Data calibration and processing

Stefano Lacaprara

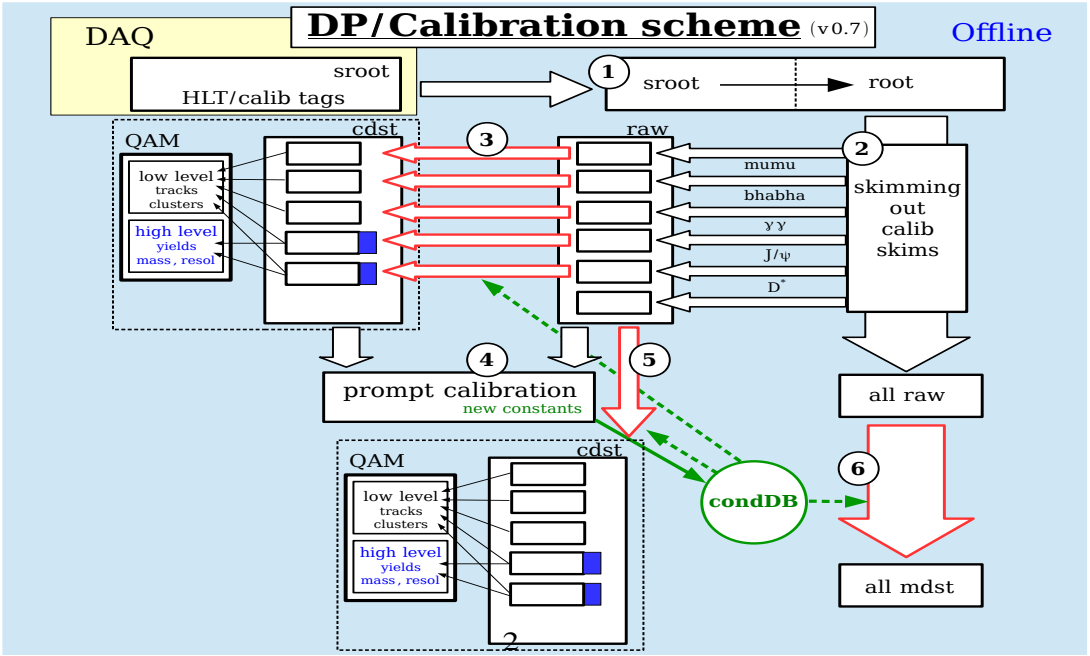
stefano.lacaprara@pd.infn.it

INFN Padova

13th annual Belle PAC Review meeting,
KEK, 12 February 2019

Outline

- General Data Processing and Calibration scheme
- Data processing for phase II
- cosmic global run (exp5)
- Calibration workflow
- Toward phase III


DP/Calibration scheme (v0.7)
Offline
DAQ
sroot
HLT/calib tags
1
sroot
root
QAM
cdst
low level
tracks
clusters
high level
yields
mass, resol
3
raw
mumu
bhabha
ϒϒ
J/ψ
D*
2
skimming
out
calib
skims
4
prompt calibration
new constants
5
all raw
QAM
cdst
low level
tracks
clusters
high level
yields
mass, resol
condDB
6
all mdst
2

Data Processing and Calibration Scheme

1. Data collected by DAQ **sroot**
   - **sroot** → **root** by Computing Group
   - **root** copied and registered to grid (CG) and locally at calibration center
2.0 DataProduction notified
2.1 DP produces HLT-skims (RAW) on calibration center (KEKCC)
3 DP produces cdst with running (or snapshot) GT and pass to calibration
4 Calibration Team does its magic and produce improved payload
5 (optional) DP produces new cdst with improved payload
   - iterate until everybody is happy (or not too sad, at least)
6 DP process RAW into mdst with happy GT
   - initially run at calibration center (eg KEKCC) and produce mdst locally
   - run on grid in parallel (eventually only) and produce mdst directly on grid
   - (publish locally produced mdst on the grid as plan B)
7. Quality Assurance Monitor (good run list, etc)
8. Profit (Announce to collaboration: physics)

Reprocessing Phase 2 collision data

- Phase 2 collision data reprocessed multiple times with various improvements
  - New features, including changes to software used in the online system, modifications to the fabrication system, calibration data format and algorithms, etc., were quickly implemented
  - Established calibration skims and data flow, including improvements to global tag management
  - Close collaboration with calibration experts to produce revised constants for later processing
  - More details at https://confluence.desy.de/display/BI/Experiment+3

- Phase 2 data reprocessings
  - Prod 1 (KEK), May 4 - ~3 pb^{-1}
  - Prod 2 (KEK), May 13 - ~92 pb^{-1}
  - Prod 3 (KEK), June 9 - ~250 pb^{-1}
  - Prod 4 (KEK), June 23 - ~250 pb^{-1}
  - Prod 5 (KEK), August 22 - ~500 pb^{-1}
  - Prod 6 (KEK), October 12 - ~500 pb^{-1}
  - Prod 6 (GRID), November 14 - ~500 pb^{-1}
  - Proc 7 (KEK), January 8, 2019 - ~500 pb^{-1}
  - Proc 7 (GRID), in progress! - ~500 pb^{-1}
  - Proc 8, coming soon! (improved tracking, potential other improvements)

Reprocess full Phase 2 data with good PID for October 2018 B2GM

Reprocess full Phase 2 data for CKM2018

Starting with prod6, Phase 2 data is being reprocessed with the distributed computing system!

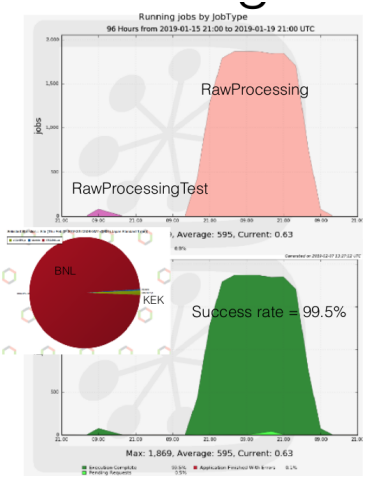
Several known issues with proc7, please move to proc8 when it is ready!

Reprocessing at KEKCC was sufficient for Phase 2, now using GRID resources to scale up for Phase 3

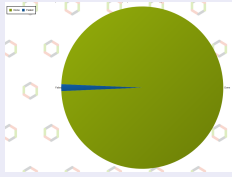
2

PhaseII reprocessing on the grid

- Improved tools to cope with “run range” definition
- Test proc7 on the grid to learn the system and stress it
- Also provide easy access to processing on the grid.
  - small test processing (few runs)
  - full processing after success of test
- Raw data hosted at KEK and BNL
  - most jobs run at BNL
  - raw data processing jobs at KEK restricted
- Success is 99.5%
  - failed jobs recoverd by automatic resubmission
  - few jobs in strange state under investigation



Success Rate 99.5%



Speed

Done in < 1 day
(at KEKCC ~ 2.5 days)

A good success, grid production can be used in phase III raw processing

Phase 2 reprocessing scheme

- General phase 2 reprocessing scheme (e.g. after new software release)
  - Pre-processing to provide cDSTs for calibration ~ 1-2 days
  - Create global tag with updated calibration payloads ~ 2+ weeks
  - Reprocess RAW data to mDST, cDST, DST ~ 1-2 days at KEKCC

Information on reprocessings

- information on the calibration for each reprocessing (what are the updated constants ?)
- Note that preprocessing are described in Experiment 3 preprocessing

proc7 (release-03-00-00/GT493): run 0-5613

- Note that starting from this processing, the name have been changed to **Proc** (as in **Proc**essing) - was **Prod** (from production) - to disambiguate with respect to ProdID used by grid-based production tools.
- **data (mdst, cdst, dst): /hsm/belle2/bdata/Data/release-03-00-00/DB00000493/proc00000007/e0003/4S/**
  - runs reprocessed (required presence of CDC/ECL/TOP):

run range	offline luminosity (bhabha) ★	offline luminosity (gamma gamma)★	offline luminosity★	Comments
0-5613⚠	503.5±0.2 pb^{-1} ⭐	499.0 ± 0.6 pb^{-1}⭐	pb^{-1}	bad runs for TOP from 2824 to 3547 (so ~135 pb^{-1}).

⭐ Luminosity has been computed by @Xing-Yu Zhou . Run-by-run results are available at https://stash.desy.de/projects/B2LOWM/repos/luminosity/browse/offline_xingyu.zhou/rslt/proc7/Run-by-run_results_of_the_integrated_luminosities.txt_good

- ★ WARNING: these luminosity are copy-pasted from prod6 one, these are NOT official lumi for Proc7
- ⚠ WARNING: due to a crash during reconstruction run 5548 is partially available for dst/mdst/cdst and NOT available at all in the skims

☑ ~~BII-4422~~ - Crash in Exp3 processing CLOSED

- **scripts: https://stash.desy.de/projects/B2P/repos/data/browse/e0003/release-03-00-00/DB00000493/Rec**
- **GT used: data_reprocessing_proc7**
- **Experiment 3 skims**

https://confluence.desy.de/display/BI/Experiment+3

Calibration Skim

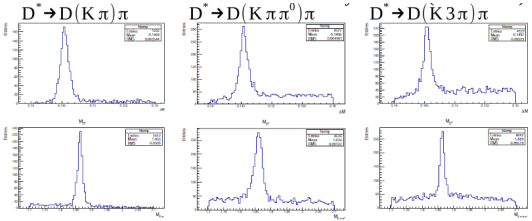
- extended detector information saved for calibration purpose
- provide only raw/cdst for experts for phase 3
  - All calibration skims will be promoted to HLT
  - To be produced at step 2 in the DP schema
- converging on new skims ($\gamma\gamma$, ECL bhabha, TOP dimuons, J/ψ, etc)
- few offline skims also (di-muon, hadronB) will be promoted
- more offline skims being developed (example D*)

Category	Skim name	Selection
HLT	hlt_mumu_2trk	[[nTracksLE >= 2] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.35] and [[P2OEbeamCMSBhabhaLE > 0.2] and [[EtotLE < 7] and [[EC2CMSLE < 1] and [maxAngleTTLE > 0.785]]]]]]
HLT	hlt_mumu_1trk	[[nTracksLE == 1] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.1] and [[EC1CMSLE < 1] and [EtotLE < 7]]]]
HLT	hlt_hadron	[[nTracksLE>=3] and [BhabhaTrk==0]]
HLT	hlt_bhabha	[Bhabha2Trk==1] **(prescale removed since prod3)**
HLT	hlt_gamma_gamma	[[nTracksLE <= 1] and [[nEidLE == 0] and [[EC12CMSLE > 4] and [EC1CMSLE > 2]]]

crucial for calibration/monitoring

used also for physics

Skim name	Selection
HadronB	[nTracks>=3] and [nCluster > 1] and [Evis > 0.2 sqrt(s)] and [|Pz| < 0.5 sqrt(s)] and [0.1 sqrt(s) < Esum < 0.8 sqrt(s)]
Dimuon	[nTracks==2] and [acollinearity <10] and [Esum <2.0] and [Esum(tracks) <2] and [Theta>45 and Theta<125]

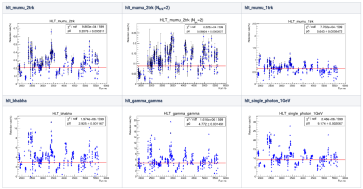


Usage of cdst

Usage of cds for detectors calibration

Detector, tasks	inputs	outputs	main samples
CDC	raw	defined	cosmics, dimuon
CDC dE/dx	cdst	defined	Bhabha, radiative Bhabha
ECL	cdst	defined	cosmics, dimuon, Bhabha, gamma gamma
TOP	cdst	defined	dimuon
BKLM	cdst	defined	dimuon, cosmics
EKLM	cdst	defined	dimuon
ARICH	cdst	defined	dimuon
PXD (non-align.)	?	?	?
SVD (non-align.)	local run	defined	local run
	cdst	defined	hadron skim (tbc)
SVD + PXD (alignment)	raw/cdst	defined	dimuon, cosmics

Retention rate is low ($\sim 1\%$)



But size not negligible

- Typical file is 1.43 GB/22 k evts = 66 kB/evt
  - ECLCalDigits: 645 MB (45%)
  - ECLDigits: 216 MB (15%)
  - ExtHits: 162 MB (11%)
  - CDCDedxTracks: 84 MB (6%)
  - RecoTracks: 80 MB (6%)

  } sum up to 83%

  - TOPDigits: 46 MB
  - SVDRecoDigits: 43 MB
  - SVDShaperDigits: 19 MB

  } but only few SVD ladders !

  - ECLClusters: 15 MB
  - TrackFitResults: 6.0 MB
  - SoftwareTriggerVariables: 5.9 MB
  - SVDClusters: 3.8 MB
  - ARICHDigits: 2.9 MB
  - ...

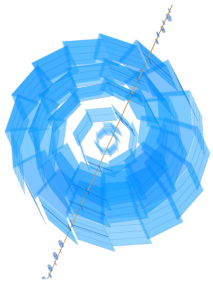
For reference: mdst size 1 kB/ev, raw: $30 - 45$ kB/ev, dst: 120 kB/ev

Will review cDST use/size soon

Might consider detector specific cDST in place of common ones

Experiment 5 - full dress rehearsal

- Processing on-going at KEKCC for cosmic runs
  - ✓ ideal playground for automation
  - ✓ train comunication channel with computing group (JIRA ticket)
  - 🔧 full automation is under test
  - **GCR5a** started immediately (with old script and payload) now obsolete
  - **GCR5b** ongoing (existing and new runs) with running GT (see later) with better script and payload
  - at least one more processing expected (updated SVD tracking)
  - RAW data copied to grid also, will test prompt reconstruction on grid also
- running documentation in confluence and JIRA ticket

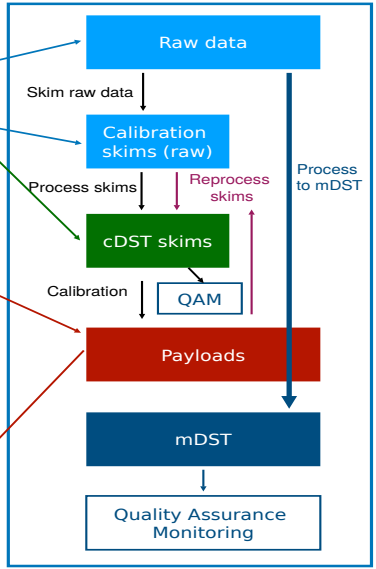


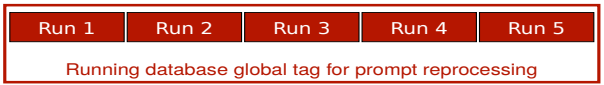
GCR5b (release-03-00-01/GT503):

- **output data type:** mdst, cdst, dst
- **Path:** /ghi/fs01/belle2/bdata/Data/Cosmic/e0005/4S/GCR5b/release-03-00-01/DB00000503/
- **Runs**:
  - 75 79 84 85 86 87 88 89 90 91 92 93 94 95 96 97 99 101 102 103 104 283 284 285 286 287 288 289 290 291 292 293 294 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 431 432 433 434 580 581 582 583 584 585 588 589 591 629 630 631 632 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699
  - 700 701 702
  - 586
  - 1136 1137 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155
- **Scripts:** https://stash.desy.de/projects/B2P/repos/data/browse/GCR5a/release-03-00-00/DB00000493/Rec
- **GT used:** data_reprocessing_prompt_snapshot_01252019
- **Logs:** /ghi/fs01/belle2/bdata/Data/Cosmic/e0005/4S/GCR5b/Log/
- **Note:**

Fast reprocessing scheme

- ROOT formatted raw data on offline system, registered, and replicated to raw data processing centers
- Calibration skims from raw data processed to cDST at “calibration center”
- Prompt calibration and QAM run at “calibration center” (includes multiple reprocessing to cDST with updated tracking for dependent calibrations)
- Calibration constants added to offline global tag
- Latest runs reprocessed to mDST
- Requires that prompt calibration algorithms are ready for automation
- When offline calibrations and/or software updates are complete, or if significant changes to prompt calibration, reprocess all available data to mDST





Calibration organization

Clear list of liason and responsible for detectors calibrations

(All) the info about the calibration group are collected here

https://confluence.desy.de/display/BI/Data+Production+Calibration

Who's who:

SVD+PXD (align):	Jakub Kandra
PXD (no align):	Maiko Takahashi
SVD (no align):	Laura Zani
CDC (tracking):	Makoto Uchida
CDC (dE/dx):	Jitendra Kumar
TOP:	UT
ECL:	Chris Harty
ARICH:	Luka Santelj
BKLM:	Jincheng Mei
EKLM:	Kirill Chilikin

Meetings:

Every other Tuesday, both at
9 am and 6 pm JST
(two sessions in the same day)

https://confluence.desy.de/display/BI/Calibration+Group+Meetings

2

Defined task via JIRA ticket

	BIIDP-1132	Processing 8 calibrations	OPEN	Umberto Tamponi
✓	BIIDP-1143	Processing 7 calibrations	RESOLVED	Umberto Tamponi
✓	BIIDP-1144	Production 6 calibrations	RESOLVED	Trabelsi Karim
✓	BIIDP-1145	Production 5 calibrations	RESOLVED	Trabelsi Karim
✓	BIIDP-1147	Production 4 calibrations	RESOLVED	Trabelsi Karim
✓	BIIDP-1148	Production 3 calibrations	RESOLVED	Trabelsi Karim
✓	BIIDP-1149	Production 2 calibrations	RESOLVED	Trabelsi Karim

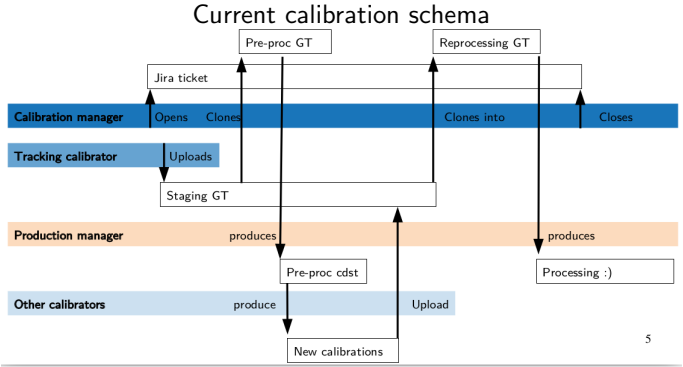
Example for proc8

Sub-Tasks +

1.		SVD alignment	OPEN	Jakub Kandra
2.		SVD non-alignment	OPEN	Laura Zani
3.	✓	CDC tracking	CLOSED	Makoto Uchida
4.		CDC dE/dx	OPEN	Jitendra Kumar
5.		TOP	OPEN	Umberto Tamponi
6.		ARICH	OPEN	Luka Santelj
7.	✓	ECL	RESOLVED	Christopher Hearty
8.		BKLM	OPEN	Jincheng Mei
9.	✓	EKLM	RESOLVED	Kirill Chilikin
10.		PXD	OPEN	Maiko Takahashi

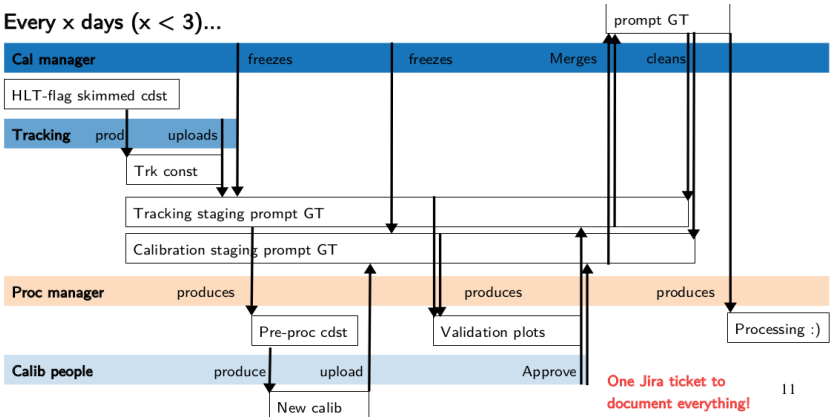
Current workflow for prompt calibration

- Gained a lot of experience in phase2 and multiple reprocessing
- also Global Cosmic Run 5 very useful to test calibration and prompt processing
  - Running GT, GT-snapshot, ...
- need for better automation for prompt calibration
  - calib must be produced asap after data taking
  - to allow prompt processing few days/one week after that
- emphasis on automation, clear workflow, communication, and reproducibility



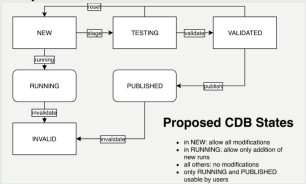
Used a s starting point to develop a better one.

Calibration plan for phase 3

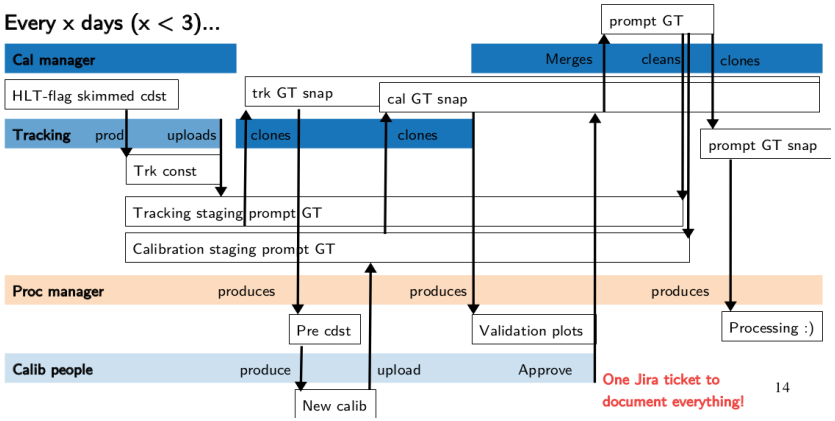


New CDB features

Some missing CDB functionality will be implmented soon



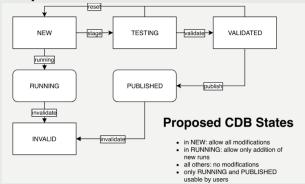
Calibration plan for phase 3



Workaround for very early data taking: more complex for Calibration manager, but based on same workflow

New CDB features

Some missing CDB functionality will be implmented soon

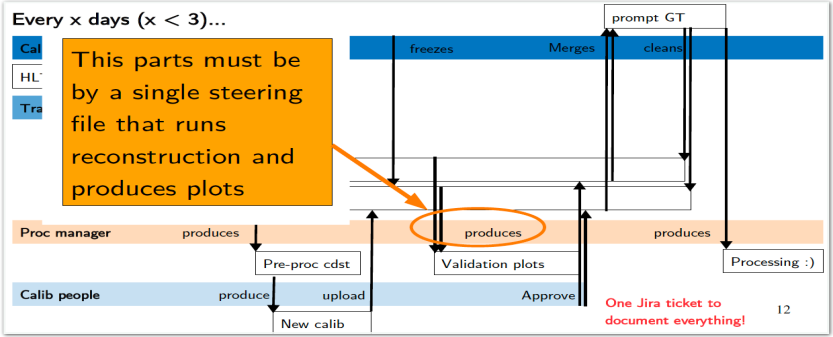


Probably none new CDB will be available for march 11^{th}, but workaround is available

Tools for prompt calibration

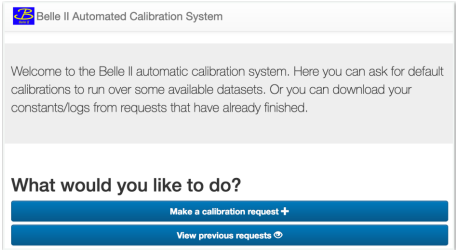
- Prompt calibration requires
  - Centralized scripts for algorithms
  - Tool for automated calibration requests, monitoring, etc.
  - Airflow server can handle this, as well as things like QAM, run conditions, etc.

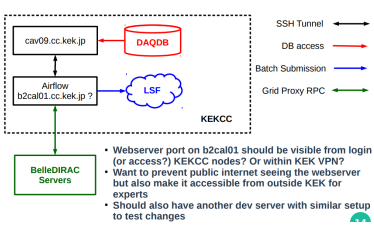
- **CAF seems to be in good shape and stabilising (for now)**
- **No features requested that are missing**
- **You may experience a bug on release-03-00-02. Fixed in master**
- **Please create JIRA issues for David Dossett if you have requests/bugs**
- **Just updated the introductory tutorial repository for release-03-00-01**



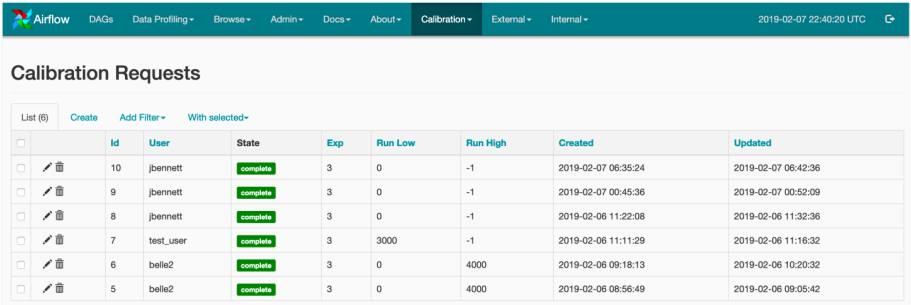
https://confluence.desy.de/display/BI/Data+Production+Calibration

Tools for Calibration: CAF/Airflow





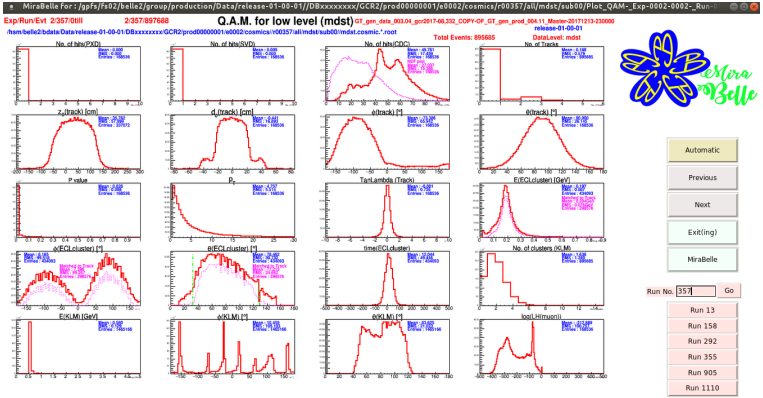
- Prod server to be placed at KEKCC
- Also possible collect info from CDB, grid metadata, etc for a user-oriented run summary DB



	Id	User	State	Exp	Run Low	Run High	Created	Updated
	10	jbennett	complete	3	0	-1	2019-02-07 06:35:24	2019-02-07 06:42:36
	9	jbennett	complete	3	0	-1	2019-02-07 00:45:36	2019-02-07 00:52:09
	8	jbennett	complete	3	0	-1	2019-02-06 11:22:08	2019-02-06 11:32:36
	7	test_user	complete	3	3000	-1	2019-02-06 11:11:29	2019-02-06 11:16:32
	6	belle2	complete	3	0	4000	2019-02-06 09:18:13	2019-02-06 10:20:32
	5	belle2	complete	3	0	4000	2019-02-06 08:56:49	2019-02-06 09:05:42

Quality Assurance Monitor

- Goal: provide tool for check quality of data processed
- Common for all detectors
- also for high level quantities
- independent on software version
- automation
- Can be used to provide list of good runs for physics



Toward Phase III

- **Prompt processing**
  - Workflow defined
    - ⋆ partial test on GCR5 (w/o real calibration, not enough events)
  - strict collaboration between CG, DP, and calibration needed (in place);
  - workflow for prompt calibration defined;
    - ⋆ CDB tools requested, but a workaround is available;
  - plan is to run mdst processing locally **and** on the grid
    - ⋆ Eventually on grid only.
  - need to review the cdst size: possibly have multiple cdst (detector specific) and/or partial processing;
- **Reprocessing**
  - Gained a lot of experience with phase II reprocessing
  - workflow is in place and well tested
    - ⋆ calibration done locally
    - ⋆ final reprocessing both locally and on grid
    - ⋆ Grid experience is good, we can envisage to do a grid only processing

Toward Phase III (2)

Tentative schedule toward LP2019

- **prompt processing** to start immediately;
  - first iteration of calibration after $\sim$ 1 week of data tacking, depending on luminosity
  - then iterate every few days
- **re-processing**:
  - according to plan (see Fabrizio's talk) for Phase 3 we expect $\sim$ 6 reprocessing in 2019 (3 with new software).
  - Luminosity and Conference driven
- **Tentative schedule toward LP2019 (5-10 Aug 2019)**
  - LP2019 (5-10 Aug 2019)
  - Belle II data taking will stop on July 1^{st}.
    - two weeke for final calibration, plus 1 week of reprocessing
    - probably not enough time for analysis
  - first partial reprocessing after Platinum Week (done by mid May)
    - large fraction of luminosity may be available early enough
    - possible top-up with full luminosity in time for LP2019

Summary

- Lot of experience from Phase 2 and Global Cosmic Run
- Plan for Phase 3 is in place both for Calibration and Data Processing
- Tentative schedule for summer conferences.
- Calib and Data Processing managers deputies are needed
  - some experienced people already showed interest: discussing now

(common last slide for $\sim$ calib and data processing slide at past B2GM)
I need a deputy
(we all neeed)

Additional stuff

Additional or backup slides

Phase2: Proc7

pre-proc7 done (14/12/2018)

- `prerelease-03-00-00b`, GT `data_reprocessing_validation_release-03-00-00`
  - improvement in tracking code, no new payload for CDC
  - produced cdst from HLT skims (prod6) `hlt_bhabha`, `hlt_gamma_gamma`, `hlt_mumu_2trk`
  - First exercise for new Data production (SL) and calibration manager (Umberto)
- Issues:
  - Most of raw data on tape on `hsm`, need to prestage them (tools available `hstage`)
  - setup of script, GT, etc via Jira ticket and PullRequest (good)
  - Processing was fast (once data on disk) about 1 day: fully at KEKCC (as expected)

proc7 done (8/1/2019) Experiment3-proc7 confluence page

- `release` 03-00-00 (2/1/2019 as scheduled)
- GT: `data_reprocessing_proc7` using pre-proc7 and release (4/1)
  - proc7 started 5/1 ended 8/1 at KEKCC
  - input RAW 613 runs, output: mDST, cDST and DST
    - only customization: Set isMC: 0 in metadata
  - also started production `proc7b` on grid (next slide)
  - feedback very fast. Tracking degradation was found BII-4359
    - intense debugging and immediately start preparation for (pre-)proc8 with fixed code/payload

Grid processing proc7b

- week after proc7, we started a reprocessing of exp3, phase2 on the grid proc7b identical to proc7
- need some time to setup processing correctly, need to learn `gb2_prod_tools` (help from Ueda-san)
- two test pre-processing (3 runs each) submitted on 16/1
- two batch of jobs because some input file are labelled "Beam" and some "Physics"

6798 Done: all looked fine, so I started the full processing
6799 Running: Transaction are Done= 6/6 but
  - (24/1): TransId:21689 Registered Done:2 Total:2 100.0% 1/1 ??

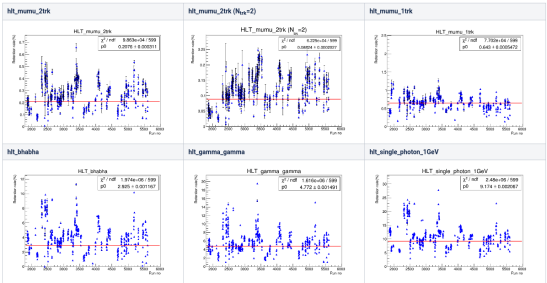
- Full production (started 17/1):

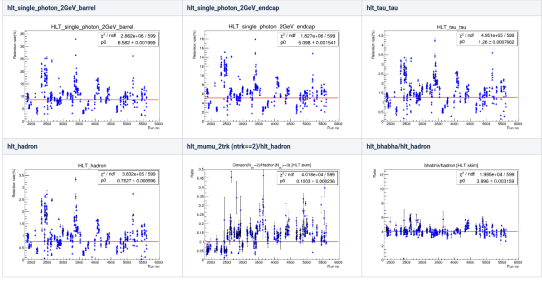
6857 Running: submitted on 2019-01-17 15:28:18, Done=400/402 (was 182/402 on 24/1)
6858 Running: submitted on 2019-01-17 15:42:48 Done=872/874 (was 187/874 on 24/1)

- NB: `proc7` took about 2.5 days at KEKCC
- Investigating with computing experts: apparently the last remaining jobs are done, but are in a strange state, and they are reported as not done

HLT skims (and offline)

- preproc7 was done starting from prod6 HLT-skims (raw)
- for proc7 we produce new HLT skims, based on code by Karim
- input **mdst/cdst** (no dst)
- output: skim **mumu_2trk mumu_1trk hadron bhabha gamma_gamma single_photon_1GeV single_photon_2GeV_barrel single_photon_2GeV_endcap tau_tau** exp3 skims
- later Karim produced offline skims **offskim_mumu, offskim_dstar**
  - do we still need all of them? Do we need more?
- retention rate stable wrt prod6 (Karim)

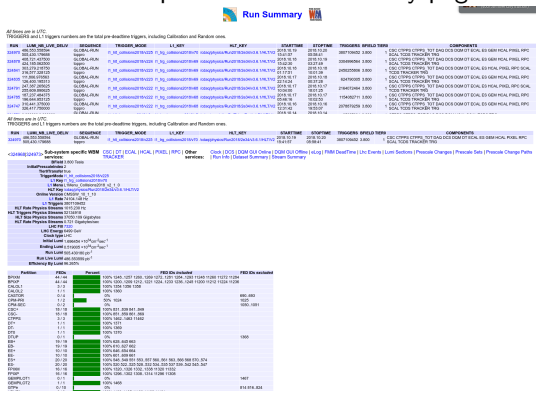




Luminosity

- Provided by Xing-Yu, based on ECL information with the Bhabha and Di-γ events [BELLE2-NOTE-PH-2018-027]
- Proc7: $(503.5 \pm 0.2)\,\text{pb}^{-1}$ and $(499.0 \pm 0.6)\,\text{pb}^{-1}$, respectively
  - was 479.8 pb^{-1} and 494.9 pb^{-1} for prod6 (correction suggested by N.Gabyshev to account for phiCMS efficiency dependence for Bhabha gives: 491.5 pb^{-1})
- also run-by-run lumi available link
- We need an offline run summary page to collect these kind of information
  - see Run+Summary confl page
  - now a txt file (extracted from DAQ run DB) link
  - Good starting point, but we need something more code and user friendly: Run+Summary
  - possibly with access via software and maybe also web interface?

An example: CMS run summary page



Experience gained (kekcc)

- Still learning the job, great help from Jake, Karim, Umberto, and many other
- ✓ Very good communication: JIRA, PullRequest, mails, chat, skype
- 🔧 Need to document in detail all phases of process, and keep up to date the documentation;
- ✓ Very good interaction with calibration team, clear definition of GT creation process helped a lot.
- some "ad hoc" tuning of standard reconstruction needed in pre-proc, which is fine(-ish)
- in real processing we need to use only standard unpacking/reconstruction/etc from release
  - ✗ what if we need a quick change? Still acceptable to have non-standard (but documented) mod in steering file or need for a fast patch-release?
  - not an issue for phaseII re-processing, but is for exp5 cosmic run, and likely (?) for phaseIII
- issue of tape vs disk storage at KEKCC:
  - hstage tool is working
  - for future (starting from proc8): keep last two dataset on disk (`gpfs` with copy to tape `ghi`)

Experience gained (grid)

- ✗ processing on grid was done (is being done) with the left hand
  - ✓ just setup (need learn about `gb2_prod_tools`), and then fire (and almost forget)
  - need more babysitting and more careful monitoring from my side: need to interact more with computing group, starting now
  - ✓ first experience: setup and submission are well structured, and can be automatized easily for phaseIII processing
  - large task can be problematic
  - coordination with Racha (skim) and Ale (MC) to set ProdID in filename: need to think a good solution

Phase2: toward Proc8

- **pre-proc8 (done)**
  - script to run `pre-proc8`
  - release-03-00-01
  - include some modifications to tracking: Change the CDC ADC threshold from Sasha Glazov
  - EKLM track matching BIIDP-1120 and BKLM additional branches BII-4458
  - GT from Umberto: `data_reprocessing_preproc8`
  - input: HLT skims `["hlt_bhabha", "hlt_gamma_gamma", "hlt_mumu_2trk", "hlt_hadron" (new)]`
  - special sub production with scan on CDC ADC threshold (0-2-4-6-8e^{-7}) Done
- **pre-proc8b (starting)**
  - Same as pre-proc but with release-03-00-02 and updated GT `data_reprocessing_preproc8b` (new payload for SVD)
  - setting up script, will start hopefully today or tomorrow
- **proc8**
  - Most likely will use release-03-00-02
  - aggressive schedule:
    - ✓ [Jan 21] First pre-proc8 with modifications to tracking settings with release-03-00-00 and data_reprocessing_proc7 Done (with delay 26/1)
    - ✓ [Feb 1] Deadline for software modifications to be used for proc8 release-03-00-02 is out
    - 🔧 [Feb 15] Updated calibration constants provided for proc8
    - 🔧 [Feb 18] Reprocessing begins

Experiment 5 Global Cosmic Run

- ✓ Started on 15/1
- ✓ good interaction with computing group (Hara-san)
- 1 Hara-san copy (by hand?) file for offline usage (sroot→root conversion). Only "Cosmic trigger data for detector performance" to be processed
- 1.1 notify that new runs are available BIIDP-1097
- ✓ Register the same files (root) on grid as well, for grid processing Done
- 2.1 DP (me) will process the runs and produce cdst/mdst/(dst)
- 2.2 once done (about 1 hour) they new runs are moved to final location
- 2.3 update confluence page https://confluence.desy.de/display/BI/Experiment+5+-+full+dress+rehearsal
  - As for exp3, need to define release, GT, input, scripts, output
  - Two processing so far (there will be more) **GCR5a** and **GCR5b** (there will be a **GCR5c** in 1-2 weeks)

Exp5 GCR5

First processing GCR5a (stopped)

- **release-03-00-00** and "wrong" GT `data_reprocessing_proc7` (done for phase II, so no SVD-PXD)
- setup script and automation, process new runs as soon as available
- processing stopped as soon as a better GT (and patch release) available
- Eventually will delete these obsolete processing

Second processing GCR5b (running)

✓ using **release-03-00-01** and snapshot of running global tag `data_reprocessing_prompt`

✗ Include better steering file to use PXD and SVD data, non standard cosmic processing (not good!)

  ✗ still no tracking using PXD and SVD (only CDC). Would have required complex mod to steering file

✗ found a (serious) problem in SVD geometry (Giulia). Need a new payload and reprocessing

- two options:

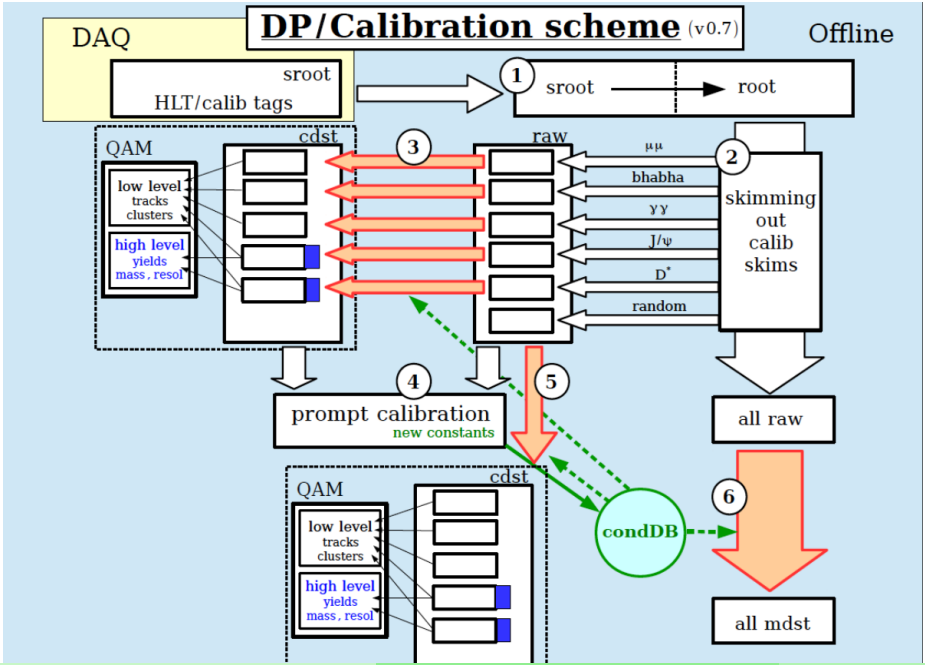
  A wait for new release (possibly with SVD+PXD in tracking) and new GT (with correct Payload for SVD)

  B start immediately GCR5c with custom payload (localDB) (custom steering **and** payload? eech!)

  - Some discussion with tracking and SVD people: we will go for [A], Nils will try to include SVD+PXD in tracking in 1/2 weeks

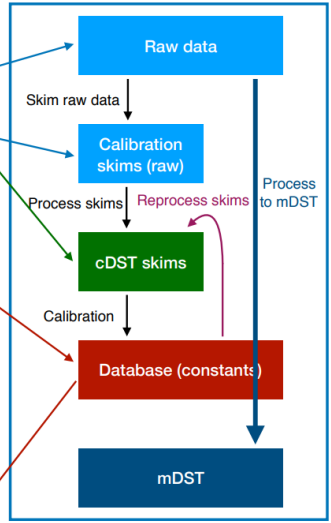
Experience from GCR5

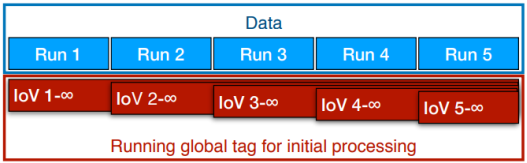
- ✓ Together with PhaseII reprocessing, a ideal playground to gain experience toward phaseIII
- ✓ Communication with computing group is good (Mail, Jira ticket, PullRequest)
- need to understand how to automatize in a robust way the submission of jobs as soon as new runs are available
  - so far I'm running on a dedicated LSF queue from a standard KEKCC node
  - still too much manual work, but scripts are in place and we can automatize most of 2.x
  - watchdog for `RunListCosmic` and submit automatically as soon as new runs are listed under test
  - eventually we will need to do this on the grid (data is already copied)
  - need to play a bit to gain some direct experience (initial thoughts later)
- as for exp3, calibration group is providing GT in a timely manner
  - issue: running GT or snapshot?
  - after this morning GT tutorial, the answer is running GT, when is implemented as described
  - in the meanwhile, I think snapshot is the way to go for reproducibility


DP/Calibration scheme (v0.7)
DAQ
Offline
sroot
HLT/calib tags
1
sroot
root
QAM
cdst
3
raw
μμ
bhabha
γγ
J/ψ
D*
random
2
skimming
out
calib
skims
low level
tracks
clusters
high level
yields
mass, resol
4
prompt calibration
new constants
5
all raw
QAM
cdst
condDB
6
low level
tracks
clusters
high level
yields
mass, resol
all mdst

Fast reprocessing scheme

- ROOT formatted raw data on offline system, registered, and replicated to raw data processing centers
- Calibration skims from raw data processed to cDST at "calibration center"
- Prompt calibration and QAM run at "calibration center" (includes multiple reprocessing to cDST with updated tracking for dE/dx and TOP calibration)
- Constants added to running offline global tag for initial reprocessing
- Latest runs reprocessed to mDST
- When offline calibrations and software updates are complete, reprocess full data to mDST





What we are doing now

Phase II - reprocessing

- HLT-skim for RAW produced (reuse previous production)
- cdst of selected HLT-skim processed
- validation and calibration from cdst
- production of PayLoad (to condDB)
- bug fix need new patch release
- reprocessing of cdst with better code/payload
- validation and further improvement
- re-processing of full phase II data with state-of-the-art calib and code (until next iteration)
- iterate

GCR Exp5 - prompt processing

- produce cdst/mdst from RAW data (no HLT skim)
- validation and calibration from cdst
- new PayLoad (to condDB) and patch
- include new GT (snapshot from calib manager) and patch into steering script (if possible - no patch release)
- re-process all runs and produce cdst/mdst. Stop previous processing
- (removal of obsolete cdst/mdst not done yet)
- Kind of ok since the data collected is not too much, but we need to move closer to processing schema (or re-discuss it)

Toward Phase III (2)

what do we need for step 2.

- Setup of steering script
  - DP responsibilities: need to be standard reconstruction as defined in release
    - what if we need small/quick change?
    - (it has already happened for CGR5)
    - for cdst processing I think that this is kind of ok (not ideal, but we need some flexibility)
    - otherwise we might need fast patch release for data processing (better but slower and possible latency)
  - GT provided by Calibration manager (prompt_data_processing)
    - see next slide: running GT vs snapshot
- **step 2.1** HLT Skim
  - In the schema, it is the first step and is to be done locally
  - current steering is mostly I/O and fast. Steering script mostly ready from phaseII (Karim), likely to need update (eg monopole skim?)
  - skim also for mdst not in the scheme: we had them for phase2 and widely used.
  - when: before or after mdst processing? For phase II now are done after but we are using the processing-1 ones for cdst ...
  - where: mdst will be eventually produced on grid, so need to produce skim there as well, but:
    - yet an other step before profit
    - most of raw skim already done at calibration center for cdst: duplication

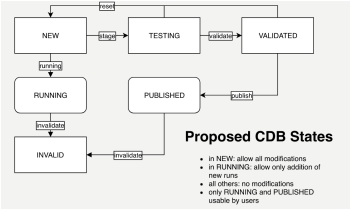
Grid processing

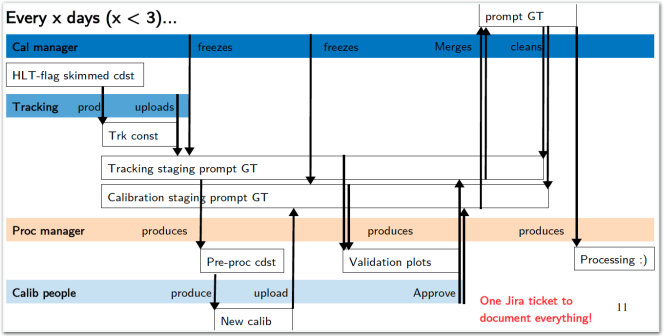
- Not plan A for initial phase III data taking.
- exploit limited luminosity to use local (kekcc) fast processing and re-processing to achieve a reasonably stable operation (unpacker/reconstruction/calibration)
- in parallel run on grid to gain experience for a smooth transition
  - Start already with GCR5
- some issue from my limited experience so far:
  - watchdog for new runs to appears (which is some time after the copy/replication has started)
  - need to develop tool to create `json` with runs to be processed when they are available (should be easy)
  - ProdID for each run or set of runs? Bigger is not better!
  - automation might have some issue (eg to submit I need a voms proxy, which expires in 24h)
  - submission is done via personal grid certificate, which last 24 hours: not possible to setup a fully automatic process (need to renew the certificate every day) certificate renewal service is possible (up to 1 week), still ...
  - other solution/idea?
  - need to use the available monitoring tool to find problems asap (eventually an important task for Data Production shift)

Calibration manager: @Umberto Tamponi

Calibration status and plans

- Gained a lot of experience in phase 2
- Now preparing automation for prompt calibration
- Some missing conditions database functionality will be implemented soon
  - Implement running GTs for prompt reprocessing
  - Workaround planned for early phase 3





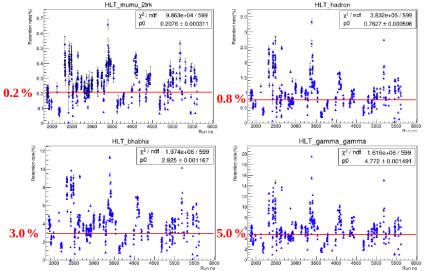
https://confluence.desy.de/display/BI/Data+Production+Calibration

Calibration skim manager: @Trabelsi Karim

Calibration skim development

- Important to reduce computing requirements for calibration and enable prompt calibration
- Converging on new skims (γγ, ECL Bhabha, TOP dimuons, J/ψ, etc.)
- Will promote new (offline) skims to the HLT
- Plan to only provide raw/cDST for experts in phase 3!
  - Will review cDST use/size soon

Category	Skim name	Selection
HLT	hlt_mumu_2trk	[[nTracksLE >= 2] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.35] and [[P2OEbeamCMSBhabhaLE > 0.2] and [[EtotLE < 7] and [[EC2CMSLE < 1] and [maxAngleTTLE > 0.785]]]]]]]
HLT	hlt_mumu_1trk	[[nTracksLE == 1] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.1] and [[EC1CMSLE < 1] and [EtotLE < 7]]]]
HLT	hlt_hadron	[[nTracksLE>=3] and [Bhabha2Trk==0]]
HLT	hlt_bhabha	[Bhabha2Trk==1]
HLT	hlt_gamma_gamma	[[nTracksLE <= 1] and [[nEidLE == 0] and [[EC12CMSLE > 4] and [EC1CMSLE > 2]]]]
HLT	hlt_single_photon_1GeV	[[G1CMSBhabhaLE>1.0] and [Bhabha2Trk==0] and [GG==0]]
HLT	hlt_single_photon_2GeV_barrel	[[G1CMSBhabhaLE>2.0] and [Bhabha2Trk==0]]
HLT	hlt_single_photon_2GeV_endcap	[[G1CMSBhabhaLE>2.0] and [Bhabha2Trk==0] and [GG==0]]
HLT	hlt_tau_tau	[[nTracksLE >= 2] and [[P1CMSBhabhaLE < 6] and [[EtotLE < 9] and [VisibleEnergyLE < 9]]]]



cDST size:

- Typical file is 1.43 GB/22 k evts = 66 kB/evt
  - ECLCalDigits: 645 MB (45%)
  - ECLDigits: 216 MB (15%)
  - ExtHits: 162 MB (11%)
  - CDCDedxTracks: 84 MB (6%)
  - RecoTracks: 80 MB (6%)

  } sum up to 83%

  - TOPDigits: 46 MB
  - SVDRecoDigits: 43 MB
  - SVDShaperDigits: 19 MB

  } but only few SVD ladders !

  - ECLClusters: 15 MB
  - TrackFitResults: 6.0 MB
  - SoftwareTriggerVariables: 5.9 MB
  - SVDClusters: 3.8 MB
  - ARICHDigits: 2.9 MB
  - ...

reference: mdst size is 1 kB/evt, raw 30-45 kB/evt, dst 120 kB/evt

https://confluence.desy.de/display/BI/Experiment+3+skims

Skim definition

Category	Skim name	Selection	Comments
HLT	hlt_mumu_2trk	[[nTracksLE >= 2] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.35] and [[P2OEbeamCMSBhabhaLE > 0.2] and [[EtotLE < 7] and [[EC2CMSLE < 1] and [maxAngleTTLE > 0.785]]]]]]]	
HLT	hlt_mumu_1trk	[[nTracksLE == 1] and [[nEidLE == 0] and [[P1OEbeamCMSBhabhaLE > 0.1] and [[EC1CMSLE < 1] and [EtotLE < 7]]]]]	
HLT	hlt_hadron	[[nTracksLE>=3] and [Bhabha2Trk==0]]	
HLT	hlt_bhabha	[Bhabha2Trk==1]	no more prescale from prod3
HLT	hlt_gamma_gamma	[[nTracksLE <= 1] and [[nEidLE == 0] and [[EC12CMSLE > 4] and [EC1CMSLE > 2]]]]	
HLT	hlt_single_photon_1GeV	[[G1CMSBhabhaLE>1.0] and [Bhabha2Trk==0] and [GG==0]]	
HLT	hlt_single_photon_2GeV_barrel	[[G1CMSBhabhaLE>2.0] and [Bhabha2Trk==0]]	
HLT	hlt_single_photon_2GeV_endcap	[[G1CMSBhabhaLE>2.0] and [Bhabha2Trk==0] and [GG==0]]	
HLT	hlt_tau_tau	[[nTracksLE >= 2] and [[P1CMSBhabhaLE < 5] and [[EtotLE < 9] and [VisibleEnergyLE < 9]]]]	

Running global tag

- running GT is, by construction, open (NEW)
- namely can change after being used to process a given run range.
- update can be forward update, namely valid from a given run to infinity (by policy, by design, or by gentle-person agreement?)
- scenario 1.
  - we produce cdst from run X to run X+10
  - calibration team analyze them, and come up with update payload, IoV $X - \infty$
  - we wait news from calibration team for a possible updated payload before producing mdst for physics
- scenario 1.2
  - calibration a week later: no wait, we do have even better payload for IoV $X - \infty$
  - calib upload payload to running GT with IoV $X - \infty$
  - we have mdst produced with not-up-to-date for $X - X + n$ runs.
  - we don't care (but it would be hard for analysis to understand what happened)
  - we do care, and reprocess run $X - X + n$ runs (and DP goes crazy pretty fast)
- processing with snapshot of runnig GT would guaranteed to know precisely what have been used for processing that run
- would need to be updated regularly by calib coord (which might go crazy . . .)
- not clear to me.

Bibliography I