Data Processing

BPAC review, 28/10/2019

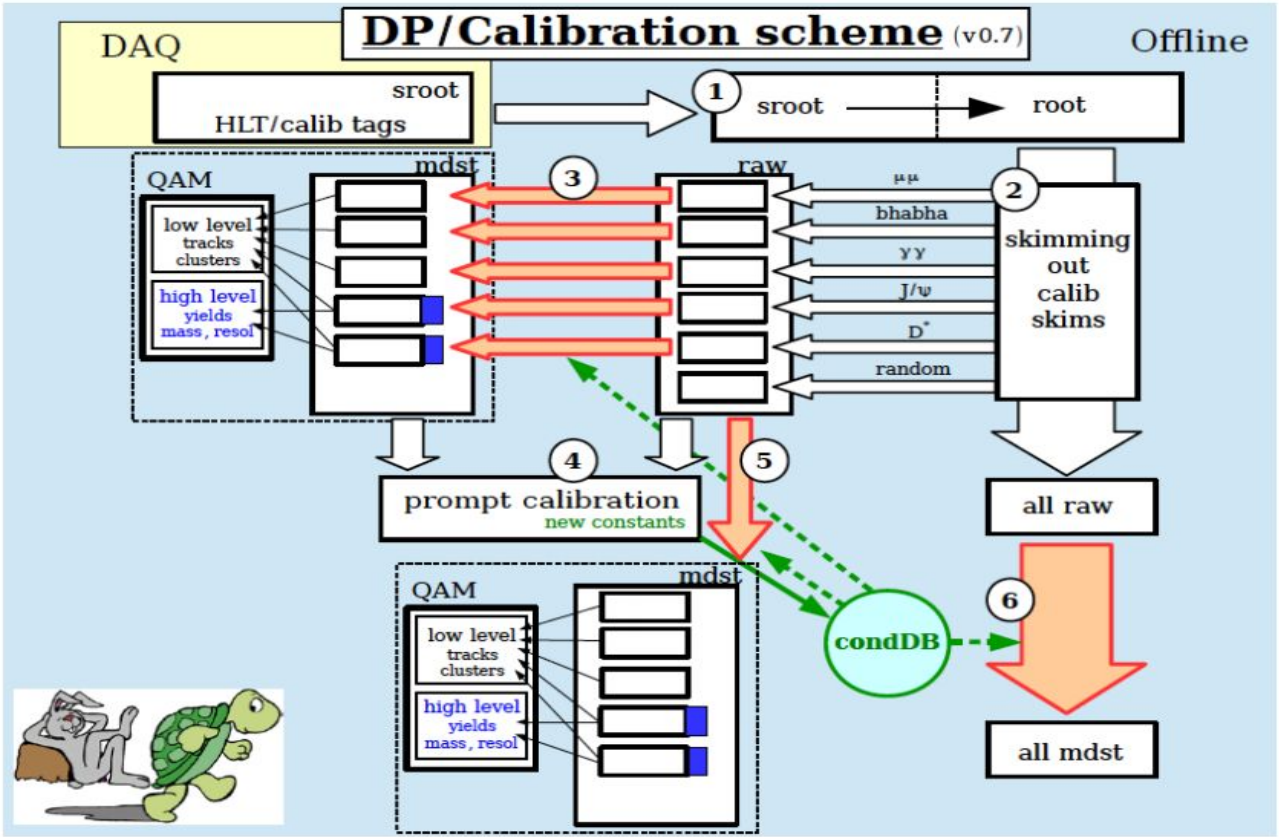
Stefano Lacaprara
INFN Padova

Outline

- Experience from 2019a run
- Data processing schema
- Plan toward winter/summer 2020 conferences

Experience from 2019a run

Processing and calibration schema



Official Phase 3 data processing 2019a

- Current state of the art: **Proc9 + Prompt.**

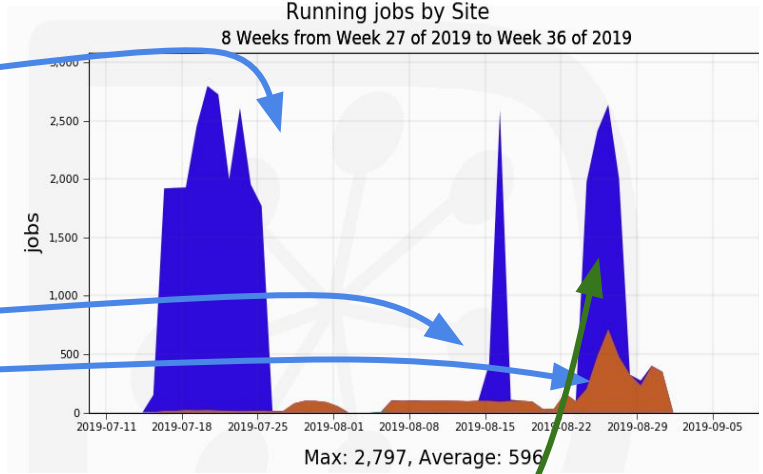
[1] arXiv:1910.05365

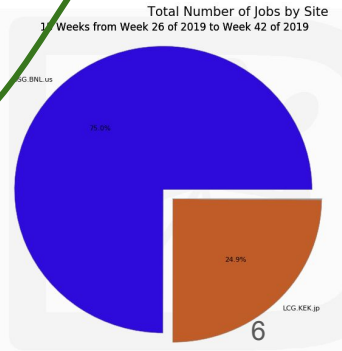
Exp	Energy	Run range	Lumi good runs	Total good lumi	Lumi all runs	Campaign
3	4S	529-5613	**496.7±0.3±3.5/pb[1]**		-	Proc 9 August 2019
7	4S	909-4120	**555.62 /pb**	**5.15 fb-1 (4S)** **6.01 fb-1 (total)**	642.8 /pb	
8	4S	43-1022, 1036-1554	**1827 /pb**		1982.3 /pb	
	Continuum	1703-1835	**39.02 /pb**		39.02 /pb	
	Scan	1025-1031	**826.79 /pb**		827.0 /pb	
	4S	1836-3123	**2764 /pb**		2973 /pb	Prompt (Bucket7) Sept 2019

- Proc9 and Bucket7 both at KEKCC and on GRID
  - All info at: https://confluence.desy.de/display/BI/Phase+3+data
  - Next iteration **Proc10** is underway

Proc9/Bucket7 on the grid

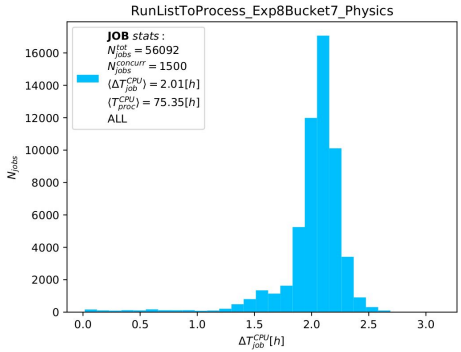
- Start on 13/7: fast start at BNL (and KEKCC)
  - BNL finish after 10 days, KEKCC goes on slowly
  - KEKCC stop due to shutdown (August)
- KEKCC restart (slow)
  - Some runs with more than 1000 jobs not properly handled by DP tools: fixed. Immediately run at BNL
  - Increase of resources at KEKCC to finish
- Overall, took >1 month to complete!
  - At KEKCC ~10 days
- Bucket7 started and finished quickly

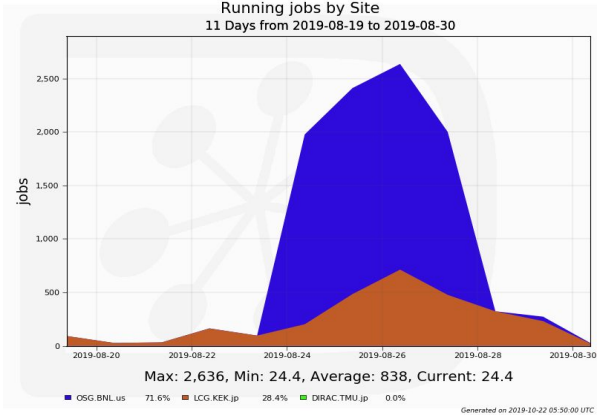




Lesson learned

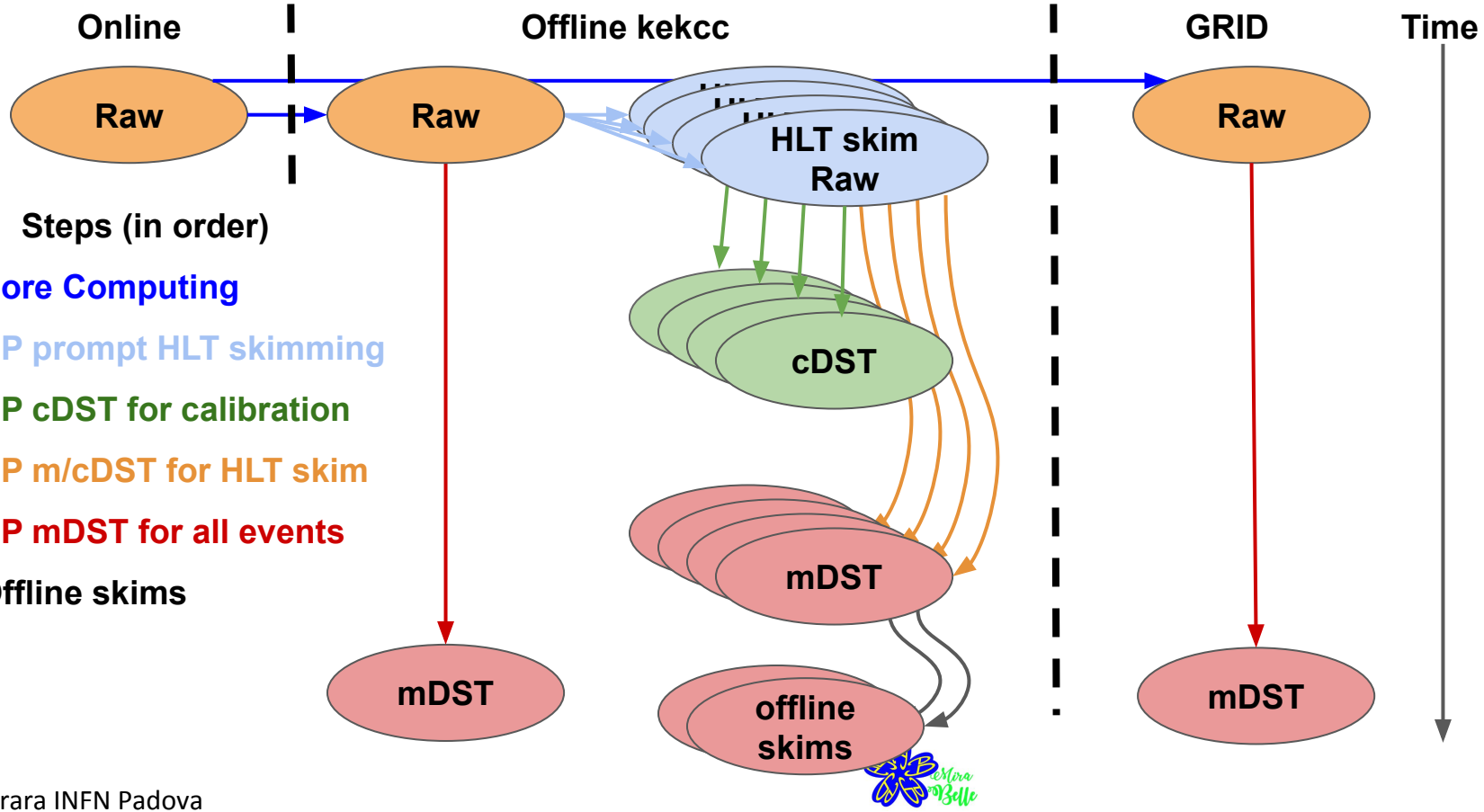
- DP tools used successfully
  - A new processing can be setup in very a short time
  - Bookkeeping, troubleshooting, monitoring, stats, etc very well automated
    - Discussion for sharing with MC/analysis skim
- Major issue with Proc9 on grid was **raw data staging**
  - All data staged, but only partially at BNL, the rest at KEKCC
    - KEKCC was limited to stress test BNL
  - Jobs created and scheduled according to data availability, so many at KEKCC
    - Should have waited to complete BNL staging
    - Introduced step to check staging before submission
- Fixed for Bucket7 processing:
  - Processing 100% done in 3.5 days
    - At KEKCC 3 days



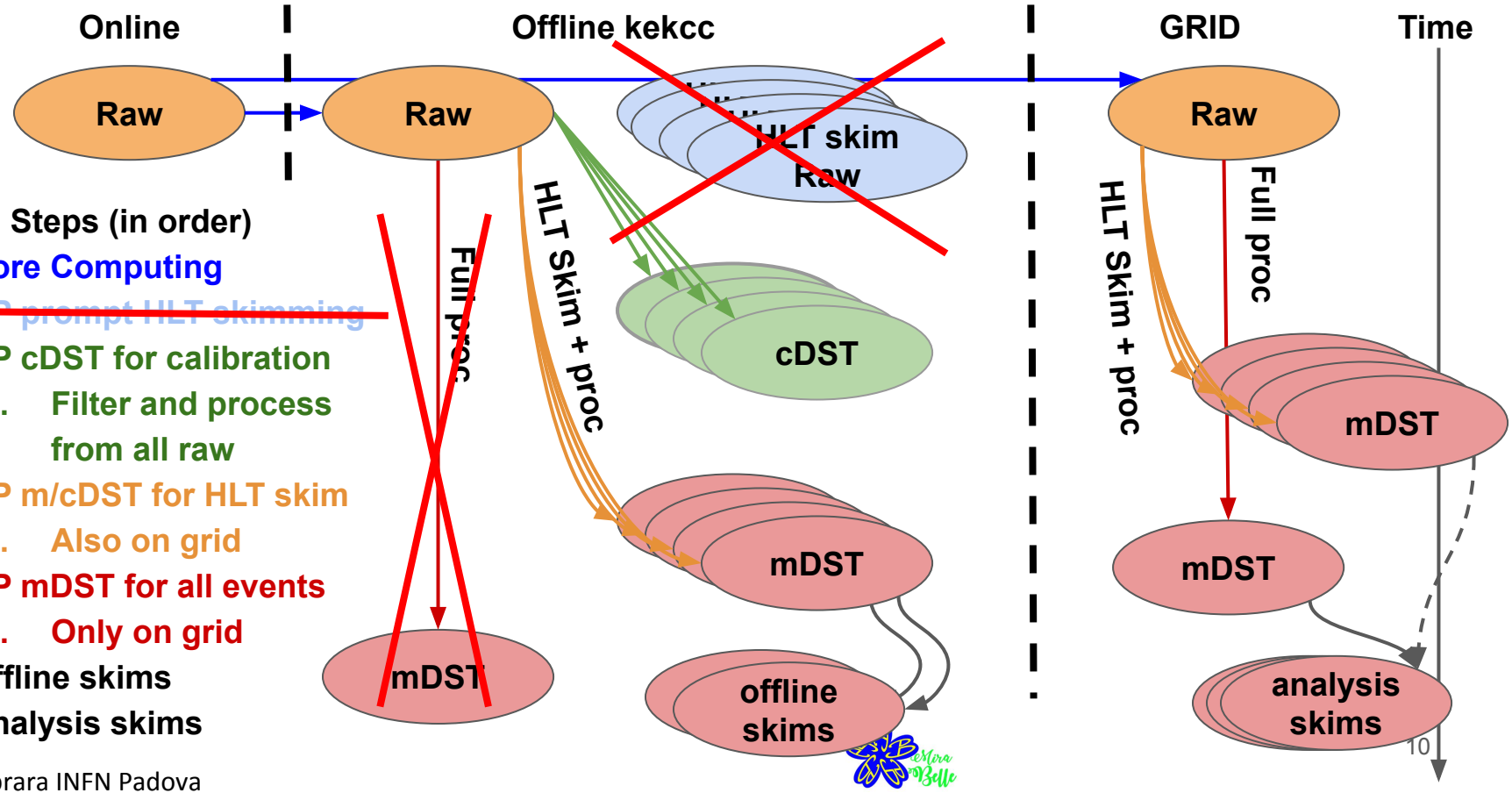


Data Processing schema

Data Processing schema (so far)



Data Processing schema (future)



Steps (in order)

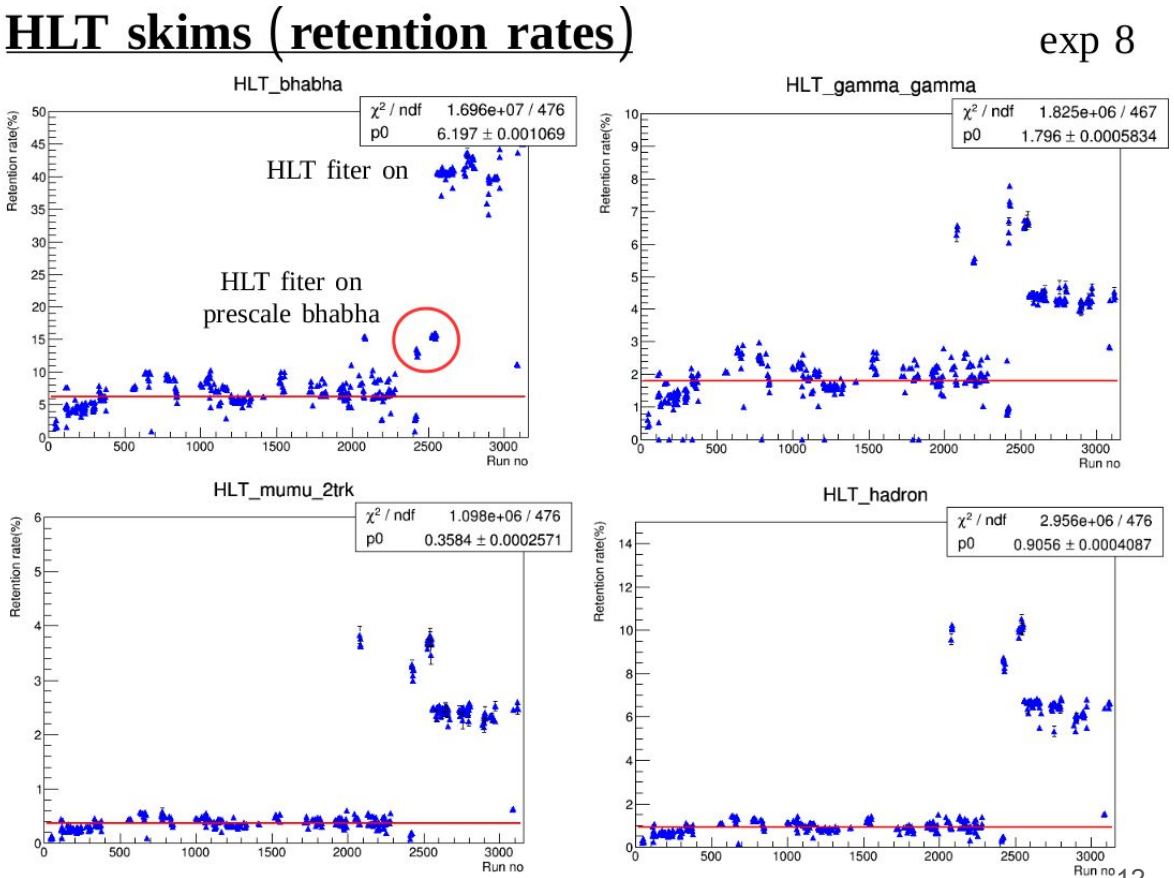
- Core Computing
- ~~DP prompt HLT skimming~~
- DP cDST for calibration
  a. Filter and process from all raw
- DP m/cDST for HLT skim
  a. Also on grid
- DP mDST for all events
  a. Only on grid
- Offline skims
- Analysis skims

The role of skims: HLT and Analysis

- Selection based on HLT objects computed online for calibration
  - Selection w/o reconstruction: fast, just I/O
    - `bhabha, digamma, mumu_2trk, hadron`
  - Other (`bhabhaecl, radee, tautau`) will be added
- Very popular, especially hadrons (for physics)
  - Good enough with limited luminosity
- **offline** skims (`D*, mumu, mumu_top`) for **Quality Assurance Monitor (MiraBelle)**
- Will start providing Analysis Skims for Data from next reprocessing
  - Offline skims targeted for analysis: one for WG
    - Already done for Bucket7, only 1 WG requested them
  - Will encourage more general usage
    - Faster production, little loss in retention, less data to process by users

HLT filtering effect

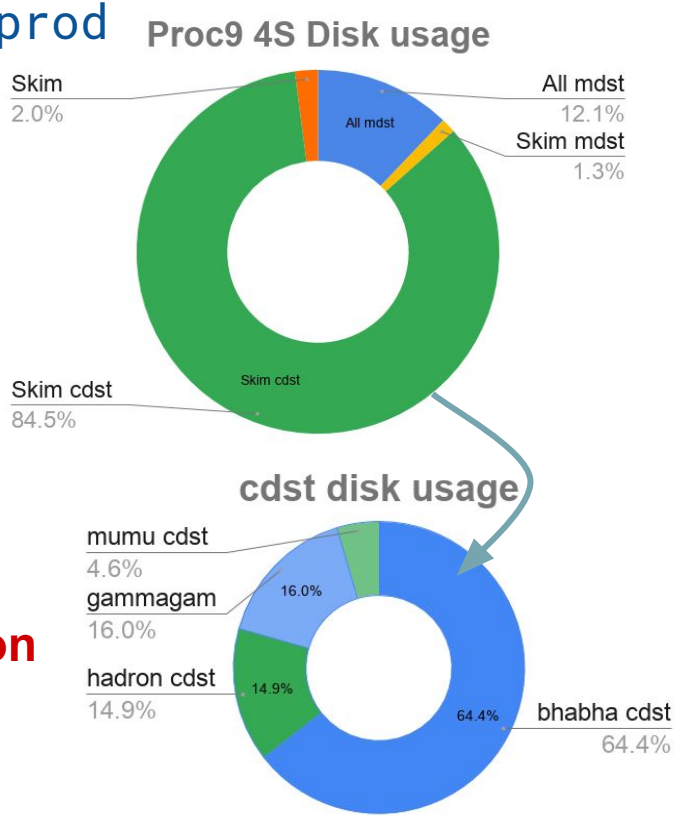
- Retention rate was ~10%
  - Dominated by bhabha (~6%)
  - With HLT in tag-mode
- Now HLT filter on
  - Retention rate is much higher
    - **~50%**
- Dataset size ratio between HLT skims and all events is about **2x**
  - Was **10x**



Disk space (local kekcc)

- Local staging issue from `/ghi` solved with `/dataprod`

- Data processing need about 10 TB/ fb^{-1}
  - Including cdst for HLT skim and mDST for all events
  - Mostly cDST (85%), mostly bhabha
- `/dataprod` has 1.5 PB of space
  - Currently we are using about 0.5 PB
  - Will do some cleanup after Proc10 (will be announced)
- No issue of disk space for Proc10 and 2019c
- **Local resources will be used only for calibration and detector studies**



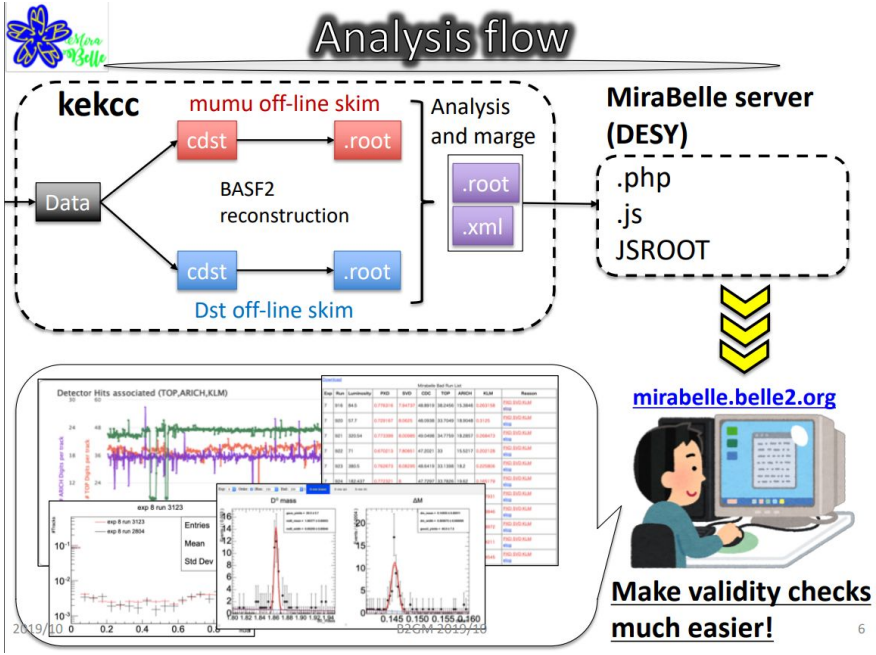
Time budget with current resources

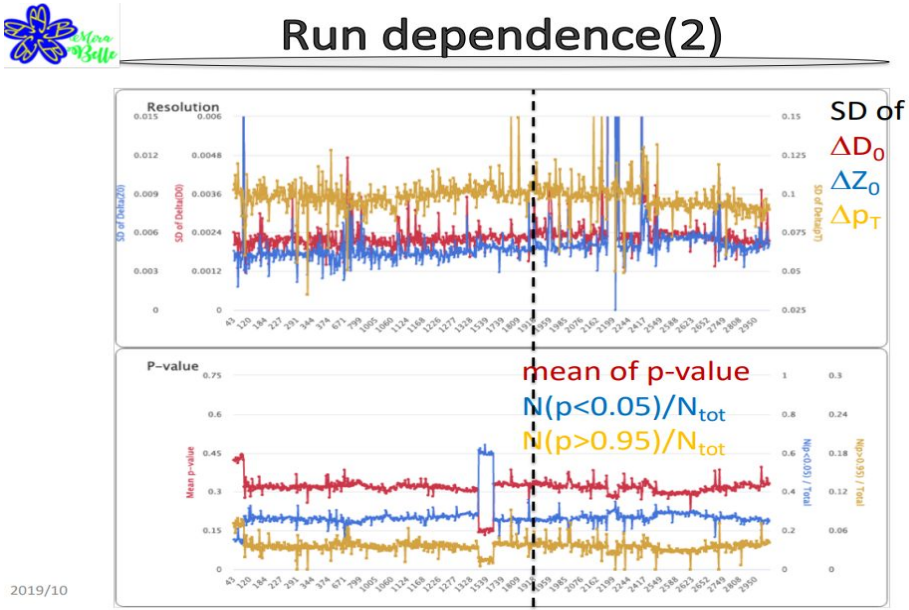
	HLT tag	HLT filter	Note:
mDST from HLT skim for calibration	**0.3 day/fb^{-1}**	**0.3 day/fb^{-1}**	Local processing
mDST from HLT skims	**0.3 day/fb^{-1}**	**0.3 day/fb^{-1}**	Same time as cDST
mDST all events	**1 day/fb^{-1}**	**0.5 day/fb^{-1}**	On GRID

- Improvement with **more resources** and **prioritization**:
  - Can run first on physics relevant HLT skims
  - e.g. HLT hadron is 6% of full dataset
    - Speed up 16x: 2 fb^{-1}/day -> 30 fb^{-1}/day

QAM: Mirabelle

- Run Quality looking at several detector/physics quantities
  - Use **OfflineSkims**
  - QAM also with online DQM, detector experts, etc

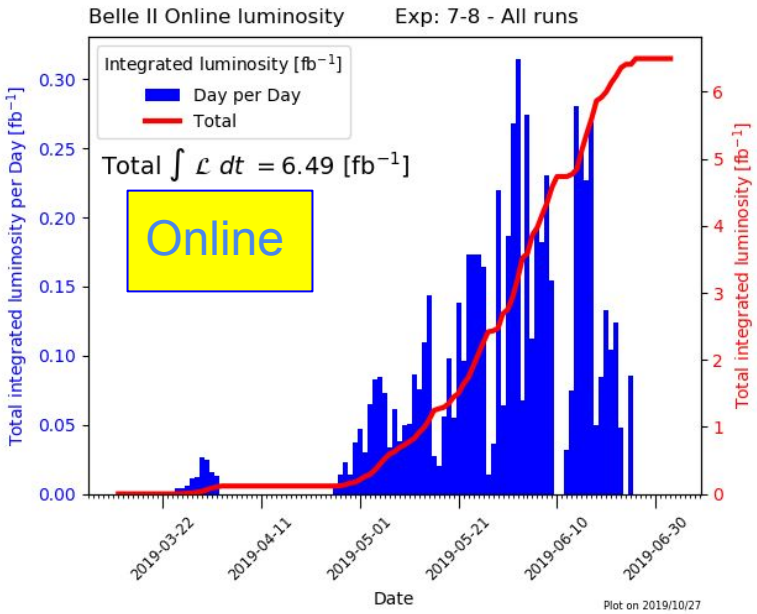


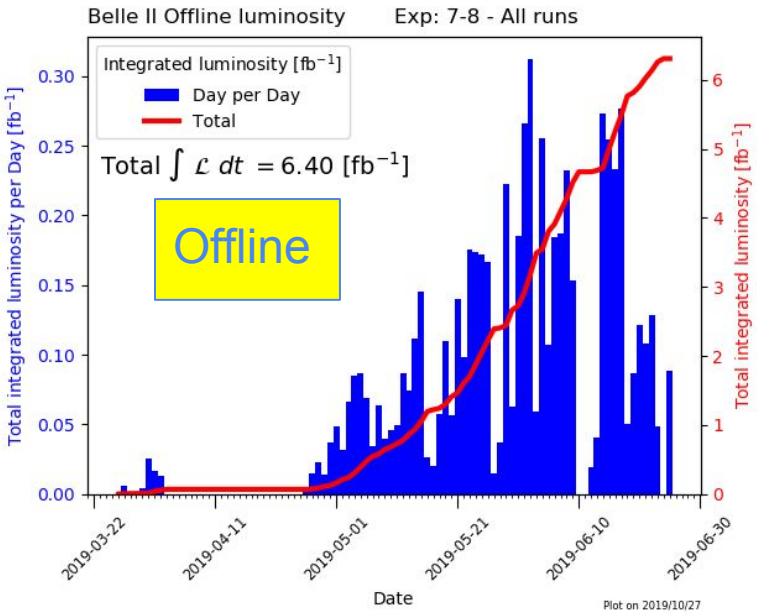


- Provide **GoodRun** list for physicists
  - Dedicated path (symlink) .../proc9/exp8/4S/**GoodRuns**/ …
- Will use RunRegistry asap
  - See Martin's talk later

Luminosity

- Last step of processing is provide offline luminosity
  - Done by luminosity group as soon as mDST data is available
  - Need mDST for all events to reduce systematic uncertainty
    - Will use also bhabha skims for first, fast determination, with larger uncertainty



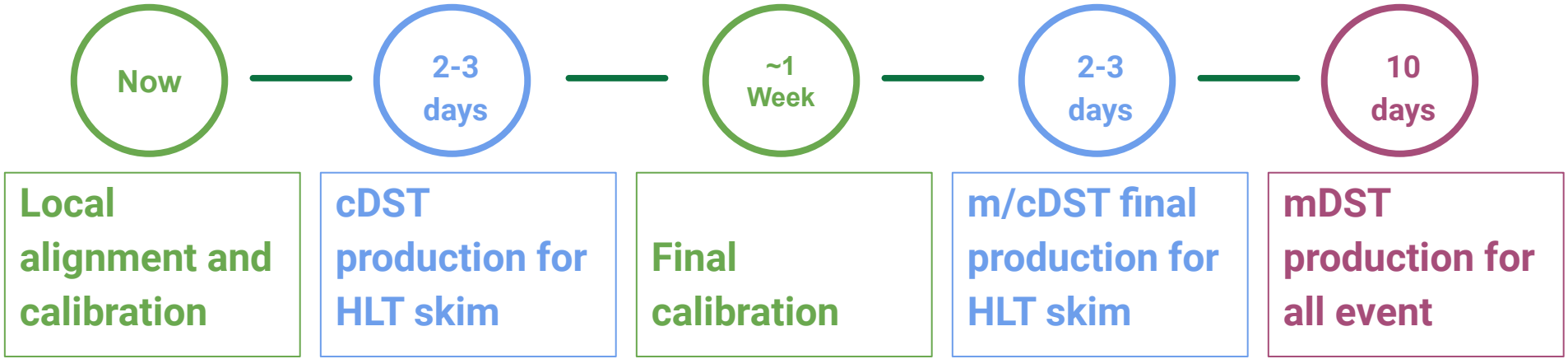


Offline GoodRuns 6.01 fb^{-1}

Plan toward winter/summer conferences

Proc10: full 2019a reprocessing

- Release4 - only phase 3 data, exp 7+8 (~6 fb^{-1})
  - No phase2 - exp3 reprocessing



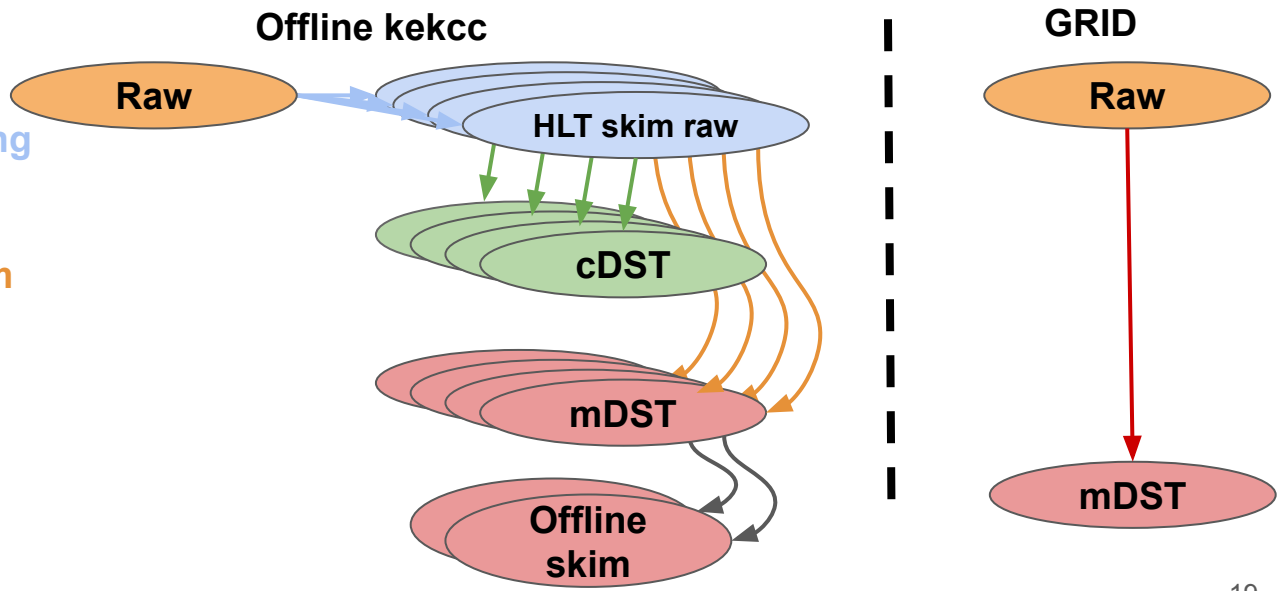
- Final step on grid
  - Also at KEKCC if no conflict with 2019c processing
- **Should be ready by late-mid november**

Readiness for 2019c run (fall)

- Release 4 tested on exp8 data and exp10 cosmic
  - Some issues found and fixed in patch release 04-00-01
- **Processing schema as in spring data taking:**
  - We can provide a **fast "unofficial"** uncalibrated (namely using old calibration) if needed

Steps

- DP prompt HLT skimming
- DP cDST for calibration
- DP m/cDST for HLT skim
- DP mDST for all events
  a. Grid only
- Offline skims



Plan for Winter conferences.

- Winter 2020 L~20 + 6 fb^{-1}
  - End of data taking mid december: **5 weeks available**
- Goal full reprocessing by late January
  - cdst processing at KEKCC **~8 days**
    - Second processing in parallel with full reprocessing
  - mdst processing on grid: **15 days**
  - Schedule fits, but not much room for delay or issues
    - Can prioritize HLT skim usable for physics for significant speedup
- **Will be a crucial scale test of readiness for calibration and processing for spring run**

Summer conferences 2020

- KEKCC replacement and schedule for summer conferences require full processing finished by the end of July
  - Start on late spring (?)
  - Plan to use release-4 (or minor release)
    - See Martin's talk
  - Unlikely that we can officially reprocess all of the data collected in June
- **Goal: officially process as much data as possible, plus prompt calibration and processing for the rest**
  - Precise timeline for processing will depend on experience gained with automated calibration tools

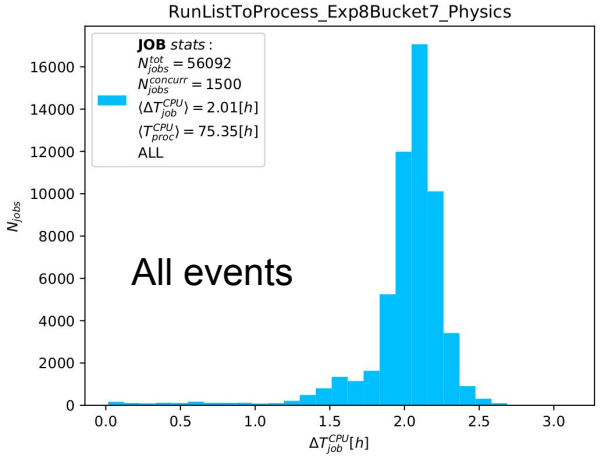
Summary

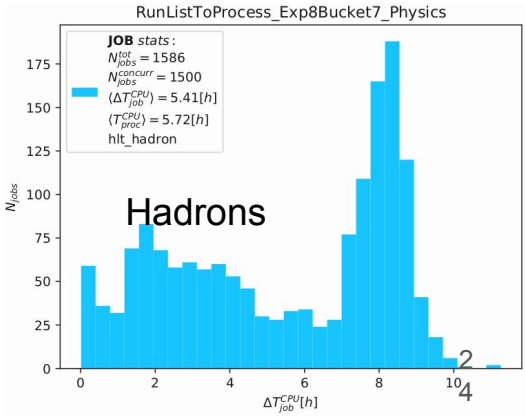
- **Data processing in 2019b experience invaluable**
  - **We did provide Processed Data to the collaboration in a timely manner**
  - Many lessons learned
    - DP tools stable and satisfactory
    - Local processing at kekcc very stable
    - When properly done, works very well also on grid.
- **Schema of data processing is established**
  - Evolving as we gain experience
- **Plan for 2019c run is in place**
  - Focus on continuous processing
  - reprocessing for winter conferences
  - Initial plan for summer 2020

Backup slides

Proc9 and Bucket7 processing

- Processing at KEKCC
  - RAW HLT skim as soon as data copied to offline
    - Bhabha, gamma-gamma, hadron, mumu_2trk
  - cdst from hlt skim for calibration
  - mdst and cdst for HLT skims
    - ~11'000 CPU-h/(fb^{-1}):
    - With 400/1000/1500/2000 job slots: 1.2/0.5/0.3/0.23 day/fb^{-1}
  - mdst for all events
    - ~1 day/(fb^{-1}) with 1500 slots: 3x HLT_skims
    - Disk space (total) ~10TB/(fb^{-1}) - mostly cdst
  - Offline hlt
    - D*, mumu, mumu_top
- In parallel processing of mdst for all event on the grid
  - Next slide





Run Registry: https://rundb.bell2.org

Collect all infos about each run from online and later (QAM, luminosity, …)

Run List

Previous 1 2 3 4 5 6 7 8 9 10 11 Next

Experiment First Run Search

	Exp	Run	Run Type	Start Time	Run Time	Stop Reason	Detectors	Energy in GeV	Magnet in T	Triggers	Events	Luminosity Online	Luminosity Delivered	Luminosity Recorded
Details	10	2594	null	2019/10/23, 11:28		PAUSED	PXD | SVD | CDC | TOP | ARI | ECL | TRG	HER: 7.01 LER: 4.00	1522202.00	In: None Out: 16114	Total: None Hadronic: None Bhabha: None	0.18		
Details	10	2593	physics	2019/10/23, 11:22		STOPPING	PXD | SVD | CDC | TOP | ARI | ECL | TRG	HER: 7.01 LER: 4.00	1522201.00	In: None Out: 58077	Total: None Hadronic: None Bhabha: None	0.19		
Details	10	2592	null	2019/10/23, 11:07		STOPPING	PXD | SVD | CDC | TOP | ARI | ECL |	HER: 7.01 LER: 4.00	1522201.00	In: None Out: 53811	Total: None Hadronic: None Bhabha:	0.19		

Quality Flag Summary

	PXD	SVD	CDC	TOP	ARI	ECL	KLM	TRG
Shifter	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET
Expert	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET	NOT SET

Show Quality Flags History Edit

Detectors

Detector	Status	HV
PXD	RUNNING	PEAK
SVD	RUNNING	PEAK
CDC	RUNNING	PEAK
TOP	RUNNING	STANDBY
ARI	RUNNING	PEAK
ECL	RUNNING	None
KLM	OFF	PEAK
TRG	RUNNING	None

Magnetic Field

Detail	Value
Solenoid Field	1522202.0
Focussing left Current	None
Focussing right Current	None
Compensation left Current	None
Compensation right Current	None

Statistics

Detail	Value
Incoming Triggers	None
Outgoing Triggers	16114
Stored Events	None
Hadronic Events	None
Bhabha Events	None
Online Luminosity ECL	0.0
Online Luminosity Belle II	0.1820011066955084

Active and collecting info from online: feedback about what else to add/change

Will be used by QAM to mark Good/Bad runs

HLT_skims

Skim name	Selection
hlt_bhabha	[Bhabha2Trk==1]
hlt_gamma_gamma	[[nTracksLE <= 1] and [[nEidLE == 0] and [[EC12CMSLE > 4] and [EC1CMSLE > 2]]]]
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_hadron	[[nTracksLE>=3] and [Bhabha2Trk==0]]