Grid usage best practice

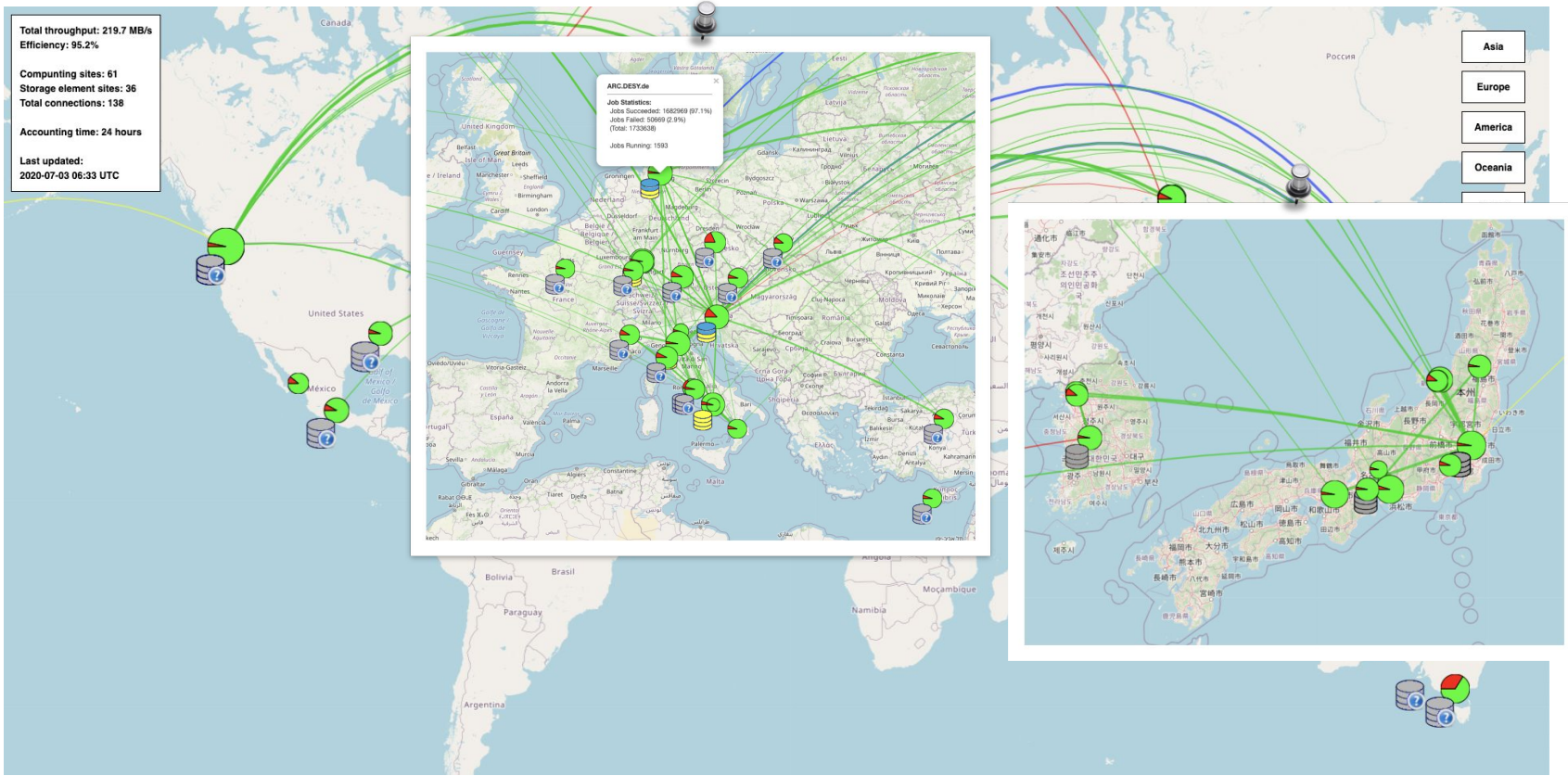
Physics Week
Valencia (ES)
30 Nov 2022

Stefano Lacaprara, Umberto Tamponi,
Michel Villanueva
for the DP/DC team

Outline and goal

- This is not a tutorial on grid usage
- **1.** Bird's eye presentation of distributed computing (grid)
- **2.** And data production
  - Emphasis on how and where get more info
- **3.** A set of suggestions and recipes to improve your grid experience
- We do expect a lot of question from you, based on your grid and analysis experience
  - Michel and I will try to answer to all of them and/or take your feedbacks home to improve tools and documentation and whatnot in the future

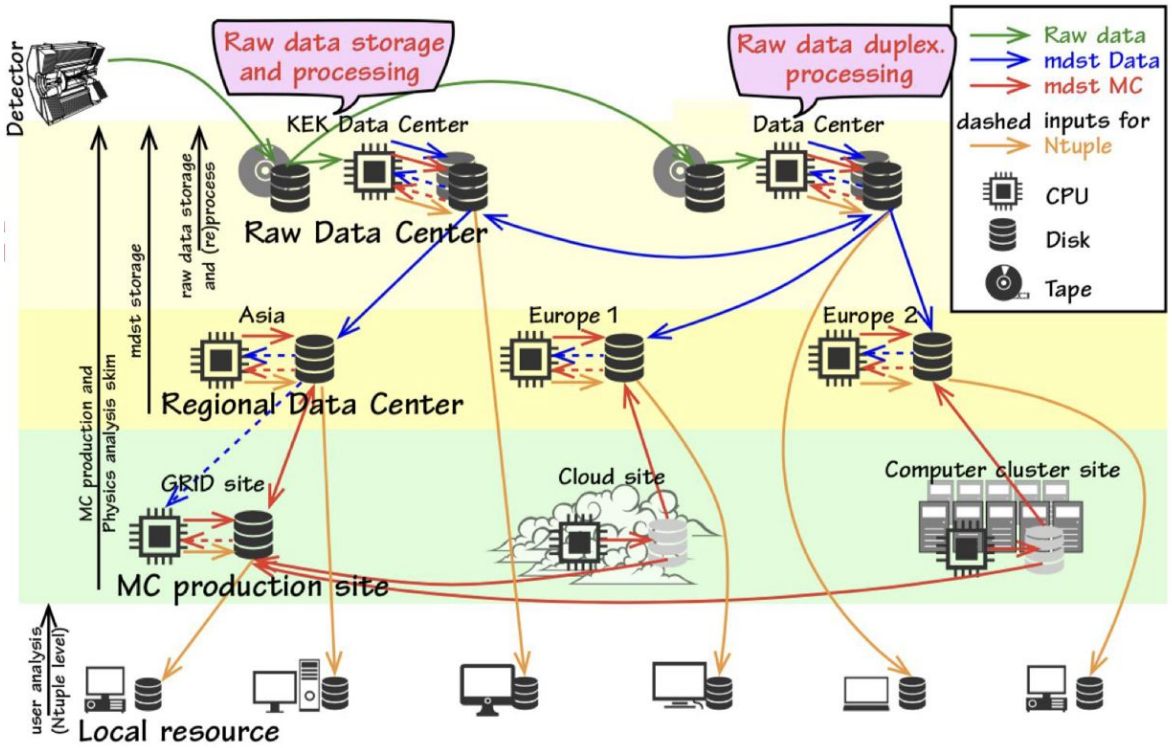
First… what is the grid?



Michel, Stefano, Umberto

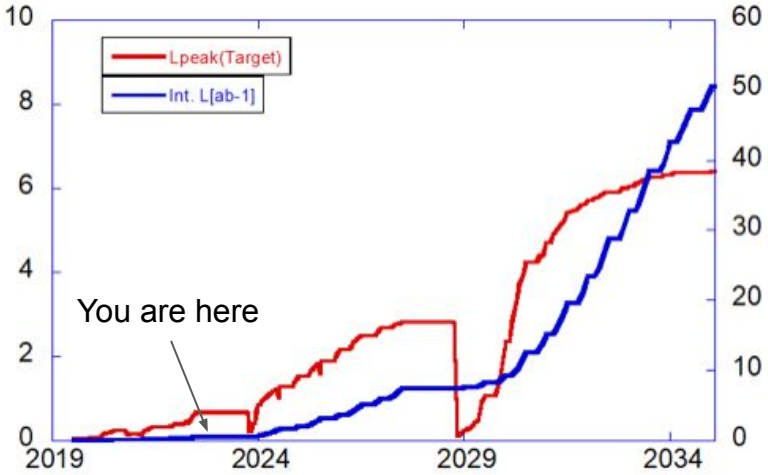
First… what is the grid?

- Many networked loosely computers.
- Data centers keep two copies of the full raw data set.
- Raw data is staged, reprocessed, skimmed and distributed over sites.
- Analysers access data **sending jobs to the grid** and **downloading the output.**



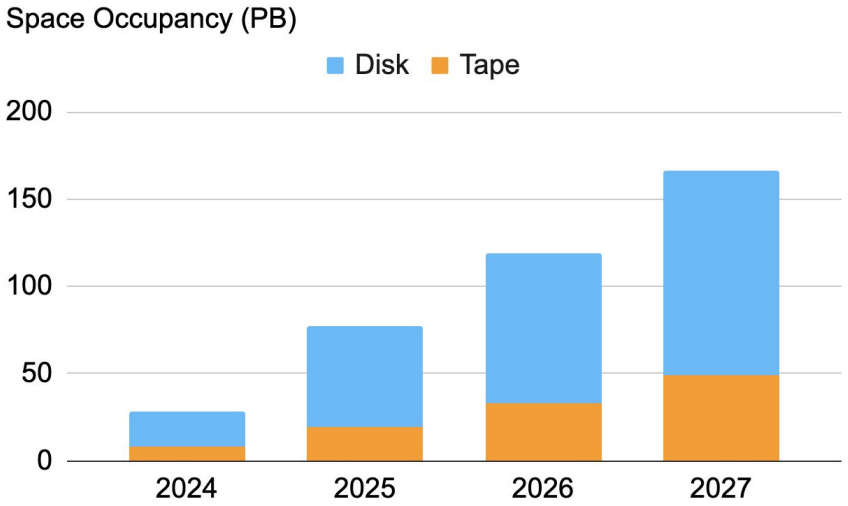
Why grid?

- Integrated luminosity expected by the end of the experiment is 50 ab^{-1}



Current RAW size: ~6 PB

- Estimated size of the dataset collected by the experiment is **O(10) PB/year.**
- Data must be distributed and analyzed by > 1000 collaborators around the world.

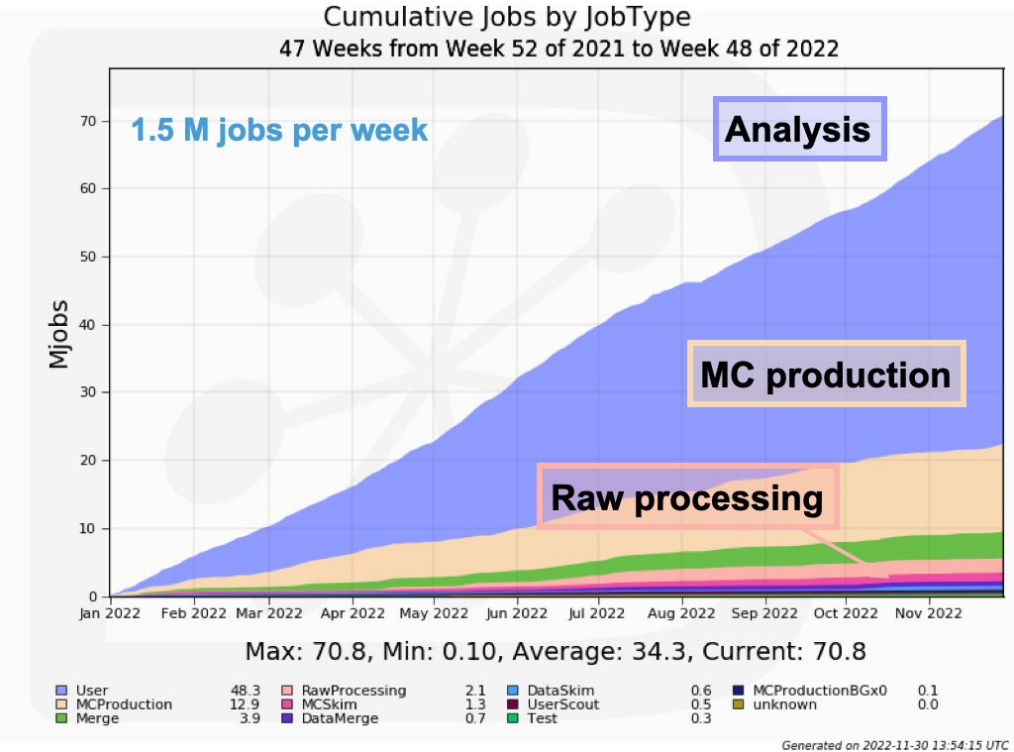


Michel, Stefano, Umberto

Some metrics

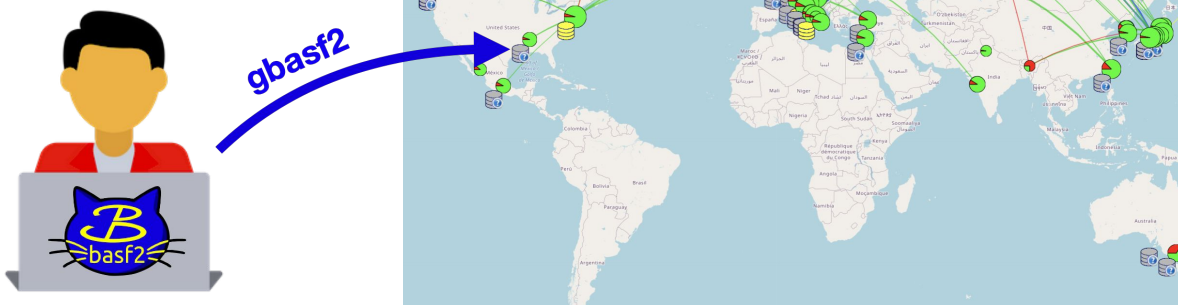
- Operations in 2022
  (including all replicas)
  - **Raw data**
    6 PB (4M files)
  - **Moriond 2023 / ICHEP**
    Proc13 + buckets: 417 TB (1.4M files)
    MC15: 870 TB (2.2M files)

- CPU usage
  - MC production: 72%
  - Data processing: 6%
  - Skimming: 1%
  - **User analysis 22%.**



gbasf2

- Gbasf2 is a tool for submission of jobs. **From your desktop to the grid.**
- The same steering files used with basf2 works with gbasf2 on the grid.
- The usual workflow is:
  - Developing a basf2 steering file at first.
  - Testing it locally.
  - Submit the jobs to the grid with the same steering file.

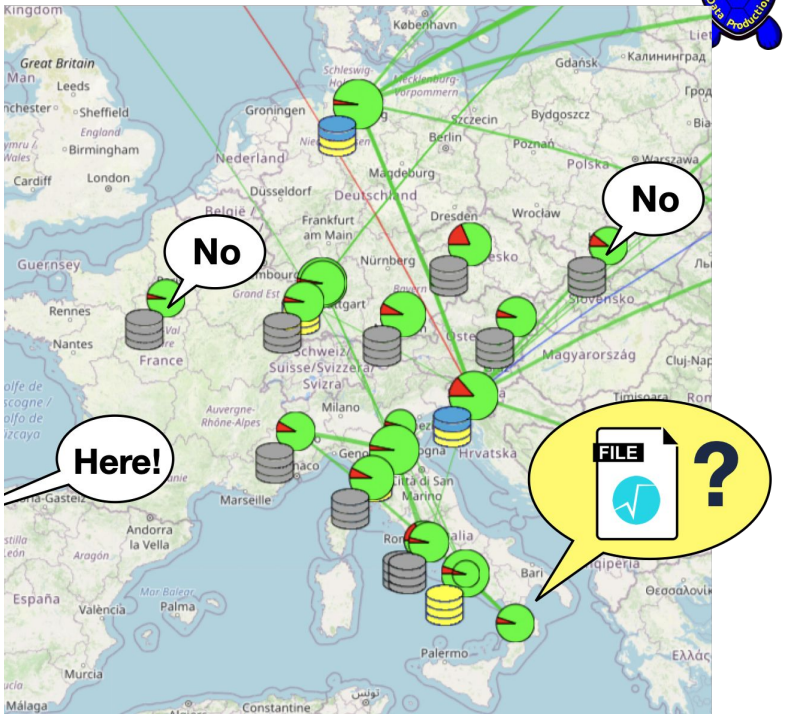


Logical File Names

- Files are stored around the world in the different storage elements.
- A logical file name (LFN) is the unique identifier of a file in the Belle II grid in the form of a unix-like file path:

  /belle/data_type/more_directories/file_name

- **Rucio File catalog** resolves the LFN, and provides the information of where to find the file(s).
- Then, you only need to provide the LFN(s) of the datasets which are relevant for your analysis.
  - `/belle/MC/release-01-00-03/DB00000294/MC10/prod00004770/s00/e0000/4S/r00000/mixed/mdst/sub00`



Collections

- Bonus: integration with Rucio allow the definition of collections
  - a single reference for a group of datasets of interest.

- Collection is Immutable
  - Analysis reproducibility is ensured.

- Decrease in pre-work for gbasf2 job submission.
- Huge decrease in gbasf2 submission time for large project submission.

datablock (subXX)
dataset
collection 1
collection 2

Where do I find help?

- The information is located at Confluence:
  - The gbasf2 manual: gbasf2.belle2.org
  - The gbasf2 FAQ.
  - The gbasf2 troubleshooting.

- Please join comp-users-forum@belle2.org
  - It is the official channel of support for gbasf2.
  - You will also receive announcements on new releases, system issues, etc.

- Use questions.belle2.org. People will be happy to help (please also try to help others).

- Many gbasf2 tutorials are available.
  - Try to follow the most recent one.

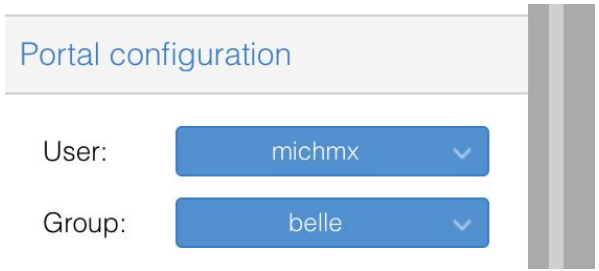
How can I ask for help effectively?

- Describe your problem
- Provide the problematic project name, jobid or dataset.
- Provide your **grid username**

gb2_proxy_init
gb2_proxy_info

```
Generating proxy...
Enter Certificate password: ******
Added VOMS attribute /belle
Uploading proxy..
Proxy generated:
subject      : /DC=org/DC=terena/D
issuer       : /DC=org/DC=terena/D
identity     : /DC=org/DC=terena/D
timeleft     : 23:53:59
DIRAC group  : belle
path         : /tmp/x509up_u47126
username     : slacapra
properties   : NormalUser
VOMS         : True
VOMS fqan    : [u'/belle']
```

https://dirac.cc.kek.jp:8443/DIRAC/ -> Settings



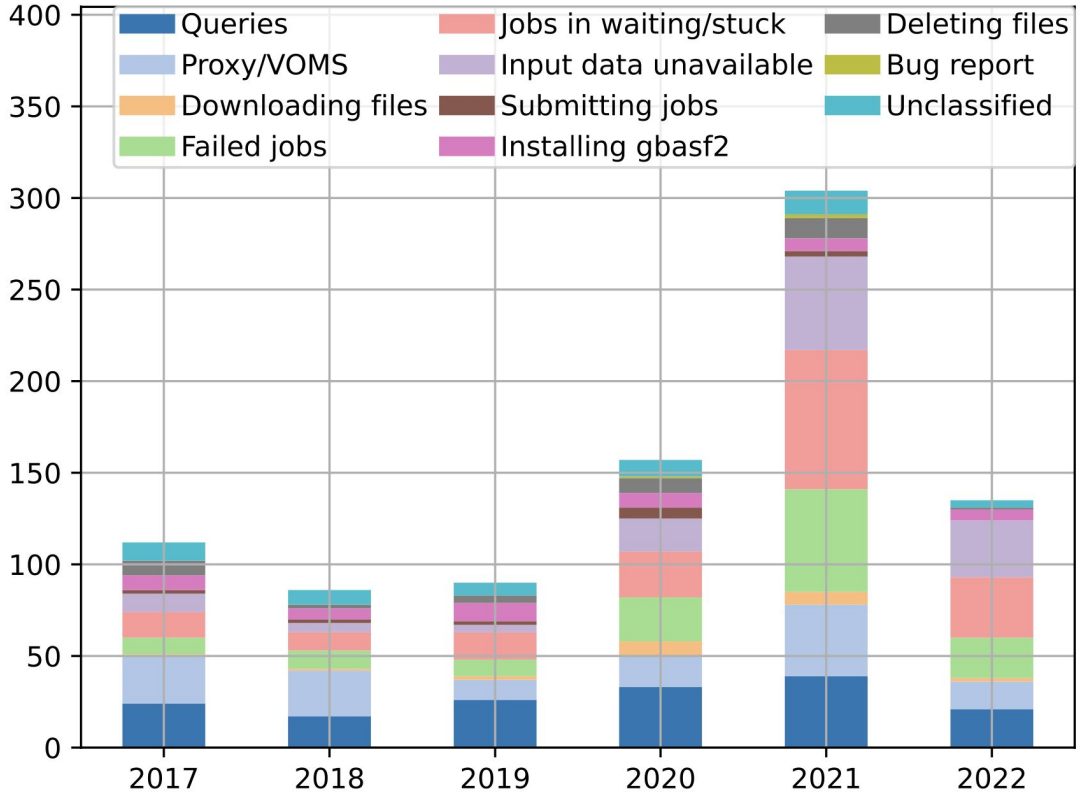
Michel, Stefano, Umberto

Most common issues

- In the last year most of the issues are related to
  - Issue in an specific site.
  - Load on central services.

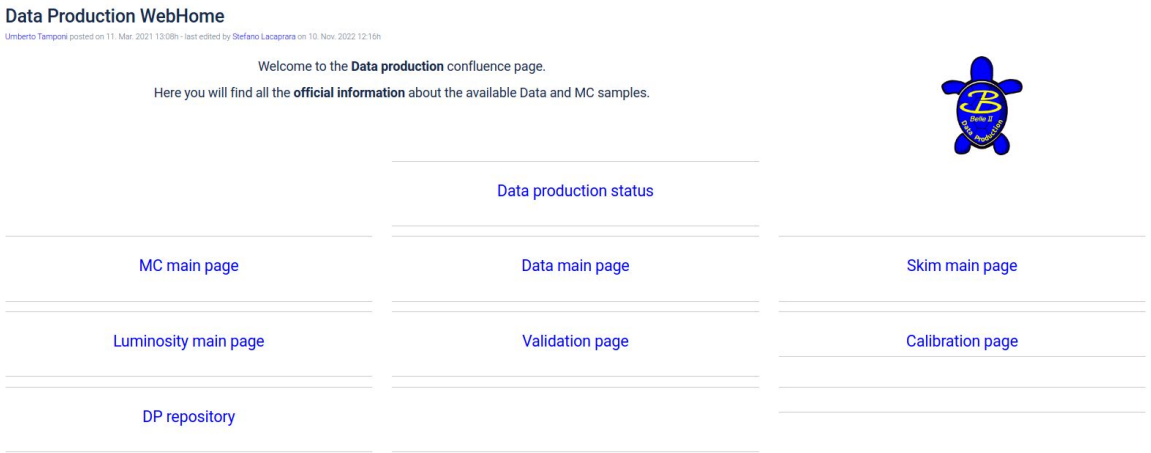
- A clear trend of higher usage on the grid with less issues :D

- Reports in the comp-users-forum:



Where to find information on what to run

- General DP page: https://confluence.desy.de/display/BI/Data+Production+WebHome
- Data Page https://confluence.desy.de/display/BI/Data+Production+Status
- MC Page https://confluence.desy.de/display/BI/MC+main+page
- SKIM Page https://confluence.desy.de/display/BI/Skim+main+page



Michel, Stefano, Umberto

Data: what we produce

- Prompt reconstruction:
  - Produced few weeks (~4-6) after the data taking, after prompt calibration is performed
  - Produced in dataset of about 20/fb selecting a set of runs (DP jargon: bucket)
- Full reprocessing:
  - All data up to a given experiment is re-calibrated and reconstructed with latest greatest basf2 release
    - **procXX** (as in “prok”)
- To have full dataset you need to combine procXX and prompt for the remainder of data.
  - Eg: proc13 (for exp7-18) + prompt (for exp 20-26)
- Output mDST
  - For ALL events
  - For HLT-hadron events (about 10% of total)
    - Contains all high multiplicity events (eg BB-bar)
    - [[nTracksLE>=3] and [Bhabha2Trk==0]] https://confluence.desy.de/display/BI/HLT+Skims

Michel, Stefano, Umberto

Data Page: how to read it

Moriond 2023 dataset	All	4S	/belle/collection/Data/proc13_all_4S_**v2** **UPDATED** /belle/collection/Data/Moriond2023_prompt_all_4S_v1	proc13 (exp 7,8,10,12,16,17,18) + prompt (exp 20, 22, 24, 26)
		4S_offres	/belle/collection/Data/proc13_all_4S_offres_v1 /belle/collection/Data/Moriond2023_prompt_all_4S_offres_v1	proc13 (exp 8,12) + prompt (exp 25)
		5S_scan	/belle/collection/Data/Moriond2023_prompt_all_5Sscan_10657_v1 /belle/collection/Data/Moriond2023_prompt_all_5Sscan_10706_v1 /belle/collection/Data/Moriond2023_prompt_all_5Sscan_10751_v1 /belle/collection/Data/Moriond2023_prompt_all_5Sscan_10810_v1	prompt (exp21)
	Hadron	4S	/belle/collection/Data/proc13_had_4S_**v3** **UPDATED** /belle/collection/Data/Moriond2023_prompt_hadron_4S_v1	proc13 (exp 7,8,10,12,16,17,18) + prompt (exp 20, 22, 24, 26)
		4S_offres	/belle/collection/Data/proc13_had_4S_offres_v1 /belle/collection/Data/Moriond2023_prompt_hadron_4S_offres_v1	proc13 (exp 8,12) + prompt (exp 25)
		5S_scan	/belle/collection/Data/Moriond2023_prompt_had_5Sscan_10657_v1 /belle/collection/Data/Moriond2023_prompt_had_5Sscan_10706_v1 /belle/collection/Data/Moriond2023_prompt_had_5Sscan_10751_v1 /belle/collection/Data/Moriond2023_prompt_had_5Sscan_10810_v1	prompt (exp21)

You want to use only collections, unless you really know what you are doing!
MANDATORY: specify collection name in your B2Note for reproducibility.

Luminosity

		Total	
all	4S	364.436 +/- 0.020	361.589 +/- 0.020 ⭐
	4S_offres	42.329 +/- 0.007	42.279 +/- 0.007
	4S_scan	0.038 +/- 0.000	0.038 +/- 0.000
	5S_scan	19.702 +/- 0.004	19.702 +/- 0.004
	All beam energies:	426.506 +/- 0.021	423.57 +/- 0.021 ⭐

- Data Offline+Luminosity+Page
- Also every collection has lumi info
  - For Data and for MC

```
~> gb2_ds_search collection --get_metadata /belle/collection/test/MC14ri_ccbar_1abinv_v1
########## Metadata of Collection ###############
dataLevel: mdst
description: Collection MC14 ri for ccbar - 4S
campaign: MC14ri_d,MC14ri_a
dataType: mc
skimDecayMode:
int_luminosity: 1000.0 /fb
generalSkimName:
################################################
```

- Also command line tool for Data (within any recent basf2 release)
  - **WARNING: do rely on officially released luminosity values for dataset**

```
b2info-luminosity --exp 7-26  --good --what offline
b2info-luminosity --exp 8 --runs 1954-3123  --good --what offline
```

MC: what we produce

- Two types of MC Run Independent (**RI**) and Run Dependent (**RD**)
- **RI**
  - Detector condition are stable, “averaged” over data taking
  - Background are coming from simulation, and also (statistically) same for all events
- **RD**
  - Detector conditions are the same as the one during data taking, run-by-run (hence the name)
  - BG is taken from data with random trigger and overlaid to simulation
- **Samples**
  - Generic:
    - qq-bar (q=udsc) tau pairs aka continuum
    - bb-bar: mixed (B0-B0bar) charged (B+B-)
  - Low multiplicity
    - mumu, ee, mumuee, 4mu, 2tau, 4tau, pipiISR, K0K0barISR, llXX, hhISR, etc
  - Signal(s)
    - later

MC

- MC Page https://confluence.desy.de/display/BI/MC+main+page
  - Similar to Data: collection are provided once the production is done
- RI:
  - You have 1/ab for qqbar/taus
    - FYI: Additional 2/ab of bbbar are being produced in these days
  - 0.1-10/ab for low multiplicity dataset
- RD:
  - Each collection has a luminosity, which depends on the sample generated (qqbar, low mult, etc)
  - It is **close** to 4x for qqbar/tau's, 2x gg, 1x for low mult, 0.1 ee BUT NOT EXACTLY
- For production in progress, no collection is provided, need to query DSS
  - gb2_ds_search dataset --campaign MC15rd_b --data_type mc --exp_low 7 --exp_high 7 --mc_event uubar
- What MCrd is available? What is running ?
  - Details on https://confluence.desy.de/display/BI/MC+run-dependent+details

Signal

- Signal is produced generally for MC Run Independent
  - Now also for Run Dependent
- RI signal list:
  - https://confluence.desy.de/display/BI/MC+run+independent+signal+production
- RD signal list:
  - https://confluence.desy.de/display/BI/MC+run+dependent+signal+production
- To access query DSS:
  - `gb2_ds_search dataset --campaign MC15ri_b --data_type mc --mc_event 1110062100 --bkg_level BGx1`
- Only signal explicitly requested by WG are produced!
  - Do not assume that all signal for which a dec file is present in basf2 are produced
- How to ask for signal production?
  - Check if not produced already
  - Ask your DP liaison in your WG
  - Signal can produced also if dec file is new (not yet in basf2)
  - Default is RI, signal in RD are prioritized: if you need it, it can be done
    - Contact your WG conveeners

Skim

- In Belle II we have two “skims”
  - HLT-skims aka general skim (eg hlt-hadron)
  - analysis-skim (aka uDST)
- An analysis-skim is a subset of a dataset (Data or MC) with all events passing a set of (loose) selection
  - Idea is to provide a smaller dataset for faster processing for specific analysis
  - Retention rate 1-10%: can start from all or hadron events
- Eg: **B2Charmless WG**
  - **Skim description**: Skim list for all charged B to charmless modes with 3 tracks and 1 Pi0.
  - **Skim name**: BtoHad3Tracks1Pi0
  - **Skim LFN code**: 19130310
- Full documentation at:
  - https://confluence.desy.de/display/BI/Skim+main+page

Reconstructed decay modes:

- $B^+ \to K^{*+} K^+ K^-$
- $B^+ \to K^{*+} K^+ \pi^-$
- $B^+ \to K^{*+} \pi^+ \pi^-$

Cuts applied:

- `5.20 < Mbc < 5.29`
- `abs(deltaE) < 0.5`

Skim: what is available?

- https://confluence.desy.de/display/BI/Skim+Production+Status
- About 70 different skims produced in 20-30 combined ones
- No collection yet: done on-demand
  - We will have just too many collections!
    - 70 skims * (14 MCri+14MCrd+1Data)
  - Working on improvement in collection to solve this issue
- Now: query DSS

By WG :

- Systematics : 3
- SL&ME : 5
- EWP : 4
- TDCPV : 2
- BToCharm : 19
- BToCharmless : 4
- Quarkonium : 4
- Charm : 14
- Dark : 11
- Tau : 3

Examples:

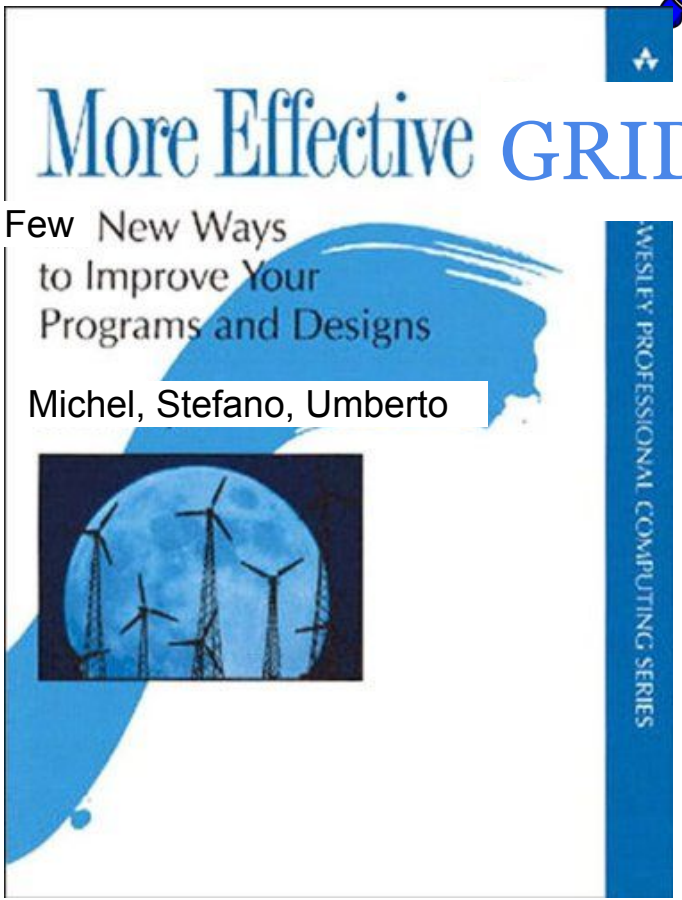
- MC15ri FEI Hadronic skims:
  - gb2_ds_search dataset --data_type MC --data_level udst --skim_decay 11180500 --campaign MC15ri_b
- proc13 TDCPV_ccs skims:
  - gb2_ds_search dataset --data_type data --data_level udst --skim_decay 13160200 --general_skim all --campaign proc13
- All proc13+prompt TauThrust skims:
  - gb2_ds_search dataset --data_type data --data_level udst --skim_decay 18570700 --general_skim all --release release-06-01-10

warning: if you don't specify the release or the campaign, the search will also return results for last campaign (MC14/proc12)

Grid how to use effectively

Some suggestion and recipe to improve your grid user experience

Frequently Asked Questions



How can I test my code locally?

- You most certainly WANT to test your code locally before submitting thousands jobs to the grid! Thanks!
- Get one file from grid via gb2_ds_get
  - You can get the grid file path via
  - `gb2_ds_search collection --list_datasets /belle/collection/Data/proc13_had_4S_v3`
  - `gb2_ds_search dataset --campaign MC15ri_b --data_type mc --mc_event 1110062100 --bkg_level BGx1`
  - gbasf2 has scouting enabled, to protect against this kind of massive failures, but you don't want to submit O(10k) jobs just to find afterward that you forget a comma in your steering.
    - Please test locally!

- Do you have a set of data and generic MC already available at KEKCC so I can use them instead of downloading them by myself?
  - That's a great idea, and we are actually looking for a volunteer to do that!

How can I reduce the number of jobs to run?

- Can you use hlt-hadron instead of all events?
  - Eg: for Moriond Data: all mDST requires ~60k jobs; Hadron mDST: ~10 k jobs
- Can you use analysis skim?
  - How can I check it?
    - Run the skims (https://confluence.desy.de/display/BI/Skim+Information+for+Analysts) on your signal and figure it out
  - But I don't have skim for signal!
    - Right, but you can run it by yourself and save the uDST output
    - Or run it and save to your ntuple

```
# Check the skim output
from skim.WGs.tdcpv import TDCPV_ccs

skim_ccs = TDCPV_qqs(udstOutput=False)  # pass udstOutput=False to disable udstOutput
skim_ccs(path=my_path)
var.addAlias("skim_ccs", f"{skim_ccs.flag}")

# then save "skim_ccs" in your final tree
```

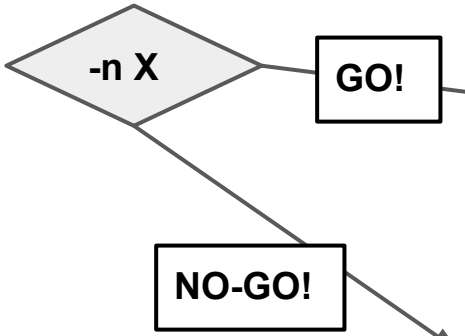
How can I reduce the number of jobs to run?

- Can I run multiple input files for a single job? `gbasf2 -n 2(3…4…1000)` ?
  - **It's complicated.**
  - For technical reason, there is a limit on the total input size you can read in one job
  - Furthermore, you cannot run on two files which have been produced with different GT
    - And some productions are actually done with different GT set
- So, you can use -n 2 (or greater) only for some dataset
  - Check Data and MC confluence page
  - It is fine for MCri sample
    - MC15ri ccbar is 5780 files
- For skims you can run on higher number of input files
  - -n 5 is ok
    - Yet another reason to use skim

How to test -n X ?

- gbasf2 printExtraInfo.py -i /belle/collection/MC/MC15ri_uubar_1abinv_v1 -n 2 -p pippo -s release-06-01-10 **--dry**

All is fine: go ahead




```
Requesting metadata of input files (6370 files). Please wait.
************************************************
*************** Project summary ****************
** Project name: pippo
** Dataset path: /belle/user/slacapra/pippo
** Steering file: printExtraInfo.py
** Job owner: slacapra @ belle (23:44:26)
** Preferred site / SE: None / None
** Input files for first job: ['LFN:/belle/MC/release-06-00-08/
t']
** Number of input files: 6370
** Number of jobs: 3191
** Processed data (MB): 11105954
** Processed events: 1586243574 events
** Estimated CPU time per job: 641391 min
************************************************
```

-n 2 is fine

```
Requesting metadata of input files (6370 files). Please wait.
Maximum input per job exceeds limit (5 GB per job)
Max size of the group of files: 5.15 GB
Please try to reduce the number of input files per job (option -n).
```

-n 3 is NOT fine

WARNING: this will not test against mixed GT issue!

Data processing: GT overview

- Proc13 mDST
  - Exp 7-18 are produced with the same GT set
    - You can try -n X
- Prompt mDST
  - Exp 20-22 are produced with 4 different set of GT
    - Using online snapshot instead of just online
    - DO NOT use -n 2 unless you run on just one bucket
  - Exp 24-26
    - Produced with same GT set
    - Can try -n X
      - UNLESS you process together with exp20-22 (eg if you process all prompt together)
    - Sorry about that!
- Should we document all of this?
  -

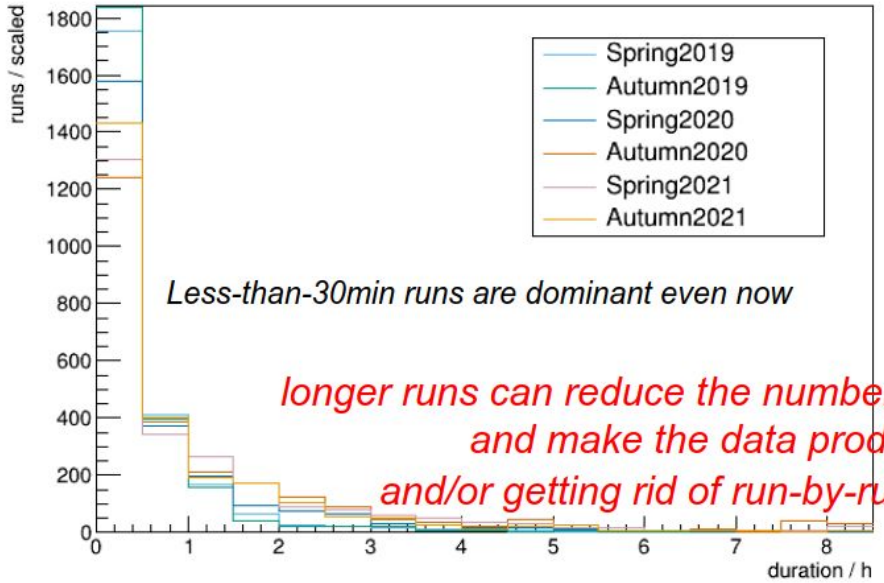
Distributed computing : Scalability

Production system

Our system is optimized with an assumption that 8-hour runs would be dominant after the first couple years of the experiment

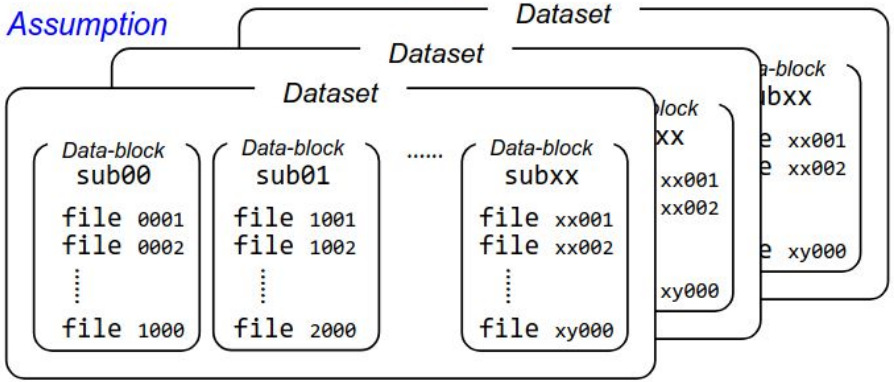
→ But in reality, so many short runs are there even after three years have passed

Sizable number of runs makes prodcution overhead non-negligibe

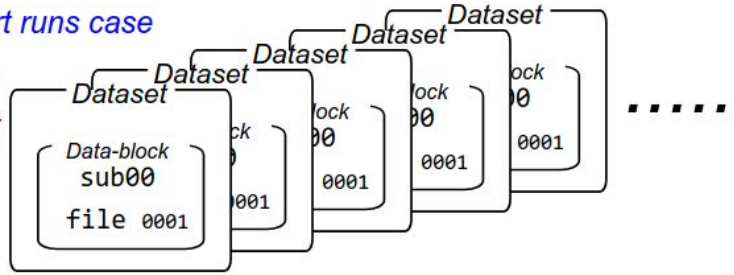


longer runs can reduce the number of short runs and make the data production efficient and/or getting rid of run-by-run datasets ?

Assumption



So many short runs case



What are we doing to improve this?

- Improved merge for future production
  - No more run boundary as up to now
  - So we will be able to merge mDST and uDST from different runs
    - This is particularly relevant for old experiment, where a lot of short runs are present
      - Eg for exp7-8 the reduction in number of files will be O(100)
    - Both for Data and MC-RD, where the run structure is followed
    - Even more effective for uDST
  - A very recent addition, in production but we already produced mDST/uDST…
- Additional merge step for current production
  - We will do a “fake” production, just to do this improved merge step
  - Working on it, will start in the next days from had-mDST of old experiments

How to run multiple channels/ntuples together

- You look at many channels at once:
  - $B^0 \to \eta'$ Ks with $(\eta' \to \eta(\to \gamma\gamma)\pi^+\pi^-)/(\to \eta(\to \pi^+\pi^-\pi^0)\pi^+\pi^-)/(\to \rho(\to \pi^+\pi^-)\gamma)$
- you do not need to run on data once per channel (so 3 times)
  - Pseudo-code

```
import basf2 as b2
my_path = b2.create_path()
…
ma.reconstructDecay(decayString='B0:ch1 -> eta':gg K_S0:merged",cut= 'Mbc > 5.1 and abs(deltaE) < 0.3', path=my_path)
ma.reconstructDecay(decayString='B0:ch2 -> eta':3pi K_S0:merged", cut='Mbc > 5.1and abs(deltaE) < 0.3', path=my_path)
…
ma.variablesToNtuple(decayString='B0:ch1', variables=vars_B0ch1, filename="B2EtapKs.root", treename='B0ch1', path=my_path)
ma.variablesToNtuple(decayString='B0:ch2', variables=vars_B0ch2, filename="B2EtapKs.root", treename='B0ch2', path=my_path)
…
b2.process(my_path)
```

How to download your output effectively

- Reduce the number of output (see before)
- Get your output close to the final destination
  - Replicate before copy
    - gb2_ds_rep -d <destination_SE> /belle/user/myUser/myProject
    - gb2_ds_rep_status /belle/user/myUser/myProject
    - NB: you can do this while your jobs are running, no need to wait them to be done
  - destination_SE is the name of the storage element (“grid disk”) which is close to you
    - gb2_se_list
      - DO NOT choose any ***TAPE* *DATA* or KEK-TMP-SE** (will not work!)
- Copy from SE to SE is optimized, that from SE to your pc is typically not, so doing a fast “long leg” is beneficial

- gb2_ds_get --new
  - significantly faster than standard get

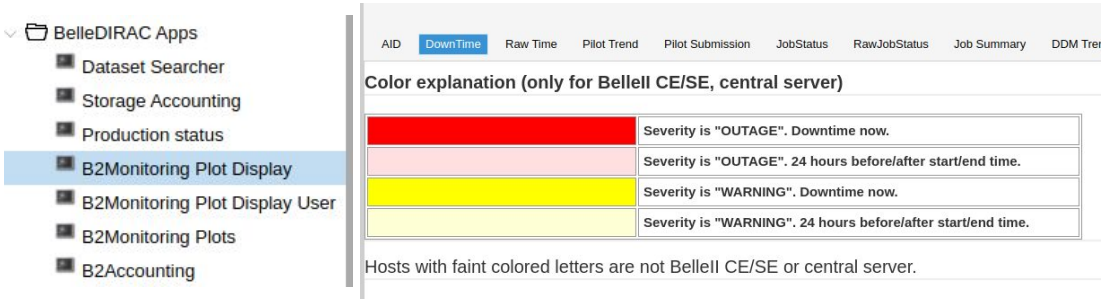
```
ccw01:~>gb2_se_list
SE                      Read     Write    Remove   Check
=================================================
Australia-DATA-SE       Active   Active   Active   Active
Australia-TMP-SE        Active   Active   Active   Active
BNL-CALIB-SE            Active   Active   Active   Active
BNL-DATA-SE             Active   Active   Active   Active
BNL-TAPE-SE             Active   Active   Active   Active
BNL-TMP-SE              Active   Active   Active   Active
```

User output lifetime

- Once you produce your output, this lives on GRID
  - Meaning that it is sitting on some disk on some server on some computing center
  - Disk is cheap, but not free, electricity is not free (and not cheap these days), keeping your ntuples has a cost.
    - Please consider this
    - Remove your old stuff once you don’t need it anymore.
- If you don’t, we will eventually do!
- Reminder: user output file has a lifetime of 1 month
  - After that your files **might** be deleted at any time
    - Meaning they are eligible to be deleted, but they are not actually deleted
  - An announcement will be done in advance, be aware
- Also: soon you will have a quota on grid!

Help: I cannot download my files!

- It's complicated (again), and typically requires expert help
- It might be that the site where your files were produced has an issue
  - It can be a scheduled maintenance: you can check by yourself on DIRAC
  - If so, just wait



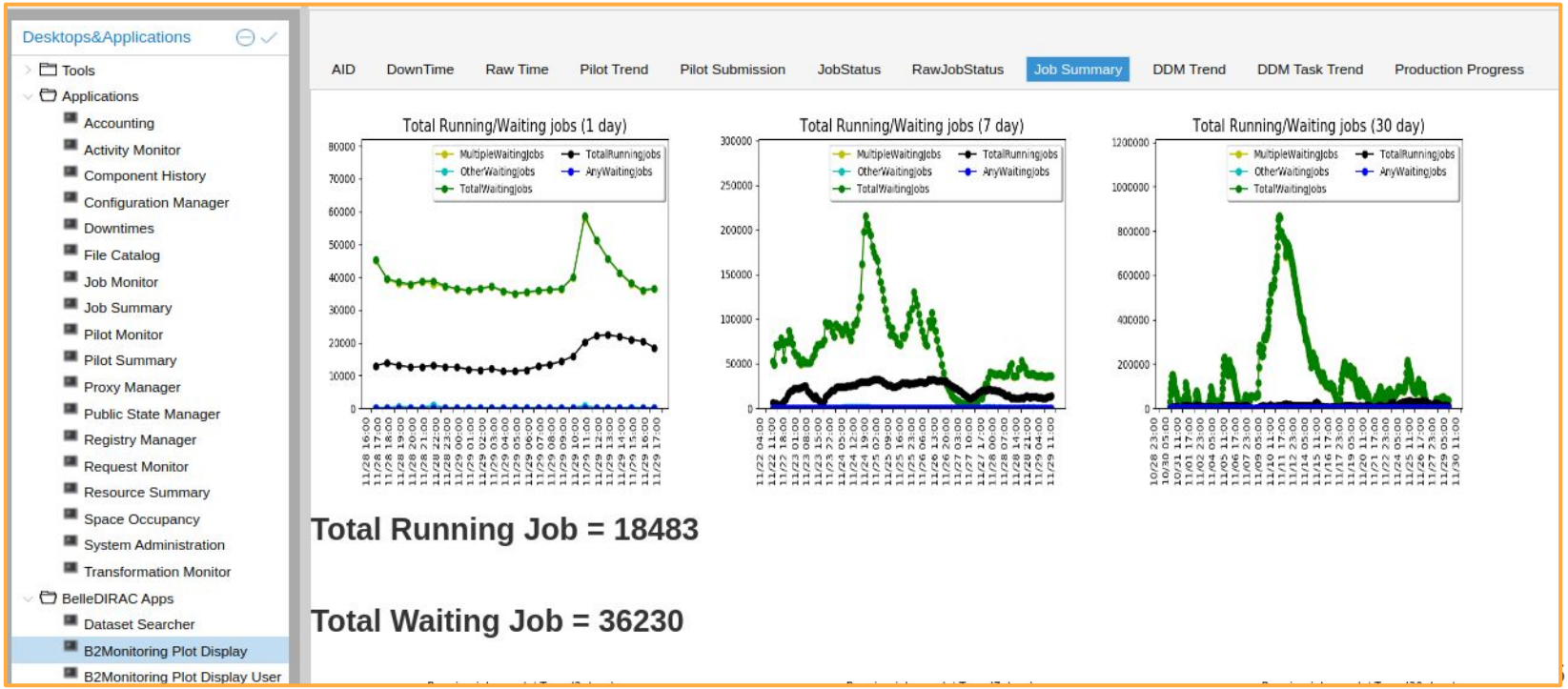
- If not, it is worth to ask comp-users-forum@belle2.org
  - Often the problem has been seen by the Data Production shifter and reported
    - DP shifters are easy to do, low intensity, but extremely useful to spot as early as possible these kind of problems.
    - Have you registered for DP shift? NO? What are you waiting for?

Why are my jobs failing?

- Have you tested them locally?
- Do you need any file in addition to your steering that you forgot to ship?
  - Eg imported modules, weight files, etc …
  - gbasf2 -f, --input_sandboxfiles
- Check the project status
  - gb2_job_status -p <project_name>
- Have a look at some of the ouptut
  - gb2_job_output -j <JobId>
    - JobID is provided by the first command

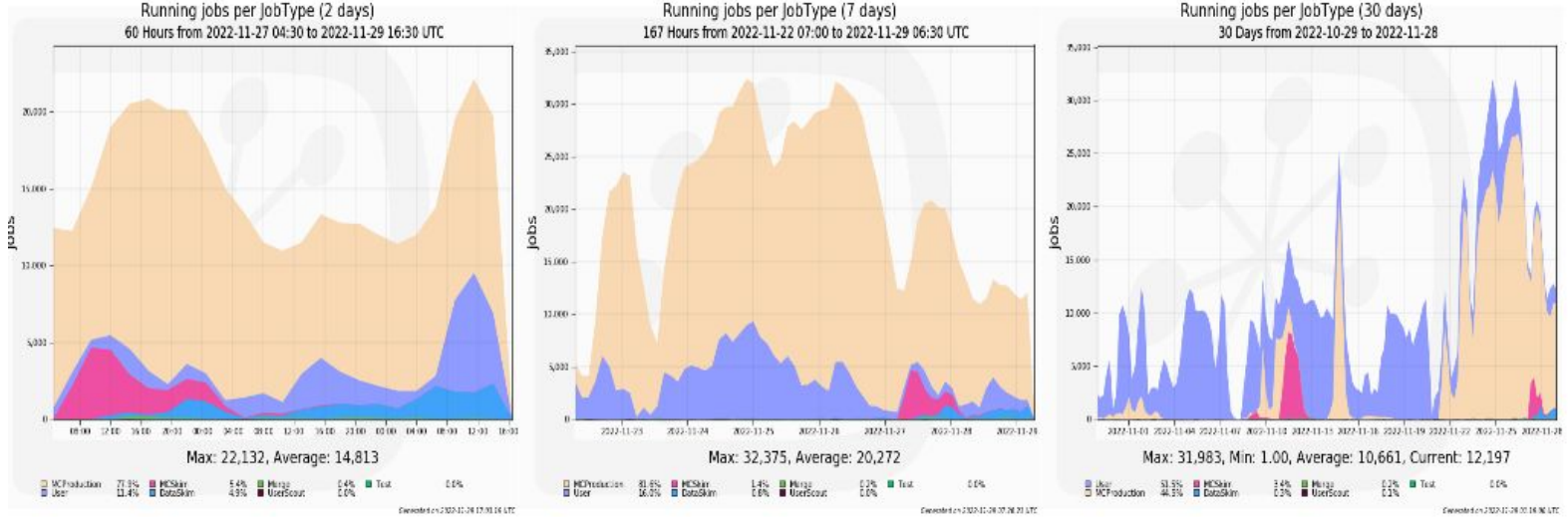
How to check general grid status

- Use DIRAC at https://dirac.cc.kek.jp/DIRAC/
  - You need your grid certificate to be loaded into your browser
  -



Michel, Stefano, Umberto

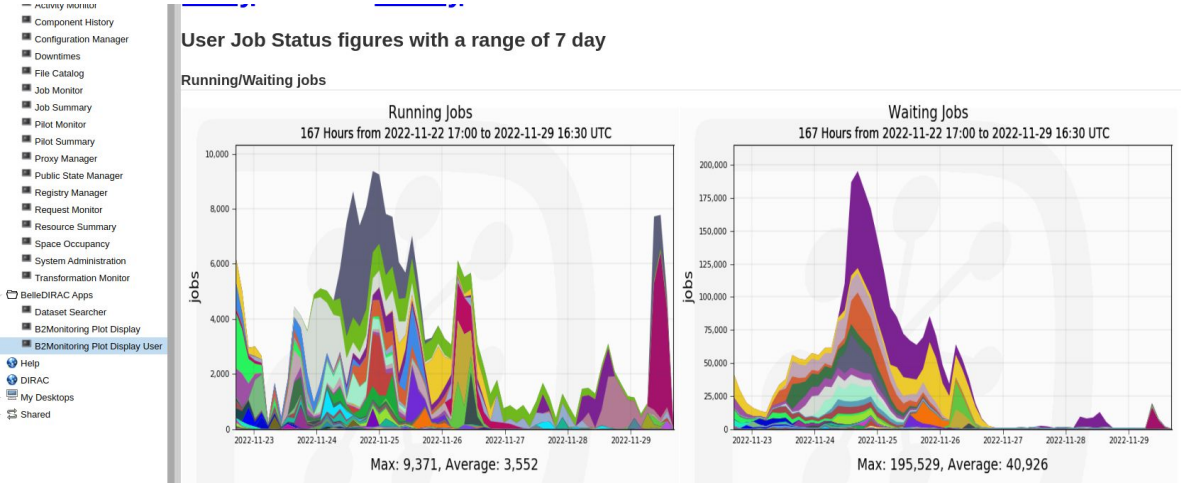
What is running

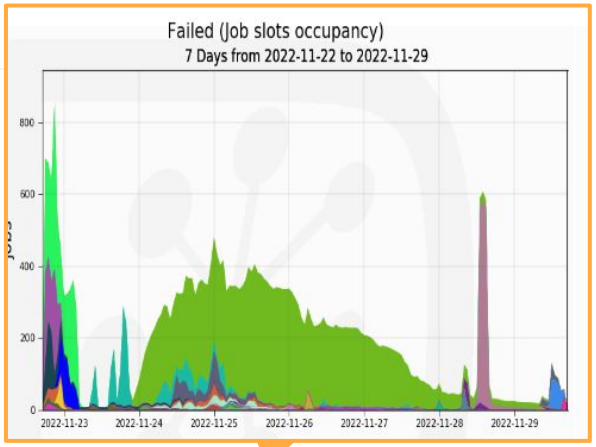


- Light orange (?) is MC
- Blue is users
- Other colors MC/Data skims
  - No raw processing in the last month

Michel, Stefano, Umberto

B2Monitoring Plot Display User





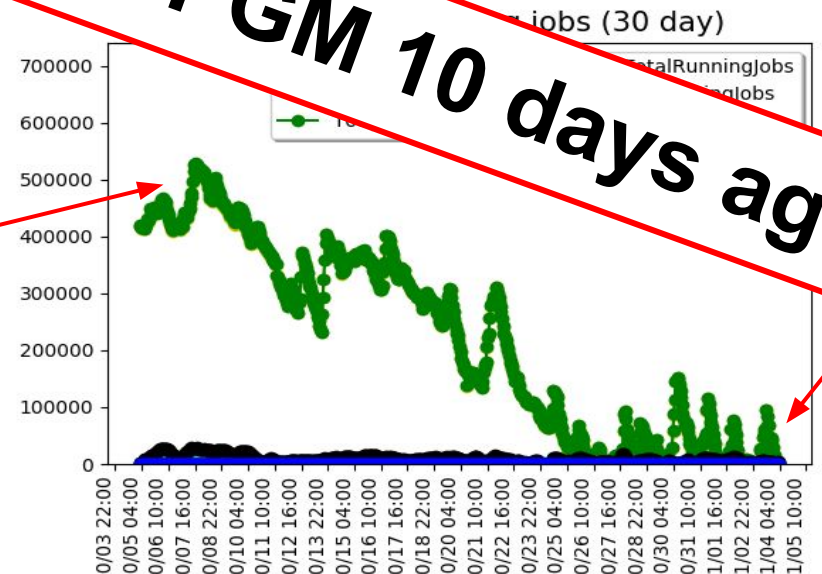
- Running/Waiting/Failing/etc divided by user
  - Very useful to spot massive failure of your jobs

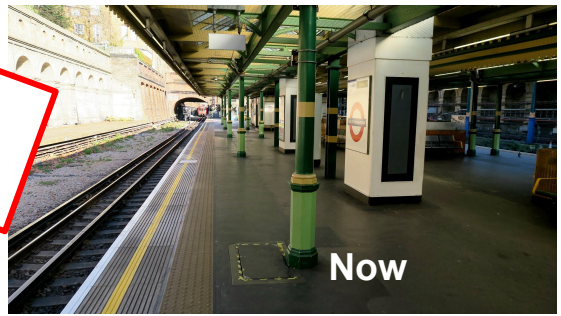
Grid usage

- Very little analysis jobs running on grid right now
- If you need to submit your analysis job on grid **now is a good time!**
  - Don't wait for rush hours

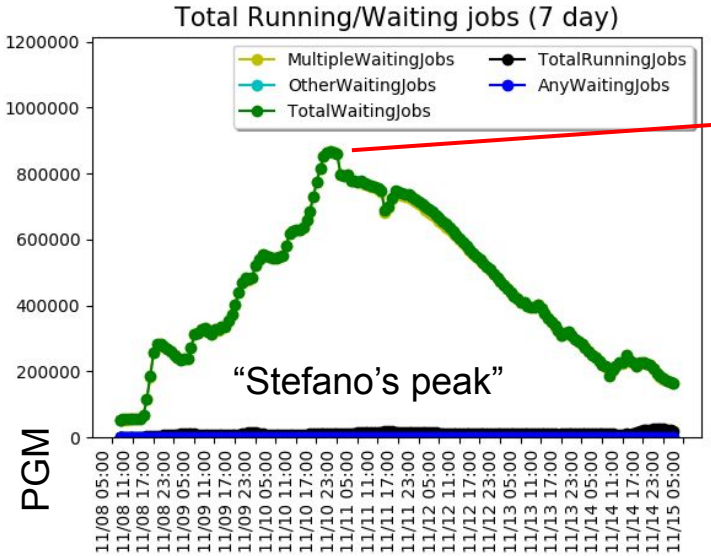
Last PGM 10 days ago

Just before B2GM

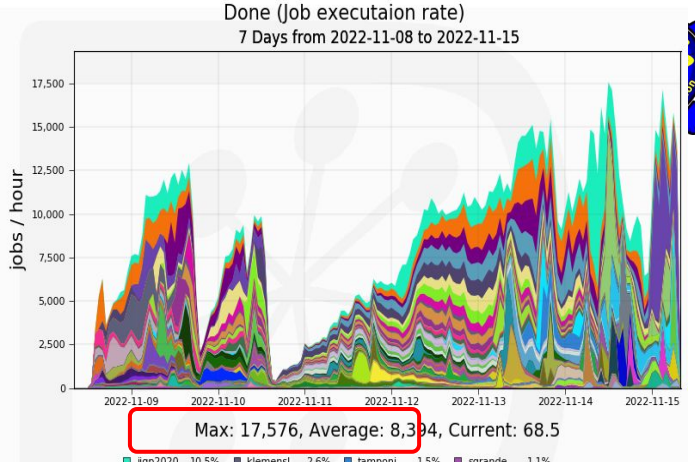




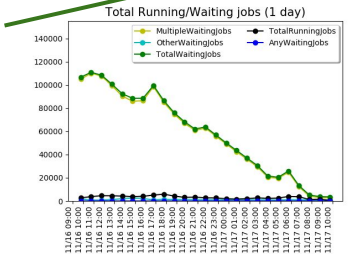
Since last PGM

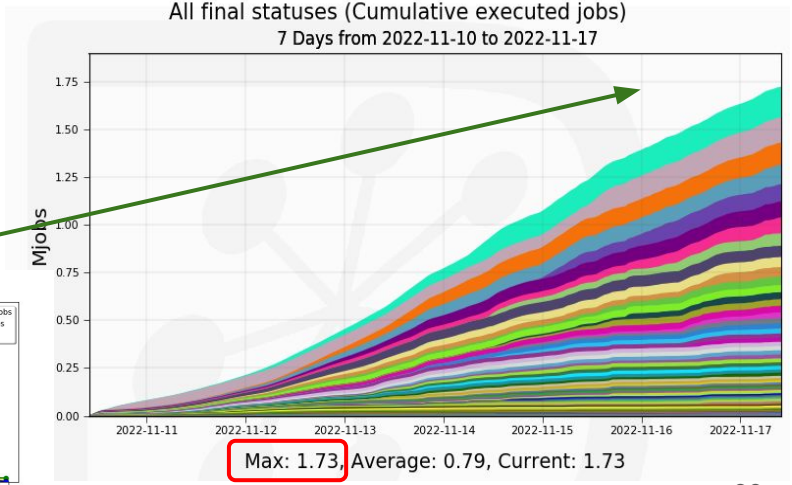


>850k waiting jobs
New all time record



- My call really had an impact
- **And the grid worked really well**
  - 10-15k jobs done/h
  - 1.73 M jobs done in last 7 days
  - Limited # of failed jobs: thanks!
  - Today only 38k jobs waiting
- **Great work by DC!**
- Still a good time to submit then!





Michel, Stefano, Umberto

We need your help!

- Computers are not so smart. Sometimes, they fail.
  - **"Sometimes" x Huge Resources = "Often"**
- The computing system need 24 hour x 7 day care.
- Please help us as a Data Production Shifter. You can book at shift.belle2.org
  - A nice manual is already prepared.
- If you have some experience as data production shifter, please become an Expert Shifter.
  - The Expert Shifter training course is open.
- Looking for a task? We have several for you!
  - Data production tasks
  - Computing tasks