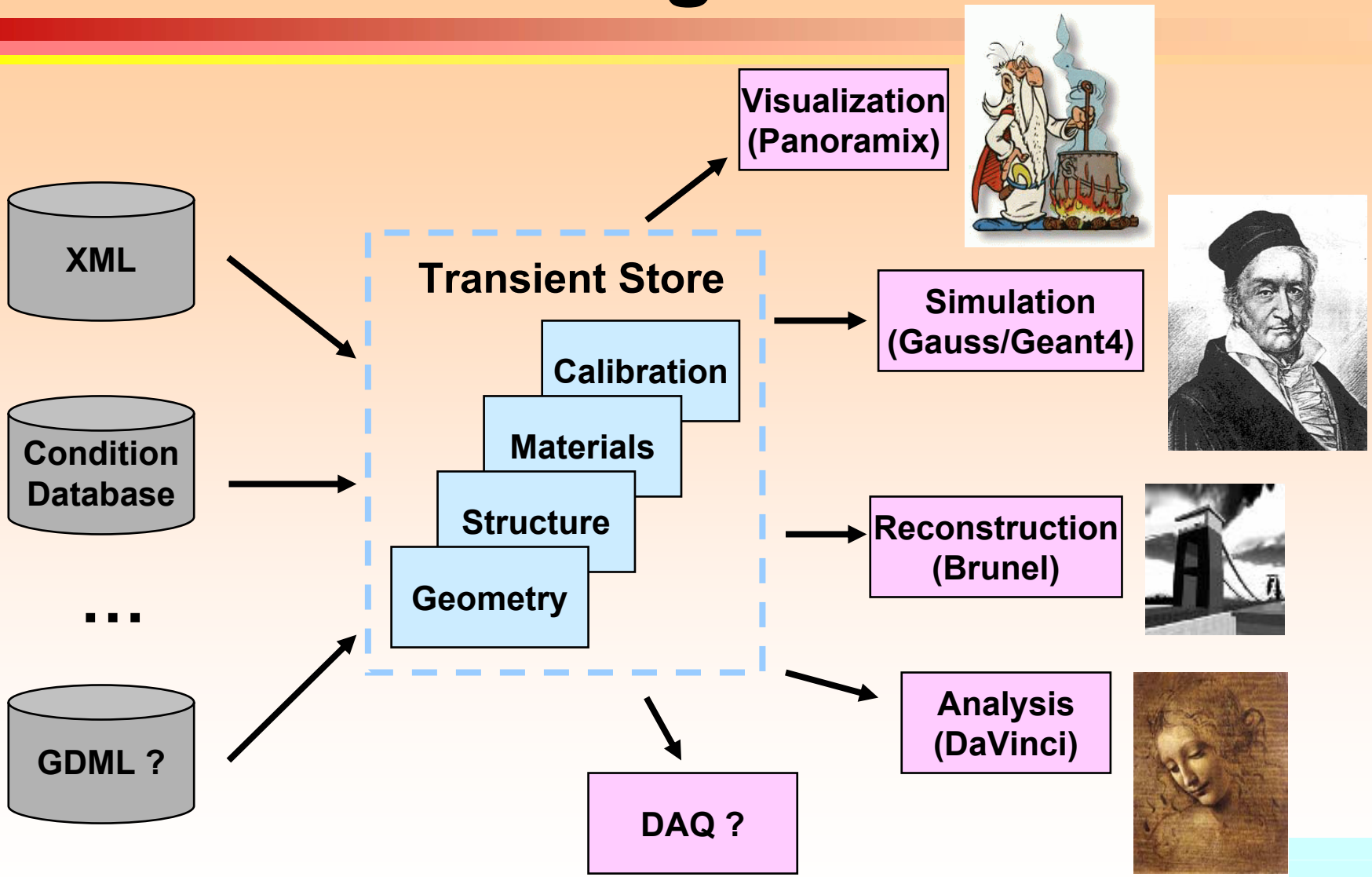
Distributed Databases in LHCb

- Main databases in LHCb Online / Offline and their clients
- The cross points
- Where db replication is expected
- What we expect from db replication

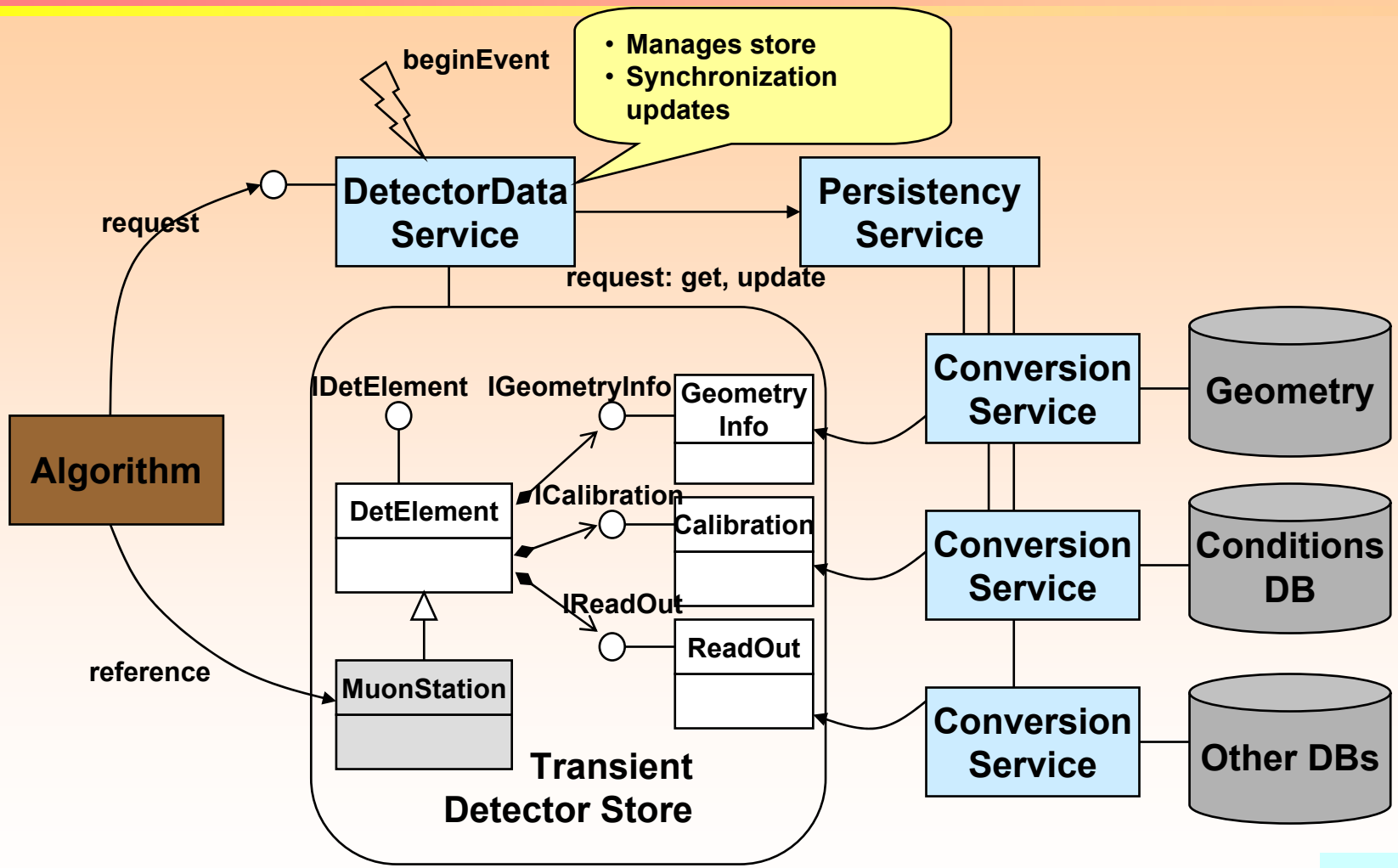
Design Goals

- **Distribute as few data as possible, but as many data as necessary**
- **Keep online and offline as loosely coupled as possible**
  - Learn from the BaBar experience
- **Try to achieve a clear hierarchy/information flow**
  - Only the master copy(s) may be data sinks
    - Minimize replication trouble
  - Allow as few active writers as possible
    - Minimize concurrency

Data Processing Architecture



Common Detector Data Access



The Inventory (not to be read)

	Application/Data Type	Distribution [none/fan out/gather]	Tier	Source/Producer	Volume[GB/site]	# of clients / site	acces mode[r/w/u]	owner [1-user/n-user/1-site	write/update rate [MP	Max. Latency	RAL us
(1)	**PVSS**	none	online	Control system	5000	1	rw	1u	4300		n	y
	Configuration DB	none	online	Control system	10	1	rwu	1u	5		n	y
(3)	**Online Conditions**	fanout	online	Control system	500	1	w	1u	400	1d	n	p
		fanin		offline/Calibrations	100	HLT	r	1u	70	1d		
	Offline Conditions	fanin	0	Control system	500	1	r	1u	400	1d	n	p
		fanout		offline/Calibrations	100	1	w	1u	70	1d		
			1	T0 copy	600	all T1	r	0		30	n	p
			2	T1 slice	50	WN	r	0			n	
			3/4	T1 slice	5	WN	r	0			n	
(2)	**File catalog**	gather	n	Worker nodes	15	WN	wu	1u	50		n	n
			n	Distrib.replica catalogs	5?	WN	rwu	1s	?	?	?	?
	Bookkeeping	fanout	0	T0 master	50	~20	wu	1u	150	1d	y	y
			1	T1 copy								

http://lcg3d.cern.ch/DataInventory/LHCb-Inventory261004.xls

(1) Databases in the Online

- **Detector and DAQ Controls (PVSS), Online Configuration database**
  - Stay at the PIT and never go out there
  - "Plug network off and still works"
  - Backup'ed but not replicated
- **Large data volume**
  - Detector controls: ~250.000 "sensors"
    Temperatures, trigger rates, detector configuration tag, …
  - ~0.5 MByte/second
  - ~5 TByte/year

(1) Databases in the Online

- **Database is accessed by relatively few tasks**
  - These provide the necessary information for
  - High Level Trigger (HLT) processes
  - Prompt reconstruction
  - Online calibration processes

- **HLT farm has no database connection**

(2) Databases in the Offline

- **File Catalogue**
  - Used by POOL
  - Implemented/Accessed by Grid middleware
    - Not discussed here: courtesy of gLite/EGEE/…
    - If replication is necessary, we inherit gLite/EGEE requirements
  - Each worker node needs at least a slice containing all input data
  - Possibly not a database (XML Catalogue)
  - Implementation: gLite
  - Final capacity: ~15 x 10^6 files/year

(2) Databases in the Offline

- **Bookkeeping & Job Provenance**
  - Allow eventually clones at Tier1 centers
  - Interactive access at Tier1s; Simple&Stupid replication
  - ~50 GByte/year

  - Possible future requirement:
    Access to provenance data from WN

(3) The Gray Area

- **Online / Offline Conditions**
- **Main connection point between**
  - Online and
  - Offline
- **Keep online and offline as loosely coupled as possible**
- **Needs separate model**

(3) Conditions: Writers

- **Online clients are likely to be tasks summarizing**
  - Detector controls data
  - Online calibrations

- **Offline clients are likely to be human (with some interface) feeding explicitly offline calibrations**

(3) Conditions: Readers

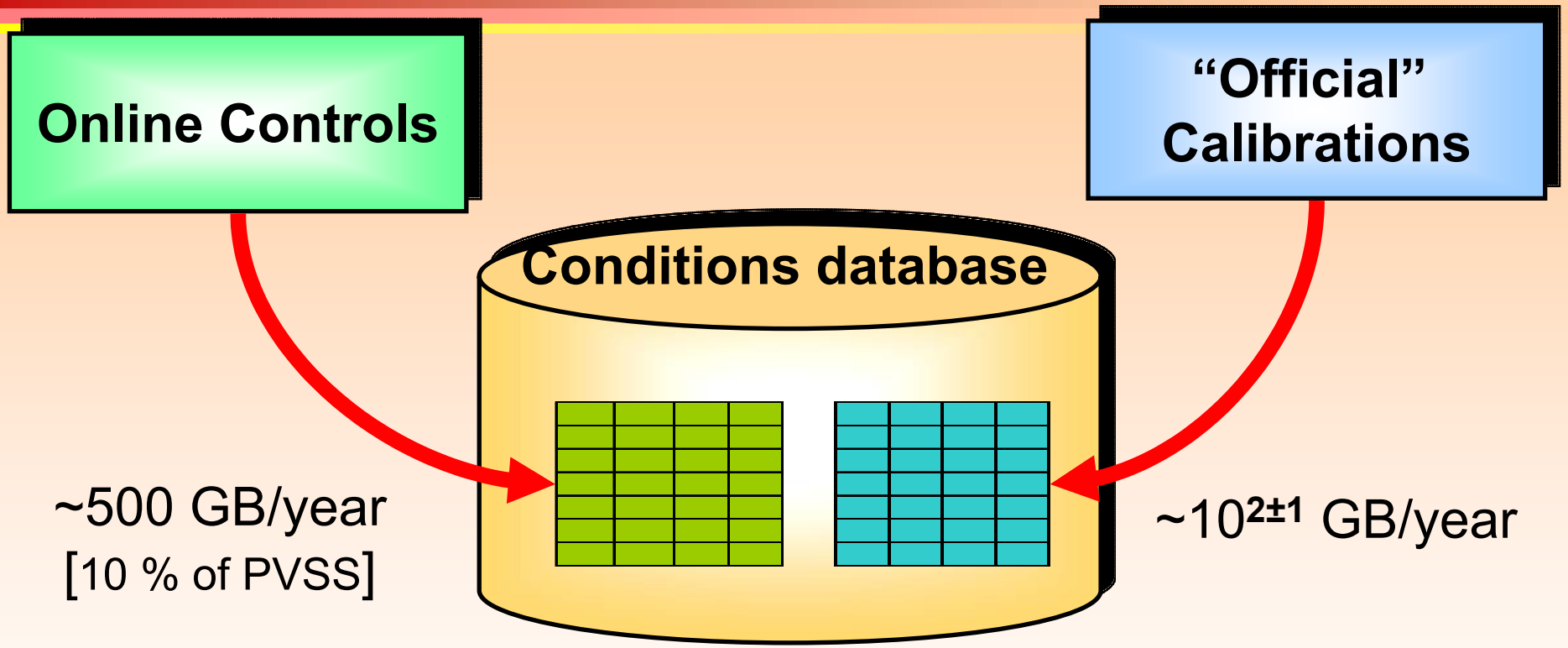
Online:

- **Tasks providing the trigger farm with the conditions needed by:**
  - HLT
  - Calibration tasks (Readers and writers)
  - Prompt reconstruction
  - …
- **All done at PA8**

Offline:

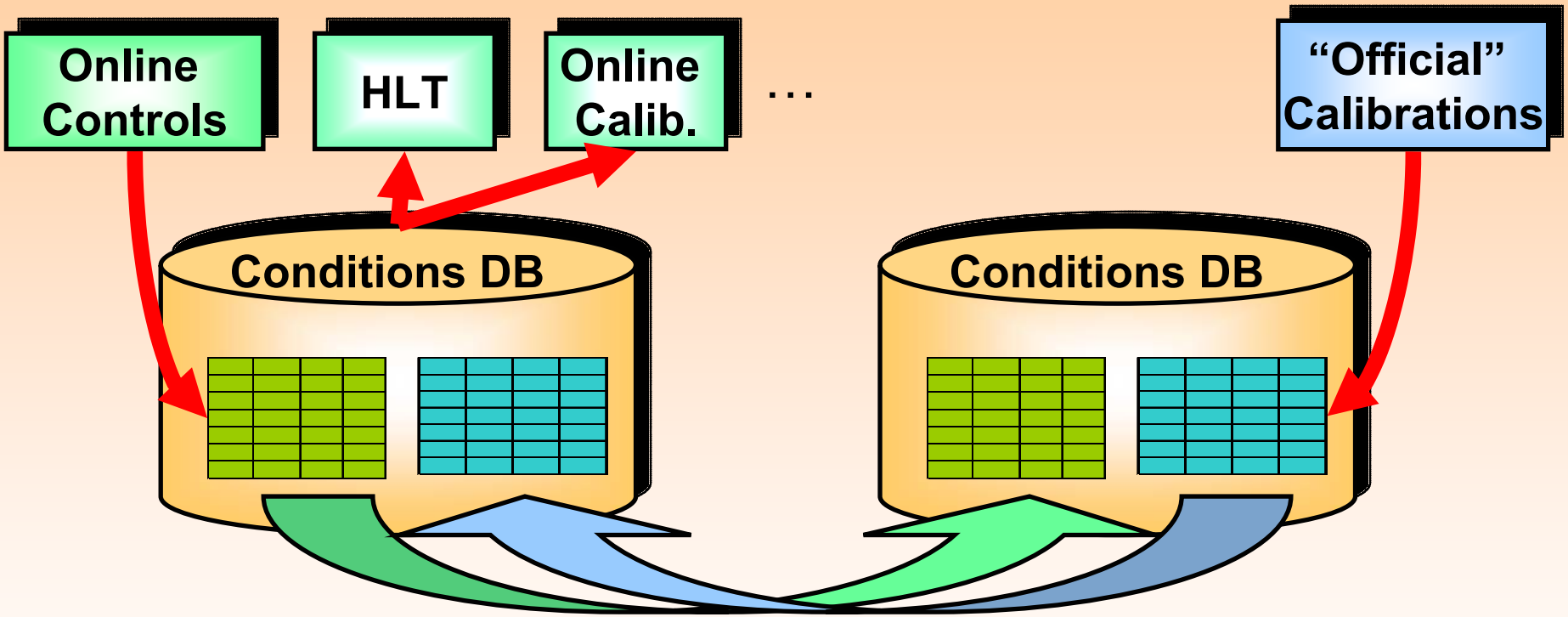
- **Any data processing task**
  - Physics analysis
  - Reprocessing
  - …
- **Anywhere in the world**

Online / Offline Conditions



- Clients see 2 very loosely coupled schemas
- Single logical database
- 2 instances: "Online" and "Offline" instance

The Online Model

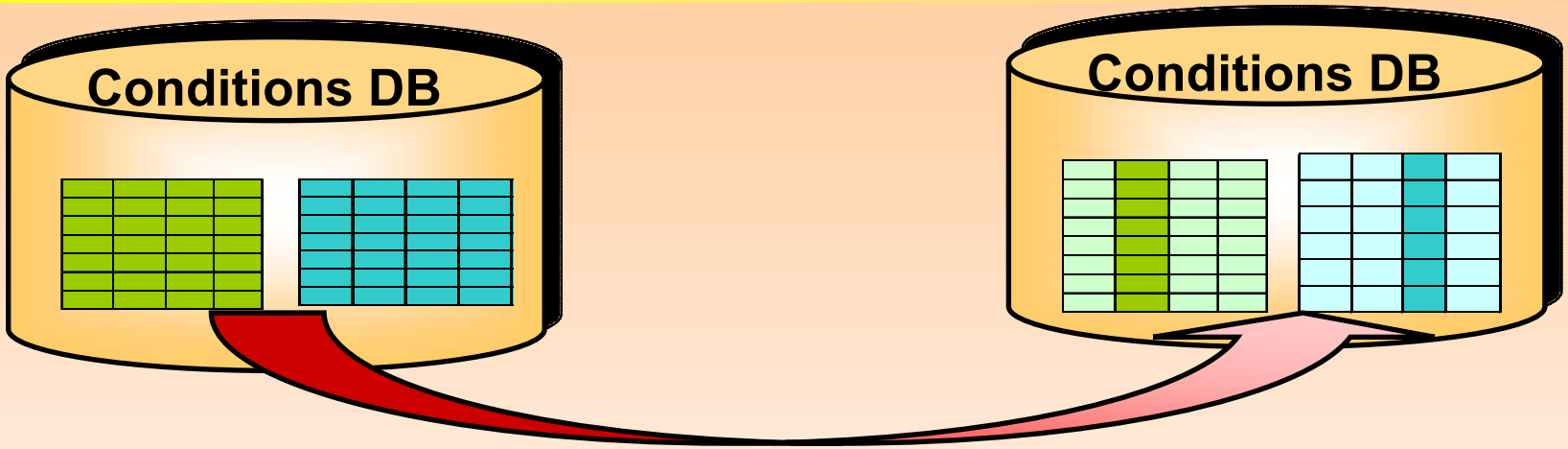


Database replication

PA8 [LHCb Pit]

CERN Computer Center Tier 0

The Offline Model



- **Worker node needs (fast) access to a valid database slice according to**
  - Time intervals
  - Item tag(s)

The Offline Model

- **We expect to have a usable solution of the conditions database provided by POOL including:**
  - Efficient database slice creation
  - Efficient access optimization “on the way”
    - Tier0 -> Tier1 -> Tier2 replication / slicing
- **What we do not need:**
  - Write access at TierN (N>0)

Summary

- **Online databases stay where they are (PA 8)**
  - Except PVSS extraction into online conditions
- **Offline databases must be accessible from worker nodes**
  - Conditions database slices
  - File catalogue
    - Depending on grid middleware
  - Optionally bookkeeping/job provenance information is replicated