氏　　名	平　舘　　優
学位（専攻分野）	博士(学術)
学位記番号	総研大甲第307号
学位授与の日付	平成１０年３月２４日
学位授与の要件	数物科学研究科　加速器科学専攻 学位規則第４条第１項該当
学位論文題目	Study of an Event Builder System using Switching Network for High Energy Physics Experiments
論文審査委員	主　査　教　授　加藤 直彦 教　授　近藤 健次郎 教　授　黒川 眞一 教　授　菅原 龍平 教　授　渡瀬 芳行 教　授　藤井　啓文（高エネルギー加速器研究機構） 助教授　能町　正治（大阪大学）

The author studied and developed an event builder system using switching network, which is an essential ingredient of a data acquisition system in a large-scale high energy physics experiments. Event builder is the collection stage of a whole event data from many data sources dedicated to each detector element.

Various commercial network switches have been investigated. However, a special functionality is required to use it in the event builder system which must handle a coherent data traffic from many data sources to a single destination. Such system can not be analyzed theoretically. A large-scale simulation has to be carried out for designing and evaluating the system.

My objectives are to design and build a prototype of an event builder system with "Global Traffic Control", and to show its performance predictable and scalable.

This system composed of typically the same number of data source nodes with the number of the destination nodes where a whole event data is collected. The data in the source is transferred to the destination through a serial link. A whole event data are collected in the destination node from a different source node by changing the link connection successively. A number of events are collected at the different destination nodes in parallel. The data transfer and the switching the link connection are controlled externally by a controller. Then, the data flow in this system can be congestion free even in the heavy load. No one has studied such system in detail.

In this study, I have analyzed theoretically the event builder system with queuing theory and built the prototype to study the whole behavior of the system. I show that the system is congestion free, scalable, and applicable to the data acquisition system in the KEKB Belle experiment.

The summary is as follows:

1. System Modeling with Queuing Theory

I have modeled an event builder with "Global Traffic Control" and analyzed the system using queuing theory. The results are shown quantitatively on traffic intensity (ρ) dependence, the ratio: packet size (*S_pkt*) / average event fragment size (*S_evf*) dependence, and switch size (number of I/O channel; N) dependence of average event fragment number (L) in an input queue.

1) L is almost independent of N and proportional to D/M/1 system prediction by queuing theory. 2) This system has a characteristic that L increases rapidly if ρ exceeds 0.8. According to the larger dispersion of event fragment size, this increase starts from the lower ρ. 3) In the case of *S_pkt* < *S_evf*, L is not dependent of the ratio: *S_pkt*/*S_evf*, but in the case of *S_pkt* >> *S_evf*, L is represented as a product of ρ and *S_pkt*/*S_evf*. Simulation results were in agreement with the calculation results.

2. System Design of the Prototype Event Builder

I developed hardware composing the prototype system (Switch, Transmitter, Receiver and External Controller) and software to control them. Data are transmitted through the G-link serial line at the rate of up to 1.17 Gbps. The switching overhead including software one is smaller than 90 μsec. I have tested a 2x2-event builder system. The total throughput to build an event is 10 MB/sec(max.).

3. System Analysis

The comparison of experimental and theoretical results about L was evaluated. 1) In the large S_pkt/S_evf, the experiment at result agreed with the theory, but in the small S_pkt/S_evf, experiment was larger than the theory. This difference caused by a switching overhead becomes remarkable for smaller S_pkt/S_evf. Hence L has minimum value near the point of S_pkt/S_evf=1. It shows that a packet size should be the same as that of the average event fragment size for saving buffer memory size.

2) Near the point of S_pkt/S_evf=1, the experiment at result agreed with the theory if an event fragment size is fix. For the larger dispersion of the event fragment size, however, the experiment has the larger value of L than the theory. This effect is caused by the software implementation for easy operation. It does not affect to the result of 1).

Experimental results are almost in agreement with theoretical one on the both 1x1 and 2x2 system. It is considered that this agreement will keep on the larger system because the switching overhead is independent of the switch size(number of nodes).

4. Application to BELLE Experiment

The maximum event size and the maximum trigger rate of BELLE experiment are assumed about 30 KB and 500 Hz, respectively. Number of source node and destination node is 12 and 6-8, respectively. This experiment requires 15 MB/sec of total throughput. If 12x6-event builder system is constructed, the total throughput is expected to be 22.5 MB/sec(max.) from this study.

The average event fragment size on this experiment is about 4 KB. If a packet size is set the same as it, the system will run at the traffic intensity 0.2-0.4. In this case, the numbers of event fragment (L) in the input queue are 1 and 9 for the average and maximum, respectively.

In this study, an event builder system with "Global Traffic Control" is analyzed in detail by queuing theory for the first time. As a result of experiment on the prototype system, it is shown that the performance of the system is predictable by the theory and scalable to the larger system. The performance of the prototype system satisfies the requirements of BELLE experiment.

（論文審査結果）

平館優君の博士論文内容は、大規模な高エネルギー物理学実験におけるオンライン実験データ処理システムの一部である「イベント・ビルダ」の部分にグローバル・トラフィック・コントロールと呼ばれる方式のスイッチング・ネットワークを用いることに関する研究であり、待ち行列理論を用いて解析を行うと同時に実際のハードウェアを製作し、それらを用いての測定による検証も行っている。KEK Bファクトリーやlhcのような大規模な加速器を用いた実験においては、実験装置そのものも大きく複雑になり、従って検出器が多数のサブシステムによって構成される。また、取り込まれたデータも多数の演算処理装置（ＣＰＵ）によって構成される、いわゆるオンライン・コンピュータ・ファームにより処理され、結果は大規模記憶装置に貯えられる。つまり、あるイベントに関するデータは複数の検出器によって収集されるが、それらを一つのイベントとして処理するためには、データを一台の演算処理装置に送り込まなければならない。そのためには効率の良いデータ転送切替えのための高速スイッチング・ネットワークが必要となる。そこで、種々のスイッチング・ネットワークの形態とその特性を評価し、最終的にバレル・シフタを用いたN×Mのトランスペアレントなスイッチング・ネットワークを提案している。このスイッチング・ネットワーク・システムを用いることにより、このイベント・ビルダ・システムはスケーラブルになり、小規模から大規模に亘るシステムの設計も容易に行えるようになった。

このデータ収集のためのイベント・ビルダの特性を決める要素としては、データ収集サブシステムからのデータがネットワーク・スイッチを通る前のバッファ・メモリの大きさ、スイッチを通った後ワークステーションに渡される前のバッファ・メモリの大きさ、また転送される際のパケットの大きさ、それにスイッチの切替え時間も含めた転送時間などがある。これらがこのシステムにおいてデータのロスが起きる可能性を決めている。そこで、この論文では、待ち行列理論を用い、要求される性能を満たす条件を求めている。

データの発生間隔が指数関数分布をし、データの長さが指数関数分布をする場合と一定分布をする場合について、トラフィック・インテンシティ（平均入力データ・レートと最大出力データ・レートの比）や、パケットの大小及びバッファ・サイズを変化させた時の平均待ち行列の長さを解析値とシミュレーションによって得られたデータとの比較を行い、両者がよく一致を確認している。これを実際のシステムへの適応を行うため、KEK Bファクトリーでの応用を目的としたハードウエアシステムを開発し、実験を行った。

実験では、検出器のサブシステムからのデータを各種のトリガーレートとイベントデータサイズの分布に従って発生させ、ＶＭＥバスのメモリーモジュール上で設定されたキュー（待ち行列）に転送する。キュー上のデータは、１．４Ｇｂｐｓのシリアル送信モジュールから、外部コントローラで制御されたトランスパレント・スイッチを通して、受信モジュールに伝送される。一つの事象に関する全ての検出器からのデータは一つの受信モジュールの対応した受信キューに蓄積されるように制御され、事象構成（イベントビルド）が行われる。測定は２ｘ２のシステムで行われ、待ち行列の長さ、スループットなどは、実験条件を入れた待ち行列理論からの予想値とよく一致しており、非常に見通しのよいシステムである事を実証した。

以上の研究は数物科学研究科加速器科学専攻の博士学位論文としての内容に値し、更に、スケーラブルなイベント・ビルダの提案、計算機によるその特性のシミュレーション、そのハードウェアによる実現という専門的にも総合的にも極めて優秀な研究業績を上げていると判断した。