1. 素核研コンピューティング・グループの背景

昨今の素粒子原子核実験では加速器の高エネルギー化・高輝度化や検出器の大型化・細密化が進み、そこから得られるデータ量が飛躍的に増加している。そのためこれらのデータを効率よく処理し、物理結果を迅速に導き出すためには、計算機が必要不可欠となっている。LHC 実験が始まって以降、それまでの標準的はデータ処理ならびに解析環境である一極集中型計算モデルから、他地点の計算機資源を高速ネットワークで有機的に接続し分散処理を行う分散計算モデルが主流となっている。日本においても Belle II 実験の開始に伴い、KEK をホスト機関とした分散計算モデルを採用することとなった。ここで用いられているハードウェア、ソフトウェアのインフラや計算機技術は欧米で開発されたものが多い。しかし、計算機資源の規模、国際間の高速ネットワーク環境の違いなど、欧米とは異なる環境下で大型素粒子実験のための分散計算モデルを構築することは、既存のインフラを利用するだけでは実現できず、この環境に適合した新たな枠組みの形成、計算資源の確保、ソフトウェアの改良など多岐にわたる開発が必要となる。一方、Belle II 実験以外でも同様の分散計算モデルを採用している ILC 計画や ATLAS 実験などは利用するネットワークや計算機技術を共有しているものもある。そのため Belle II 実験で開発した計算機技術は他でも利用できる可能性がある。また世界的に見ると素粒子原子核実験の枠を超え、SKA などの天文分野でも分散計算モデルが浸透し始めている。そこで、国外の計算機・ソフトウェア環境の動向を考慮しつつ、国内の素粒子原子核実験における計算機環境の方向性を示し、国内外の機関と協力して実現に向けた活動を行うため、素核研では 2018 年 5 月より、コンピューティング・グループを立ち上げた。グループとは言え現在は職員一名での活動となっており、職務内容は Belle II 実験の為の分散計算モデルの構築ならびに運営、ソフトウェアの整備などを中心とした活動を Belle II 計算機グループと共同で行う一方、DPHEP(Data Preservation in High Energy Physics)[1]や HSF(HEP Software Foundation)[2]などには KEK で行われている実験・計画全体の対外的窓口として参加している。

2. Belle II 分散計算モデルの運用ならびに改良

Belle II 実験における分散計算モデル導入のための活動は素核研コンピューティング・グループ発足前の 2008 年から開始した。その後、ソフトウェアやミドルウェアなどのテクノロジーの選択を行い、当時 LHCb で開発され使われ始めた DIRAC[3]をワークロードならびに分散データマネージメント・システムとして採用、Belle II 実験に適合するよう拡張モジュールやユーザー・インタフェース、各種モニターなどを整備した。これらを組み合わせシステムとして構築し、2013 年 3 月、KEK を中心に CYFRONET(ポーランド)、DESY および GridKa(ドイツ)、SiGNET（スロベニア）、UA-ISMA（ウクライナ）、CESNET（チェコ共和国）、KISTI（韓国）、PNNL および FNAL など（アメリカ）が参加し、第一回目の大規模 MC 作成テストを行った。その後、何度も大規模 MC 作成テストや長期間のストレステストなどを通して改良を行い、国内外におけるより多くの計算機資源を取り込むことで徐々に規模を拡大、現在では約 60 研究機関が計算機資源を提供し、約 25k コア相当、16PB のディスク、12PB のテープ領域を有する分散計算システムとなっている。図 1 に Belle II 分散計算システムを立ち上げてから現在に至るまでの約 10 年間における同時処理したジョブ数の推移を示す（色

の違いは CPU を提供した国を表す)。年を経る毎に処理可能なジョブ数が増え、また連続的に運用し続けていることが分かる。国際ネットワークについても国立情報学研究所の協力の下、日米、日欧間の直通高速回線の実現に寄与した。この過程の中で、2018 年より Belle II 計算機グループから独立する形で素核研コンピューティング・グループを発足し、それ以降は Belle II 計算機グループおよび KEK 計算科学センターと共に分散計算システムの運用ならびに改良を行っている。なお、Belle II 分散計算システムの詳細については文献[4]に記述している。

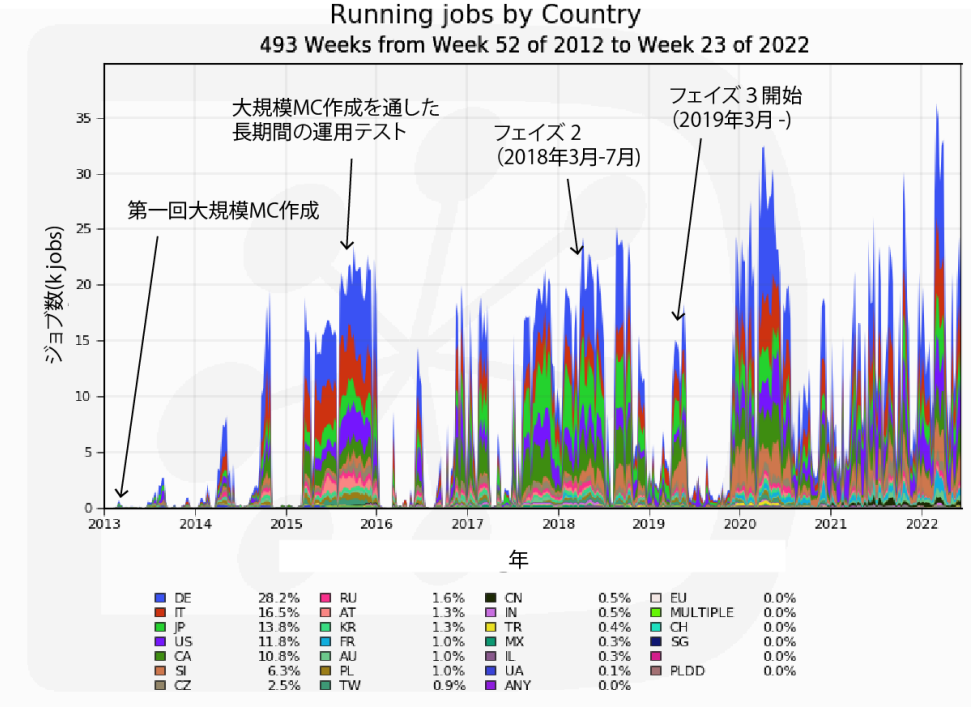

図 1：Belle II 分散計算システムを立ち上げてから現在に至るまでの約 10 年間における同時処理したジョブ数の推移。色の違いは CPU を提供した国を表す

最近では恒常的な Belle II 分散計算システムの運用に加え、より効率的な分散データ管理が可能となる Rucio[5]と呼ばれるソフトウェアを 2021 年初旬に導入した[6]。これはもともと ATLAS 実験で開発され、その後 CMS 実験など他の素粒子実験でも導入されているもので、これにより今後 SuperKEKB 加速器の性能向上によるデータ量の増加や不必要なファイルの自動消去、データの使用頻度に対応したファイル管理などが実現できる。現在までほぼ 1 年以上 Rucio は大きな問題も無く稼働し、最近 1 年間で計 9PB、2,300 万ファイルを処理した。

3. オンライン-オフライン間 RAW データ転送ならびにその分散管理

Belle II 実験では DAQ グループが実験に特化した Sequencial ROOT 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側 HLT ディスクユニットに記録する[7]。Belle II 計算機グループはこの HLT ディスクユニットから RAW データをオフライン側計算機（KEKCC：中央計算機システム）に転送し、速やかに解析に適した標準 ROOT フォーマットに変換後、データ保全のため KEKCC のテープシステムに記録する。さらにこれら RAW データのレプリカを国外の大規模計算機センターである BNL（アメリカ）、CNAF（イタリア）、DESY・GridKa（ドイツ）、CC-IN2P3（フランス）に転送し、

分散して保存する。これらオンライン-オフライン間 RAW データ転送システムならびに RAW データの分散データ管理システムの詳細については文献[8]に記述している。初期のシステムは動作確認やトラブルシューティングのため、24 時間体制で手動によるデータ転送の制御やクオリティーチェックなどを行っていたが、徐々に自動化を進め、現在ではほぼ自動でオンライン側からのデータ転送、フォーマット変換、国外計算機センターへのレプリカ作成を行えるようになった。図 2 にここ 1 年間にオフライン側で保存した RAW データ総量の推移を示す。Belle II 実験では期間ごとに exp 番号が割り振られる。図 2 はそれぞれの実験期間が始まると DAQ 側から RAW データを吸い上げ、滞りなくオフライン側で保存されたことを示している。現在 (2022 年 6 月 14 日)の段階で総量は 2.5PB に達し、蓄積された Physics ランのデータも 2PB（約 400 fb^{-1} に対応）を超えている。

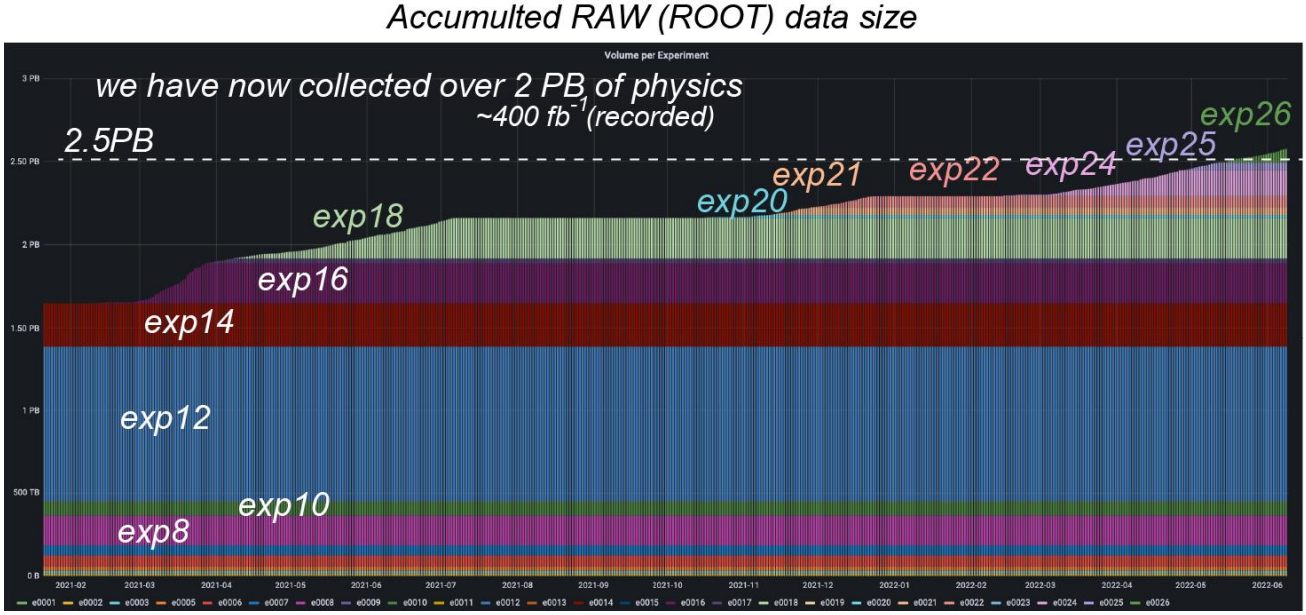

図 2：昨年 6 月からオフライン側で蓄積、保存した RAW データ総量の推移

4. さらなる改良を目指して

Belle II 分散計算システムは定常運用に入っているが、これで終わりではない。Belle II 実験においては益々SuperKEKB 加速器の性能が上がり、得られるデータ量も一桁ほど上がると期待される。それに伴いオンライン-オフライン間 RAW データ転送のスケーラビリティーも向上させなくてはならない。また、計算機技術の移り変わりは激しく数年前まで標準だったものが気づけば陳腐化している、ということはよくある。そのため外国研究機関の専門家と情報交換を密にし、技術のトレンドに逆らうことなく Belle II 分散計算システムに取り入れていく努力を常にしなくてはならない。Belle II 実験は 2015 年 4 月より LHC 実験で必要な計算資源を整備するために作られた WLCG (World-wide LHC Computing Grid)[9]にオブザーバという形で LHC 実験以外では初めて認められ、参加している。将来の分散計算システムにおける認証技術や次期 OS などの情報を世界と共有し、Belle II 分散計算システムに反映させるだけでなく、将来的に日本で行われる大規模素粒子原子核実験で使用される分散計算システムを構築できるよう、そのための基礎基盤作りも兼ねている。またその一環として若手の人材育成にも力を入れ、KEK 計算科学センターらと協力し毎年『粒子物理コンピューテ

ィングサマースクール』を開催しており、今年も 8 月 1 日から 8 月 5 日の開催を目指し準備中である。

5. 参考資料

[1] : Data Preservation in High Energy Physics : https://dphep.web.cern.ch/

[2] : HEP Software Foundation : https://hepsoftwarefoundation.org/

[3] : F.Stagni, A.Tsaregorodtsev, L.Arrabito, A.Sailer, T.Hara, X.Zhang, "DIRAC in Large Particle Physics Experiments", *J. Phys. Conf. Ser.* 898 (2017) 092020

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

[5] : M.Barisits *et al.*, "Rucio – Scientific data management", *Comput. Softw. Big Sci.* 3 (2019) 1, 11

[6] : C.Serfon *et al.*, "Integration of Rucio in Belle II", *EPJ Web Conf.* 251 (2021) 02057

[7] : 高エネルギーニュース 2014. 33. 3 伊藤領介，中尾幹彦，山田悟，鈴木聡，今野智之，樋口岳雄. Belle II 実験のデータ収集システム

[8] : M.Barrett, T.Hara, M.H.Villanueva, K.Huang, D.Kalita, P.Kettunen, P.Shingade. "The Belle II Online-Offline Data Operations System", *Comput. Softw. Big Sci.* 5 (2021) 1, 1

[9] : "Worldwide LHC Computing Grid", http://wlcg.web.cern.ch/