■研究紹介

Belle II 実験データ読み出しシステムのアップグレード

KEK 素粒子原子核研究所
山 田　悟
satoru.yamada@kek.jp

2024 年（令和 6 年）8 月 4 日

1　はじめに

2019 年より本格的なデータ収集を開始した Belle II 実験は, 2022 年 7 月からの Long shutdown 期間(LS1)を経て 2024 年 2 月に物理ランを再開した。LS1 では様々なアップグレードがなされたが，本稿では LS1 に作業が完了した Belle II 実験の読み出しシステムのアップグレードについて紹介したいと思う。

2　Belle II DAQ と読み出しシステム

読み出しシステムの定義は実験によって多少変わると思うが, Belle II 実験では図 1 に示す Belle II データ収集(DAQ)システム[1]において, フロントエンド電子回路(FEE)と下流の High Level Trigger(HLT)を結んでいる部分を読み出しシステムと呼んでいる。FEE からデータを読み出して処理を行い，イベントごとにまとめて HLT のサーバ群に送る役割を持つ。Belle II 検出器は七つのサブ検出器からなっているが，最内層のピクセル検出器についてはデータ量が他の検出器と比べて膨大になるので特殊なデータ収集系が構築されており，それ以外の六つの検出器については共通の読み出しシステムを用いている。

Belle II DAQ システムで最初に準備した読み出しシステムは COPPER(COmmon Pipelined Platform for Electronics. Readout)ボード[2]を用いたシステムである。このボードは先代の Belle 実験の後半から使われてきたもので, 実験での用途に応じて FINESSE と呼ばれるメザニンボードを搭載して使うことの出来る汎用性の高いボードである。Belle 実験では TDC ボードをメザニンボードとしてつけていたが, Belle II 実験では検出器信号のデジタル変換は検出器近傍に置かれている FEE で行うため，FINESSE としては FEE との高速データ通信を担う HSLB(High Speed Link Board)が搭載されている。FINESSE に加えて受信したデータの処理を行う Intel Atom CPU を持った CPU ボード，クロック受信に用いる TTRX(Trigger and Timing Receiver)ボードも搭載されており, また COPPER ボード自身についても新バージョンのボードを製作し，オンボードの PHY デバイスをギガビットイーサネット対応とすることで, 下流への高速データ転送を図っている。この COPPER ボードをベースとした読み出しシステムで，Belle II 実験開始以来データ収集を行ってきた。

3　アップグレードに向けて

2019 年の Belle II 実験のデータ収集開始以降も COPPER 読み出しシステムは順調にデータを処理できており，特に大きな問題は起きなかった。時折 COPPER ボードの CPU がダウンするというトラブルがあったため，クレートの電源容量を増強したり，ボードの冷却不足などの原因を取り除くなどして改善を行った。それでも 1 か月に数回のペースでは起こってはいたものの許容できない頻度では無く，概ね安定して運転が行えていたといえる。

そうした中で一つの懸念として，COPPER ボードを作っ

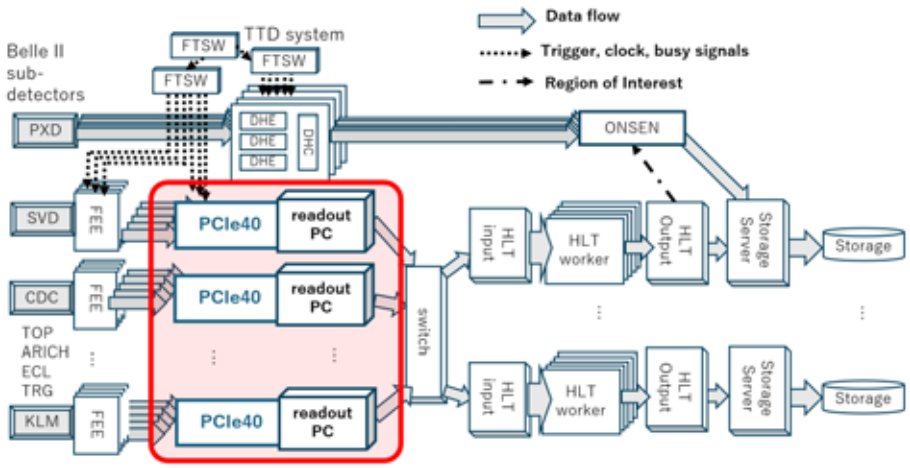

図 1　Belle II 実験データ収集システムのダイアグラム。枠で囲った部分がアップグレードした読み出しシステム。

てくれた業者さんからの生産中止部品の連絡が徐々に増えていたことがあった。Belle II 実験が今後 10 年あるいはもっと長く続くことを考えると，ハードウェアの故障が頻発してきた際に修理が難しくなるのではないかという心配である。これは COPPER ボードだけでなく，3 種類のメザニンボードについても同じことが言えた。

パフォーマンスという点からみても，従来の読み出しシステムは Belle II 実験で想定している最大トリガーレート 30kHz でも動く設計ではあるが，COPPER ボードで使われている通信規格は PCI バスやギガビットイーサネットであり, 実験開始当初においても既に PCIExpress のボードや 10G イーサネットのネットワークボードが出回っていた状況からすると刷新による改善の余地は大きいように見えた。また COPPER の FINESSE として最大 4 枚の HSLB を搭載できるが, このボードは 1 枚に一つの SFP 光モジュールを使って FEE と通信する形になっており, これを Firefly や Minipod などのトランシーバーにすることで 1 枚のボードで今より多くの FEE からのデータを受け取れる可能性があった。

こういったこともあり，まだ本格的なデータ収集が始まる前の 2017 年ごろから Belle II DAQ グループ内で，次の読み出しシステムに関わる検討会を開くようになっていた。そこで分かったことは, 読み出しシステムを, FEE の FPGA からのシリアル通信を下流の PC ファームにイーサネット経由で送るインターフェース，ととらえると異なる実験においてもかなり似通ったものが使われているということだった。これらの議論を踏まえて 2018 年頃から本格的にアップグレードの準備を進めることになった。

4　読み出しハードウェアについて

この読み出しシステムは Belle II DAQ システムの一部であり，DAQ グループが責任を持つところなのでグループ内で刷新プロジェクトを進めてしまってもよいのだが，その場合 Belle II コラボレーション内の技術力, パーソンパワーや資金をあまり活用できない。DAQ グループにおいても読み出しシステム以外に担当するコンポーネントも多く，人手も足りていなかったため， DAQ upgrade committee というのを作って Belle II コラボレーションワイドにプロジェクトを広げて提案を公募する形をとった。この committee のチェアをしてくれた方がシステム開発の知見がある人だったので，手続きは割とフォーマルな形で進められ，まずは DAQ グループから要求仕様の文書を発表し, 続いてプロポーザルの公募，そして各提案グループから要求仕様を満たすという Letter of Intent を提出してもらって審査するという流れになった。ここで留意したことは，このプロジェクトは既に Belle II 実験で動いている DAQ 内のシステムをアップグレードするものであるため，アップグレードする部分以外に極力影響を及ぼさないような外部インターフェースにする必要があるということであった。

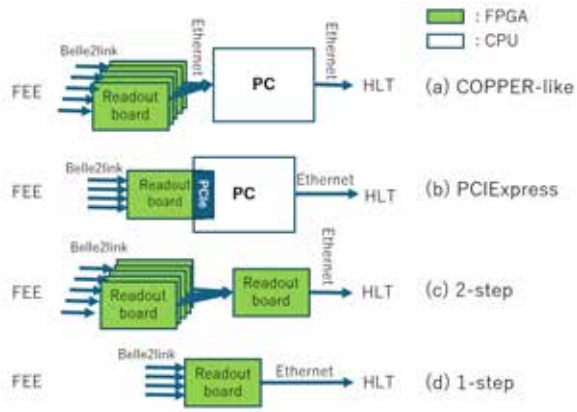

図 2　FEE と HLT をつなぐ新読み出しシステムの構成案。

読み出しシステムにおける外部とのインターフェースの一つは FEE との通信である。 Belle II 実験ではこの通信を Xilinx の FPGA 間シリアル通信の上に Belle2link[3]という中国 IHEP と KEK で共同開発された独自のプロトコルを構築することで実現しており, 同じ伝送ラインで FEE からのデータの受け取りと，装置の設定やモニターを行うスローコントロールの為の FEE のレジスタ読み書きを並行して行えるのが特徴になっている。要求仕様には新システムでもこの Belle2link を維持することを明記した。二つ目のインターフェースは下流の HLT サーバ群との TCP/IP 通信になる。これらの境界条件を考えると，新システムの構成は図 2 にあるように PCIExpress ボードか, あるいはボード上の FPGA が TCP/IP スタックを持っていてイーサネットに接続するタイプのいずれかになることが想定された。三つ目のインターフェースは Belle II 実験の TTD(Trigger and Timing Distribution)システム[4]というクロック信号やトリガー信号の分配を行うシステムとの接続で，この部分は接続するコネクタや通信プロトコルにおいて実験ごとに特色が出てくる箇所になっていると思う。

これらを考慮して 2018 年に策定した読み出しハードウェアの要求仕様は基本的には COPPER システムと同等の能力を最低限として要求するもので，主な内容は以下のとおりである。

- FEE との通信 ：Belle2link プロトコルによる FPGA 間の高速シリアル通信を行う
- データ処理 ： データフォーマッティング，データチェック，partial event building などを行う
- 下流の HLT サーバとの TCP/IP 通信を行う
- Belle II トリガー/タイミング分配システムとの通信，バッファあふれの際の BUSY 信号送出が出来ること
- トリガーレート最大 30kHz までの処理が可能であること
- 長期にわたり安定して運転可能であること

この仕様書を作成してBelle IIコラボレーション内でプロポーザルを募った結果，四つの提案が出た。ハードウェアの名前を列挙すると以下のようになる。

・FELIX (PCIExpress ボード) [5]
・PCIe40(PCIExpress ボード) [6]
・CPPF(MicroTCA ボード)[7]
・Aurora2PCIe(仮称) (PCIExpress ボード)

FELIX ボードに関しては ATLAS 実験の DAQ システム，PCIe40 ボードに関してはLHCb実験のDAQシステム, CPPFボードはCMS実験のトリガーシステム用に開発されていたものである。Aurora2PCIeボードについてはは KEK DAQ グループで新規ボードとして提案したもので，Xilinx の Ultrascale FPGA で FEE からのデータ受信と処理を行い，PCIExpress 経由でホスト PC にデータを送るというものであった。またバッファあふれによるデータロスを防ぐために，ボード上にバッファ用のメモリも載せるという構想になっていた。各案を図 2 に当てはめると CPPF ボードは(d)の 1 step の方式で，それ以外はすべて(b)の PCIExpress ボードになる。

DAQ upgrade committee で各プロポーザルを審査する中で，搭載している FPGA が Xilinx(AMD)社製か Altera(Intel)社製かという点も論点の一つになった。Belle II 実験の DAQ やトリガーシステムでは Xilinx FPGA を使っていたため，開発しやすさを考えると Xilinx を選ぶほうが良いのではないかという意見も出た。ただ Belle II 検出器の FEE の一部では Altera の FPGA を使っているものもあったため決定的な論拠とはならなかった。結局ハードウェアは最終的に Altera の FPGA を搭載したボードを使うことになるのだが，使用する IP コアの仕様が少し異なる点などを除けばそれ程大きな問題とはならなかったように思う。

プロポーザル書面だけで審査してハードウェア案が決まれば早かったのだが，結局どれも要求仕様を満たしているという結論になったので選定作業はあまり先に進まなかった。そこで，もう一段階 R&D フェーズというものを設定して各プロポーザルのハードウェアがある程度動くことを確認したのちに最終的な判断をすることになった。この段階でKEK DAQグループから提案を出していたAurora2PCIeボードについては R&D フェーズのためにまずプロトタイプを作る必要があったが，出揃った提案を見てみるとメモリを載せている点以外は他のボードと似かよった物になっており，また新規で作るため費用は他よりもかかるという状況であった。そこで KEK DAQ グループとしてはこの際自提案は取り下げて, 他の提案されたシステムの R&D 試験に協力する立場をとることにした。実際 R&D フェーズにおいては Belle II DAQ システムとのインターフェース開発において KEK DAQ グループとの共同開発が不可欠になるため，自前の提案があると他の提案グループとの作業がやり辛くなるという考えもあった。結局 R&D フェーズに進んだのが，

・FELIX ボード
・PCIe40 ボード
・CPPF ボード

で，次の段階としてそれぞれがテストベンチで動くようにファームウェアとソフトウェアの簡単な開発と試験を行うことになった。半年ほどに設定された R&D フェーズで三つの提案グループとの作業を並行して行いながらとりあえずハードウェアが動くようにするのはなかなか大変であったが，最終的には 2019 年 10 月の期限までにどの提案も一応動くという形になった。となると R&D フェーズを設定した思惑の一つである技術的な順位付けもまた難しくなり，どの提案でも問題ないという答申結果になってしまった。

図 3　PCIe40 ボード。

議論の結果，最終的には開発や保守におけるサポート人員やボード製作費用の負担主体などを総合的に判断して PCIe40 案を採用することになった(図 3)。これを提案したのは仏 IJCLab のグループで，もともと高エネルギー物理分野では LHC での実験をメインとしていた研究所だが，数年前にフランスの研究所として初めてBelle II実験に加わったこともあってこのアップグレードに積極的に協力姿勢を示してくれたことも大きかった。LHCb 実験に参加している IJCLab の研究者の方も Belle II コラボレーションに技術サポートという形で入ってもらい, アップグレードのプロジェクトに参加してもらった。

5　新読み出しシステムの開発

従来の COPPER 読み出しシステムは読み出しボードがクレートに入っており，図 2 の(a)にあるように Ethernet で readout PC に接続された形だったが，新しい読み出しシステムは図 2 の(b)のにある readout PC 内に読み出しボードがインストールされる形となる。PCIe40 ボードについてはLHCbアップグレードのマスプロダクションに加わる形でBelle IIに必要な 21 枚に予備を加えた 30 枚程を作って貰う算段がついたので, あとは PCIe40 の Arria10 FPGA のファームウェアと readout PC のソフトウェアの開発に集中する形となっ

た。図 4 に PCIe40 と readout PC のファームウェアおよびソフトウェアコンポーネントを示す。PCIe40 のファームウェアについてはハードウェア抽象化層にあたる LLI(Low-Level Interface)という枠組みが LHCb 実験グループで開発されており, FPGA の I/O 信号はそこを経由して得ることができるので，主な開発部分はファームウェアではデータを処理するユーザーロジック部分に加えて, Belle2link, TTD インターフェース，PC への DMA 転送といった外部とのインターフェース部分，ソフトウェアでは読み出しソフトウェアとスローコントロール部を作っていくことになった[8]。

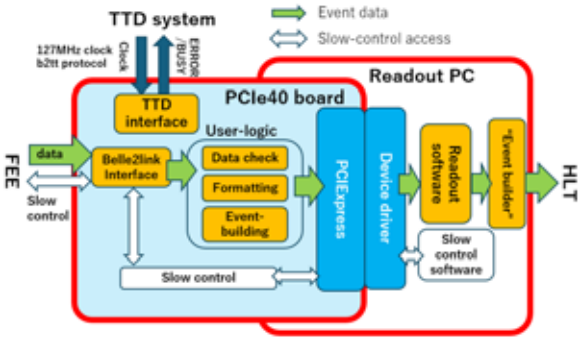

図 4　新システムのブロックダイアグラム。

まず FEE との通信部分の開発であるが, Belle2link の通信速度についてはオーバーヘッドを考慮しないラインレートは 2.54Gbps であり，COPPER システムではボードの処理速度を考慮して 1.27Gbps で設定されていた。この速度は FEE からの 1 チャンネルあたりの転送レートとしては現在も十分であり, FEE 側のファームウェアを変更しない方針もあって，新システムでも同じ速度にしている。これについては今後必要があれば高速化を図ることも可能である。通信機能としては，Belle2link ではパケットに付けられたヘッダの値によって FEE からの検出器のイベントデータとスローコントロール用の FEE のレジスタ読み書きを区別することができ, 同一ラインで両方の通信が可能になっている。スローコントロールの通信には FEE のレジスタ読み書きだけでなく，ファームウェアや設定データなどのやや大きいファイルをストリーミングで送るという機能も含まれる。COPPER ボードにおいては FEE からのイベントデータはボード上の FIFO チップに書き込まれ，一方でスローコントロール用のレジスタ読み書きデータは COPPER のローカルバスを経由して CPU メザニンボードと通信するという形だった。PCIe40 ではよりシンプルに，FPGA 内の異なる FIFO に二つのデータタイプを振り分けて扱う形にした。新システムのスローコントロール通信については図 5 に示すように，FEE からのイベントデータ量が増えても FEE のレジスタへの最大アクセス頻度が劣化しにくいという性能向上が得られている。

ユーザーロジック部においては COPPER ボード上の CPU ボードでソフトウェアによって行われていたデータ処理を PCIe40 の FPGA で行うことにした。ここでは各チャンネルのデータのサニティーチェックの為に, イベント番号, データの区切りとして付けているマジックナンバー, CRC(Cyclic Redundancy Check)値の照合を行い，また複数の FEE のデータで重複しているヘッダ情報(ラン番号，イベント番号，クロック情報)はまとめてひとつのヘッダにしてデータ量を減らすことも行う。その後各チャンネルのデータをまとめる partial event-building を行い，読み出しボード単位のヘッダとフッタを付与してデータのフォーマットを作る。PCIe40 では最大 48 チャンネルのデータ受信が可能となり, COPPER での 1 枚あたり最大 4 チャンネルという数字からは大幅な増加になる。データはスループットを考えて 256bit 単位でクロックごとに入力 FIFO からユーザーロジック部に読み出す形にしている。Belle II のデータフォーマットの基本単位は 1 ワード=32bit 単位であり, 256bit で割り切れない中途半端なデータを切り貼りしてつなぎ合わせるために，色々と場合分けの処理が必要となるが，なるべくクロック数を使わないようにロジックを組んだ。PCIe40 で使われている Arria10 FPGA のリソースはユーザーロジック部に使うには十分で，残ったリソースはバッファあふれの可能性を低くするために各チャンネルからのデータを一時的に貯めておく FIFO に費やしている。

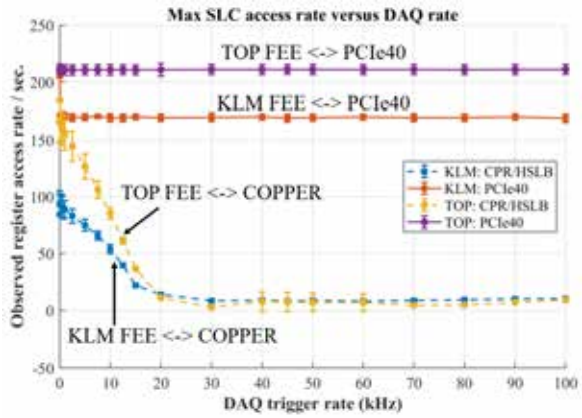

図 5　Belle2link でデータを FEE から読み出しながらスローコントロール(SLC)で FEE のレジスタにアクセスした際の最大アクセス頻度の比較。

PCIe40 でまとめたデータは PCIExpress を通じて readout PC に DMA 転送される。実は LHC の ALICE 実験のアップグレードでも CRU(Common Readout Unit)と呼ばれる部分に PCIe40 ボードと同仕様の読み出しボードが採用されており，IJCLab 経由で CRU の DMA 転送用に作られたファームウェアを使わせてもらうことができた。これは基本的に Altera の Avalon Memory Mapped 構成で DMA 転送を行うのだが，DMA コントローラーについては独自のものを使うというもので, 複数回の DMA 転送データの転送先を PC のメモリ上でまとまるようにできるなど，readout PC 上でデータを読

み出しやすい形にできる。またユーザーロジック部と DMA エンジンの間にはFIFOが入っており, DMA転送とユーザーロジックからのデータ流入をパイプライン化できる形になっている[9]。このコンポーネントをAlteraのシステム統合ツールを用いて PCIe40 ファームウェアに組み込んだ。

次に TTD システムとのインターフェースであるが, Belle II DAQ では TTD システムを通じて SuperKEKB 加速器の RF クロックを分周した127MHzのマスタークロックが分配されており，加えてトリガー情報やエラー情報のやりとりのために，このクロックを用いて 254Mbps で通信を行う形になっている。FEE から来るデータが多くなり DAQ システムの処理能力が追い付かない場合にはTTDシステムはトリガーを分配しないことでバッファあふれを防ぐが，それは読み出しボードのバッファの使用状況から判断した BUSY 信号をTTDシステムに送ることにより成り立っており, DAQ システムにとって非常に重要な部分になっている。まず考えるべきことは物理的な接続であるが，TTD システムでは主に CAT7 カテゴリの LAN ケーブルを用いて通信を行っており，一方で PCIe40 ボードには RJ45 コネクタが搭載されていない。幸いなことにPCIe40ボードは FPGAのGPIOポートにつながる 18pin の Molex コネクタを持っており，この Molex コネクタと RJ45 とのアダプタを作成することで PCIe40 と TTD システムを接続することができた。

TTD システムとのインターフェースの為のファームウェアについては基本的にはCOPPERシステムからの移植になるが, COPPER システムではタイミング調整のために Altera の FPGA には存在しない Xilinx の機能を使っていたため移植については少し苦労があった。その他に大きく変わる点としては読み出しボードのチャンネル数があり, 今までボード当たり最大4チャンネルだったところが48チャンネルに増えるので, これら48チャンネルそれぞれのBUSYやエラー信号(リンクダウンなど)といったステータス情報をまとめてTTD システムとやり取りする必要がある。従来の COPPER 用のファームウェアではボードあたり20チャンネルまでしか考えられていなかったため, 内部で 3 分配と 16 分配を組み合わせることで 3×16=48 チャンネルの処理をひとつのファームウェアの中に組み込んだ。またバッドチャンネルのマスク情報も 48 チャンネル分やり取りする必要があり，TTD システムにおいて読み出しボードと各チャンネルのアドレッシングの変更を行った。

PCIe40 ボードから読みだされたデータは readout PC 上の読み出しソフトウェアで更に処理される。最初に開発したデータ処理のスキームでは, 基本的にPCIe40 FPGAのファームウェア内でデータのフォーマッティングやチェックを行うため，読み出しソフトウェアの役割は PCIe40 からデバイスドライバ経由でデータを読み出したのち，データを再度チェックして下流の HLT サーバに送ることになる。この部分については後節で述べる LS1 期間中の改良で大きく変更を加えることになった。

読み出しソフトウェアからのデータは event-builder ソフトウェアを経由して下流のHLTサーバに送られる。COPPERシステムでは event-builder ソフトウェアの名の通り複数枚の COPPER のデータを event-building していたが[10]，新しい構成ではPCIe40一枚分のデータを下流にある複数のHLTサーバユニットに振り分けるのが主な役目となる。ここでも HLT のソフトウェアに変更が必要とならないように，インターフェースは極力キープする形をとった。

FEE からのイベントデータ処理のソフトウェアに加えてスローコントロールソフトウェアの開発も行った。Belle II DAQシステムではコンポーネントごとに作られたスローコントロール用のデーモンプロセスが PC 上で走っており，それらはKEKで開発されたNSM2(Network Shared Memory2)というフレームワークを通じてネットワーク通信を行っている。新読み出しシステムでのスローコントロールはPCIe40ボード全体の情報をまとめるpcie40controldとチャンネル毎に操作やモニターを行う pcie40linkd というデーモンプロセスを走らせて，PCIe40 の Belle2link 経由での FEE レジスタ読み書き，チャンネルごとの状態モニターやバッドチャンネルのマスクなどを行うようにした。

読み出しシステムから下流の HLT サーバをつなぐネットワークの増強も行った。これまで readout PC から HLT サーバへの出力はギガビットイーサネット 1 本もしくは数本をまとめる形をとっていた。平均的にはデータ送信はこれで問題なかったが，様々な理由でデータ量が一時的に増えた際に，しばしばここがボトルネックとなっていた。加えてPCIe40 の導入で readout PC 1 台で処理する FEE の数が増えることもあり，将来的なデータ量の増大も考慮して読み出しシステム側についてはスイッチも含め 25Gイーサネットで整備した。下流の HLT サーバに関しては CPU によるイベント再構成の処理能力に合わせて現在は 10G イーサネットで受けているが，今後 25G イーサネットに更新することも検討している。

6　システム入れ替えと物理ラン運転

Belle II DAQの読み出しシステムは検出器の隣にあるエレクトロニクスハットに置かれている。エレクトロニクスハットは 2 階建ての建屋で，1 階のフロアの約 4 分の 1 程度のスペースを現行のCOPPER読み出しシステムが占めている。検出器上の FEE からエレクトロニクスハット内の COPPER ボードまではデータ通信のため光ファイバーが配線されている。この配線を一から張り直すのは大変なのと，しばらくの間は新旧システムを素早く切り替えられるようにして旧システムをバックアップとして使いたかったこともあり，図6の写真にあるようにCOPPERクレートが入っているラッ

クの上部にパッチパネルを設置して，FEE からの光ファイバーをパッチパネル経由で新システムへつなぐという形をとった。これによって新旧読み出しシステムのハードウェア的な切り替えは光ファイバーを COPPER ボードとパッチパネルの間で挿し変える作業で行うことができる。

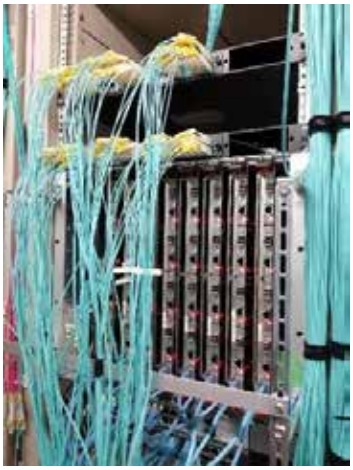

図 6　FEE からの光ファイバーを COPPER ボードからその上に設置された PCIe40 へつながるパッチパネルに挿し直したところ。

システムの入れ替えについては，FEE との間の読み出しテストやスローコントロールの開発などをそれぞれのサブ検出器グループと協力して進める必要があり，一度に入れ替えた場合 DAQ グループの人手が足りないことが予想されたため，何段階かに分けて行うことになった。このやり方だと入れ替え完了までの期間は旧システムと新システムが混在した形でデータ収集が行われることになるのだが，新システムの外部とのインターフェースを旧システムとコンパチブルにしておいたことがここで役立つことになった。新システムを導入する最初のサブ検出器を決めるのはトラブルのリスクが高いので頼みにくいことではあったが，アップグレード準備段階から試験等で協力してくれていたチェレンコフ光を用いた粒子識別検出器 TOP(Time Of Propagation)と最外層の KLM(Klong- Muon)検出器についてまず移行することにした。

動いている実験のシステムを入れ替えるので，チャンスは SuperKEKB 加速器が止まる夏の 3 か月ほどのシャットダウン期間と冬の 2 か月ほどのシャットダウン期間の年 2 回になる。当初は 2020 年夏に TOP および KLM 検出器に関する入れ替えを行う予定であったが，PCIe40 のボードがマスプロダクションを終えて KEK に届くのが 8 月末になり，またファームウェアやソフトウェアの準備もまだ不十分だったため，延期することになった。次のチャンスである冬のシャットダウン期間でもほぼシステムは出来上がっていたものの，システムコントロール用の GUI の作成や，PCIe40 と TTD システムとの接続の細部がまだ完成していなかったため，もう一回リスケジュールして 2021 年夏のシャットダウン期間に入れ替えを行うことになった。従来の COPPER 読み出しシステムがまだ余裕をもって物理ランのデータを処理できていたので，Belle II 実験自体に影響を与えることにならなかったのは幸いであった。さすがに次もリスケジュールするのは心苦しいので，2021 年春のランでは 2 週間に 1 度加速器保守の為にビームが止まるメンテナンス日を利用して，朝から TOP と KLM 検出器 FEE からの光ファイバーを COPPER から PCIe40 に付け替えて何時間かダミートリガーでのデータ収集試験を行い，夕方にまた元に戻すことで Belle II 検出器実機を用いた試験を行い，動作確認を繰り返した[11]。

2021 年夏の停止期間で TOP と KLM 検出器の読み出しシステムの入れ替えを行い，2021 年秋のランからは新システムでのデータ収集を開始した。一週間ほどは順調に読み出しができていたが，11 月 2 日にトラブルが発生した。その日の朝に readout PC でスローコントロールのソフトウェアライブラリをアップデートしたのだが，その後引き続き物理ランをとっているとラン開始から数分から長くても 1.5 時間くらいで PCIe40 のスローコントロール用の pcie40linkd プロセスが死んでしまうという状況が起こった。朝のアップデートが原因かと思い，古いバージョンに戻しても症状が治まらない。原因が分からないままデータ収集がすぐ止まってしまう状態が一晩続いたため，TOP 検出器の読み出しについてはこういうこともあろうかとバックアップとして残しておいた旧読み出しシステムに戻すことになった。切り替え作業は 4 時間ほどで終了して物理ランが続けられる状態になったので，さっそく原因の調査を行った。死んだデーモンプロセスのコアダンプファイルを見るとメモリ関連のエラーのようだったが，KEK の計算科学センターのエキスパートの指摘により，プログラム中の read システムコールが中途半端なデータ受信量で戻ってきた際の残りサイズの処理にバグがあってバッファオーバーランを起こしていたことが分かった。これは旧システムの時から使っていたコードだったのだが，新システムにしてから読み出しボードあたりのチャンネル数が増えて read システムコールが頻繁に呼び出されたことでバグが表に出てきたということかもしれない。今まで動いていたものでも使用環境が変わる際には十分な試験が必要になるというよい教訓になった。

原因が分かったので，早急にまた新システムに戻しておきたかったが，このトラブルで物理ランに 9 時間程のダウンタイムを出していたため，もう大丈夫ということを関係各方面に説明して納得してもらい，何とか 11 月末に再び新システムに戻してデータを取り始めることができた。その後は安定してデータ収集が行えたので，冬のシャットダウン期間には ARICH(Aerogel Ring Imaging CHerenkov)検出器も新システムに移行した。その後の 2022 年春のランでも読

み出しシステムは特に問題がなく，2022 年夏からの LS1 期間で残っている四つのサブシステムの読み出しについても入れ替えることになった。

7　さらなる改良とパフォーマンス測定

2022 年 7 月から 2023 年までは SuperKEKB 加速器の長期シャットダウンの期間となり，Belle II 検出器においてもピクセル検出器の入れ替えや TOP 検出器 PMT の一部交換などが行われた。DAQ についても予定していた残りの検出器について COPPER ボードからの PCIe40 ボードへの切り替えを行い，ダミートリガーを用いて長期のデータ収集試験を行った。これらの作業はビームラン中から準備を進めていたこともありすぐにめどがついたので，今までの新システムでのオペレーションで分かった問題点を踏まえた改良をこの時期に行うことにした。最も大きな変更点は，event-building を PCIe40 FPGA のファームウェアで行う方式から，readout PC 上のソフトウェアで行うように変えたことである[12]。これらの改良により読み出しシステムの性能も大きく向上し，2024 年の LS1 明けのランからは新方式でデータ収集を行っている。これについてこの節で少し詳しく説明したい。

そもそもの問題としては，最初の設計では PCIe40 ボードで event-building の為に FEE からのデータを一時蓄えるバッファは FPGA のオンチップメモリを利用しているため，サイズが限られるということから始まっている。PCIe40 について初期の LHCb 実験の資料などを見るとプロトタイプではメモリがついているものもあったようだが，基板上のスペースの問題もあったのかその後廃止したようである。小さいバッファサイズで発生しうる問題としては，

1. FEE から非常に大きなイベントがくる場合
2. いくつかの FEE からの event が遅れる場合

がある。1.についてはキャリブレーション用に波形データを FEE から PCIe40 に送ったりしない限り，通常の物理ランではそう起こることではない。問題になりそうなことが分かったのはむしろ 2.のほうで，この場合遅れたチャンネルからのデータを待つ間に他のチャンネルのバッファが埋まってしまうことが起こりうる。実際に TOP 検出器の FEE から送られてくるデータについて，チャンネルによって最大 18ms 程度の到着時間の遅れが出ることがあると分かった(図 7)。これはビームバックグラウンドが大きいイベントの場合に FEE での処理に時間がかかる為のようだが，今後より高いデータレートになると PCIe40 FPGA 内の入力 FIFO があふれる事態が想定される。

この問題を改善するため，event-building は readout PC 上のメモリを利用してソフトウェアで行うように改造をすることにした。最初のデザインでは COPPER メザニンボード上の CPU でやっていたことを読み出しボードの FPGA に移

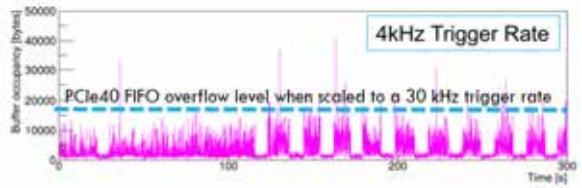

図 7　TOP FEE からのデータ到着時間のばらつきにより生じる PCIe40 バッファ占有の状況。メインリングへのビーム入射に同期した構造が見える。点線はトリガーレート 30kHz に換算した際にバッファが溢れるレベルを示したもの。

すというコンセプトであったため，この改造はその逆コースになってしまうが，現在の高エネルギー実験のデータ収集の流れとしてはむしろデータはとりあえず下流に流してそこでCOTS(Commercial Off-The-Shelf)デバイスであるサーバとスイッチの処理能力を生かしてソフトウェアで対応するというのが一般的のようである。

Software event-building 方式にした場合の読み出しシステムのブロックダイアグラムを図 8 に示す。この場合ファームウェアでは event-building をする必要がないので，早くデータが PCIe40 に到着したチャンネルから readout PC のメモリにどんどん転送していくわけだが，最大 48 チャンネルのデータを一つまたは二つの PCIExpress のポート(PCIe40 は PCIExpress Gen.3 2 ポート×8 レーンというスペック)で DMA 転送するので，交通整理は必要になる。そのため入力 48 チャンネルを 3 段のマルチプレクサ(MUX)に通して，どのチャンネルのデータを先に readout PC に送るかを制御する形となっている。最初のバージョンでは MUX の出口としては PCIExpress のポート一つのみを使っており，MUX を駆動するクロック周波数は FPGA 内のルーティングしやすさなどを考えて少し低めの 210MHz に設定していて，FPGA 内のデータ処理における理論的な最大速度は 6.7GB/s となっている。

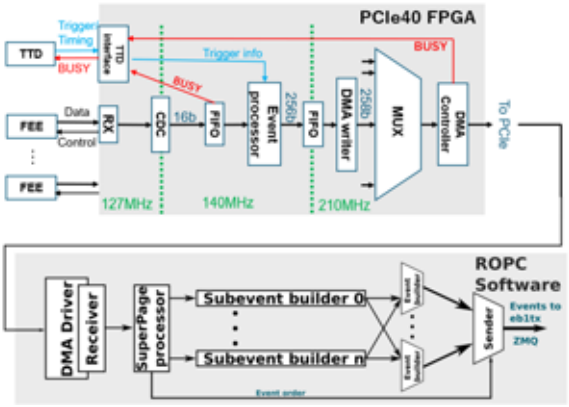

図 8　Software event-building 方式をとった場合の読み出しシステムのブロックダイアグラム。FEE からの各チャンネルのデータは先に来たものから MUX で選択して DMA で PC のメモリに送られる。

PCIe40 のファームウェアでのデータ処理方式の変更に伴い，readout PC 上の読み出しソフトウェアでも改造が行われた。software event-building 方式では異なるチャンネルのデータが PC のメモリ内にある DMA バッファにバラバラに記録されている。そこで各チャンネルに sub-event builder というスレッドを用意して，そこで DMA 転送で複数ページにまたがって送られてきたデータをイベント単位にまとめ，その先のスレッドによって複数のチャンネルのデータをまとめるという形をとっている。

この改良によって PC のメモリをバッファとして使えるようになった効果は大きい。極端な仮定として 48 チャンネル全部が Belle2link のラインレートでデータを受け取っていると考えてみても，512MB の DMA 用バッファが FEE からのデータで埋まるまでの時間的余裕は約 40ms あることになり，現実的なイベントサイズであれば先ほど述べた TOP FEE からのデータ到着の遅れも十分吸収することが可能になった。この software event-building の方式は LS1 中にダミートリガーを用いた試験を行った後，2024 年 2 月からの LS1 明けのランから実際の物理データ収集に使われており，現在まで安定して動いている。

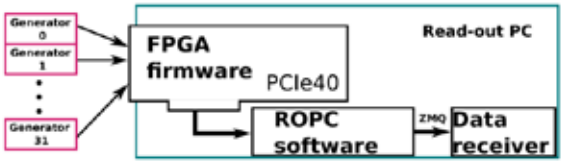

図9　ダミーデータを用いたパフォーマンス測定のセットアップ。

アップグレードされた新読み出しシステムのパフォーマンスについても述べておきたい。パフォーマンス評価においては旧システムで用いていた COPPER ボードをデータジェネレーターとして使い，そこから送信したダミーデータを PCIe40 で受けるという形で処理速度の測定を行った(図 9)。また PCIExpress の使用ポート数については LS1 明けの 2024 年の春のランでは 1 ポートのみを利用する設定を使っていたが，その間に 2 ポートを利用して readout PC にデータを送るファームウェアとソフトウェアの改造も完了したため，このパフォーマンス測定でも 2 ポート×8 レーンを使うバージョンを用いている。readout PC は実際に Belle II DAQ システムで使っているものと同様に Intel Xeon Silver 4214R CPU (12 cores)を用い，6 枚で合計 48 GB(2400 MHz)のメモリを搭載している。このセットアップでデータを処理してみたところ，当初の測定結果は readout PC 1 台あたり 0.94GB/s となった。ボトルネックは CRC 値を readout PC のソフトウェアで再計算させている部分であった。そこで今まで CRC 値の計算に用いていた Sarwate のアルゴリズムの部分を改善できないかと調べたところ，Slicing-by-16 algorithm [13]というアルゴリズムが有望であると分かり，これを実装したところスループットを 4.85GB/s と大きく向上させることができた。加えて CRC 値計算を行っている event builder スレッドの数を変えて最適化したところ，図 10 に示すように readout PC あたり 7.3GB/s の処理速度が達成できた。現在のシステムの処理速度についてはサーバ内のメモリアクセスの速度で律速しているのではないかと推察している。

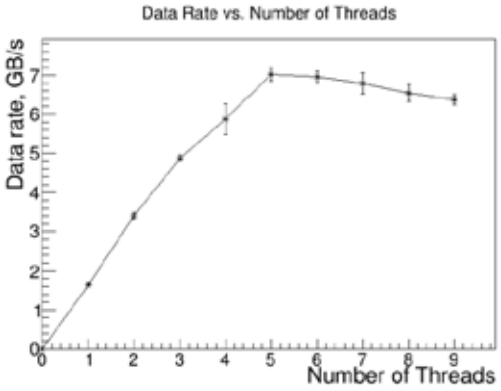

図 10　測定した読み出しシステムのスループットと event-builder スレッド数との関係。

これらの数値が Belle II 実験での最終的な目標ルミノシティに対して十分なのかという見積もりは現状ではビームバックグラウンドの予測が絡むため簡単ではないが，イベントサイズが大きいシリコンバーテックス検出器(SVD)からのデータはトリガーレート 30kHz で 3.6GB/s 程度になるという試算があり，現在 5 台の PCIe40+readout PC で SVD からのデータを処理していることを考えると，十分に余裕のある処理速度を実現できているといえる。

8　今後の発展

2024 年現在においては物理ランで FEE から流れてくるデータ量はアップグレードした読み出しシステムの処理能力よりもまだまだ小さいが，今後ルミノシティの向上に伴ってトリガーレートやイベントサイズが増えてきた時に，読み出しシステムにおいて律速することがないような形にすることができた。新システムの処理能力をもっと生かすデータ収集の方法としてはレベル 1 トリガーを用いないトリガーレス読み出しなどが考えられるが，Belle II 実験では現在レベル 1 トリガーが優秀で B 中間子対生成イベントについては 100%近い効率で収集できており，直ちにレベル 1 のハードウェトリガーからサーバ上で動く柔軟で高度なソフトウェアトリガーへと移る緊急性はない様子である。読み出しシステムのボトルネックが解消したことで，今後は Belle II DAQ システムにおけるもう一つのボトルネックである HLT でのイベント選別の速度向上についても Belle II ソフトウェアグループと協力して鋭意進めていきたいと考えている。

検出器グループからもアップグレードによって向上したパフォーマンスを活用しようという提案がなされている。これは近年の加速器実験で問題となっている FEE の放射線耐性の話と関連するのだが，TOP 検出器の FEE でデータ処理に使っている Xilinx Zync の PS(Processing System)で起こる single event upset によるエラーを避けるために, PS で行っているデータ処理作業を下流の読み出しシステムに移せないかというものである。この場合 TOP FEE から読み出しシステムに送られるデータ量が 2 倍に増えるが，新システムにおいては十分処理できると想定しており，現在共同で準備を進めているところである。また現状の DAQ システムでは崩壊点検出器の一つであるピクセル検出器についてはデータ量の大きさから Region of Interest を用いたデータ削減機能を持つ専用の読み出しシステムが使われているが，2028 年以降に予定している SuperKEKB 加速器の Long shutdown 期間(LS2)で入れ替えが検討されている新しい崩壊点検出器については，アップグレードで処理能力が上がった新システムを用いて他の検出器と同様に読みだすことが出来るかもしれない。他の近々の改良の予定としては, TTD システムと PCIe40 のリンクのノイズ耐性を高めるため，PCIe40 の SFP+ポートを使って, 現在の CAT7 ケーブル接続を光ファイバーに変更しようという計画も進んでいる。

意外なところとの関連では，今までのシステムは 13 台のラックに 203 枚の COPPER ボードと 43 台のサーバが入っていたが，新システムでは 21 枚の PCIe40 ボードとそれを搭載する 21 台のサーバが 2 台のラックにおさまるという形になっており(図 11)，この規模の加速器実験としては大分コンパクトな構成となった点がある。また新読み出しシステムで物理ラン中の消費電力量を調べたところ，全体で 3.9kW の電力消費となっており，以前のシステムでは概算で約 12kW の電力を消費していたことを考え合わせると，この高度化でおおよそ 1/3 に電力消費を減らすことができたことになる。特にこういった方向性を目指していた訳ではないが，期せずして省スペースで省電力という環境に優しい変更になったようで，KEK の年次環境報告のトピックの一つとしても報告文を寄せさせてもらったところである。

図11　インストールが完了した新読み出しシステム。

9　終わりに

以上 Belle II 実験の読み出しシステムのアップグレードの顛末を紹介させていただいた。本稿についてはアップグレードのプロジェクトを代表して書かせていただいたものであり，内容については個々に名前を挙げていないものの Belle II 実験 DAQ グループに参加している KEK や IJCLab，山東大学，ハワイ大学，東大 Kavli IPMU ほかの研究者の方々の多大な努力で行われたもので，この場を借りて感謝を申し上げたい。現代においてヒト・モノ・カネのリソースが最も投入されているのは IT や半導体技術の分野のようであるが，この分野での技術革新をうまく使って物理実験の世界でもまた面白いことができればと考えている。

参考文献

[1] 伊藤領介，中尾幹彦，山田悟，鈴木聡，今野智之，樋口岳雄，高エネルギーニュース, **33-3**, 232 (2014)

[2] 伊藤領介，田中真伸，高エネルギーニュース, **26-3**, 232 (2007).

[3] D.Sun, Z.Liu, J.Zhao, H.Xu, Phys. Procedia **37**, 1933 (2012).

[4] M. Nakao, J. Instrum. **7**, C01028 (2012)

[5] S. Ryu, Journal of Physics: Conference Series **898-3**, 032057 (2017)

[6] J. P. Cachemiche, Proc. Top. Workshop Electr. Part. Phys., Lisbon, Portugal, pp. 1–10, (2015).

[7] Z-A. Liu, *et al.*, IEEE Trans. Nucl. Sci. **67**, 1904 (2020).

[8] QD. Zhou *et al.*, IEEE Trans. Nucl. Sci. **68**, 1818 (2021).

[9] S. Mukherjee *et al.*, Springer Proceedings in Physics **201**, 107 (2018).

[10] S.Y.Suzuki *et al.*, IEEE Trans. Nucl. Sci. **62**, 1162 (2015).

[11] Y. T. Lai *et al.*, IEEE Trans. Nucl. Sci. **70**, 890 (2023).

[12] D. Levit *et al.*, IEEE Trans. Nucl. Sci. (Early access), https://doi.org/10.1109/TNS.2024.3462595 (2024)

[13] M. Kounavis and F. Berry, Proc. of 10th IEEE Symposium on Computers and Communications (ISCC'05), Murcia, Spain, pp. 855–862, (2005).