The effect was that a fraction (~5%) of the repacked files got wrongly mapped in the CASTOR name server
- December 2003 – May 2004
  - One CASTOR developer working full time on finding and repairing incorrectly mapped CASTOR files
  - A bit less than 50,000 files wrongly mapped out of >1 million
  - Repair applied to the CASTOR name server the 26th of April 2004
  - Affected users (L3C) were informed about the problem

New stager developments delays Unplanned grid activities

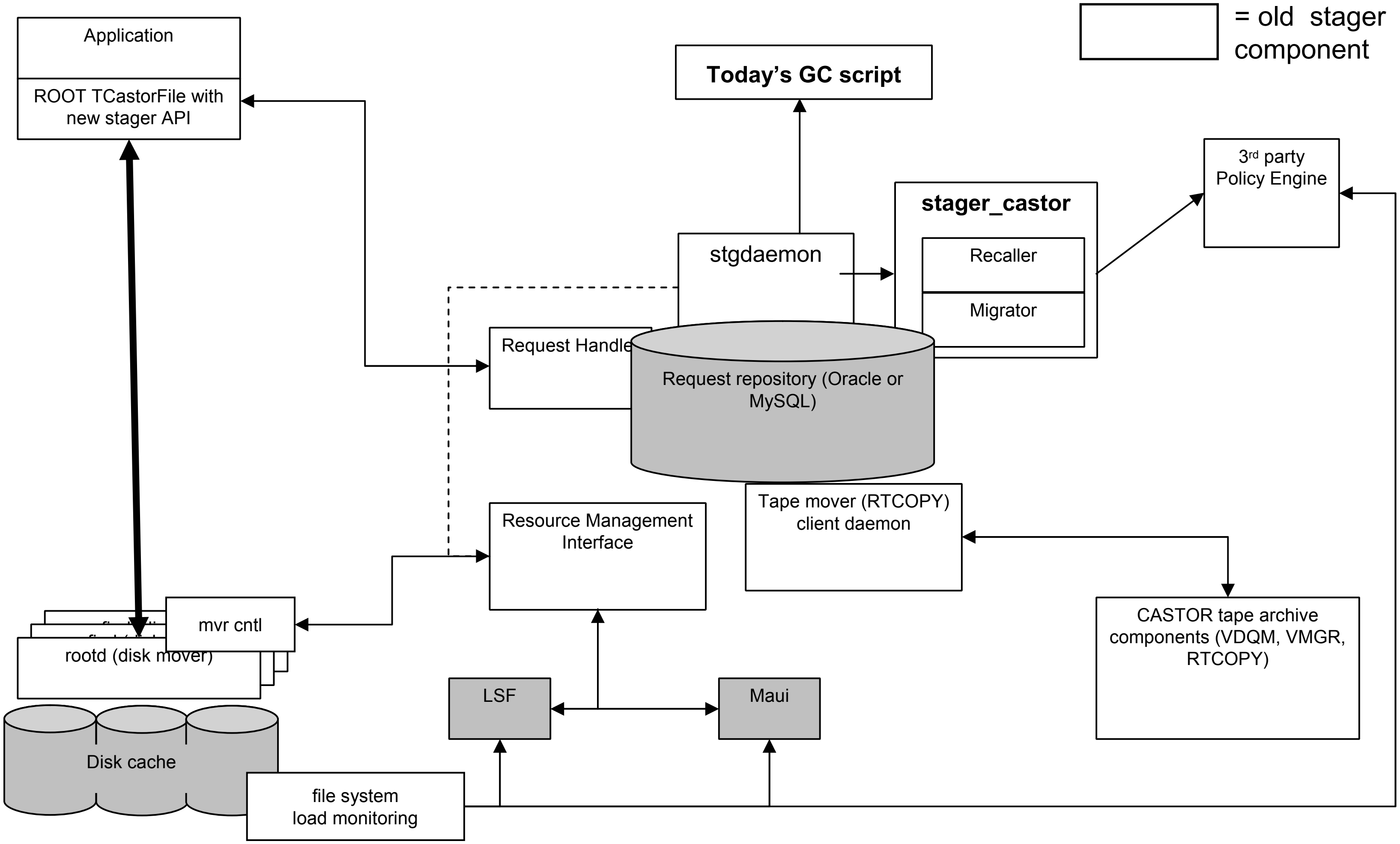
- SRM interoperability
  - Drilling down the GSI (non-)interoperability details
  - Holes in the SRM specs
  - Time-zone difference (FNAL-CERN) does not favor efficient debugging of interoperability problems
- Other grid activities: CASTOR as a disk pool manager without tape archive
  - We provided a packaged solution for LCG
  - But… support expectations pointed towards a development sidetrack
    - Castor is not well suited for such configurations
  - Decided to drop all support for CASTOR disk-only configurations and focus on the CERN T0/T1 requirements

New stager, ALICE MDC-VI prototype

New stager developments ALICE MDC-VI prototype

- Because of the delays there was a risk to miss the ALICE MDC-VI milestone
  - New stager design addresses important Tier-0 issues:
    - Dynamically extensible migration streams
    - Just-in-time migration candidate selection based on file system load
    - Scheduling and throttling of incoming streams
  - ALICE MDC-VI the ideal test environment. Could not afford to miss it…
    - The features were ready but the central framework did not exist
    - Decided to build a hybrid stager re-using a slimmed-down version of the current stgdaemon as central framework

New stager developments
ALICE MDC-VI prototype

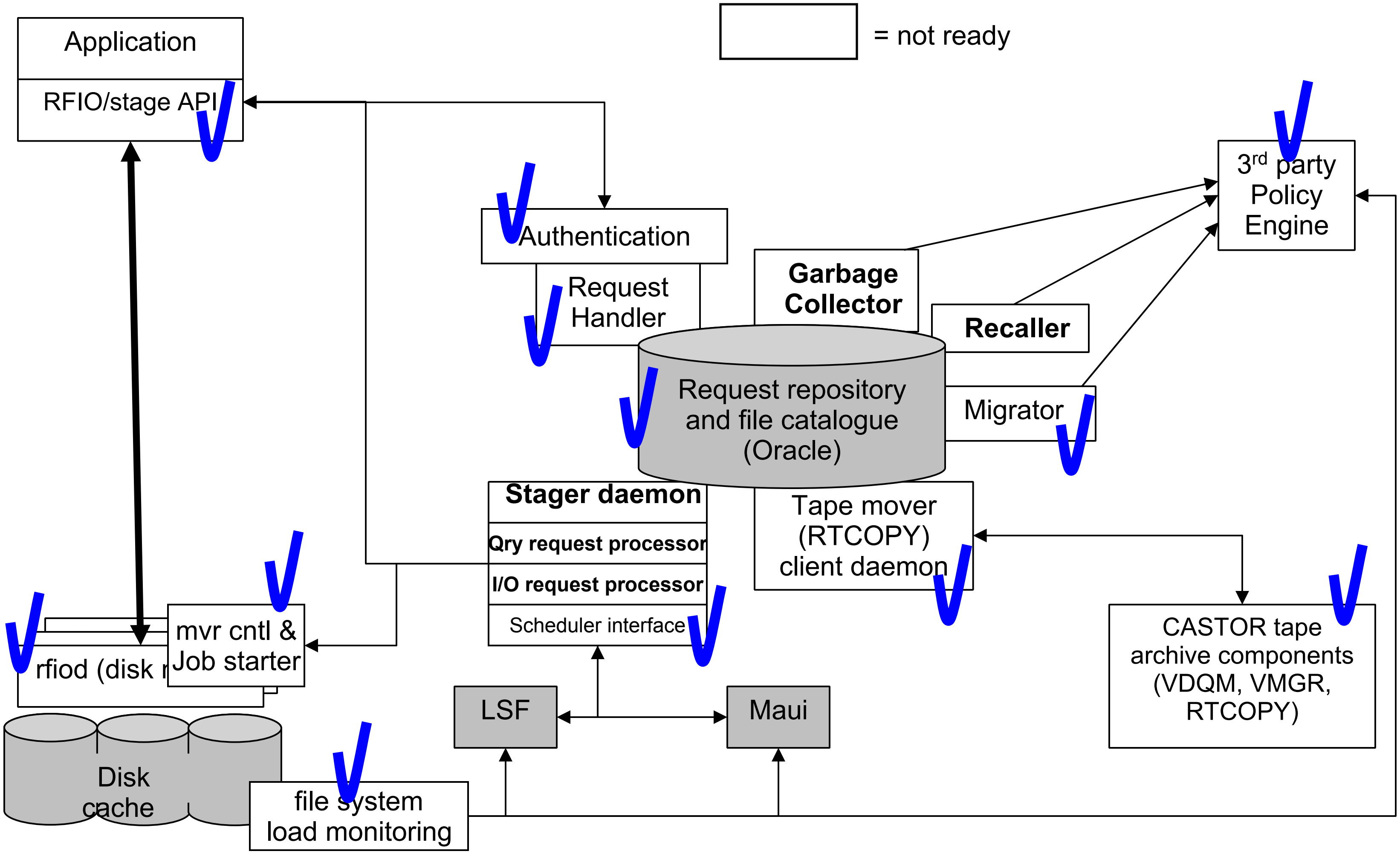

= old stager component
Application
ROOT TCastorFile with new stager API
Today's GC script
3rd party Policy Engine
stager_castor
Recaller
Migrator
stgdaemon
Request Handler
Request repository (Oracle or MySQL)
Tape mover (RTCOPY) client daemon
Resource Management Interface
CASTOR tape archive components (VDQM, VMGR, RTCOPY)
mvr cntl
rootd (disk mover)
LSF
Maui
Disk cache
file system load monitoring

New stager developments
Testing ALICE MDC-VI prototype

- The prototype was very useful:
  - Tuning of file-system selection policies
  - The designed assignment of migration candidates to migration streams was not efficient enough →redesign of catalogue schema
    - Migration candidates initially assigned to all tape streams
    - The migration candidate is ‘picked up’ by the first stream that is ready to process it
    - Slow streams (e.g. bad tape or drive) will not block anything
- Also found that the disk servers used for our tests were not well tuned for competition between incoming and outgoing streams

New stager, status

New stager developments
Current status

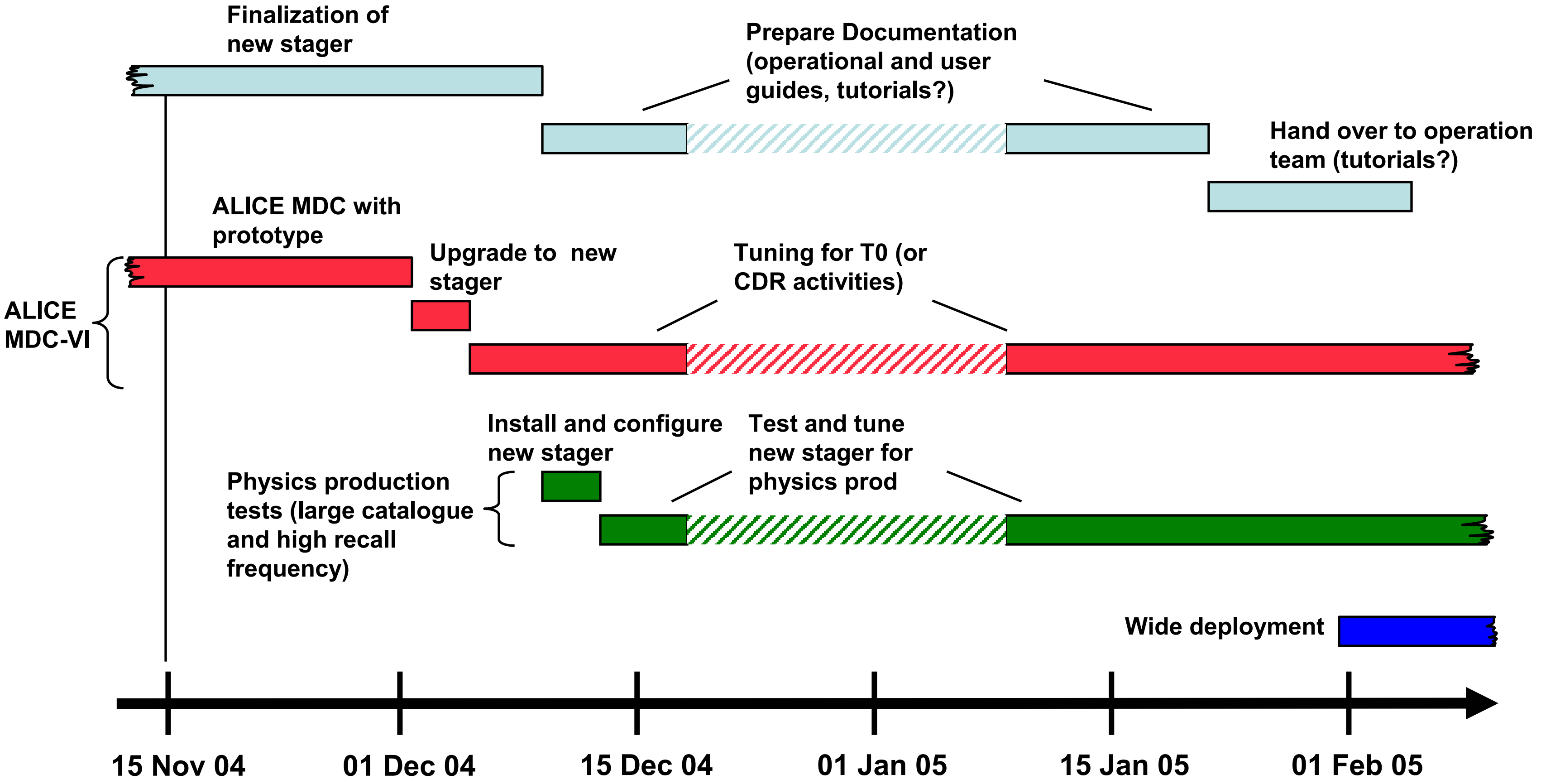


New stager developments Current status

- Catalogue schema and state diagrams are ready
  - Code automatically generated
  - Only ORACLE supported for the moment
  - http://cern.ch/castor/DOCUMENTATION/STAGE/NEW/Architecture/
- The finalization of the remaining components is now running at full speed
  - Central request processing framework (the replacement of stgdaemon):
    - New stager API defined and published for feedback (http://cern.ch/castor/DOCUMENTATION/CODE/STAGE/NewAPI/index.html)
    - I/O (stagein/stageout) and query processors: implementation started. Ready in 3-4 weeks
  - Recaller
    - Implementation started. Ready 1 – 2 weeks
  - Garbage collector
    - Implementation not started. Estimated duration ~2 weeks
- Hopefully we will be able to replace the ALICE MDC6 prototype by the final system in early December
- Would also need to test physics production type environment with large stager catalogue (millions of files) and tape recall frequency
  - Any Guinea-pigs?
  - ROOT clients using TCastorFile would need a new version of that class as well as libshift.so
  - ROOT clients using TRFIOFile would only need to upgrade libshift.so

New stager developments

Deployment plan from the developers' perspective



Finalization of new stager

Prepare Documentation (operational and user guides, tutorials?)

Hand over to operation team (tutorials?)

ALICE MDC-VI

ALICE MDC with prototype

Upgrade to new stager

Tuning for T0 (or CDR activities)

Physics production tests (large catalogue and high recall frequency)

Install and configure new stager

Test and tune new stager for physics prod

Wide deployment

New stager developments Deployment (cont)

- Security issues
  - All CASTOR services are technically prepared for strong authentication
    - http://cern.ch/castor/DOCUMENTATION/CODE/SECURITY/CASTOR_Security_Implementation.pdf
    - Kerberos-4, 5 and GSI supported
  - CASTOR security plug-ins used by other projects (LCG, EGEE)
  - A number of deployment issues remain:
    - Kerberos-5 infrastructure not yet in place
    - Batch job clients must have appropriate credentials
    - No solution yet for windows clients
    - Management of CASTOR service keys
  - Propose to do first deployment without strong authentication and upgrade when all infrastructure issues are solved
- Packaging
  - New packaging model envisaged:
    - One RPM for each CASTOR client and server
      - rfio
      - Stage
      - Nameserver
      - VMGR
      - …
    - One RPM for libraries
    - One 'devel' RPM (include files, man-pages)
- It will be possible to import disk servers from current to the new stager without having to re-stage the files

Conclusions

- CASTOR production status is OK
  - Important new features in 2004:
    - Checksum calculation/verification in production
    - Tape mover with all necessary features needed by new stager is running in production since March
    - VDQM prioritization of tape write since September
  - But, for the first time some experiments have hit the limitations of the current stager
- New stager developments
  - Important delays mainly due to high priority investigation and cleanup of repack problem
  - Prototype hybrid stager developed for the ALICE MDC-VI
  - Implementation is being finalized in coming 3-4 weeks
  - Hopefully the ALICE MDC-VI prototype can be replaced by the final system in December
  - Would also need to perform realistic tests for physics production environment with large file residence catalogue and high tape recall frequency