Repack in CASTOR

Sébastien Ponce
CERN IT DSS

CASTOR Face to face meeting - 28-30 Nov 2012

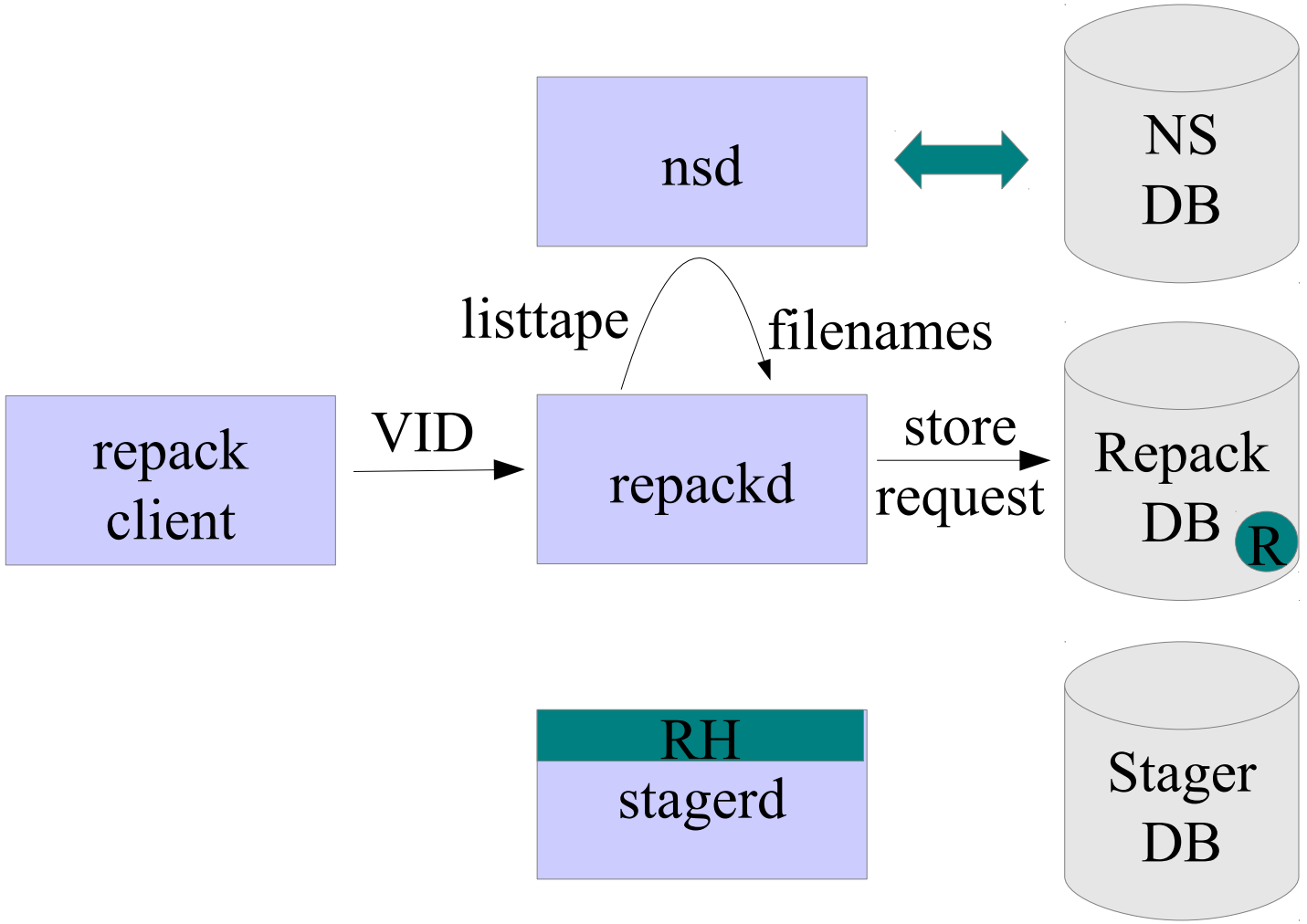
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Overall program

- Old repack (pre 2.1.12)
  - Schetch of a repack session
  - Issues
- New Repack (post 2.1.12)
  - Schetch of a repack session
  - Improvements
  - New tools
  - Gory details
  - DB schema
  - Performance numbers

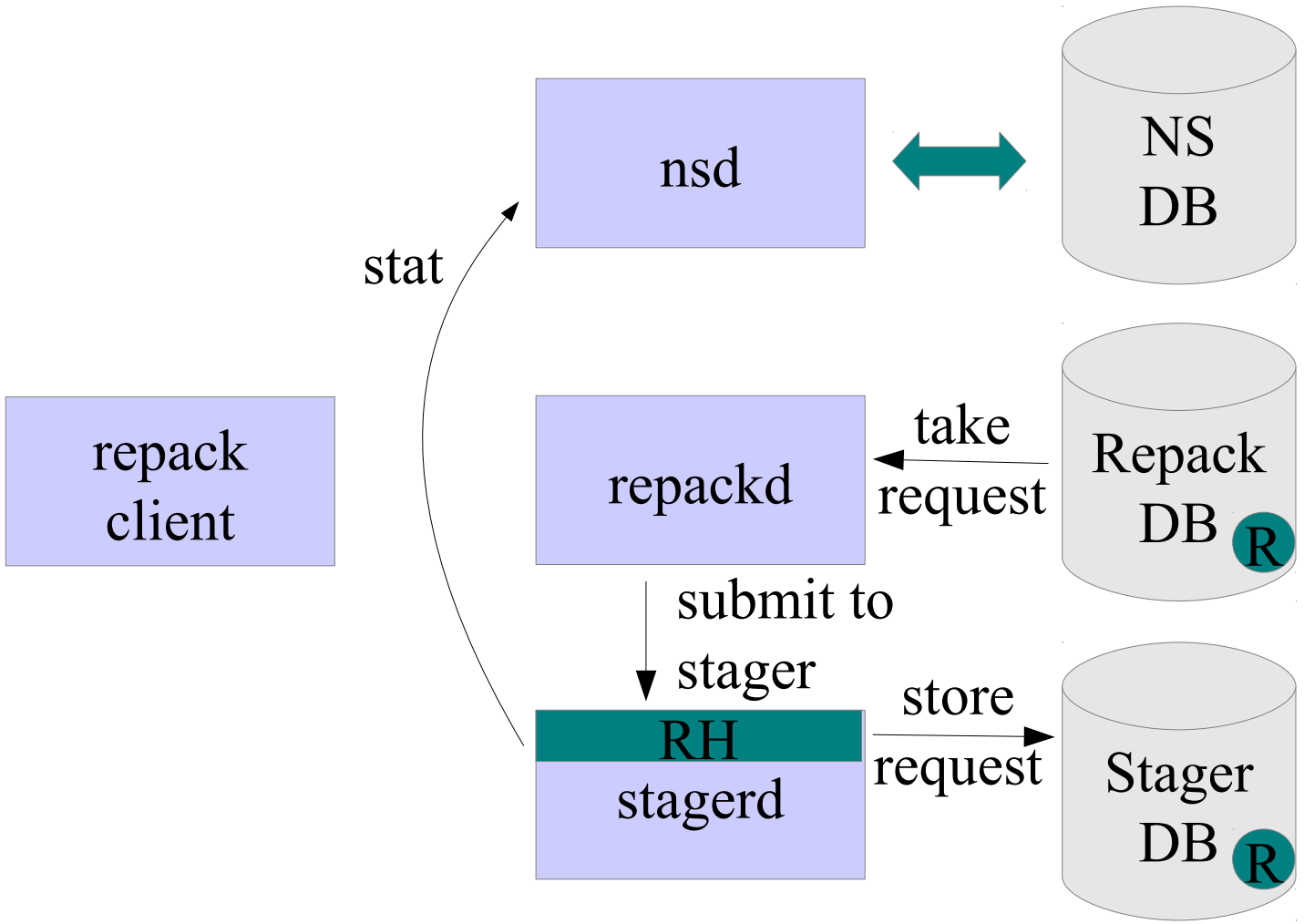
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack - start (1)



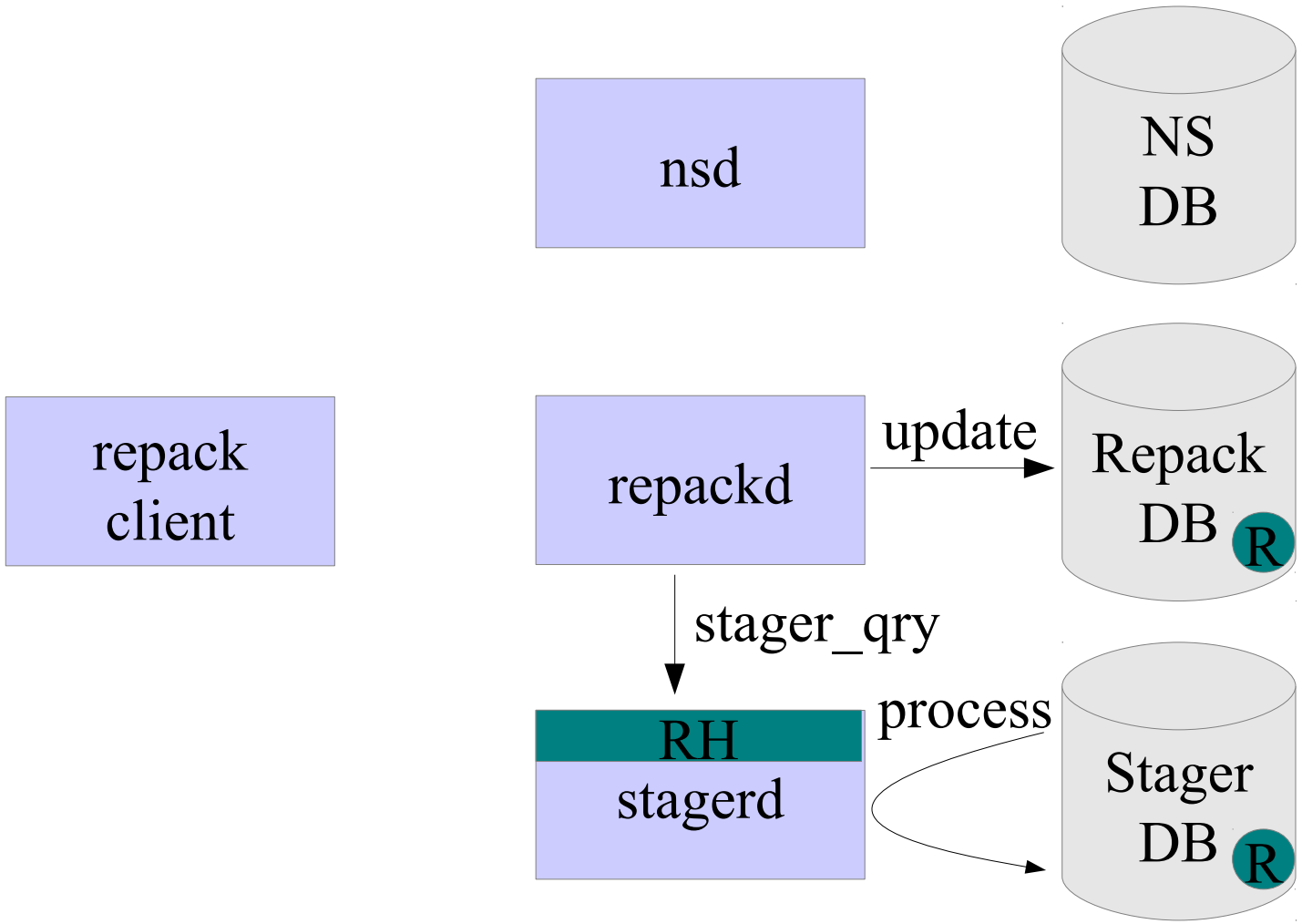
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack – start (2)



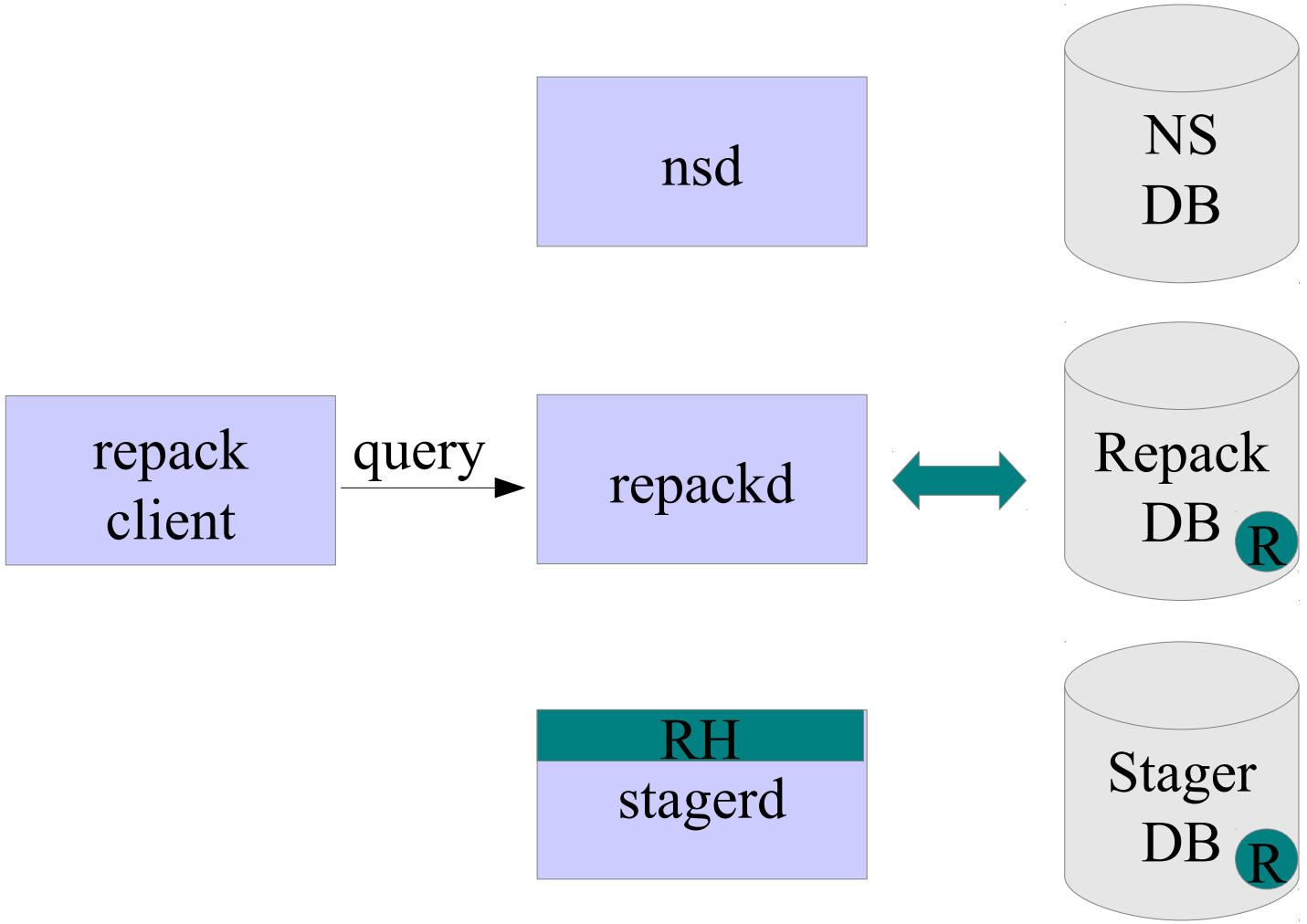
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack – stable processing



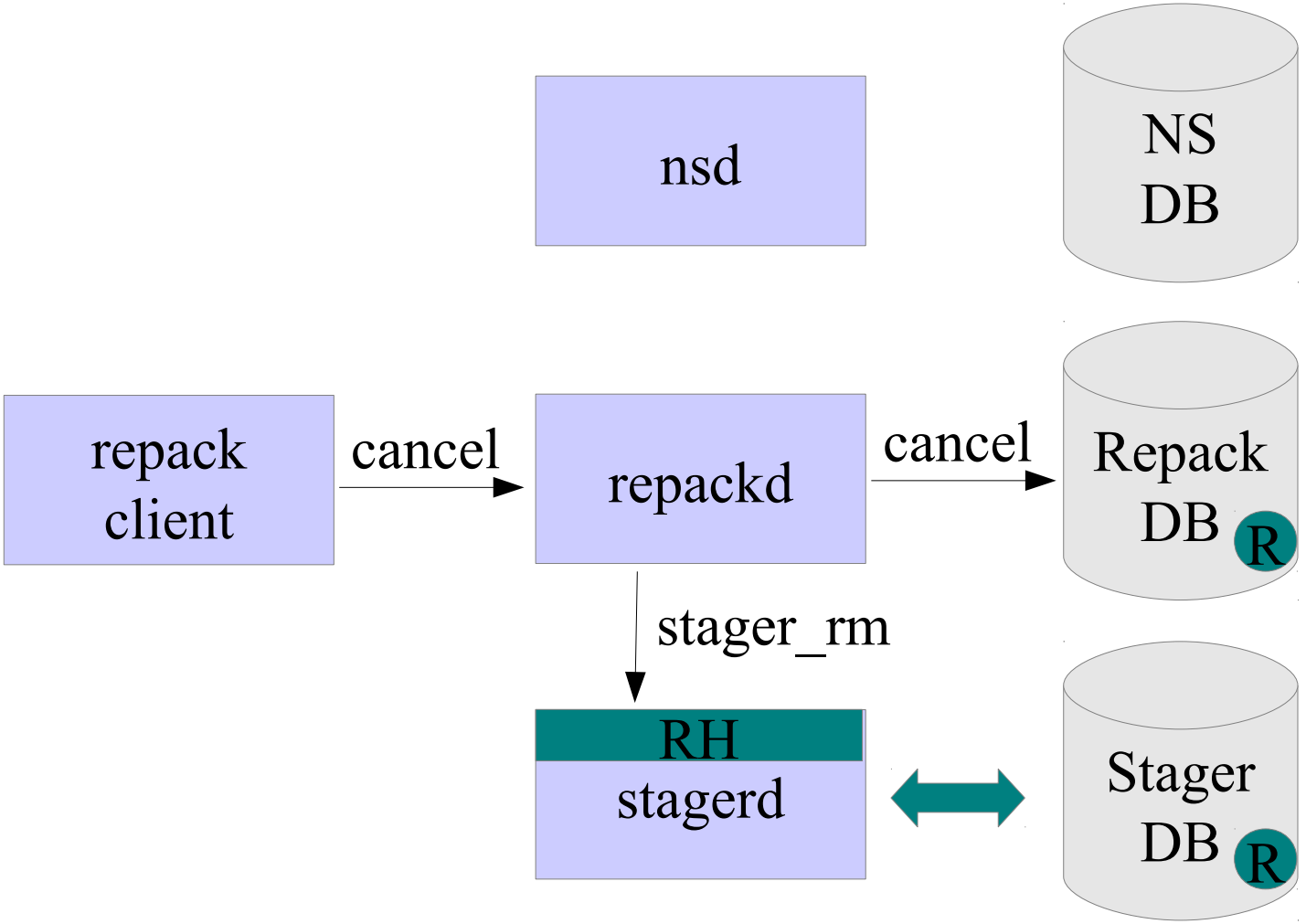
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack – query status



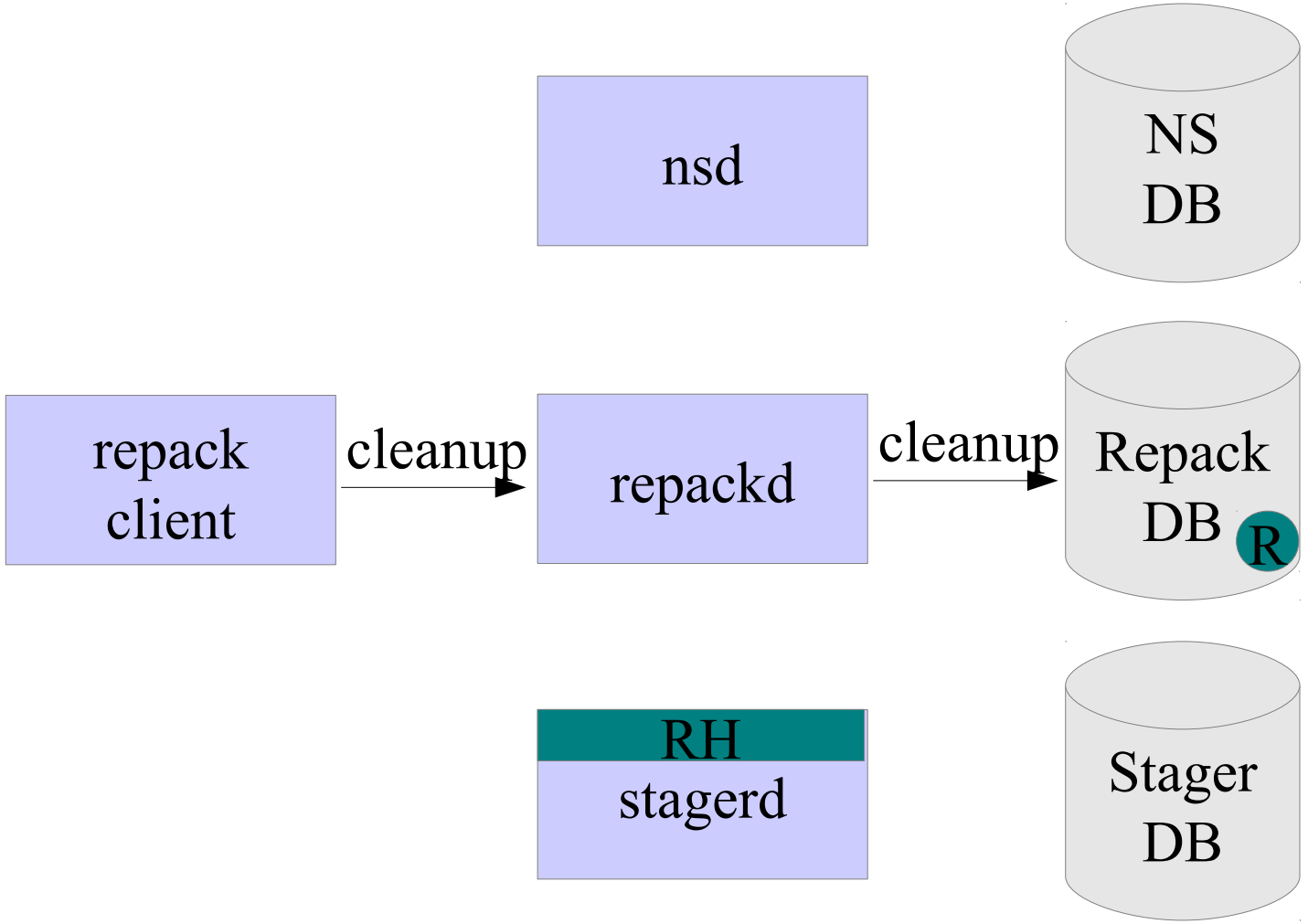
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack – cancelling



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Old repack – cleaning up

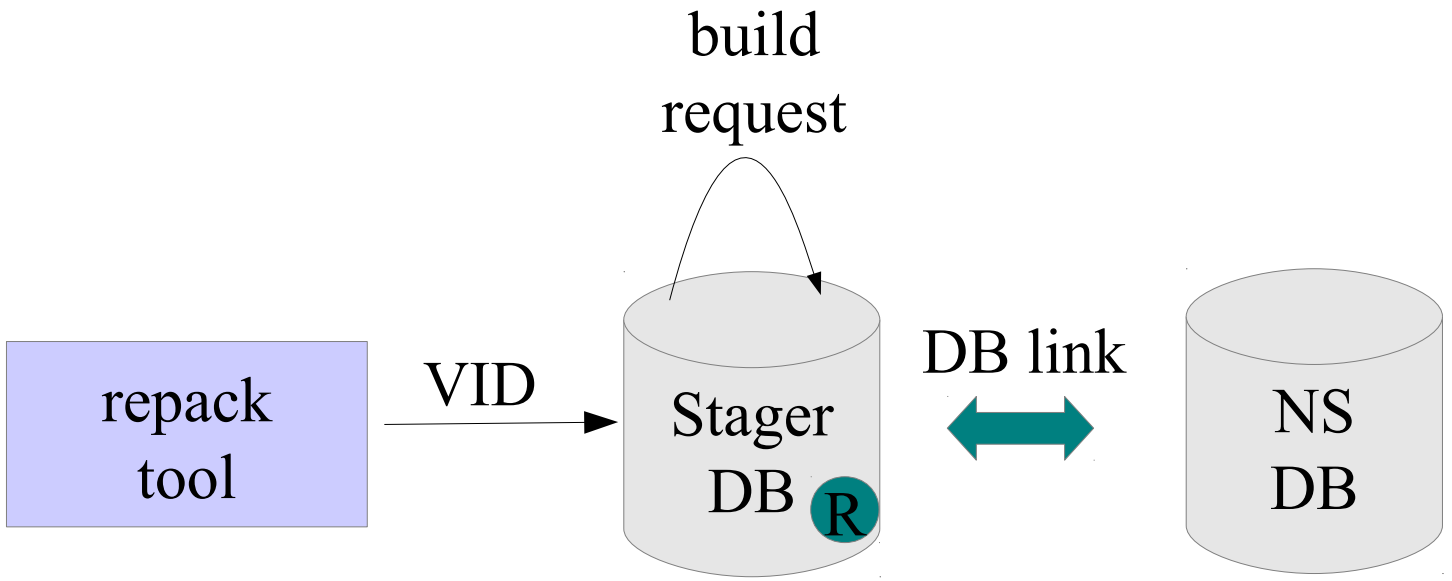


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Issues

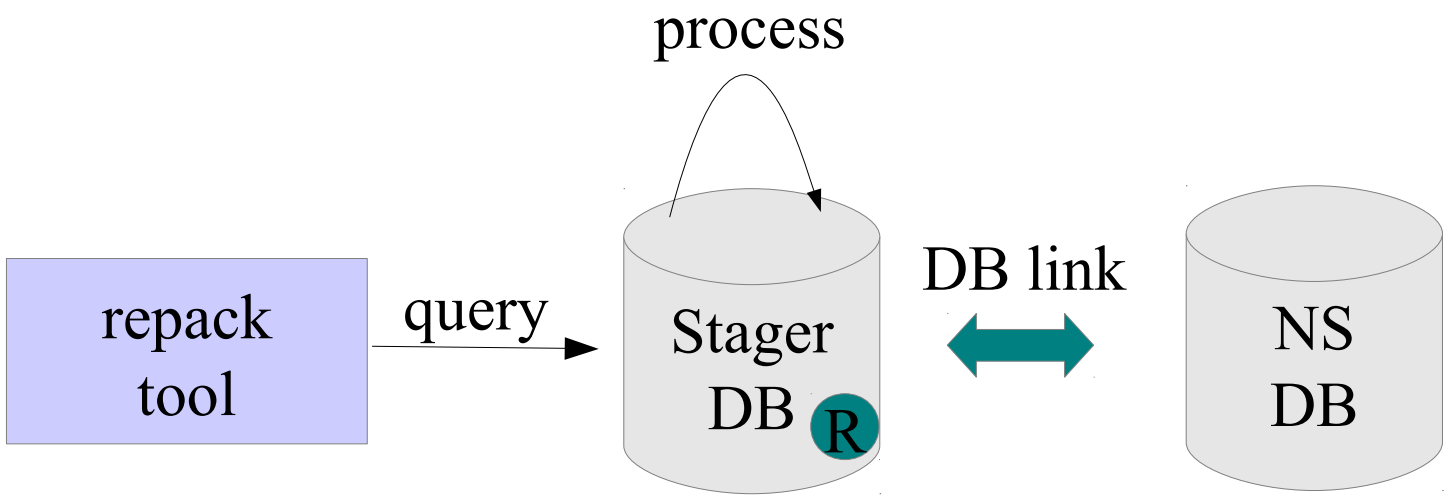
- Duplication of information
  - In repack and stager DB
  - Hard to synchronize
    - Constant stager_qry loading the stager
- Cancelling is not clean
  - stager_rm leaves migrations behind
  - So cancellation is only valid for recalls
- Slow to start
  - For each file, we go many times to NS
  - We submit huge requests to stager
- Complex code

New repack - start



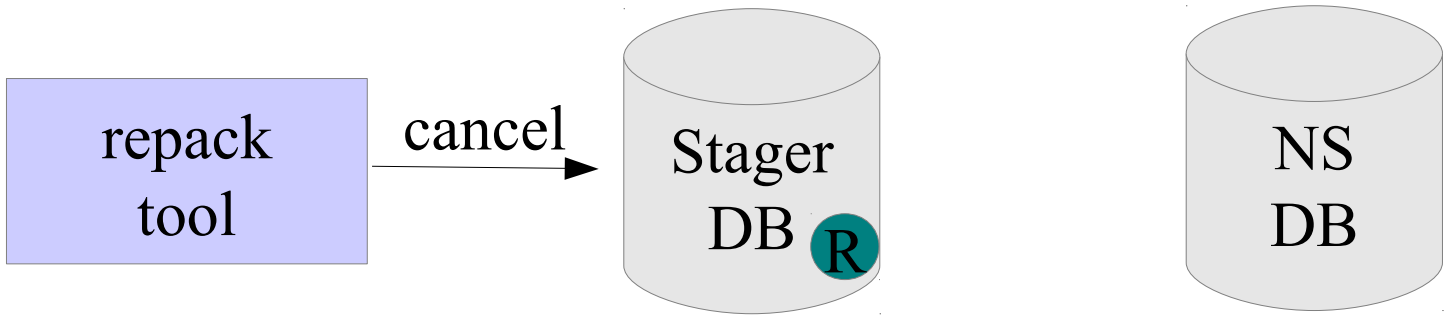
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

New repack – stable processing



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

New repack – cancelling



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Improvements

- Simplification
  - No repack daemon, no repack DB
    - any instance is de facto a repack instance
  - Single source of information : the stager
    - Simple, precise querying
- Fast start
  - Seconds before the first file is recalled
  - Up to 1M files handled per hour
- Full cancellation implemented
  - Including migrations
- Handling of corner cases
  - Missing copies recreated
  - Badly numbered copies corrected
  - Several copies on same tape handled and fixed
- Recall of the repacked copy only
  - other copies are not touched

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The repack tool

- Python script
  - Accessing directly the stager DB
- New features
  - Handling of files containing list of tapes
    - --bulkvolumeid, --bulkdelete
  - SvcClass must be given
  - Easy display of errors
- Better output

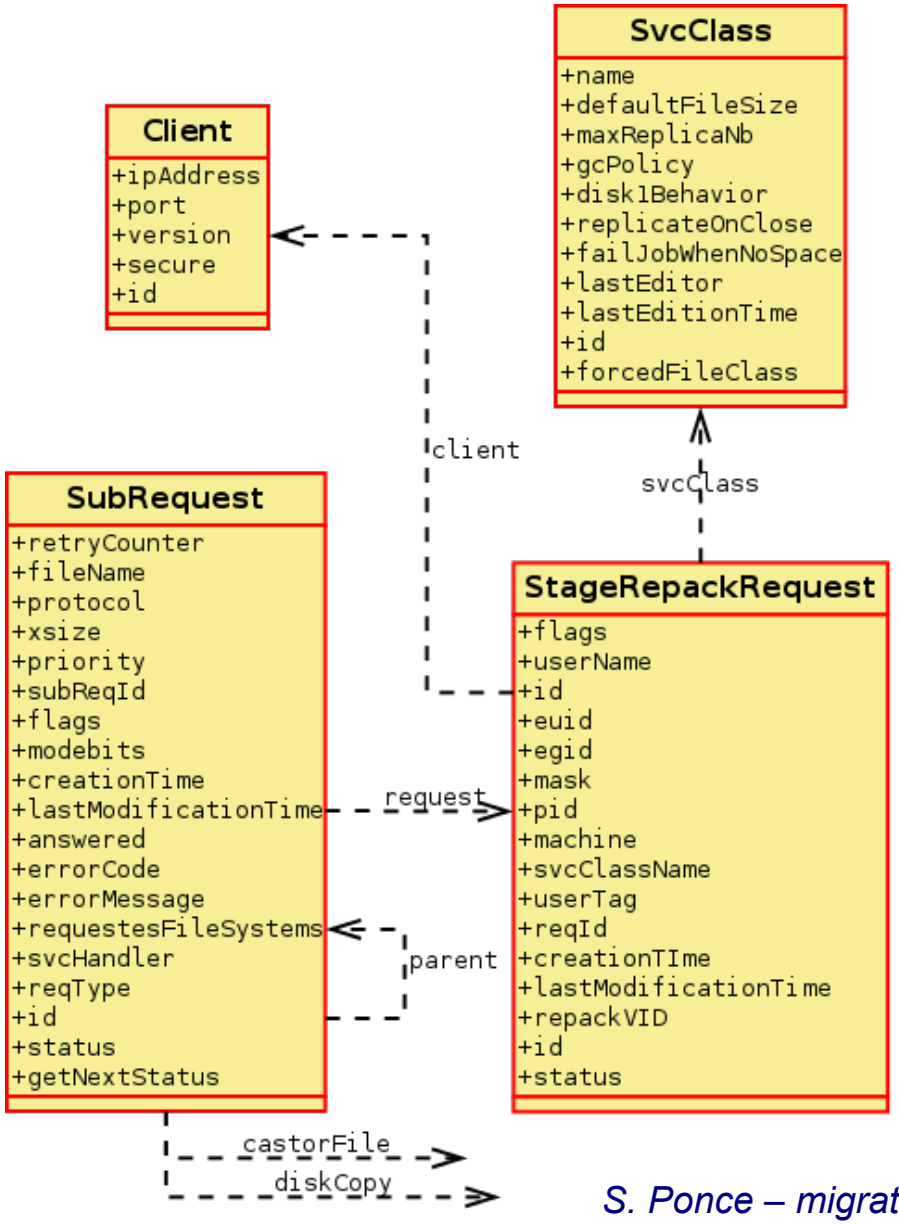
SubmitTime	RepackTime	User	Machine	Vid	Total	Size	toRecall	toMigr	Failed	Migrated	Compl%	Status
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17715	11066	810.76GiB	11064	0	0	2	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17716	9362	789.42GiB	9361	0	0	1	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17717	9615	684.10GiB	9613	0	0	2	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17718	10598	843.49GiB	10598	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17719	15512	844.94GiB	15512	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17720	12450	828.88GiB	12450	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17721	8858	715.80GiB	8857	0	0	1	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17728	23532	0.98TiB	23532	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17730	21268	569.74GiB	21268	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17731	17028	983.08GiB	17028	0	0	0	0%	ONGOING
...												
03-Oct-12 16:08	-	-	-	TOTAL	8657847	1.10PiB	6477990	3117	47	2176693	25%	ONGOING

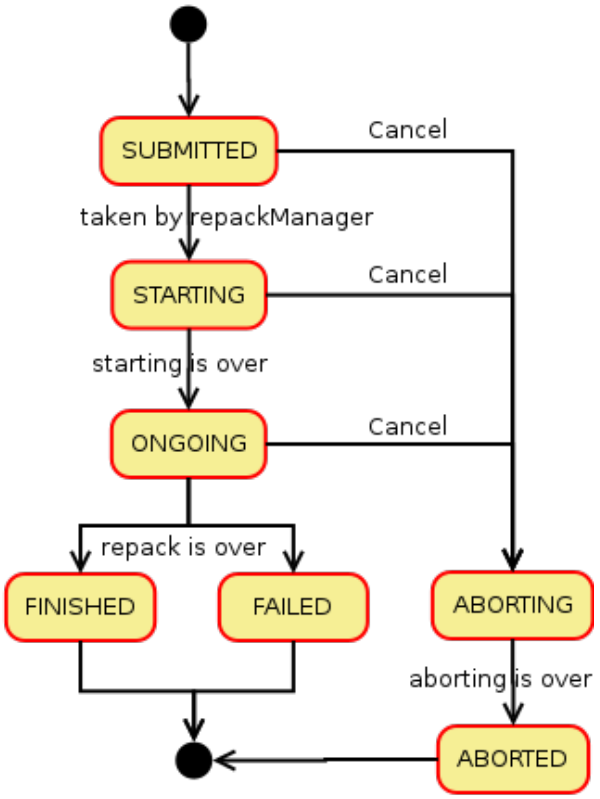
Gory details

- Stager interface
  - Is 100% PL/SQL
  - Does not go through PL/SQL
- Start and cancelling of Repack
  - Take a lot of DB CPU
    - To extract list of files from NS Database
    - To create corresponding request in stager
  - Concurrency is limited to a single start/cancellation
    - Hidden is a “queue” and a DB job processing the queue
      - repackManager
    - In practice, status “SUBMITTED” of repack requests
- handleRepackRequest
  - The key procedure, filling and starting a repack request from an empty shell containing only VID

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

StageRepackRequest





CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Performance

- Transfer speed not limited by repack
  - Should saturate drives (>220MB/s)
    - If enough dedicated diskservers (2.5/drive if gigabit network)
  - Achieved on few drives in stress test instance
- Starting speed
  - Order of magnitude : 1M files per hour
    - Effective repack starting after few seconds
    - Pretty independent of number of tapes
- Limits
  - Can handle tapes of > 1M files
  - 1000s of concurrent tapes is not a problem
  - Biggest production use so far :
    - >1400 production tapes in one go (>1PB, 8M files)
    - 25 files failed (bug in corner case handling, being fixed)

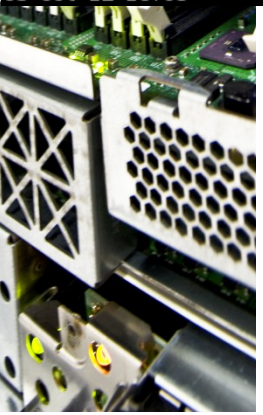
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Performance

CERN IT Department

- Transfer speed not limited by repack
  - Should saturate drives (>220MB/s)
    - If enough dedicated diskservers (2.5/drive if gigabit network)

SubmitTime	RepackTime	User	Machine	Vid	Total	Size	toRecall	toMigr	Failed	Migrated	Compl%	Status
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17715	11066	810.76GiB	11064	0	0	2	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17716	9362	789.42GiB	9361	0	0	1	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17717	9615	684.10GiB	9613	0	0	2	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17718	10598	843.49GiB	10598	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17719	15512	844.94GiB	15512	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17720	12450	828.88GiB	12450	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17721	8858	715.80GiB	8857	0	0	1	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17728	23532	0.98TiB	23532	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17730	21268	569.74GiB	21268	0	0	0	0%	ONGOING
25-Sep-12 17:17	7d 22:51	root	c2cernt3srv301.cern.ch	I17731	17028	983.08GiB	17028	0	0	0	0%	ONGOING
...												
03-Oct-12 16:08	-	-	-	TOTAL	8657847	1.10PiB	6477990	3117	47	2176693	25%	ONGOING



- Limits
  - Can handle tapes of > 1M files
  - 1000s of concurrent tapes is not a problem
  - Biggest production use so far :
    - >1400 production tapes in one go (>1PB, 8M files)
    - 25 files failed (bug in corner case handling, being fixed)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it