Belle II RAW data management: The Online-Offline data transfer system

Matthew Barrett, Takanori Hara, Michel Hernández Villanueva[A], Kunxian Huang[B], Dhiraj Kalita, Petteri Kettunen, Prashant Shingade[C]

KEK, University of Mississipi[A], National Taiwan University[B], Tata Institute of Fundamental Research[C]

30th July 2020

ICHEP 2020 | PRAGUE

40th INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS

VIRTUAL CONFERENCE

28 JULY - 6 AUGUST 2020

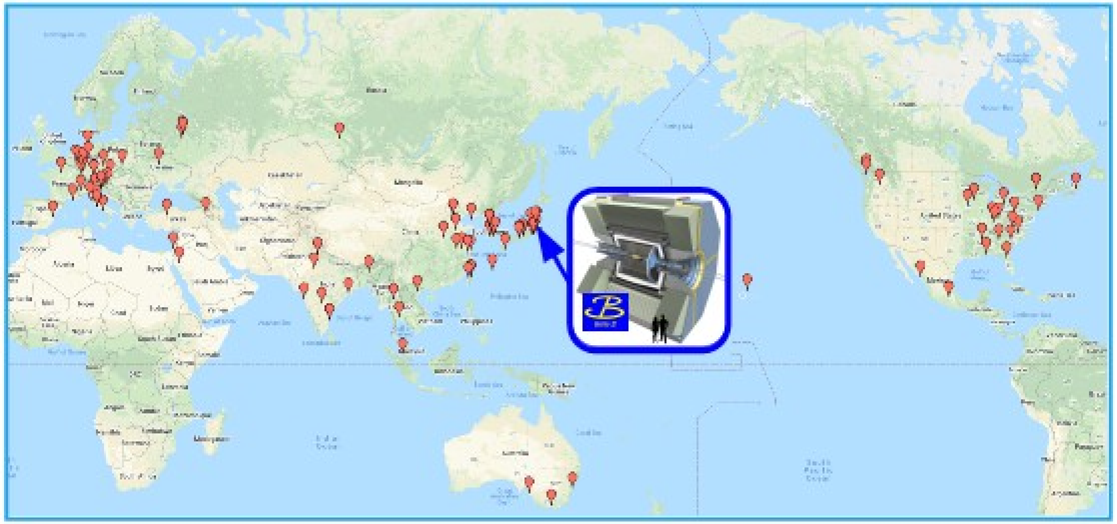
PRAGUE, CZECH REPUBLIC

Outline

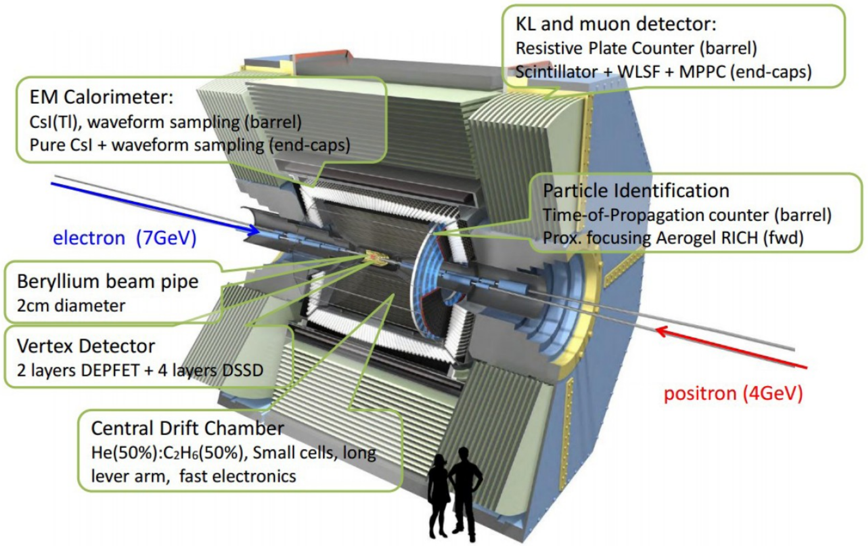
- The Belle II experiment
- Belle II data flow
- Evolution of the data transfer system
- Data transfer operations and monitoring

The Belle II experiment

- Belle II is a particle physics experiment located at the KEK laboratory in Tsukuba, Japan.
  - The Belle II experiment is the successor to the Belle experiment (ran from 1999-2010).
  - Started first physics run in March 2019.



Belle II Detector

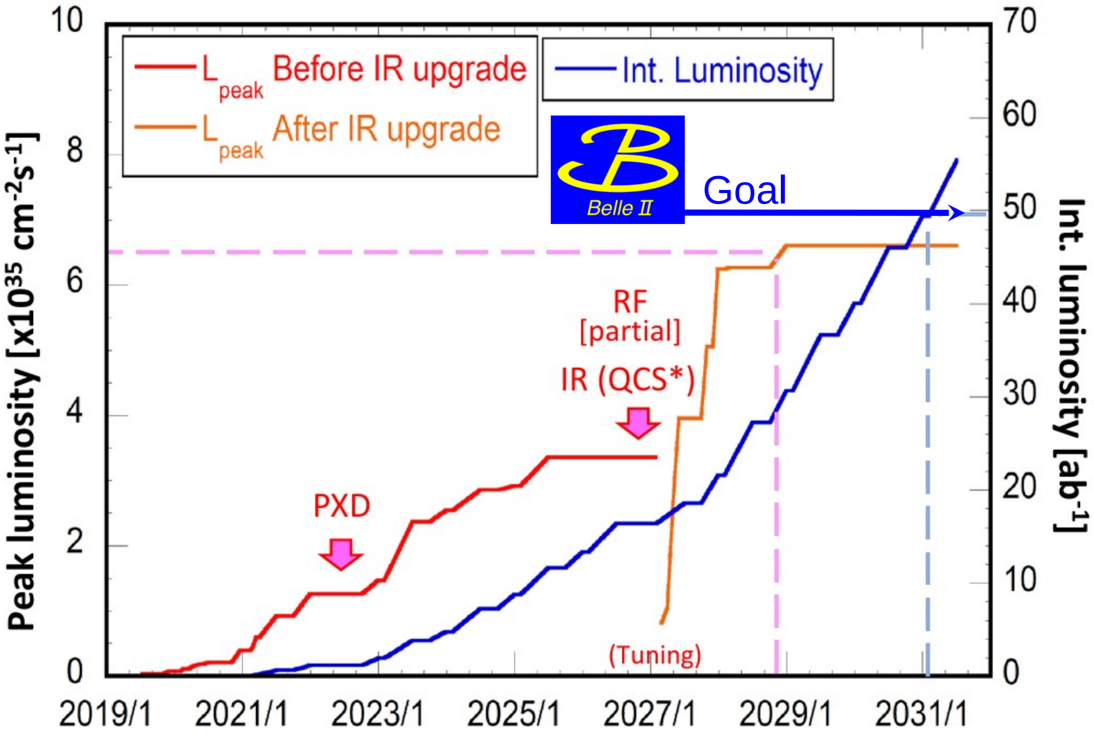

KL and muon detector:
Resistive Plate Counter (barrel)
Scintillator + WLSF + MPPC (end-caps)
EM Calorimeter:
CsI(Tl), waveform sampling (barrel)
Pure CsI + waveform sampling (end-caps)
electron (7GeV)
Particle Identification
Time-of-Propagation counter (barrel)
Prox. focusing Aerogel RICH (fwd)
Beryllium beam pipe
2cm diameter
Vertex Detector
2 layers DEPFET + 4 layers DSSD
positron (4GeV)
Central Drift Chamber
He(50%):C2H6(50%), Small cells, long
lever arm, fast electronics

- The Belle II collaboration is made up of:
  - over 1000 members
  - from over 100 universities and research institutions
  - in 26 countries.

Belle II Luminosity and Data Size

- The goal of Belle II is to collect a dataset of 50 ab^{-1} (50 times that of Belle) to hopefully discover new physics.
  - 50 ab^{-1} of raw data corresponds to approximately 60 PB.
- The rate at which the data are collected will increase over the data taking lifetime of the experiment.
- In later years the instantaneous luminosity be O(100) times what it was in 2019.
  - June 2020: Belle II set new world record instantaneous luminosity!

See talk by Kodai Matsuoka for overall Belle II status and prospects: https://indico.cern.ch/event/868940/contributions/3813745/

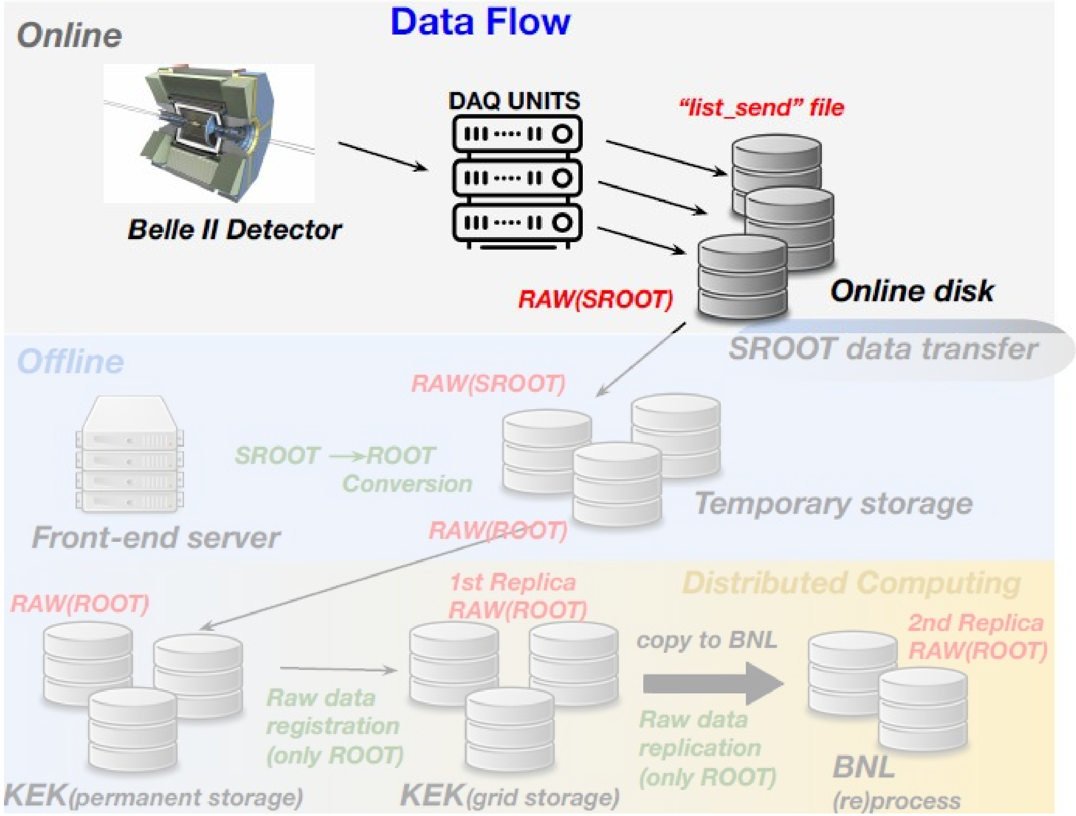


Units used in this talk: 1 PB = 2^{50} bytes, 1 TB = 2^{40} bytes.

Belle II Data flow

***Online*:**

- Data are recorded by the Belle II Data Acquisition (DAQ) System,
  - stored on servers located close to the detector.
- Written in a format called SROOT (sequential ROOT)
  - allows for serialised writing of the data with no compression.

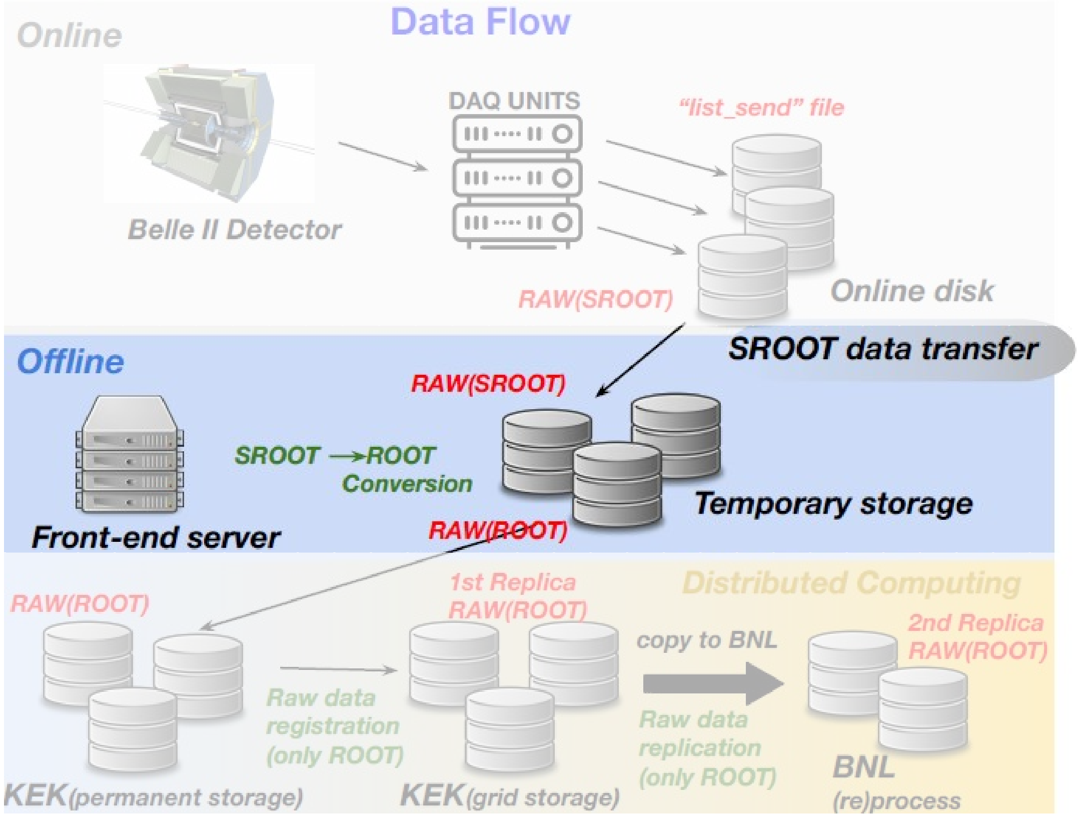


Belle II Data flow

[1] https://root.cern.ch/

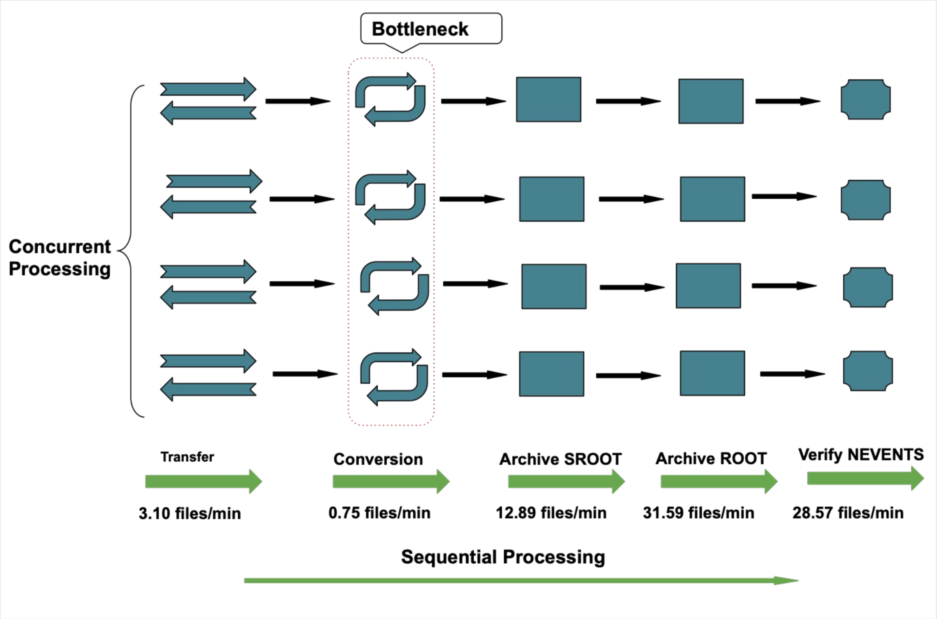
Offline:

- Data (SROOT) are transferred via a dedicated connection to the KEK Computing Research Center (KEKCRC) about 1.2 km from the detector.
    - SROOT files must be converted to standard ROOT [1] format, which is compressed, to be used for physics analysis.
    - Front-end (FE) servers at KEKCRC are used to convert the data into ROOT format.
- The raw data in ROOT format are copied to permanent storage;
    - It is enforced that at least two copies of every file must exist.



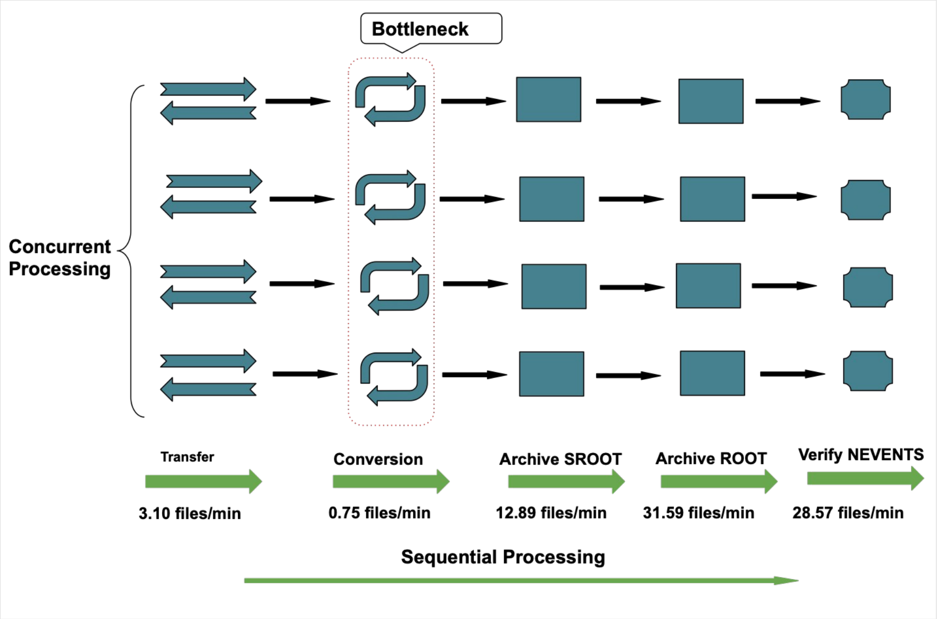
Data flow implementation – Early 2019

- Belle II first full physics run started in March 2019.
- A "`list_send`" file was created (typically once a day):
  - List of all SROOT files ready to be copied.
- The `list_send` file was copied to the FEs, then transfer of raw SROOT files to FEs was started.
  - SROOT files then converted to ROOT.
- After conversion the SROOT and ROOT files are archived.
  - Verification that #events in the SROOT files matches that in the ROOT files.



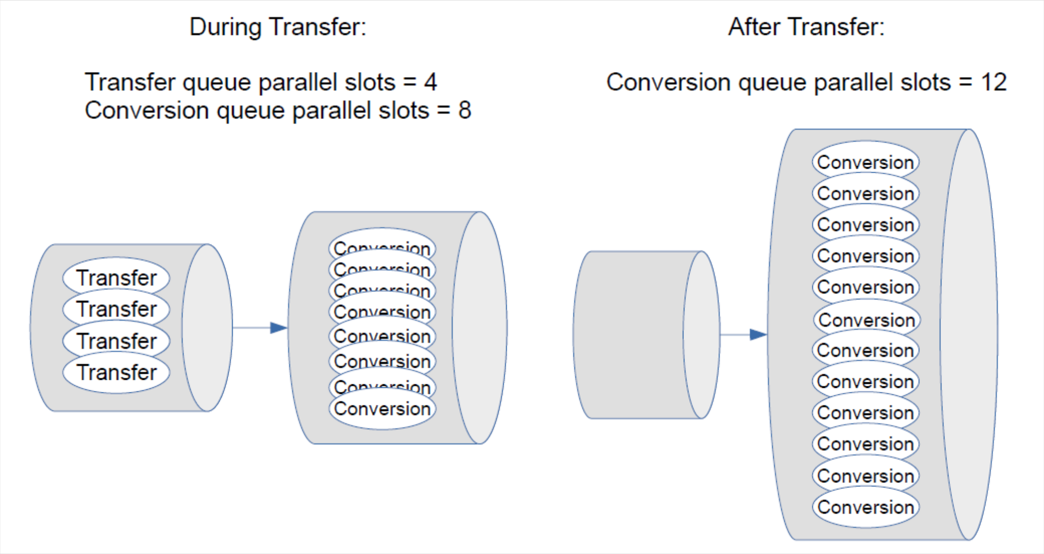
Data flow implementation – Early 2019

- After completion: data availability announcement sent to “data production” group.
- Many of the steps were initiated manually:
  - This helped to detect any anomalies,
  - but such a manual system could not scale to the higher data rates expected.
- Sequential processing under-utilised the computing resources:
  - only about 25% of available CPU was used.
  - Bottleneck: SROOT → ROOT conversion.



Automated Implementation – From mid 2019

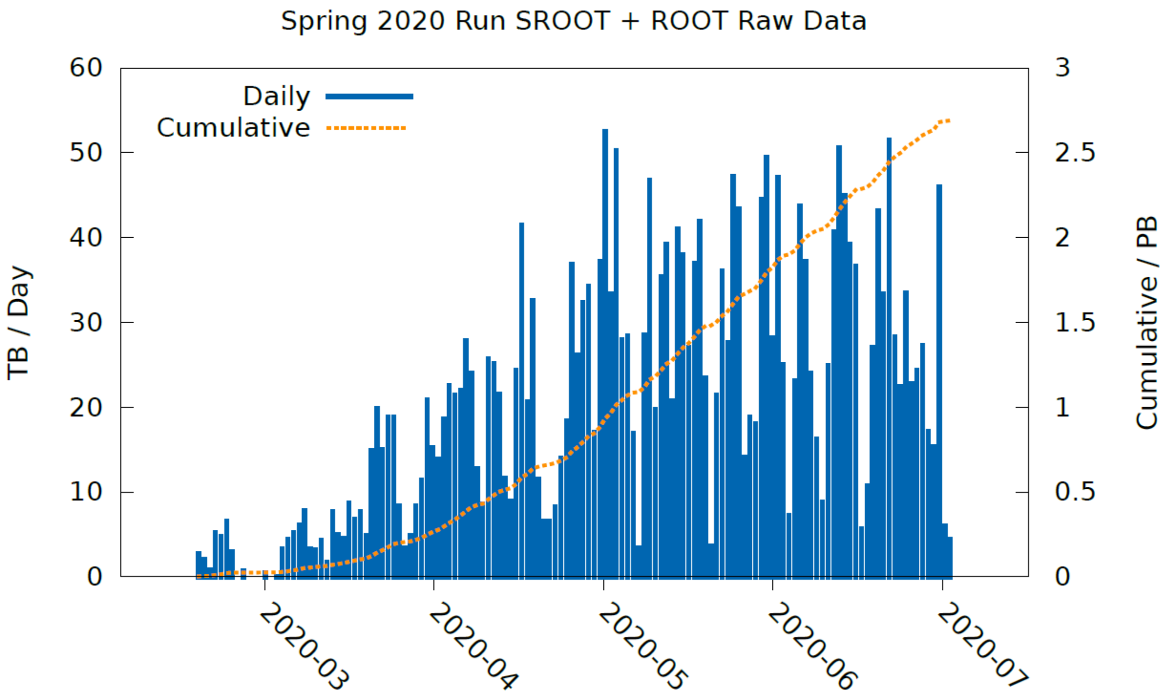
- System automatically searches for new `list_send` files.
  - If found, it spools them, and initiates transfers of SROOT files.
- Queue system (using task spooler [2]):
  - Split between transfers and conversion initially.
  - Once transfers have finished: all resources are dedicated to conversion.
  - CPU utilisation increased from ~25% to ~85%.
  - Automated copy to permanent storage and release to the collaboration then follow (if the data pass quality checks).



[2] http://viric.name/soft/ts/

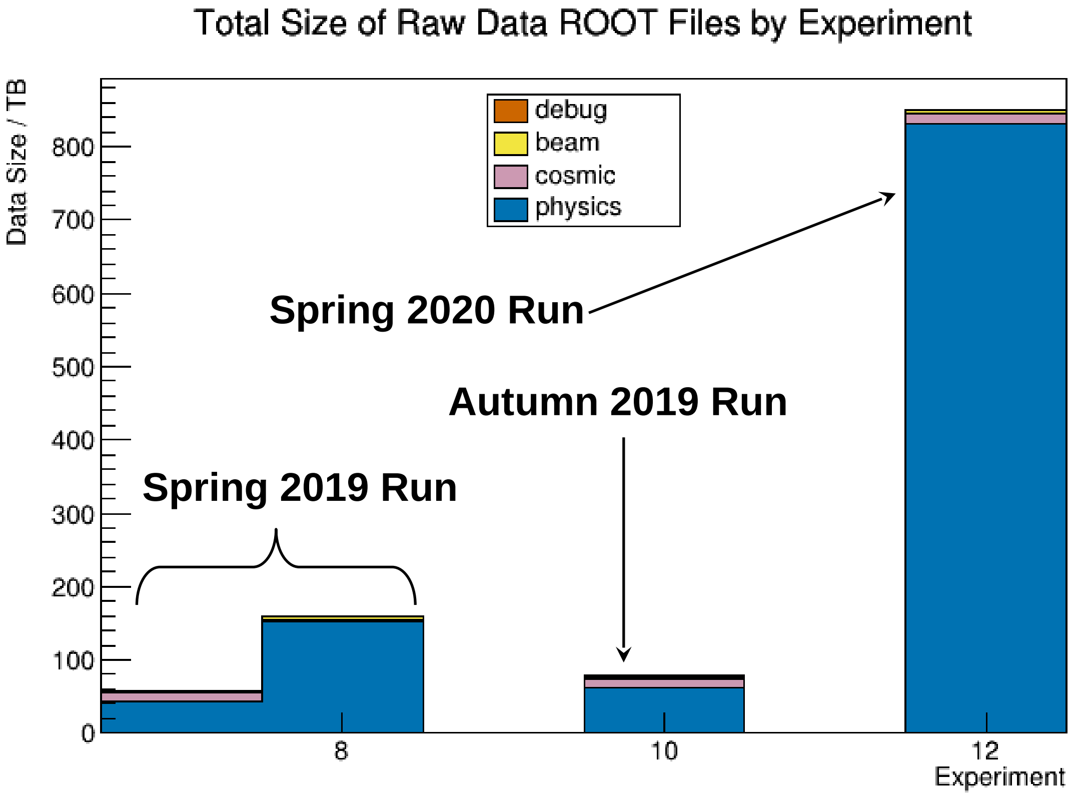
50 TB of Data Daily

- During Spring 2020 Run:
- 50 TB of (SROOT + ROOT) data produced many days.
- ROOT file typically ~45% size of SROOT file.
- Keeping SROOT files useful for understanding/debugging early data.
- We will not store SROOT files indefinitely:
  - They will be deleted.
  - ROOT will be the only RAW data.



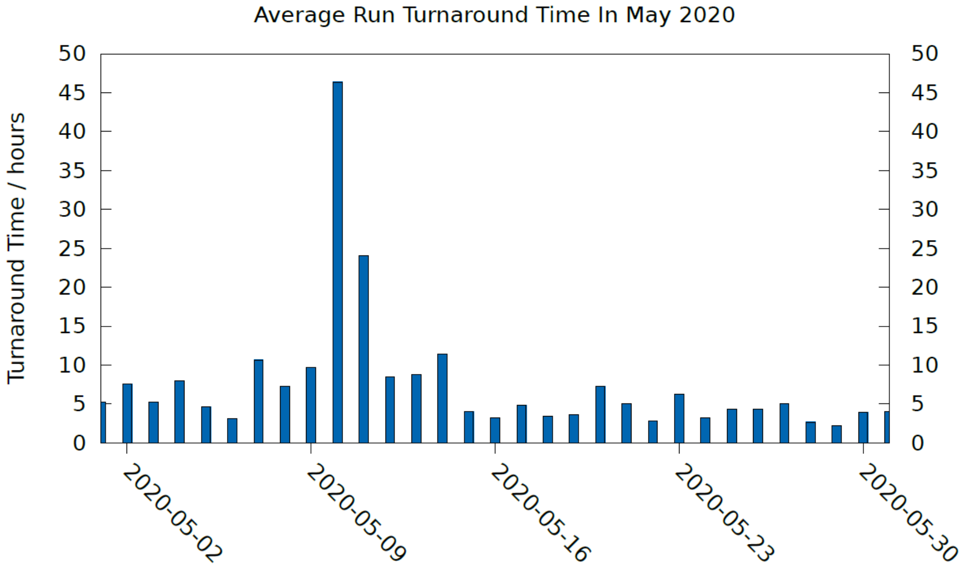
Over 1 PB of ROOT Data Recorded

- As of July 2020, the total size of ROOT files recorded so far is over 1 PB.
  - Most data are from physics runs.
  - Also beam studies, global cosmic runs, and runs for debugging.
- Luminosity will increase in future Runs:
  - Filtering at the HLT (High Level Trigger) level must be applied:
    - Will reduce data size by ~9×.
  - Maximum DAQ rate: 1.8 GB/s.
  - Data transfer system throughput over 2 GB/s demonstrated.



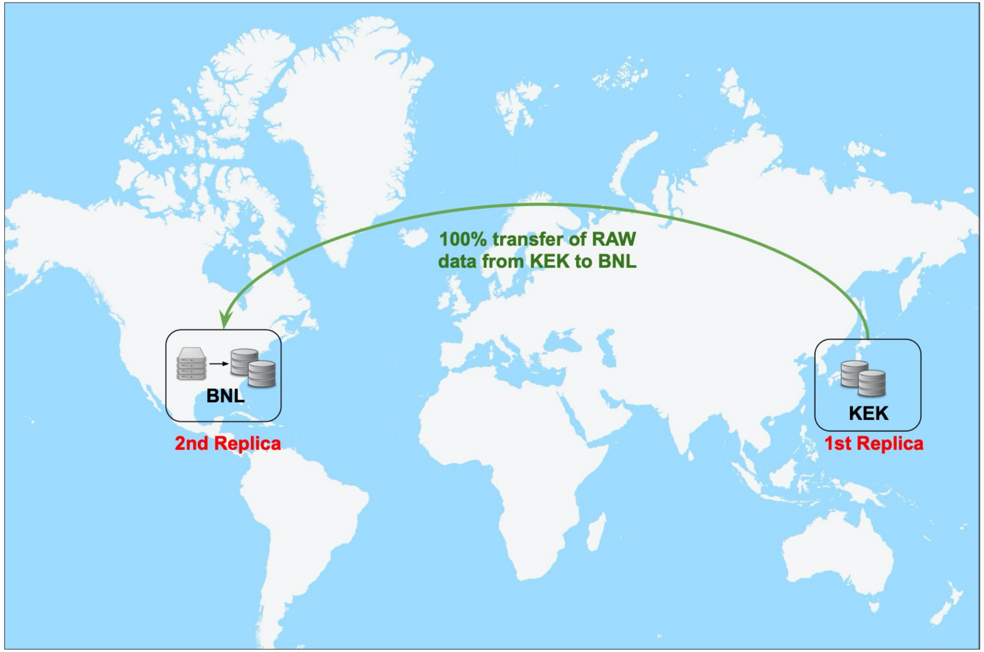
Turnaround time and the Fast Lane

- Turnaround time:
  - Time from end of run until ROOT files have been archived, and are available to the collaboration.
  - Generally 5 – 10 hours.
- Sometimes sub-detector experts need the data faster.
  - E.g. during a study with changing detector or accelerator conditions.
- We have created a “fast lane” to allow experts to request data faster.
  - For short runs, files may be available within about ten minutes.



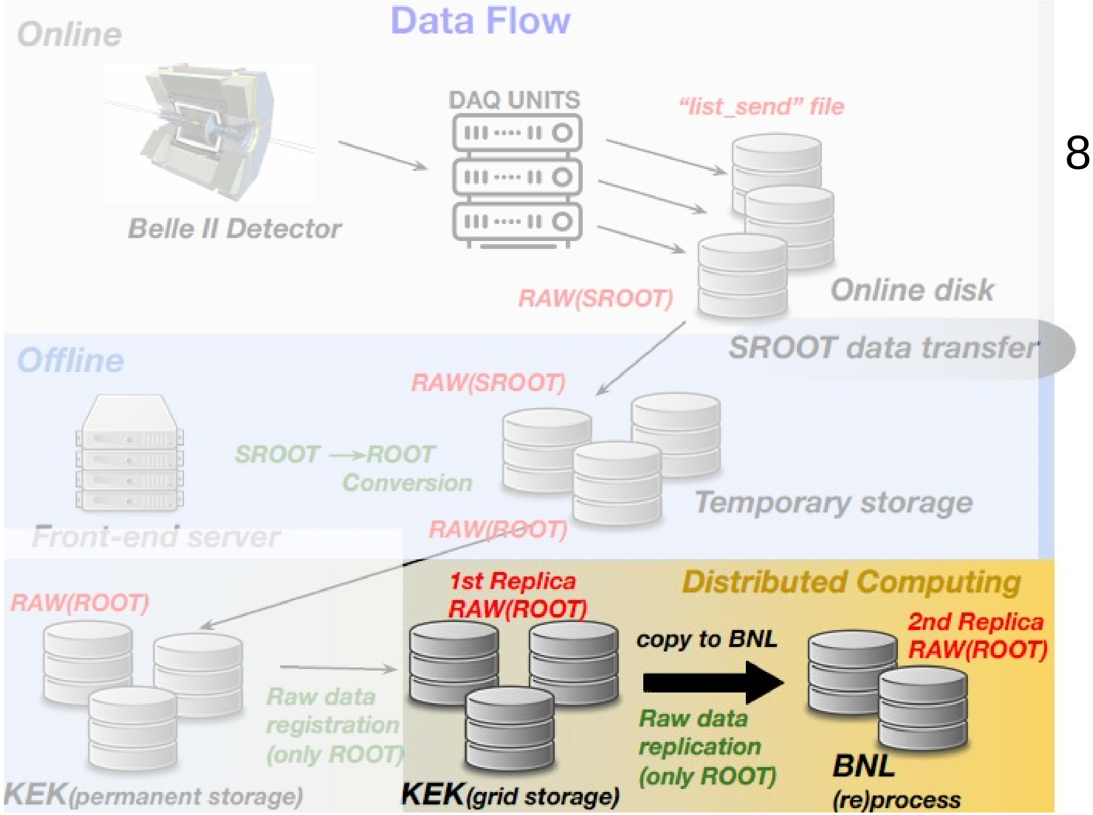
Permanent Storage of Data

- Belle II utilises the grid for distributed data analysis.
    - *DIRAC* and *BelleRawDIRAC* [3] used to distribute raw data.
- Two permanent copies:
    - One permanent copy at KEK.
    - Second permanent copy at Brookhaven National laboratory (BNL), USA.
- From 4th year of data taking operations:
    - Second permanent copy split between BNL and sites in Italy, Germany, Canada, and France.

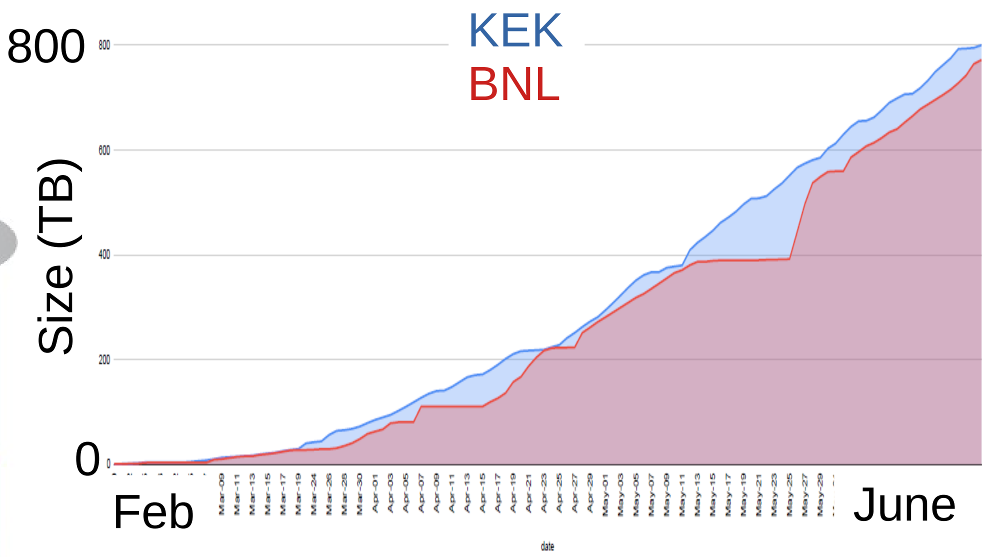


[3] https://indico.cern.ch/event/773049/contributions/3474468/

Transfer to BNL



Cumulative raw data size at KEK and BNL



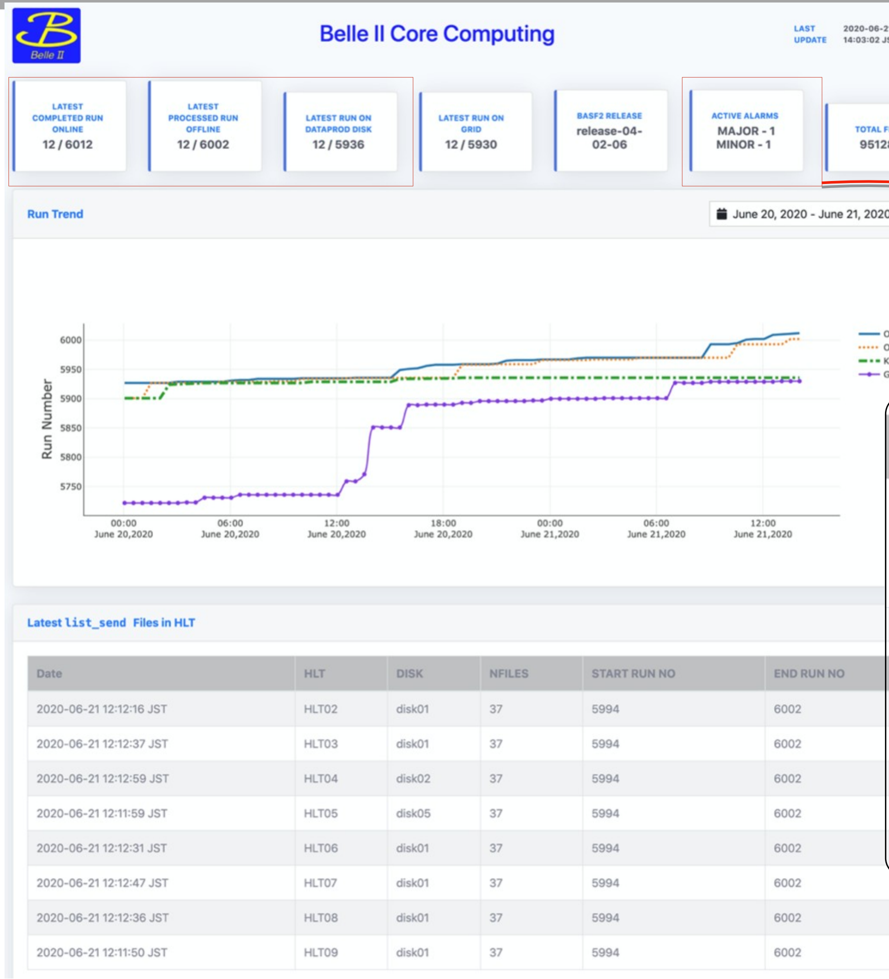
- Automated system to transfer raw data to BNL deployed during Spring 2020 run.
  - Features from, e.g. development periods, visible in above graph.

Control Room Monitoring

Shifter records the run information end of shift

Visual feedback for shifters about the progress of the runs from detector to grid.

Information about arrival of new SROOT files to be processed offline.



Belle II Core Computing

LAST UPDATE 2020-06-21 14:03:02 JST

LATEST COMPLETED RUN ONLINE 12 / 6012

LATEST PROCESSED RUN OFFLINE 12 / 6002

LATEST RUN ON DATAPROD DISK 12 / 5936

LATEST RUN ON GRID 12 / 5930

BASF2 RELEASE release-04-02-06

ACTIVE ALARMS MAJOR - 1 MINOR - 1

TOTAL FILES 951287

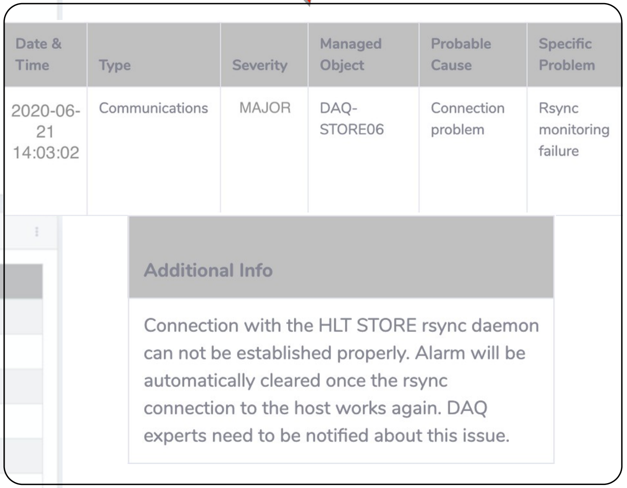
Run Trend

June 20, 2020 - June 21, 2020

Latest list_send Files in HLT

Date	HLT	DISK	NFILES	START RUN NO	END RUN NO
2020-06-21 12:12:16 JST	HLT02	disk01	37	5994	6002
2020-06-21 12:12:37 JST	HLT03	disk01	37	5994	6002
2020-06-21 12:12:59 JST	HLT04	disk02	37	5994	6002
2020-06-21 12:11:59 JST	HLT05	disk05	37	5994	6002
2020-06-21 12:12:31 JST	HLT06	disk01	37	5994	6002
2020-06-21 12:12:47 JST	HLT07	disk01	37	5994	6002
2020-06-21 12:12:36 JST	HLT08	disk01	37	5994	6002
2020-06-21 12:11:50 JST	HLT09	disk01	37	5994	6002

Alert system for shifters and experts.



Date & Time	Type	Severity	Managed Object	Probable Cause	Specific Problem
2020-06-21 14:03:02	Communications	MAJOR	DAQ-STORE06	Connection problem	Rsync monitoring failure

Additional Info

Connection with the HLT STORE rsync daemon can not be established properly. Alarm will be automatically cleared once the rsync connection to the host works again. DAQ experts need to be notified about this issue.

- Delays in data flow can reveal issues on the online side.
- Monitoring screen added in the Belle II control room to see progress of data transfers.

Summary

- Belle II is a particle physics experiment that started taking data in Spring 2019.
    - The instantaneous luminosity will increase by $O(100)$ times by the end of data taking operations compared to 2019.
- An automated data transfer system has been implemented to transfer data from the detector to (multiple copies on) permanent storage.
    - This system will scale to the higher data rates expected during later years of data taking.
    - Hardware at KEKCRC replaced every 4 years, ensuring system will meet experimental needs.
- Performance monitoring and making information easily accessible to the Belle II collaboration generally have also been key to the design of the new system.
- The new system has been operating in production since June 2019, and has performed robustly.
    - Journal paper describing the system has been submitted.