CASTOR progress report

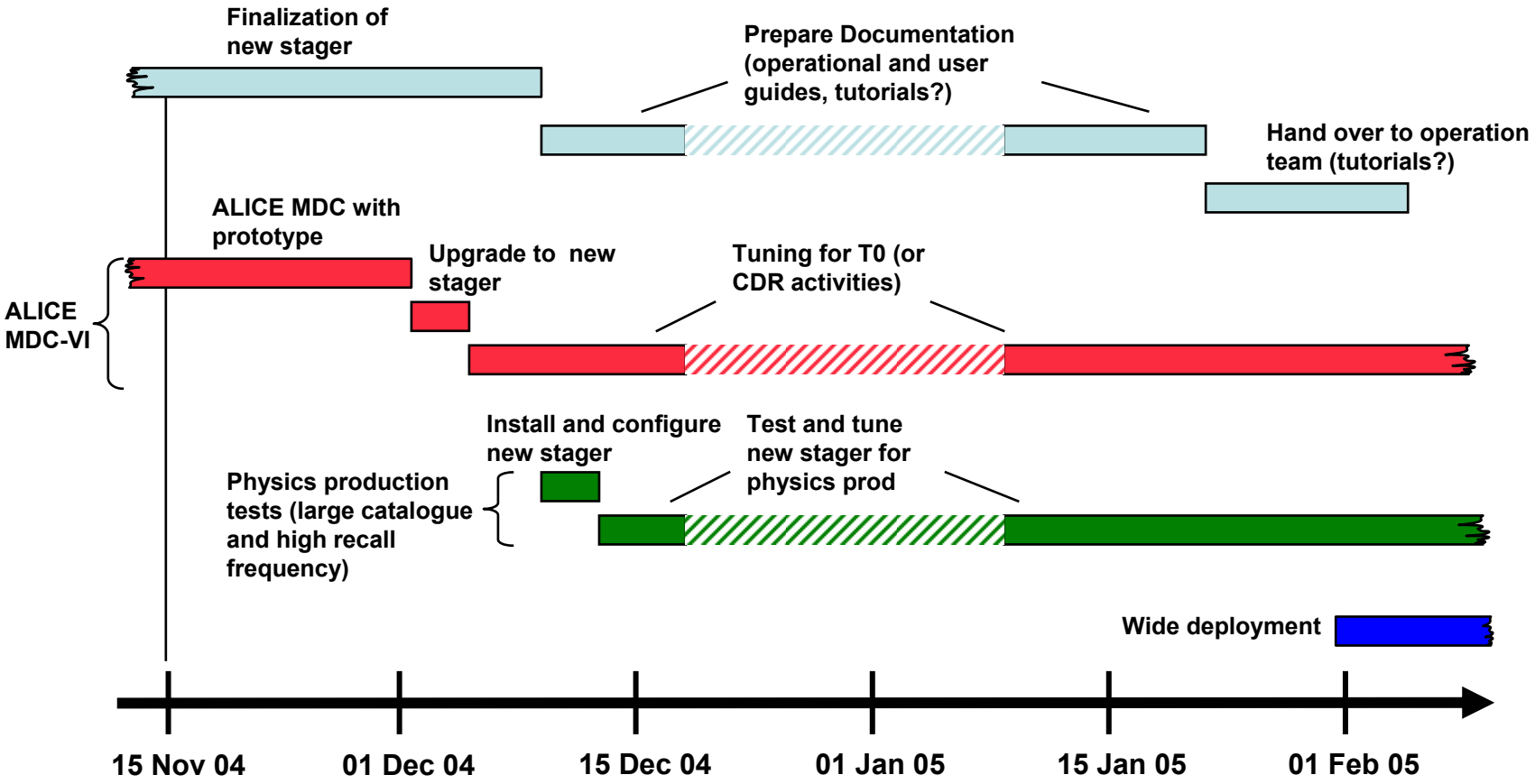
31/1/2005

Outline

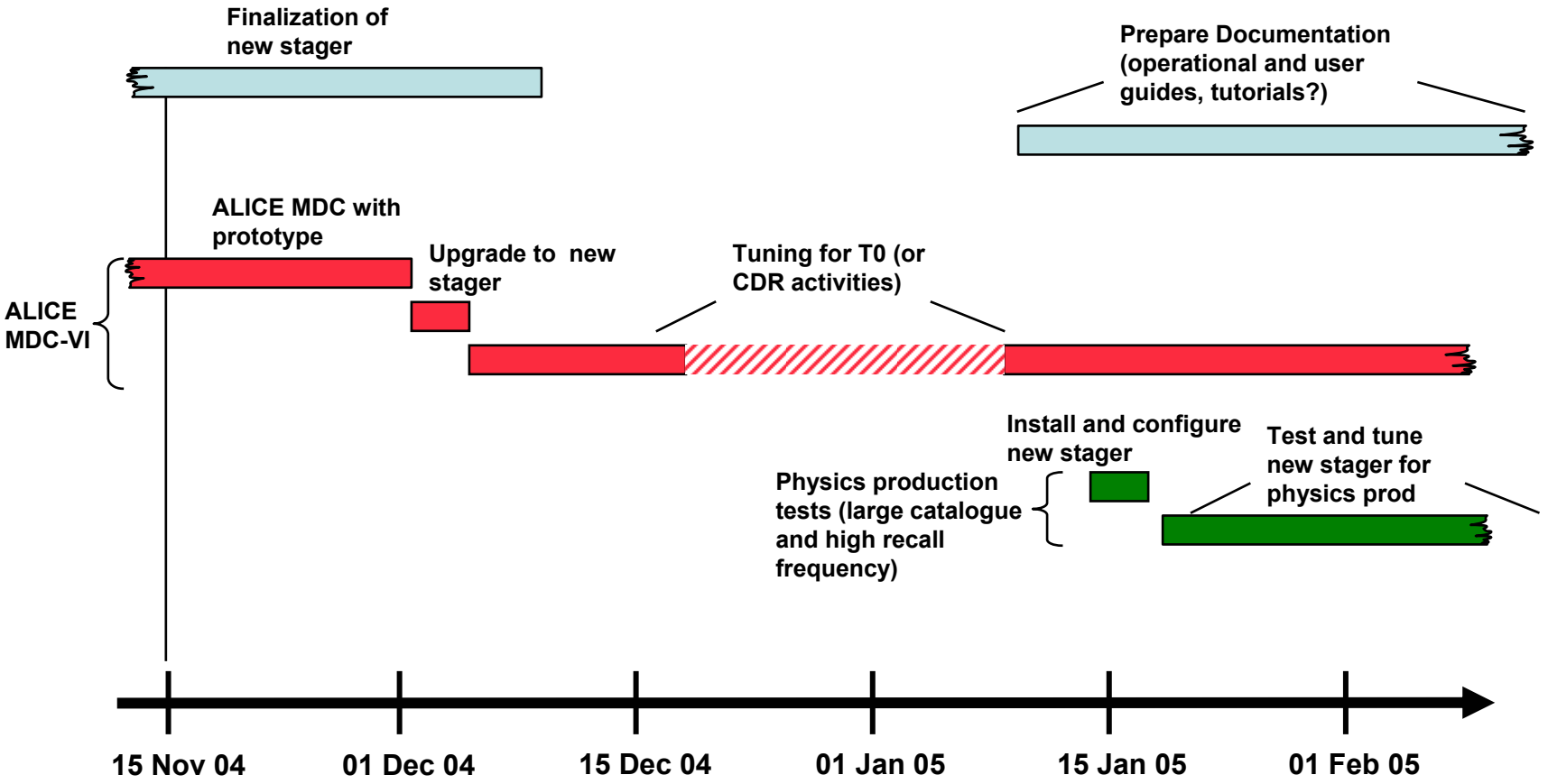
- Plan outlined at PEB 9/11/2004
- Progress
- Highlights
- Issues
- Remaining items for deployment
- Outlook

Plan outlined at PEB 9/11/2004

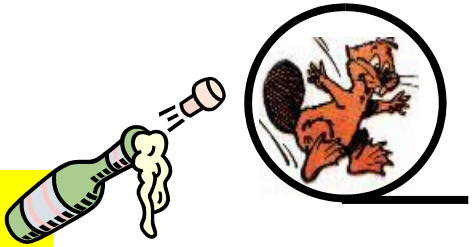
Deployment plan from the developers' perspective



Progress


Finalization of new stager
Prepare Documentation (operational and user guides, tutorials?)
ALICE MDC with prototype
Upgrade to new stager
Tuning for T0 (or CDR activities)
ALICE MDC-VI
Install and configure new stager
Test and tune new stager for physics prod
Physics production tests (large catalogue and high recall frequency)
15 Nov 04
01 Dec 04
15 Dec 04
01 Jan 05
15 Jan 05
01 Feb 05

Progress



The full software system was working for the first time the 8/12

Finalization of new stager

umentation (operational and user guides, tutorials?)

ALICE MDC with prototype

Upgrade to new stager

Tuning for T0 (or CDR activities)

ALICE MDC-VI

Install and configure new stager

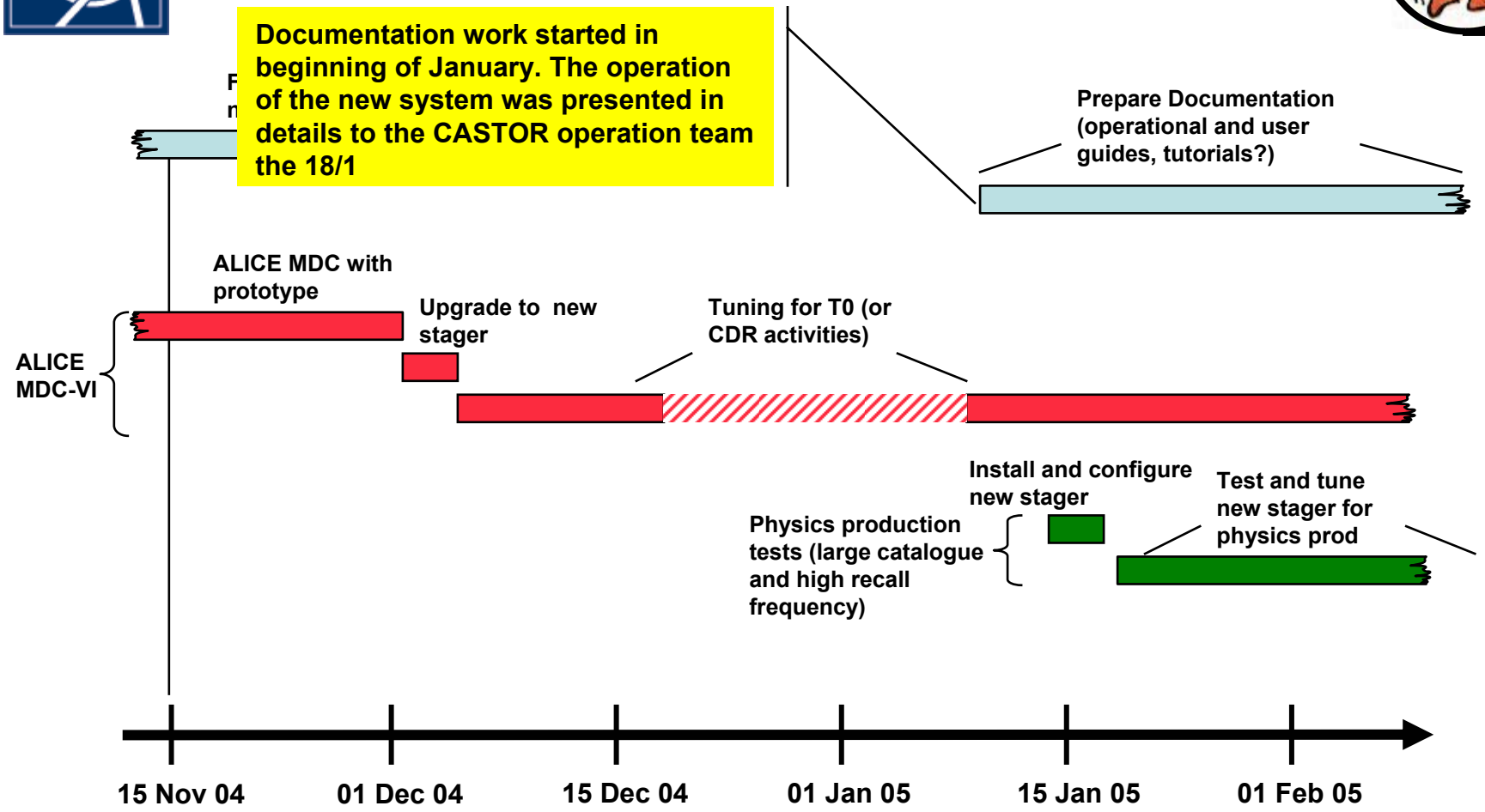
Test and tune new stager for physics prod

Physics production tests (large catalogue and high recall frequency)

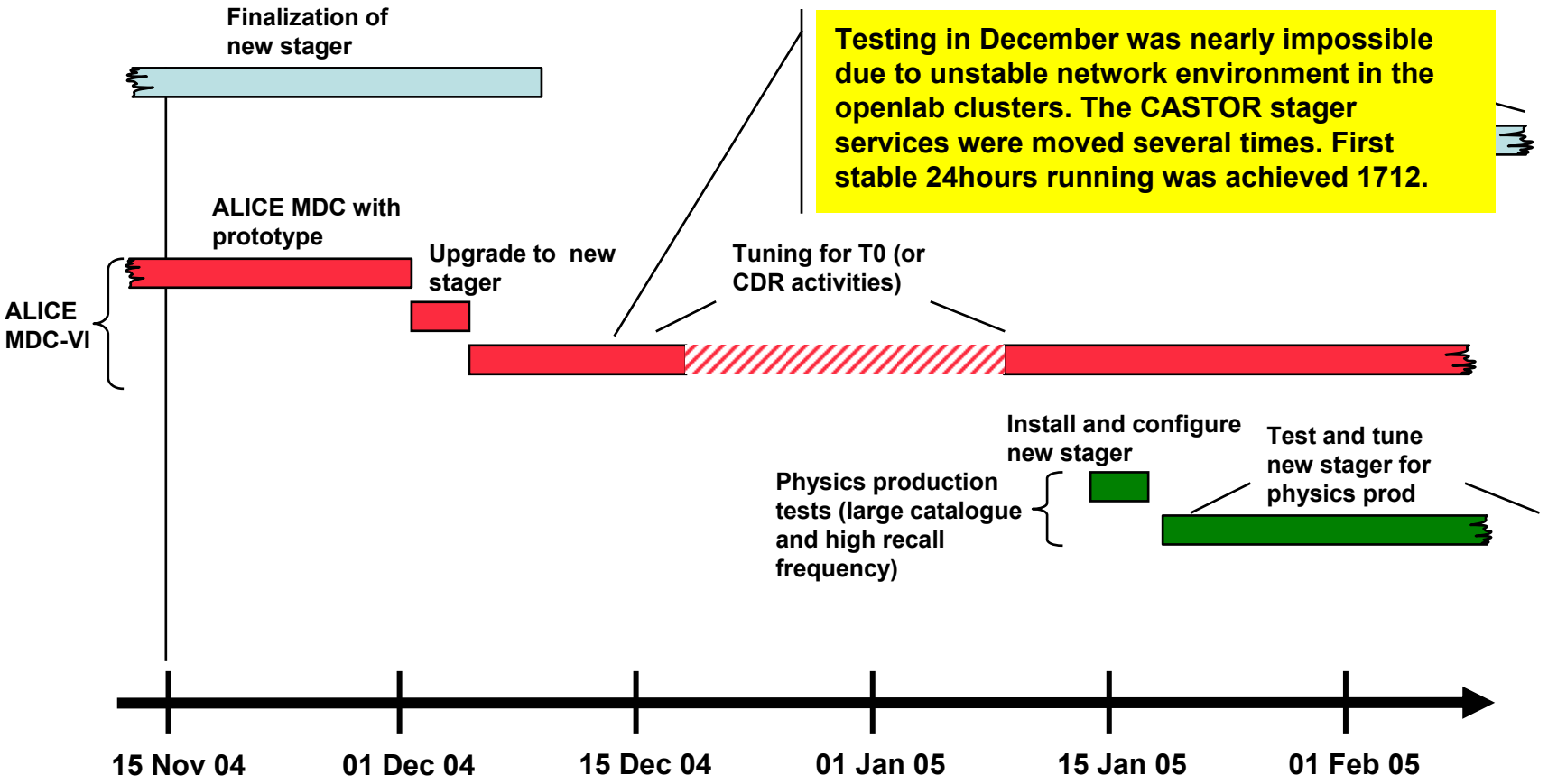
15 Nov 04 | 01 Dec 04 | 15 Dec 04 | 01 Jan 05 | 15 Jan 05 | 01 Feb 05

Progress

Documentation work started in beginning of January. The operation of the new system was presented in details to the CASTOR operation team the 18/1

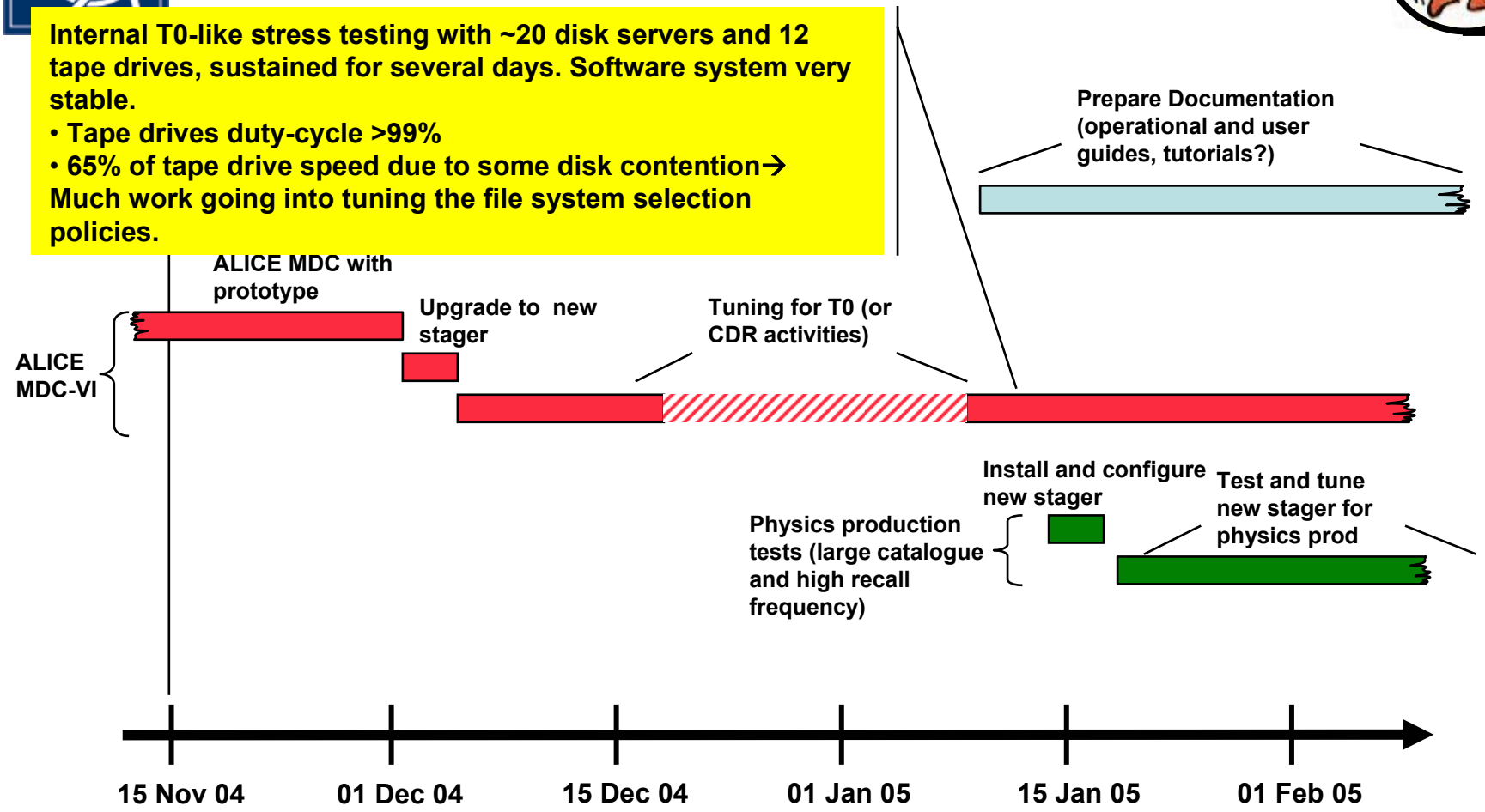


Progress

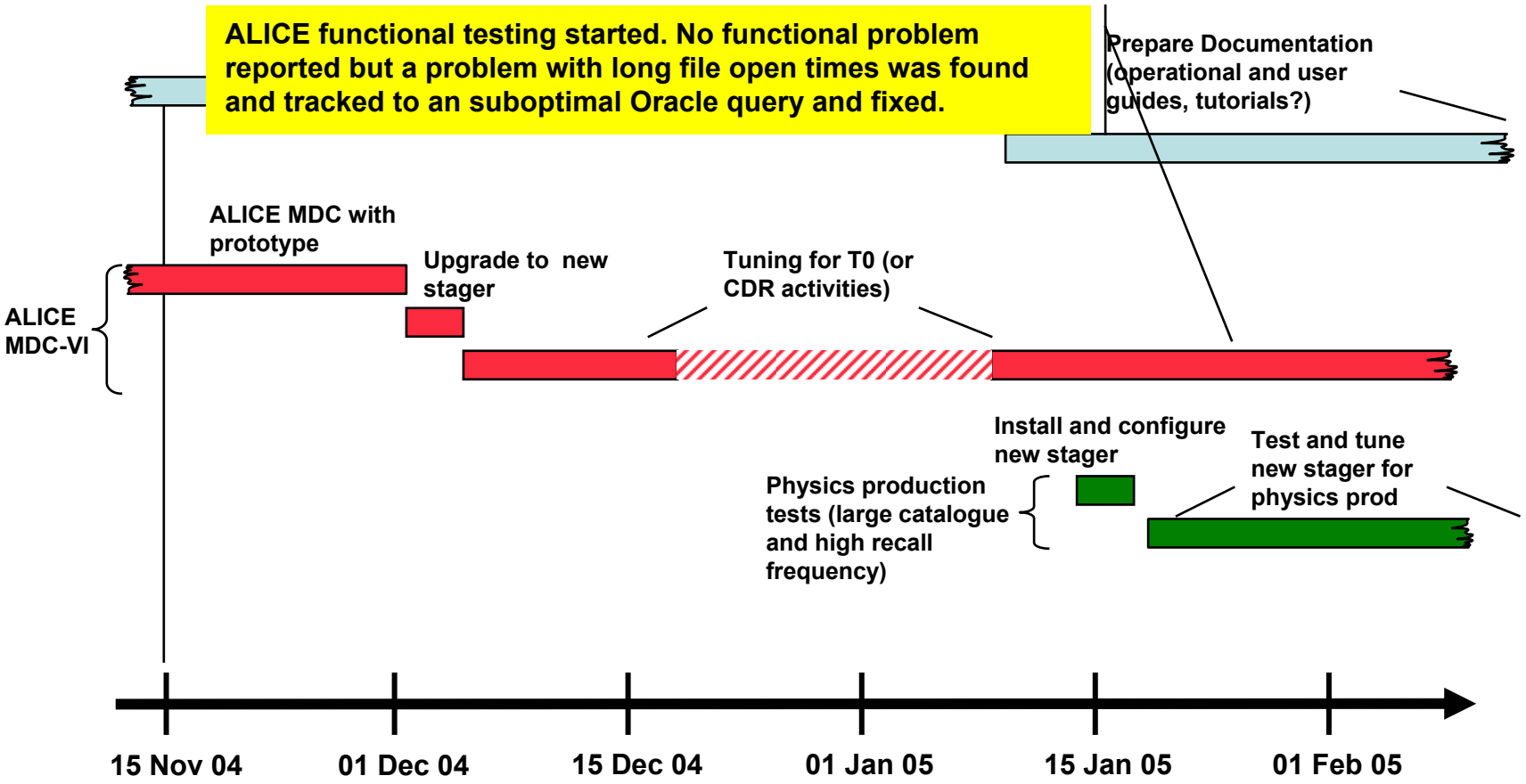

Finalization of new stager
ALICE MDC with prototype
Upgrade to new stager
ALICE MDC-VI
Tuning for T0 (or CDR activities)
Testing in December was nearly impossible due to unstable network environment in the openlab clusters. The CASTOR stager services were moved several times. First stable 24hours running was achieved 1712.
Install and configure new stager
Test and tune new stager for physics prod
Physics production tests (large catalogue and high recall frequency)
15 Nov 04
01 Dec 04
15 Dec 04
01 Jan 05
15 Jan 05
01 Feb 05

Progress

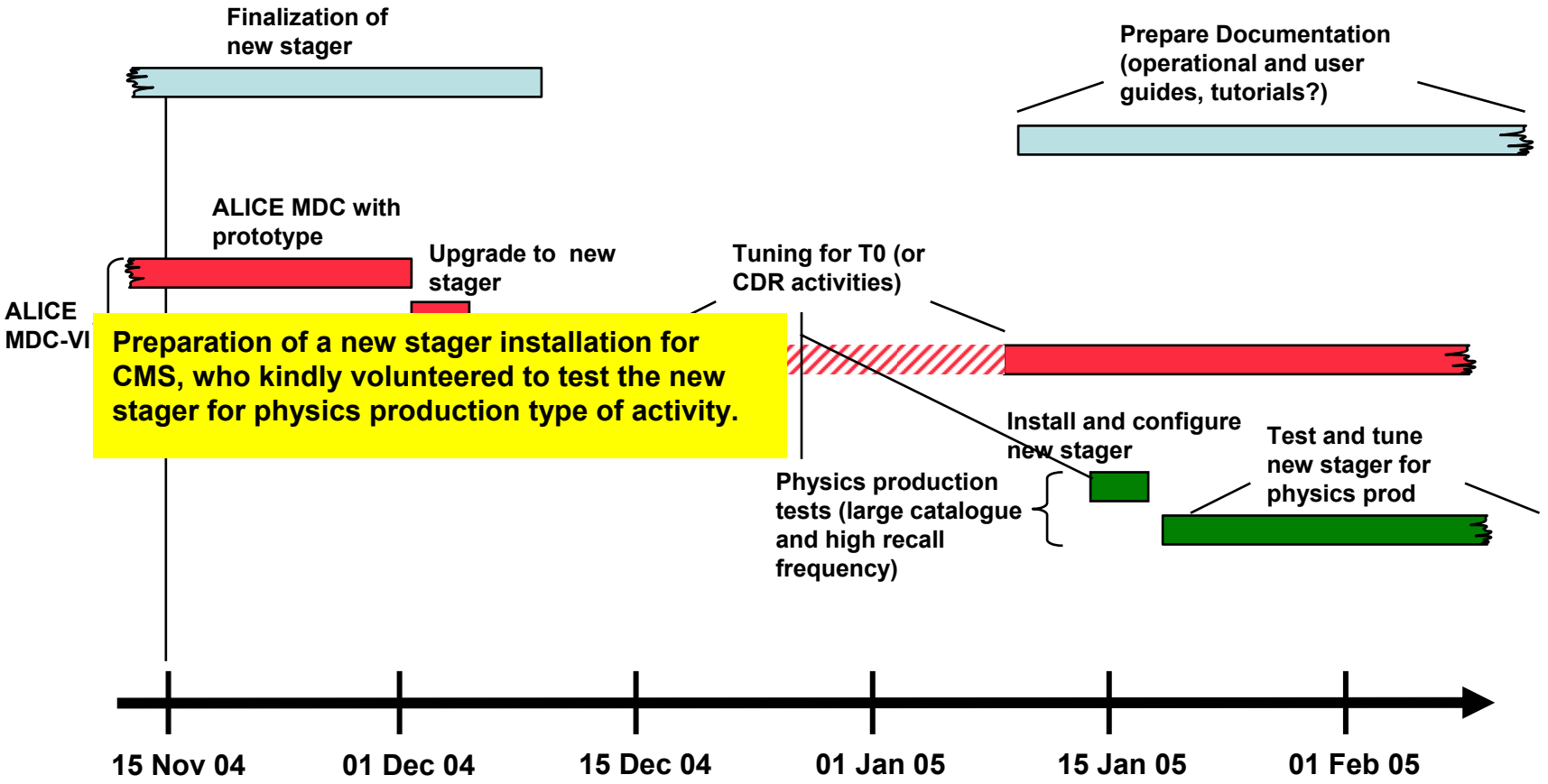
Internal T0-like stress testing with ~20 disk servers and 12 tape drives, sustained for several days. Software system very stable.
- **Tape drives duty-cycle >99%**
- **65% of tape drive speed due to some disk contention→ Much work going into tuning the file system selection policies.**



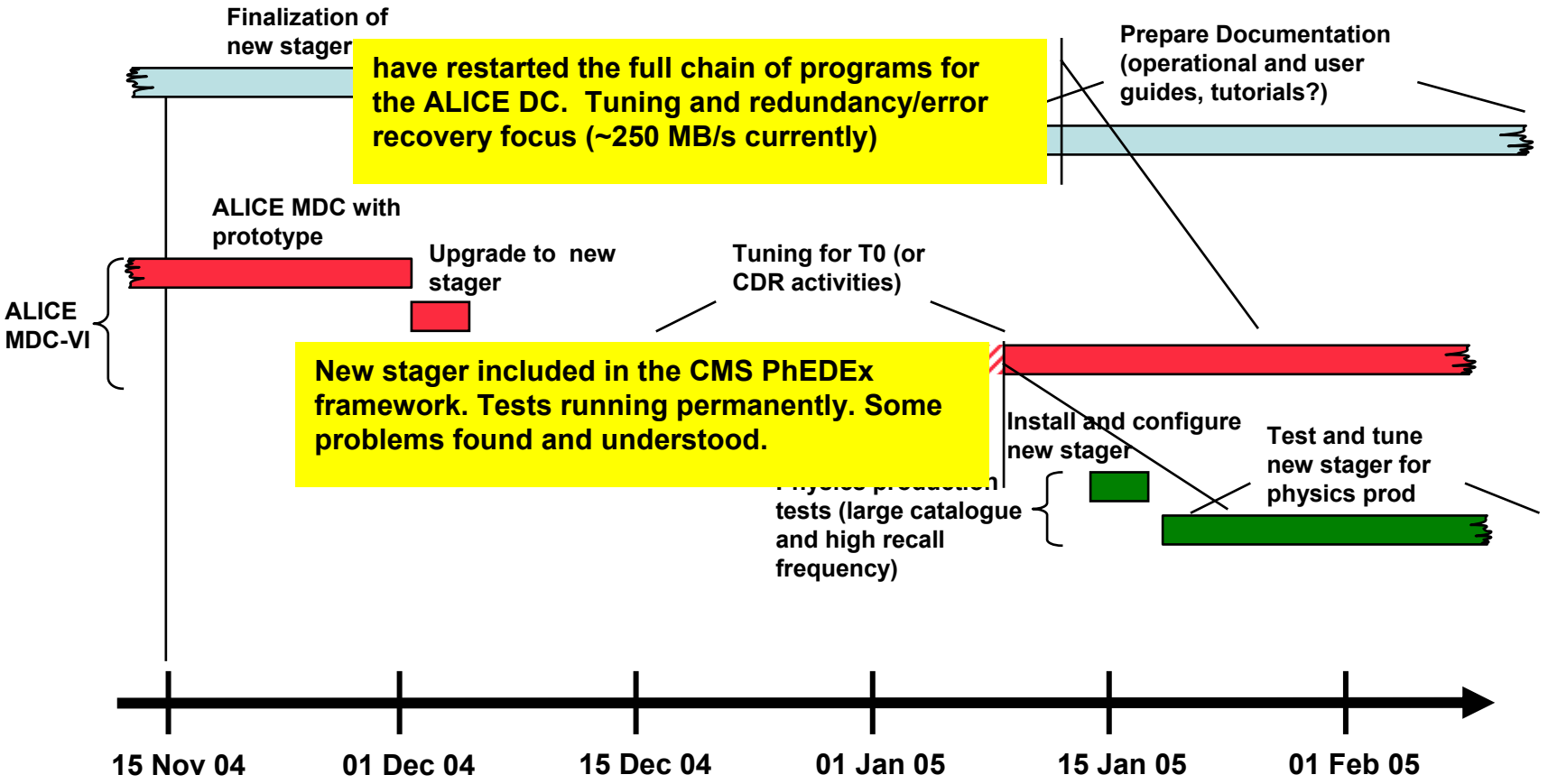
Progress


ALICE functional testing started. No functional problem reported but a problem with long file open times was found and tracked to an suboptimal Oracle query and fixed.
Prepare Documentation (operational and user guides, tutorials?)
ALICE MDC with prototype
Upgrade to new stager
Tuning for T0 (or CDR activities)
ALICE MDC-VI
Install and configure new stager
Test and tune new stager for physics prod
Physics production tests (large catalogue and high recall frequency)
15 Nov 04
01 Dec 04
15 Dec 04
01 Jan 05
15 Jan 05
01 Feb 05

Progress


Finalization of new stager
Prepare Documentation (operational and user guides, tutorials?)
ALICE MDC with prototype
Upgrade to new stager
Tuning for T0 (or CDR activities)
ALICE MDC-VI
Preparation of a new stager installation for CMS, who kindly volunteered to test the new stager for physics production type of activity.
Install and configure new stager
Test and tune new stager for physics prod
Physics production tests (large catalogue and high recall frequency)
15 Nov 04
01 Dec 04
15 Dec 04
01 Jan 05
15 Jan 05
01 Feb 05

Progress


Finalization of new stager
have restarted the full chain of programs for the ALICE DC. Tuning and redundancy/error recovery focus (~250 MB/s currently)
Prepare Documentation (operational and user guides, tutorials?)
ALICE MDC with prototype
Upgrade to new stager
Tuning for T0 (or CDR activities)
ALICE MDC-VI
New stager included in the CMS PhEDEx framework. Tests running permanently. Some problems found and understood.
Install and configure new stager
Test and tune new stager for physics prod
tests (large catalogue and high recall frequency)
15 Nov 04
01 Dec 04
15 Dec 04
01 Jan 05
15 Jan 05
01 Feb 05

Highlights

- The full system is functionally robust since more than a month
  - Some tuning problems remain (next slide)
- Both *rfiod* and *rootd* are inherently supported disk movers
  - Xrootd not supported yet because of lack of time. Ongoing discussion with xrootd devs.
- Dynamic migration and recall streams work very well
  - Drive duty-cycle >99%
  - No need for clients to order or group files for recall. This is automatically optimized by the stager (works well in the CMS tests)

Issues

- Tuning
  - To achieve 100% drive speed, the externalized filesystem selection policy require more tuning.
    - In particular the old generations of disk servers may completely kill migration performance if the same filesystem is concurrently hit by an incoming and migration stream
  - Long file open time has been observed by ALICE. The problem was tracked to a suboptimal Oracle query and fixed.
  - CMS observed that transfer performance is not constant. Bursts of good rates are followed by periods of low rates. Being investigated…
- LSF support not yet tested
  - So far the full system has been tested with the Maui scheduler due to lack of time.

Remaining items for deployment

- Garbage collection
  - Framework ready but default policy missing
  - Crude policy (remove everything that has been migrated) brings us through the ALICE MDC
  - ~ 1 person week of work
- Tape error recovery
  - Almost no tape errors are automatically retried directly by the migrator/recaller process
  - Instead another process (ErrorHunter) scans all failed tape requests and applies an external retry policy:
    - If retry → reset the tape request for migration/recall
    - If retry limit exhausted → report error to user (recall) and/or administrator (recall and migration)
  - About 1-2 person weeks to complete the ErrorHunter with a default policy

Outlook

- Deployment ready version in second half of February
  - Support for rfiod and rootd
  - Only tested with Maui scheduler
  - No SRM or Gridftp interface yet
  - ➢ SRM uses low-level stager RPM → requires an upgrade
    2-3 person weeks to port the current SRM version to the new CASTOR (LCG in March, define SRM version/functionality)
    ~2 person months to rewrite and deploy completely new version (SRM2), could reuse existing framework (→ Fermilab)
- Second release in ~ May
  - Support for xrootd
  - LSF scheduler
  - more tuning and policies
- Currently 2-3 month delay of the 2005 milestones (compared to ~9 month in 2004) and very good progress during the last month