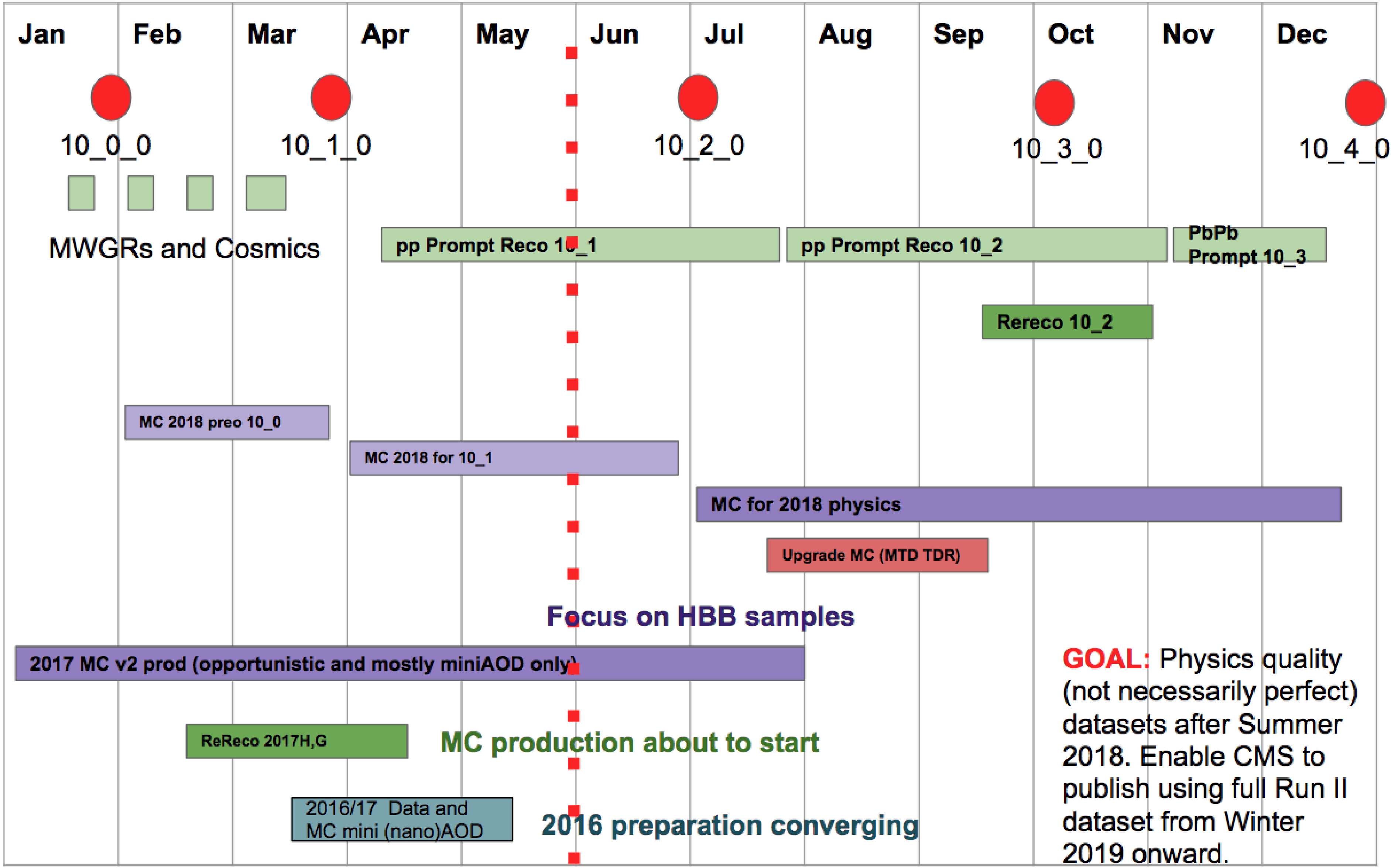
CMS REPORT – LHCC 134

Tommaso Boccali (INFN-PISA)
Liz Sexton-Kennedy (FNAL)

Outline

- Distributed computing status
- Data taking status
- Plans for 2018
- News and improvements
- Preparation for RunIII, RunIV: ongoing activities

2018 Production and Release Schedule



Focus on HBB samples

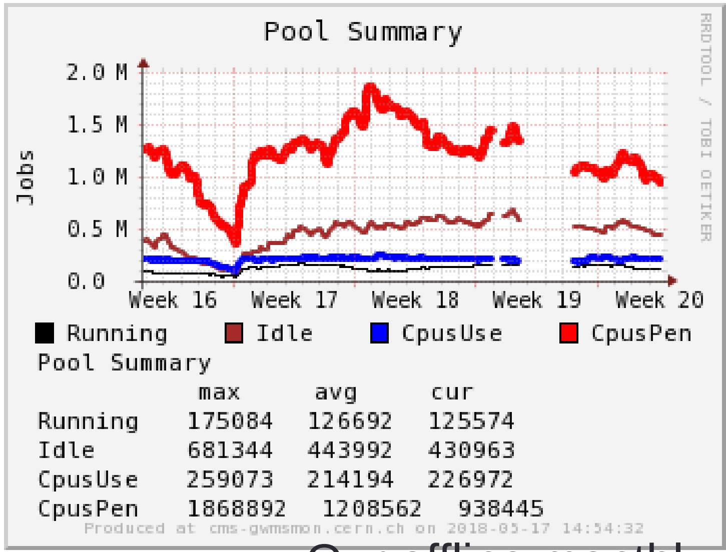
MC production about to start

2016 preparation converging

GOAL: Physics quality (not necessarily perfect) datasets after Summer 2018. Enable CMS to publish using full Run II dataset from Winter 2019 onward.

Distributed Computing Operations

- Full utilization of Distribute Resources is the norm since long
  - Including T0 and HLT in the YETS
  - Many fronts open:
    - MC2018 initial campaigns (HLT, Object calibration)
    - MC2017 (continuing MCv2, Re-MiniAOD and Re-NanoAOD)
    - PhaseII for continuing studies + Yellow report + MTD TDR
  - Record was 12k workflows injected in one day
  - Main worry at the moment is the increased load on debugging workflow problems; trying to find a solution (PH+COMP+PPD)



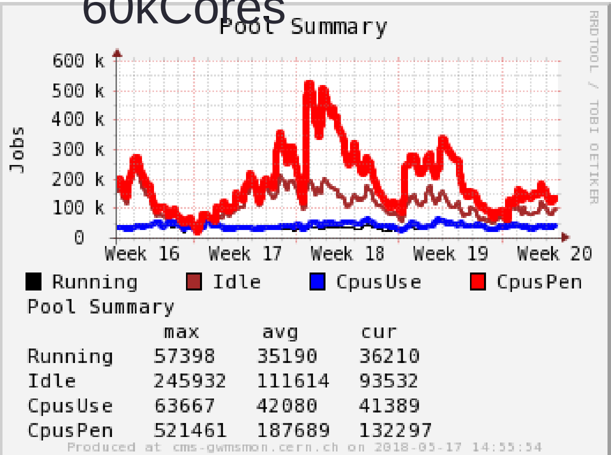
Pool Summary

	max	avg	cur
Running	175084	126692	125574
Idle	681344	443992	430963
CpusUse	259073	214194	226972
CpusPen	1868892	1208562	938445

Produced at cms-gwmsmon.cern.ch on 2018-05-17 14:54:32

Our offline monthly utilization (while in data taking) is 200-250 kCores

Analysis uses 40-60kCores



Pool Summary

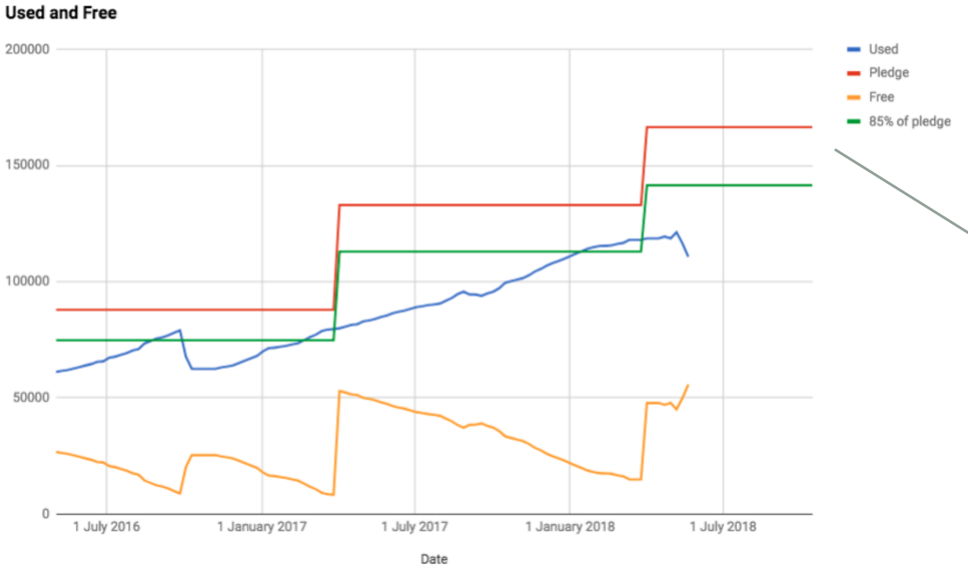
	max	avg	cur
Running	57398	35190	36210
Idle	245932	111614	93532
CpusUse	63667	42080	41389
CpusPen	521461	187689	132297

Produced at cms-gwmsmon.cern.ch on 2018-05-17 14:55:54

Some notable facts

- A new tape cleaning campaign has started, should clean O (25 PB) at Tier-0 and Tier-1s
  - As expected in the new operational mode, most of GEN-SIM (Geant4) samples are deleted after ~ 1y if produced at all
- Actual deletions not complete (sites will approve at their preferred moment – then repack!)

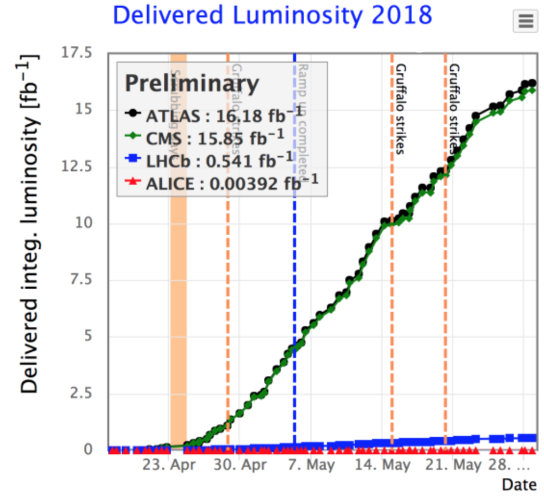
T1 tape evolution (Data Management view) 2016-now

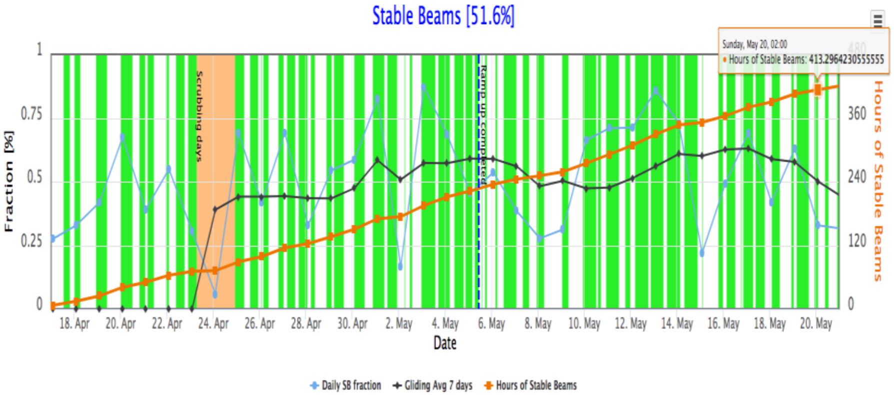


Available tape is in the middle (depends on repacking, …)

Data taking 2018

- We are at 16/fb, 490 hours of SB
  - 25+% of data taking “done”
- CMS Tier0 largely different from 2017 setup
  - Tier-0 and Tier-2@CERN merged
    - CPU and EOS
  - Agile → HTCondor
  - CPU resources in fairshare with other experiments (no static allocation)
- **Pros:**
  - Easier to run production @ CERN (no flocking from another pool)
  - No need to overflow Prompt processing to T2 explicitly (there is no separated T2)
  - Easier to manage storage areas (and to increase the Tier-0 buffers in case of problems)





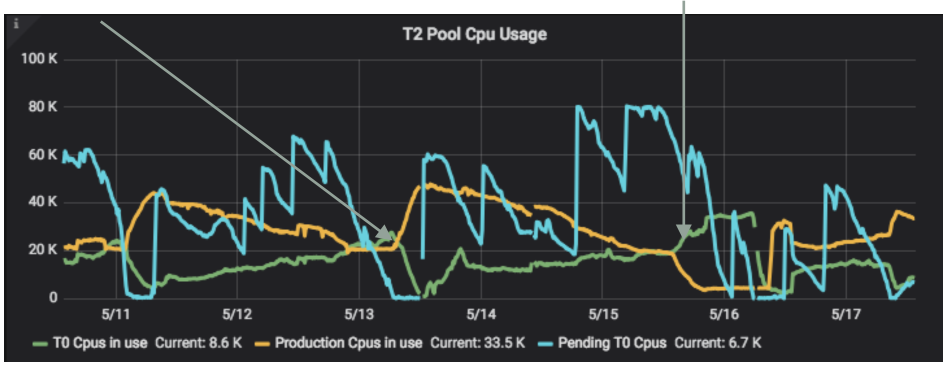
- **“Cons”:**
  - No static allocation: slower in grabbing resources for Tier-0
  - Need discipline in “T2” disk areas not to overflow in T0 buffers

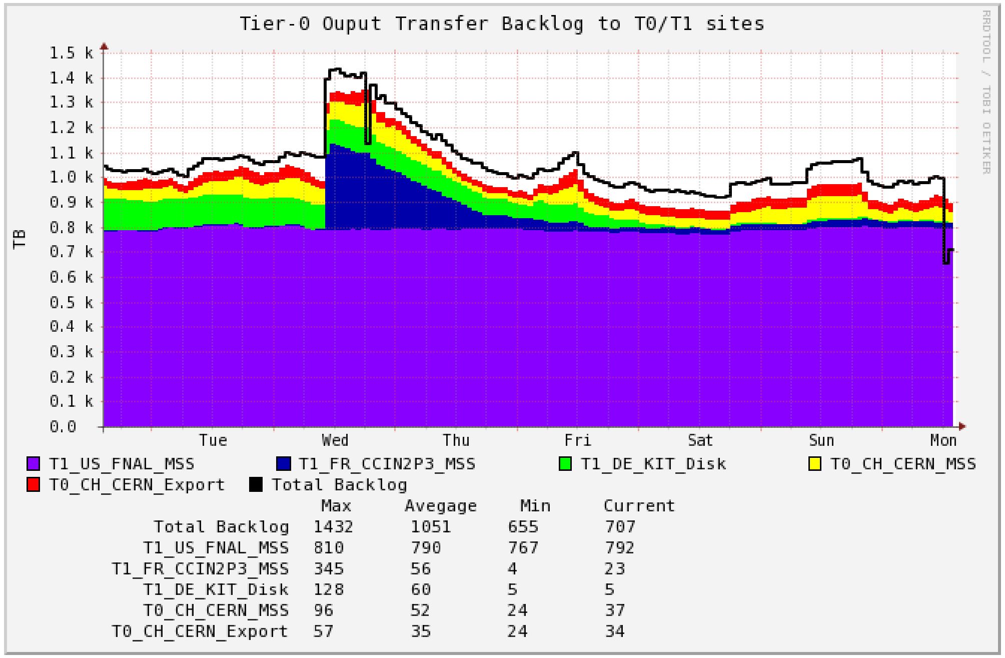
The New Tier-0

- Still tuning HTCondor settings, but basic functionality present
- Storage areas:
  - 16 PB assigned to T[0+2] main storage area
  - Only Express and Input areas separated from the main area
- Data Transfer Backlog to Distributed Sites
  - Only relevant one is a ~ 1PB to FNAL, being analyzed

The T0 is empty, production can take all

When T0 jobs pending, production goes down



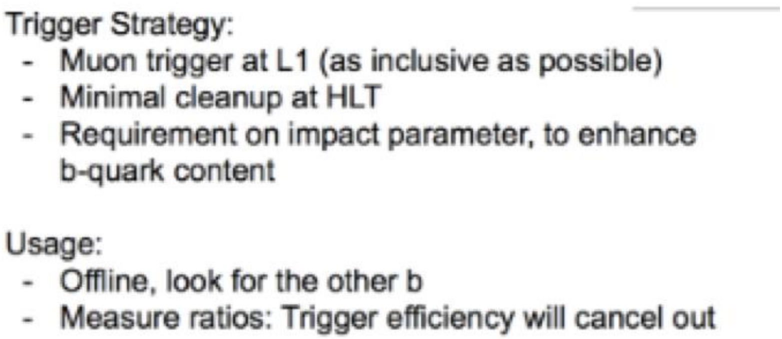


	Max	Avegage	Min	Current
Total Backlog	1432	1051	655	707
T1_US_FNAL_MSS	810	790	767	792
T1_FR_CCIN2P3_MSS	345	56	4	23
T1_DE_KIT_Disk	128	60	5	5
T0_CH_CERN_MSS	96	52	24	37
T0_CH_CERN_Export	57	35	24	34

B - parking

CMS is attempting to collect a **large dataset enriched in B physics**. One specific and one general use cases:

- Allow CMS to measure R_K and R_{K*} in a competitive way
- Prepare a O(10 B) sample of unbiased B hadron decays
  - Trigger on “the other B”
- How: on average, we need to increase our parking rate from 500Hz to 2kHz
  - This collects ~10B of Bs
- This is new: after a lot of internal discussions, green light on May 10th



Trigger Strategy:

- Muon trigger at L1 (as inclusive as possible)
- Minimal cleanup at HLT
- Requirement on impact parameter, to enhance b-quark content

Usage:

- Offline, look for the other b
- Measure ratios: Trigger efficiency will cancel out

$$R_{K^{(*)}} = \frac{\mathcal{B}\left(B \to K^{(*)}\mu^+\mu^-\right)}{\mathcal{B}\left(B \to K^{(*)}e^+e^-\right)}$$

How?

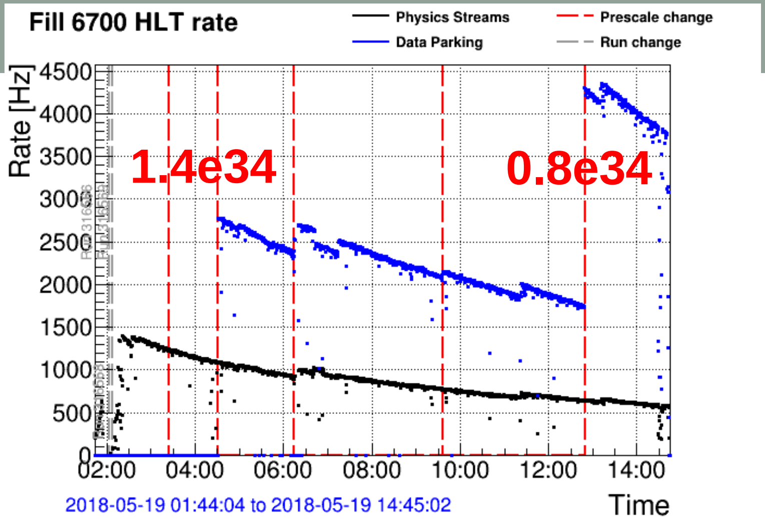
When?

Where?

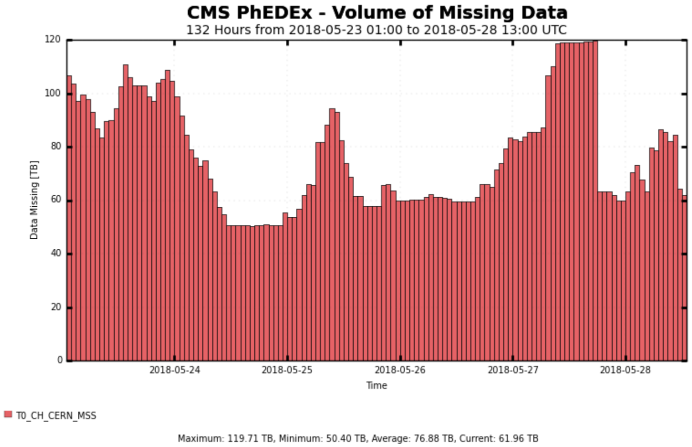
B – parking strategy and operations

- +1.5 kHz of parked trigger rate would exceed our
  - Tape @ CERN and @ Tier-1s
  - Transfer bandwidth to Tier-1s
  - EOS Tier-0 Disk Buffers
- Solution:
  - Bs have high xsec, take them @ high rate when the PU is low (second part of the fill)
- Current strategy (preliminary)
  - 0 Hz when lumi > 1.4e34
  - ~2 kHz between 1.4-0.8e34
  - ~4 kHz wshen lumi < 0.8e34
- In this way, effective rates depend on fill lifetimes; they will be monitored

HLT rates



T0 tape backlog less than a day of data taking



The added pressure on Tier-0 and DAQ needs constant monitoring of data taking buffers @ P5 and T0

- Developed a "**red button**" to switch off parking as soon as buffers become problematic
- So far, CERN tape seems to sustain the rate

B – parking data collection: impact on resources

- Final green light on May 10th; out of phase and very late with respect to computing requests via RRB
- **Idea: fit the events in the standard computing budget.**
- Main handles:
  - Take events with a **PU** substantially **lower** than the average for by using the final part of the fills.
  - **Remove** all other forms of parking; stay disciplined with Prompt rates
  - Have a **single Tape copy** @ CERN. The second copy would eventually be restored during LS2 (or not)
  - No impact on T0 CPU, apart from a few Hz of monitor triggers
  - **Defer** processing until available free CPU are present (see later)
  - Deliver **only MiniAOD** from processing
  - Additional MC samples small
    - "**Data driven**" Analyses

- All in all, CMS expects the B-parking main consequences is ad **additional load on operations** during data taking
  - **DAQ output buffer**: critical, controlled by the DAQ Shifter @ P5
  - **T0 input / output buffers**: easier to provision more space thanks to the T0-T2 merge at the expenses of group and central spaces
- **No long term additional resources required**
  - **Tape**: space for single copy @ CERN already prepared via the deletion campaign (still not executed); second Tape copy most probably not needed
  - Disk: **MiniAOD** only analyses, ~500 TB (<0.5% of CMS disk)
  - CPU: analyses will be carried on during LS2, no longer scale impact; MC requests small

In any case B-parking is understood by the Collaboration not to have the same level of data safety and priority as standard Prompt data taking

- If possible, take these data. Otherwise, back to plan A

Other notable 2018 Runs

- Low beta* (90m):
  - Somewhere in June (moving target) – with TOTEM
  - Expect to get up to 10 kHz of “small” events, reconstruction needed
- Heavy Ion
  - Plans not changed since last LHCC
    - 500 Hz of “Physics” events
    - 6500 Hz of Minimum Bias (6B events needed for HF studies)
  - Handshaking with IT done – data handling seems feasible
  - Process promptly Physics + a (small) fraction of MB
  - Tape writing only @ CERN initially, second copy established during LS2

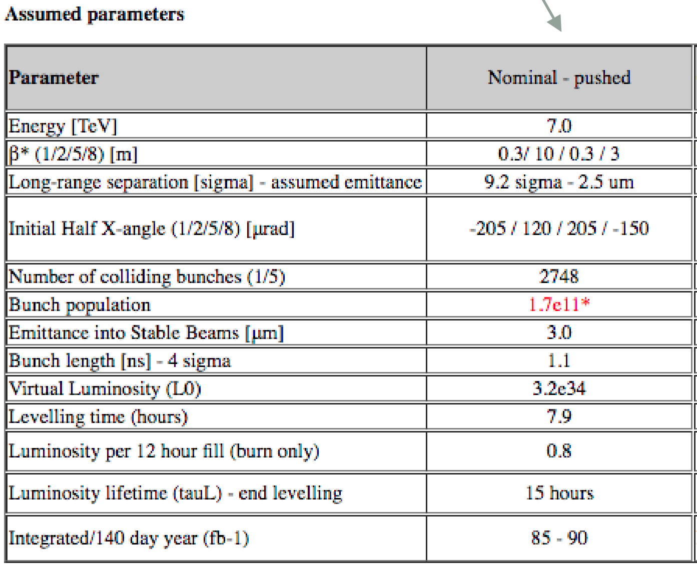
Processing of B-Parked + HI data?

- **When**? Not easy tasks, months long
- There is a window of opportunity before Legacy RunII processing starts ~ April 2019 (so dec18-mar19)
  - Depends on the critical availability of “good enough” calibrations
  - Depends on the need to reprocess 2018 Data (if prompt not good enough) for Winter conferences 2019
  - Depends on the actual availability of HLT in that period (yet unknown)
- Other creative solutions being searched for
  - HPC centers? Opportunistic resources? Partial reconstruction only for initial studies?
- Otherwise something can easily slip to 2020

Preparation for RunIII, RunIV

- We just saw the first assumptions for RunIII (2021):
  - Not extremely different from expectations – but we will know better by October
- On paper, RunIII is (still) an adiabatic extension of RunII, with
  - +1 TeV (nearly irrelevant)
  - Up to 50% of the fill time in levelling (so <PU>~55 or so)
- As Ian said @ RRB, we expect for 2021 a +50% with respect to 2018
  - **Seems still valid in this picture**

“best” scenario



Assumed parameters

Parameter	Nominal - pushed
Energy [TeV]	7.0
β* (1/2/5/8) [m]	0.3/ 10 / 0.3 / 3
Long-range separation [sigma] - assumed emittance	9.2 sigma - 2.5 um
Initial Half X-angle (1/2/5/8) [μrad]	-205 / 120 / 205 / -150
Number of colliding bunches (1/5)	2748
Bunch population	1.7e11*
Emittance into Stable Beams [μm]	3.0
Bunch length [ns] - 4 sigma	1.1
Virtual Luminosity (L0)	3.2e34
Levelling time (hours)	7.9
Luminosity per 12 hour fill (burn only)	0.8
Luminosity lifetime (tauL) - end levelling	15 hours
Integrated/140 day year (fb-1)	85 - 90

Unclear facts:

- The LHC task force will finish in October, some “much higher” numbers have been seen
- On CMS side, not yet clear if we can stay at 1 kHz of Prompt trigger rate if most of the fill is at 2e34 – studies ongoing

Work to be done in LS2

- CMS is planning reviews of major computing software stacks in LS2
  - In principle RunIII could be handled with the same tools as RunII
  - BUT: we plan to use RunIII as a testbed for new solutions / ideas
  - Use LS2 to gain experience
  - Workload management: review started on May 10th
    - Analyzing interplay between Production system (WMAgent) and Analysis system (CRAB3)
  - Data Management: first panel meeting last week
    - Scope is deciding which is the most suitable DM product for CMS (use cases, support model, …)
    - **Dynamo** (CMS/MIT) and **Rucio** (ATLAS) are the candidates under analysis

CRAB
Cms Remote Analysis Builder

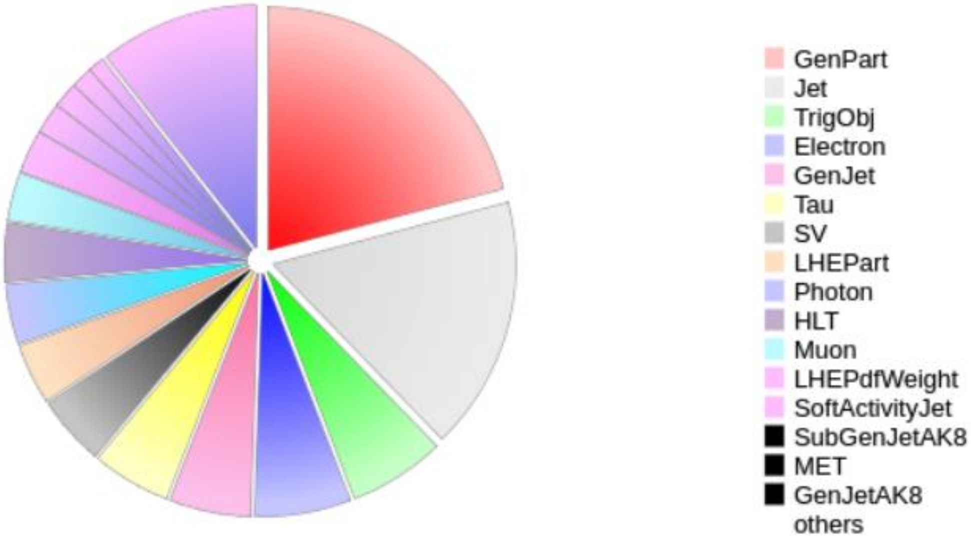
DYNAMO

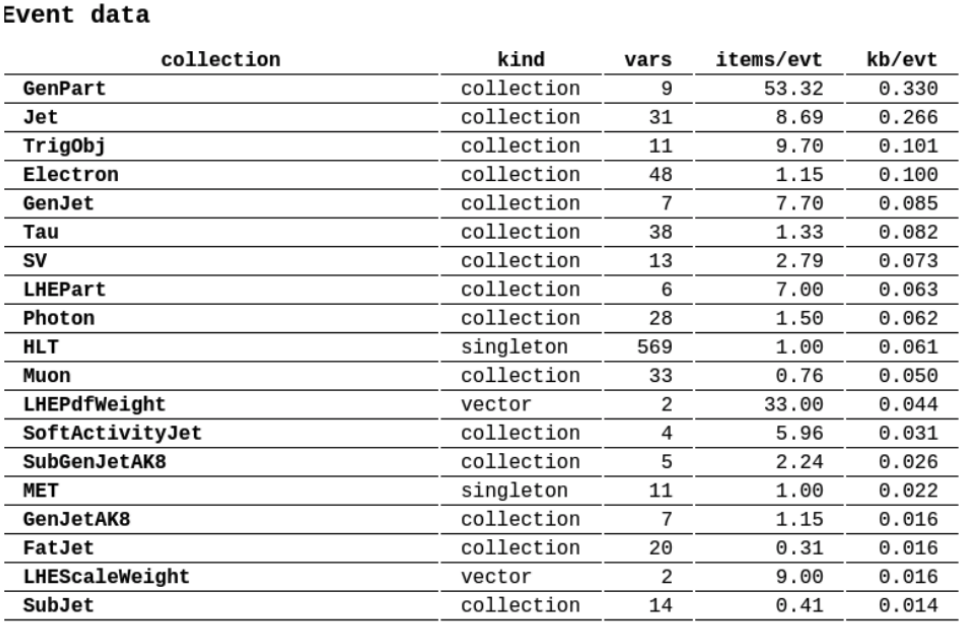
New notable developments

- CMSSW
  - Tests with **gcc 7** – positive so far
  - Tests with alpha version of **GeantV**
    - Current plan is to evaluate beta when out, decide for a switch during RunIII
  - Moving Detector Description from **DDD** (CMS, 2002) to **DD4HEP** (SFT/AIDA)
  - We have a prototype for using premixing also for PhaseII simulations
  - **Keras/Tensorflow** distributed with CMSSW
  - **CUDA** support out of the box
- Web Services
  - From Agile infrastructure to **Kubernetes**
  - From X509 to **CERN/SSO**
  - From Python to **Go** (a few overloaded services)
- CRAB3 improvements
  - CRAB3 accepts tasks reading **tape only datasets** and issues (smart) tape recalls
  - CRAB3 automatically computes the amount of work per job → **fewer shorter jobs**

NanoAOD

- Already reported at previous LHCCs
- Progressing faster than expected:
  - Already available to users **11B+ 26B** (DT+MC) centrally processed events (16/17)
    - Not counting private productions
  - Being used in analysis (LHCP2018 is the target)
  - Content still fluid and adapting for new use cases, but still below budget:
    - DT: 700 Bytes/ev
    - MC: 1000 Bytes/ev





Event data

collection	kind	vars	items/evt	kb/evt
GenPart	collection	9	53.32	0.330
Jet	collection	31	8.69	0.266
TrigObj	collection	11	9.70	0.101
Electron	collection	48	1.15	0.100
GenJet	collection	7	7.70	0.085
Tau	collection	38	1.33	0.082
SV	collection	13	2.79	0.073
LHEPart	collection	6	7.00	0.063
Photon	collection	28	1.50	0.062
HLT	singleton	569	1.00	0.061
Muon	collection	33	0.76	0.050
LHEPdfWeight	vector	2	33.00	0.044
SoftActivityJet	collection	4	5.96	0.031
SubGenJetAK8	collection	5	2.24	0.026
MET	singleton	11	1.00	0.022
GenJetAK8	collection	7	1.15	0.016
FatJet	collection	20	0.31	0.016
LHEScaleWeight	vector	2	9.00	0.016
SubJet	collection	14	0.41	0.014

And a final message …

- Please let us introduce you **Markus Klute, Professor @ MIT**
- He will serve ac Offline and Computing co-coordinator Jul 1st 2018 – Aug 31st 2020
- He is currently “Physics Performance and Dataset” co-coordinator in CMS, a group whose interactions with O+C are much more than daily
- He has a rich past in computing operations in RunI

- I want personally to thank Liz for the collaboration we had in the last year; she will not go too far anyway:
  - She agreed to serve as Chief Information Officer (CIO) at Fermilab

Conclusions

- So far, 2018 data taking and processing activities going as planned
- B-parking and HI run are putting unplanned pressure on the computing operations
  - Not yet a clear plan on final processing, depends critically on calibration availability
  - No long term impact expected on resources
- CMS is preparing for the mid(RunIII)-long(RunIV) term operations with
  - New features in CMSSW
  - Evaluation of new products (GeantV, DD4Hep, …)
  - Reviews for mission critical Computing components