素粒子原子核研究所・計算機グループは、KEK で行われている素粒子原子核実験や計画における計算機・ソフトウェアの開発・改良ならびにその運用をサポートし、また国内外の研究機関と協力して将来的な実験・計画に適応できる計算機環境を目指して活動を行っている。

現在の主な活動は Belle II コンピューティング・グループに参加し、実験におけるオンライン－オフライン間 RAW データ転送とその分散管理、さらに分散計算環境の改良およびその定常的運用である。

1. オンライン-オフライン間 RAW データ転送とその改良

Belle II 実験では DAQ グループが Sequential ROOT（以後 SROOT と表す）と呼ばれる独自のフォーマットで RAW データをオンラインストレージに記録する。それを計算機グループがオフライン側計算機（KEKCC : KEK 中央計算機システム）に転送、そこでより汎用的に使用できる ROOT 形式にデータを変換し、ならびにデータ保全のため KEKCC のテープシステムに記録する。さらに変換後の RAW データのレプリカを国外にある複数の RAW データセンター上に速やかに作成する。これら一連のワークフローはそのモニター・システムも含め 2021 年までにほぼ自動化されている。現在、Belle II 実験は 2022 年 7 月から SuperKEKB 加速器が運転停止中（LS1）のため、ビームを使ったデータ取得は行われていないが、検出器の性能評価のため宇宙線データなどが不定期に取得されている。なお、現在までに ROOT 形式に変換した RAW データは約 2.7PB 蓄積されている。このうち物理ランは 428 fb^{-1} に相当する約 2PB、残りは宇宙線などの検出器性能評価等の目的で取得したデータである。これら RAW データのオフライン側への転送や KEKCC でのアーカイブ、米独伊仏加にある複数の RAW データセンターでのレプリカの作成などデータ管理をしつつ、計算機グループでは LS1 後のビームデータ取得再開に向けたデータ転送システムの改良を DAQ グループと共に遂行している。

SROOT 形式は圧縮されておらず、データサイズは ROOT 形式の 2〜3 倍の大きさがある。このためオンライン側からオフライン側へデータ転送する際の転送量が多くなり、今後加速器のルミノシティが向上するに従って、転送負荷が大きくなると予測される。またオフライン側へデータ転送した後、形式を変換するためには CPU が必要となり、さらに SROOT を読み ROOT を書き出すためにディスクの I/O 負荷も生じる。そこで予てより、オンライン側で直接 ROOT 形式の RAW データを書き出す事が検討され、DAQ グループと協議の結果、LS1 期間中にこれを導入することを決定した。またこれと同時に、データ転送をより高速化するため、使用するプロトコルを rsync から xrootd へ移行すること、ならびにより柔軟かつ連続的なデータ転送を行うため、データベースを介した情報共有を行うことも決定した。図 1 にこれまでのオンライン－オフライン RAW データ転送のワークフローと LS1 後に開始するワークフローを示す。また同時に行う改良についてもまとめて表示している。

これらの開発は 2022 年 11 月より本格的に開始し、概念設計、各コンポーネントの開発、ユニットテストを繰り返し行ってきた。 そして、2023 年 12 月からのビームデータ取得再開に向けて、7 月にシステム全体のストレステストを行い、その後、デバッグならびに最適化などシステムの改善を継続的に行う予定である。

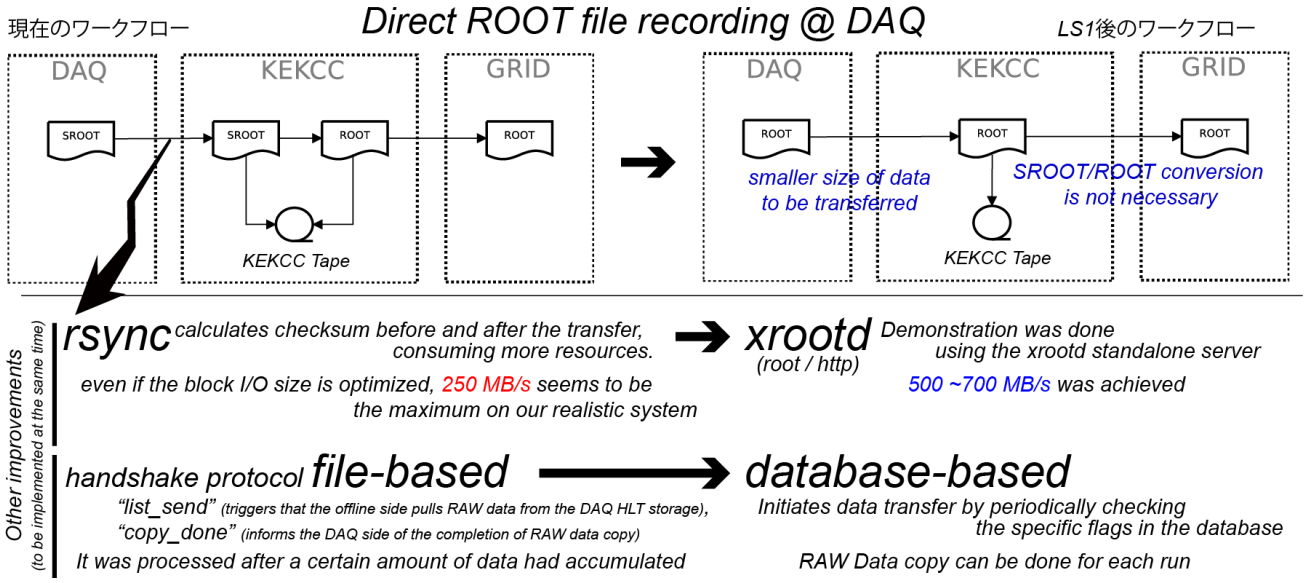

図 1：オンラインーオフライン RAW データ転送の改良点についてまとめた概念図

2. Belle II 分散計算基盤の運用と根幹システムの変更

Belle II 実験では莫大なデータ量を処理または解析するため、実験に参加する各国の機関ならびに計算センターで分散計算環境を整え、必要なデータの分散管理や同時に 30k 以上のジョブを投入できる 計算資源を備えている[1]。扱うデータは前節で触れた RAW データだけではなく、そこから派生した物理解析用データ、検出器較正用データ、ユーザーデータなど多岐にわたり、現在総計 15PB 以上のデータが世界各地の計算センター上で分散管理されている（図 2 参照）。また RAW データの 1 次処理、物理 skim データの作成、モンテカルロ・シミュレーション・データの作成、などプロファイルの異なったさまざまなタイプのプロダクションを滞りなく実行するため、日々運用を行っている。

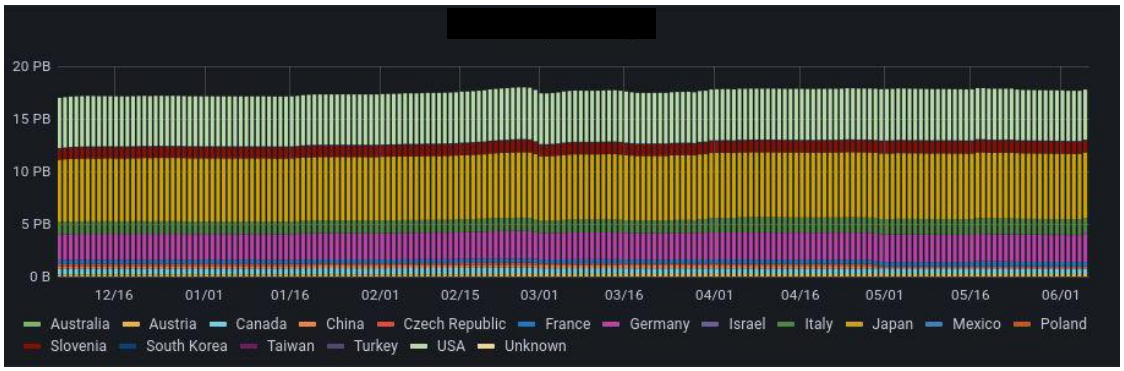

図 2：現在までに Belle II 分散計算環境下で保持しているデータの総量ならびにその分布。色の違いは分散計算環境に参加している国・地域を表す。

特に 2021 年以降、解析するべき物理データの量も増えてきたことから、分散計算環境を使ったユーザーによる物理解析が本格化している。図 3 に 2022 年 4 月から 2023 年 6 月までのユーザージョブ数を示す。常時、平均して 5000 本のジョブが流れていることが分かる。また同期間で計 300 名

ほどのユーザーがなにがしかのジョブを分散計算環境上で実行していることも示している。

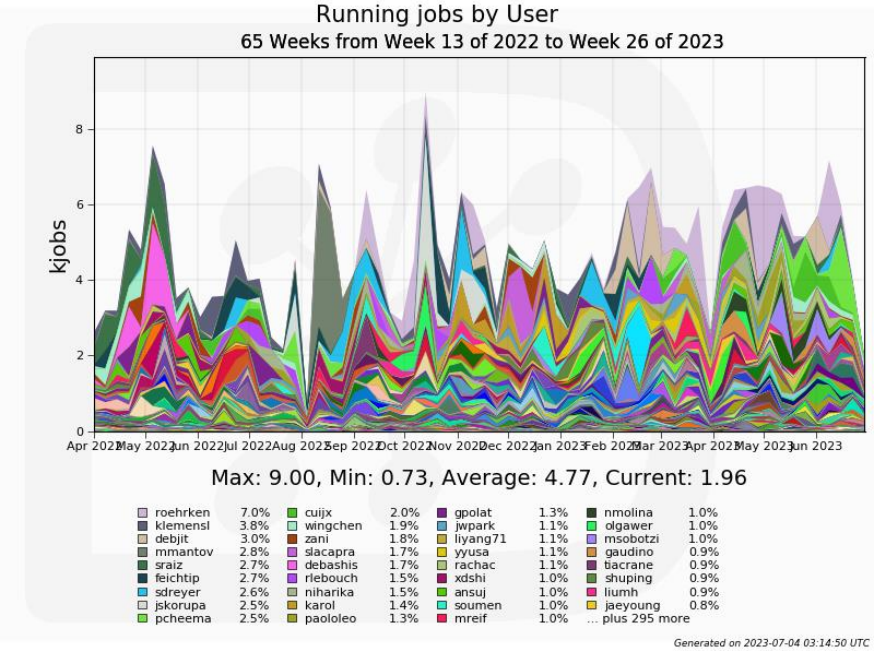

図 3：2022 年 4 月以降の物理解析ユーザーの実行ジョブ数

これら分散計算環境の定常的運用と平行して、このシステムの根幹をなすユーザー認証やサポートが切れたデータ転送プロトコルの置き換え作業も随時行っている。現在、Belle II 分散計算環境は WLCG (World-wide LHC Computing Grid)[2]が整備したグリッド・コンピューティングと呼ばれる広域分散処理システムを基盤としている。ここでは公開鍵認証基盤（x.509 認証を利用）を用い、安全性と利便性をともに担保しつつ、世界中に分散している計算機インフラを利用できる環境を実現している。しかし、この基盤が確立した 2000 年代初頭以来、クラウドなどの新しい計算システムなどが一般化し、これらの環境により適したトークンベースの認証システムへの移行が現在 WLCG で進められている。Belle II 分散計算環境は WLCG のインフラを利用している部分が多く、WLCG の移行スケジュールに従って、同様に x.509 認証からトークンベースの認証システムに移行すべく、現在、KEK 計算科学センターの協力の下、作業中である。また WLCG が Belle II 分散計算環境でも使用してきた SRM (Storage Resource Manager)ならびに gsiftp (Grid Security Infrastructure File Transfer Protocol)と呼ばれるサービスのサポートを終了するため、WebDAV (Web Distributed Authoring and Versioning) というプロトコルへ移行する必要もある。こちらの移行も 2022 年より進めており、2023 年 5 月時点で殆どの Belle II 分散計算環境に参加し、ストレージを提供しているサイトの WebDAV への対応が完了している。

3. 参考資料

[1] : T.Hara for the Belle II Computing group, “Computing at the Belle II experiment”, *J. Phys. Conf. Ser.* 664 (2015) 012002

[2] : “Worldwide LHC Computing Grid”, http://wlcg.web.cern.ch/