Research and Development of Event Building Farm for SuperKEKB

Ko Ito
Department of Physics, University of Tokyo

February 4, 2005

Abstract

As an upgrade *B*-Factory experiment following the current ongoing Belle experiment, we are planning the SuperKEKB experiment with a luminosity of $5 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, which is factor ten or more higher than the current Belle experiment. The current Belle DAQ system can not work efficiently at such a high event rate experiment.

We design a new DAQ system using an event building farm approach for the SuperKEKB experiment. We set up a prototype event building farm to study its performance. The event building farm consists of three parts, readout, distribution and full event building parts. If the number of the readout PC is eight, the prototype readout part works with the 30 kHz event rate which is the expected trigger rate at the start of the experiment. If the number of the readout PC is 20, the prototype readout part works with the 10 kHz event rate which is expected to be the maximum trigger rate. The distribution part tolerates the trigger rate of 30 kHz by increasing the number of the event builder units with the typical data size of 200 kB per event.

Based on this study, we consider the design specification and confirm that the designed event building farm satisfies the requirement at the SuperKEKB experiment. We conclude that the new DAQ system discussed in this thesis is the strong solution towards the high luminosity experiment at SuperKEKB.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most important discoveries in the modern elementary particle physics is the existence of the CP violation. The CP violation is expected to be related to the basic principle of the nature. It is expected to answer one of the most attractive questions in cosmology and in elementary particle physics, why the universe we currently live in consists predominantly of the matter. To observe the CP violation and to test the Kobayashi-Maskawa model in the B meson system, we started a B factory experiment, Belle, using the e^+e^- collider (KEKB) at the High Energy Accelerator Research Organization (KEK) in 1999.

In the summer of 2001, the presence of CP violation in the B meson system was established by the Belle and BaBar collaborations through the measurement of the time dependent CP asymmetry in the decay process of $B^0(\overline{B}^0) \to J/\psi K_S^0$ [1, 2]. The Belle experiment also proved its ability to measure a number of decay modes of the B meson and other interesting observables; the precision measurement of interior angles of the unitary triangle, the semi-leptonic FCNC processes, and the existence of a new CP phase in the penguin process and so on. To collect many such observables, the KEKB has delivered the world highest luminosity of $1.39 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$, and the Belle has collected an integrated luminosity exceeding 300 fb^{-1} data.

The KEKB luminosity is expected to be doubled in near future for many interesting physics decay modes that require larger data sample. To collect more large data sample, we are planning the SuperKEKB experiment with a luminosity of $5 \times 10^{35}\text{cm}^{-2}\text{s}^{-1}$, which is ten or more larger than that of the current KEKB.

In such a high luminosity, the level-1 trigger rate is expected to be 20 - 50 times higher than that of current trigger rate. The data size also is to be increased by a factor of six since a total channels of the Belle detector increase and so on. The number of readout modules (COPPER board), which is a pipelined readout electronics, is estimated more than 1000 if we assume that one COPPER board has a capacity of about 100 channels.

The current event building farm is not expected to work in these conditions because of a lack of the CPU power of the current system and no parallelization of data stream. Thus, it is necessary to develop a *scalable* and *tolerable* event building farm for the SuperKEKB experiment. To handle the high trigger rate, the large data size, and the large number of the COPPER boards, we design the event building farm for the SuperKEKB experiment. The designed event building farm employs the multi-stages for event building, which comprise Stage1 (readout stage), Stage2 (distribution stage), and Stage3 (full event building stage), where Stage3 consists of multiple event building units that is based on the current event building farm and the reconstruction farm. Major change from the current event building farm is Stage1 and Stage2. To collect the event fragments the large number of the COPPER boards in the high rate, Stage1 constructs the readout network for data transmission with gigabit network switches. To distribute the event fragments to the unit event by event, Stage2 constructs the network matrix for the distribution of event with gigabit network switches. To decide the design specification, it is important to validate the design and measure the performance of Stage1 and Stage2.

In this thesis, we present the performance of Stage1 and Stage2 of designed event building farm including acceptable trigger rate, transfer rate, scalability of Stage2. We describe the brief introduction of the Belle experiment and the overview of SuperKEKB in Chapter 2, the data acquisition (DAQ) system for the SuperKEKB experiment including the current DAQ system in Chapter 3, the performance study of Stage1 in Chapter 4 and the performance study of Stage2 in Chapter 5, conclusion in Chapter 6.

Chapter 2

SuperKEKB

2.1 *CP* violation

Various symmetries play very important roles in particle physics. Some of them are continuous and the others are discrete. The *CP* symmetry is one of the latter and the origin of its violations is one of the most exciting mysteries in the present particle physics. As its name indicates, the *CP* transformation is a product of two discrete operations, C and P.

Charge conjugation, C, is a symmetry between particles and antiparticles. Parity, P, is a symmetry of space. P invariance means that the mirror image of an experiment yields the same result as the original.

Until 1956, it was believed that all elementary processes are invariant under C and P transformation. Lee and Yang pointed out the possibility of the violation of these symmetries, and subsequent experiments proved that C and P symmetries are really not conserved in weak interactions. However, the product of C and P transformations, *CP* was still considered to be a good symmetry. The second impact came in 1964. An experiment using neutral K mesons showed that *CP* is also not conserved under weak interactions [3]. Neutral K mesons (K^0 and $\overline{K}^0$) are created by strong interactions. The mass eigenstates of the $K^0 - \overline{K}^0$ system can be written

$$|K_S\rangle = p|K^0\rangle + q|\overline{K}^0\rangle, |K_L\rangle = p|K^0\rangle - q|\overline{K}^0\rangle \tag{2.1}$$

(choosing the phase so that $CP \mid K^0\rangle = \mid \overline{K}^0\rangle$). If the *CP* invariance is held, we would have $p = q$ so that K_S would be *CP* even and K_L would be *CP* odd. Because the kaon is the lightest strange meson, it decays through the weak interaction. Neutral kaons can decay into two or three pions. Since pion has *CP* eigenvalue of -1, K_L always decays into two pions, if *CP* is conserved in weak interactions. The experiment performed at Brookhaven proved that a small faction of K_L decays into two pions, which means *CP* is violated in the weak interaction. In the kaon system, the order of observed *CP* asymmetry is about 10^{-3}.

2.2 Belle experiment

The primary goal of the B factory experiment [7] is to establish the CP violation in the B meson system. The KEKB accelerator [8] is an energy-asymmetric e^+e^- collider to produce B mesons. The decay products of B mesons are detected by the Belle detector.

2.2.1 KEKB Accelerator

The KEKB accelerator has two rings in a tunnel which used for TRISTAN. The total length of the accelerator main rings is 3 km. Beam energies are chosen to be 8.0 GeV for the electron and 3.5 GeV for the positron, so that the center of mass energy comes on the $\Upsilon(4S)$ resonance and $\beta\gamma \simeq 0.425$ corresponds to the flight length of the B meson decays of approximately 200 μm. Configuration of KEKB accelerator is shown in Fig 2.1. In May, 2003, KEKB has achieved the design luminosity, $10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. As of the end of 2004, the peak luminosity is $1.39 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The integrated luminosity exceed 300 fb^{-1} data.

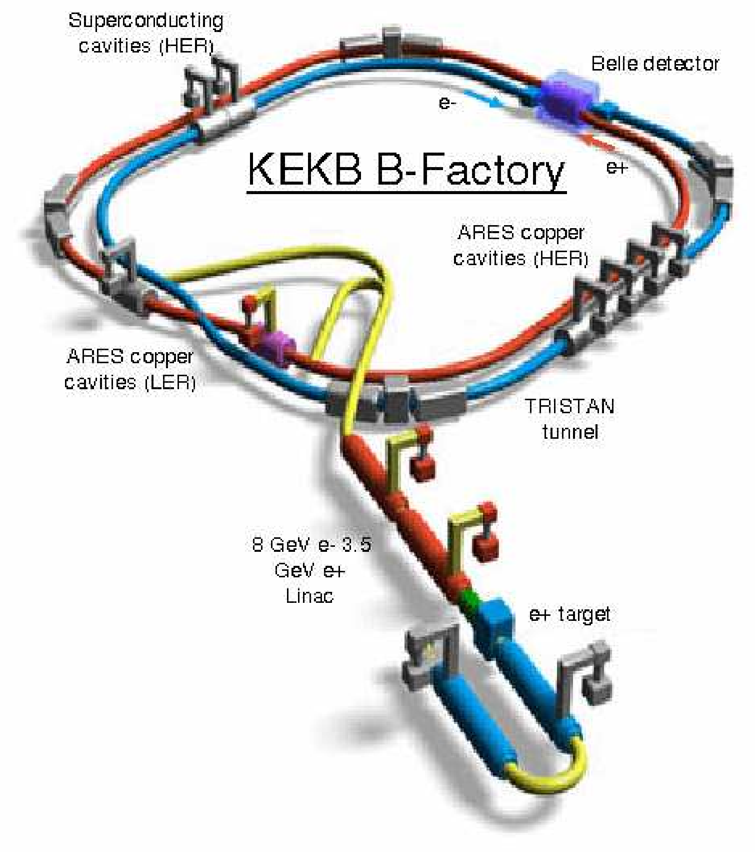

Superconducting cavities (HER)
Belle detector
e-
KEKB B-Factory
e+
ARES copper cavities (HER)
ARES copper cavities (LER)
TRISTAN tunnel
8 GeV e- 3.5 GeV e+ Linac
e+ target

Figure 2.1: The KEKB accelerator system.

2.2.2 Belle Detector

The Belle detector is a detector designed for the study of the CP violation in the B meson system. The Belle detector consists of a silicon vertex detector (SVD), a 50-layer central drift chamber (CDC), an array of aerogel threshold cherenkov counters (ACC), time-of-flight scintillation counters (TOF), an electromagnetic calorimeter composed of CsI(Tl) crystals (ECL), a K_L^0 and muon detectors (KLM), and a pair of extreme-forward calorimeters (EFC). Figure 2.2 is the schematic view of the Belle detector.

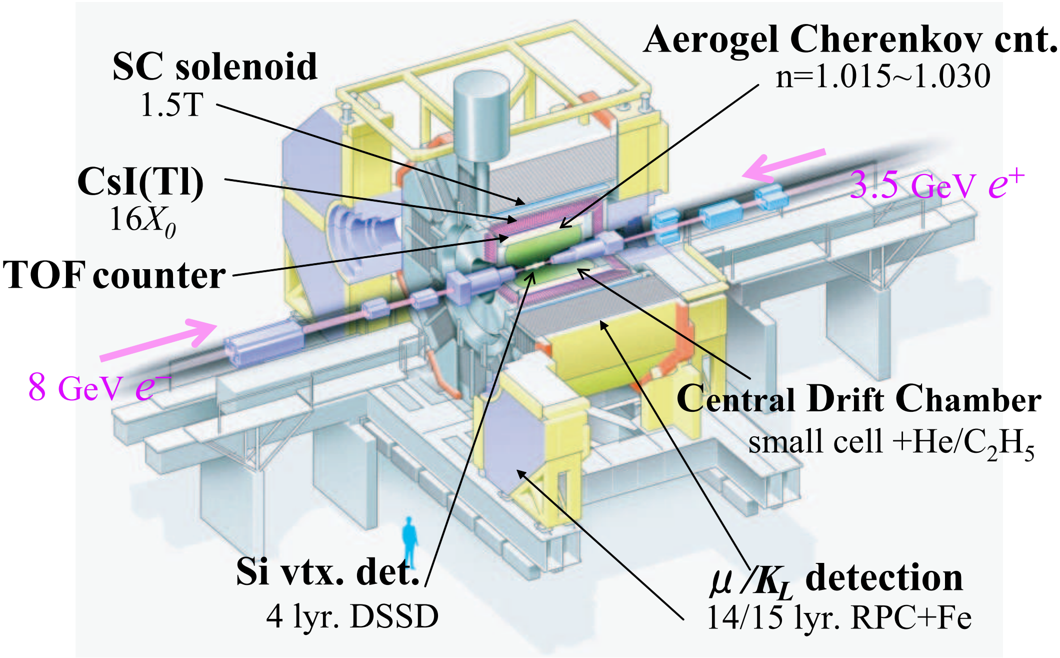


Figure 2.2: Schematic view of the Belle detector.

Sub detectors

SVD

SVD is a vertex detector with precise vertex resolution. The measurement of *CP* asymmetry parameters requires that the resolution of vertex detector is better than the average flight distance of B meson, which is about 200 μm at the KEKB accelerator.
In summer of 2003, SVD system was upgraded to SVD2 [9]. SVD2 consists of four layers of silicon ladders with covering polar angle from 17° to 150°. The momentum dependence of the impact parameter resolution of SVD2 is described by $\sigma_{r\phi} = 21.9 \bigoplus 35.5/(p\beta sin^{\frac{3}{2}}\theta)[\mu m]$ and $\sigma_z = 27.8 \bigoplus 31.9/(p\beta sin^{\frac{5}{2}}\theta)[\mu m]$ with cosmic ray muons. The number of channel is about 110,000 in total. Figure 2.3 show r-z view of SVD2.

CDC

The main role of CDC is the detection of charged particle tracks and the reconstruction of its momentum. CDC also take part of particle identification information by measuring dE/dx. The polar angle coverage of CDC cover from 17° to 150°. In summer 2003, the inner most three layers were replaced with a new chamber in order to provide a space for

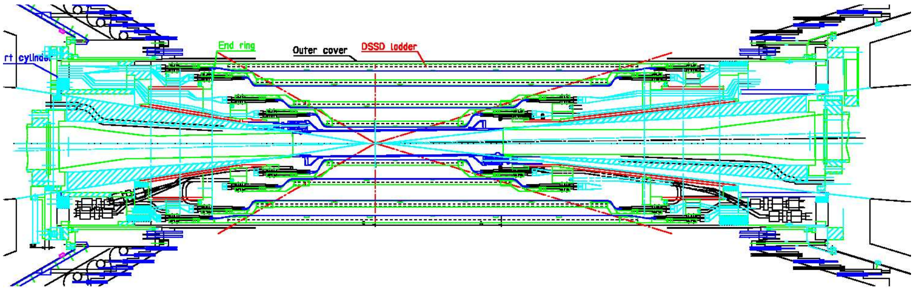


Figure 2.3: Side view of SVD2.

SVD2. The new chamber consists of small cells of 5 mm × 5 mm in dimension to reduce the occupancy. The number of channel is about 9,000 in total.

ACC

ACC is to separate kaons from pions for momentum region in $1.2 < p < 3.5$ GeV/c. The aerogel of the ACC is made of SiO_2 whose refractive index is $n \simeq 1.015$.

In general, the threshold of Cherenkov light emission in the matter with the refractive index of n is represented using the velocity of particle β as follows:

$$n > 1/\beta = \sqrt{1 + (m/p)^2}, \tag{2.2}$$

where the particle momentum p is measured by CDC, m is the particle mass. The particle can be identified whether it emitted a light or not. ACC is divided into two parts. A barrel array covers an angular range of $34° < \theta < 127°$, and a forward end-cap array covers an angular range of $17° < \theta < 34°$. The number of channel is about 2,200 in total. The side view of ACC is shown in Fig. 2.4.

TOF

TOF, which is made of a plastic scintillation counter, is also used for the particle identification. It has responsibility to identify the charged particles, whose momentum is less than 1.2 GeV/c. The relation between the measured flight time T and the particle mass is as follows:

$$T = \frac{L}{c}\sqrt{1 + (m/p)^2}, \tag{2.3}$$

where L is flight length. In the Belle, L is 1.2 m. The polar angle coverage of TOF is from 34° to 121°. Also TOF system has one another sets of scintillation counters, which are used to generate the trigger signal. The number of channel is about 400 in total. The side view of TOF is shown in Fig. 2.4.

ECL

The main purpose of the ECL is to detect photons and the identification of electrons from B meson decays with high efficiency and good energy resolution. ECL is made of CsI(Tl) crystals.

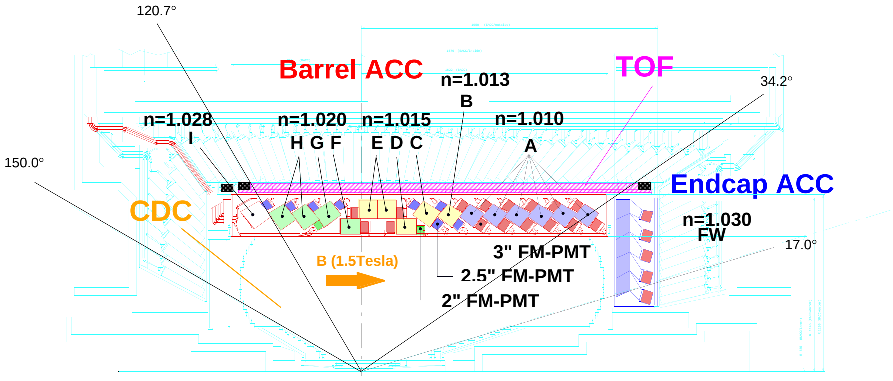


Figure 2.4: Construction of the ACC and TOF.

ECL is also used for the measurement of the luminosity by Bhabha scattering, which yields high energy electrons. Thus ECL covers the energy range widely from 20 MeV to 8 GeV. The number of channel is about 9,000 in total.

KLM

KLM is the outermost detector to detect K_L^0 and muon, and to measure their position. KLM consists of an alternating sandwich of 4.7 cm thick iron plates and resistive plate counters(RPCs) located outside the superconducting solenoid. KLM covers an angle range $25° < \theta < 145°$. The number of channel is about 45,000 in total.

EFC

EFC measures the energy of photons and electrons at the extreme forward (backward) direction outside the ECL acceptance. EFC covers $6.4° < \theta < 11.5°$ in the forward direction and $163.3° < \theta < 171.2°$ in the backward direction. We use BGO ($Bi_4Ge_3O_{12}$) crystals for EFC, because EFC is expose in the high irradiation (about 5 MRad per year) of photons from the synchrotron radiation and the spent electrons. The number of channel is about 320 in total.

Trigger

The Belle trigger system mainly consists of the Level-1 hardware trigger and the Level-3 software trigger. Figure 2.5 shows the the block diagram of Level-1 trigger system. It consists of the 6 sub-detector trigger systems and the central trigger system called the Global Decision Logic (GDL). The trigger system provides the trigger signal with the fixed time of 2.2 μs after the event occurrence. The average trigger rate of current Belle experiment is about 400Hz at the luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$.

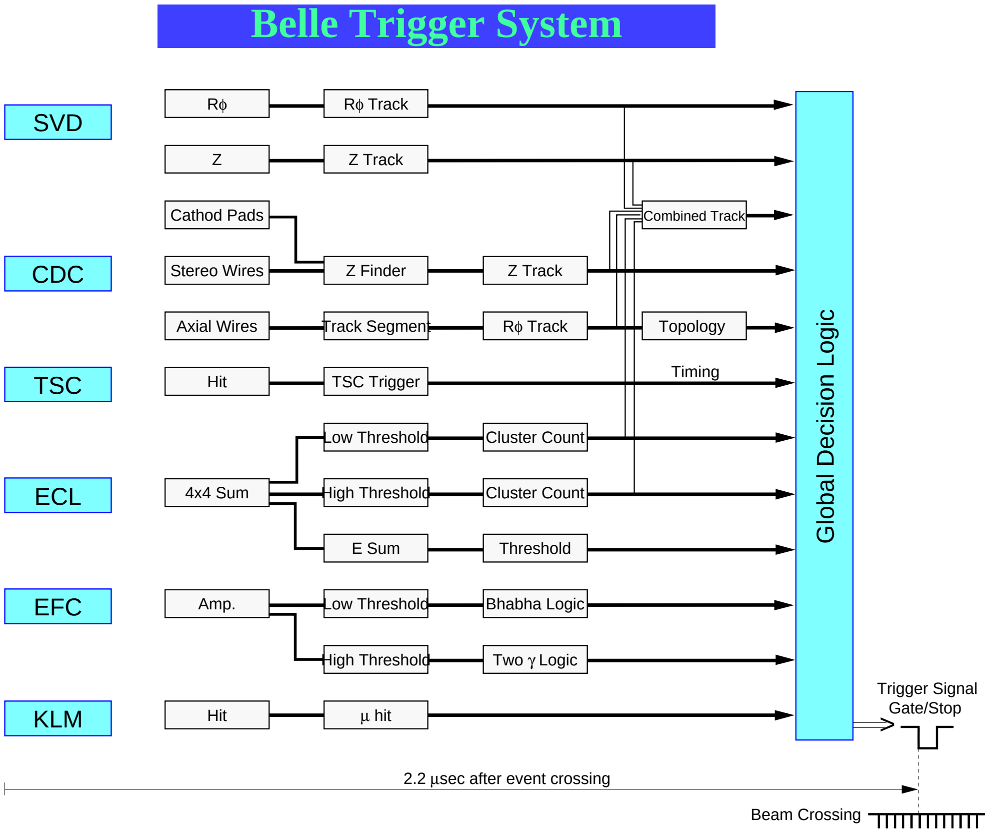


Figure 2.5: Belle Level-1 trigger system.

2.3 Overview of SuperKEKB

In order to search for a new physics beyond the Standard Model through the rare B decays, the upgrades of the accelerator and the detector are proposed as SuperKEKB project. The SuperKEKB aim 30 times higher luminosity than the current KEKB accelerator. We expect an annual integrated luminosity of 5 ab^{-1} assuming 100 days of operation. This section gives a brief description of a physics motivation, accelerator and detector of the SuperKEKB.

2.3.1 Physics motivation

We show two brief introduction of many physics motivations of SuperKEKB. The details can be found elsewhere [10].

$b \to s\ell^+\ell^-$

The $b \to s\ell^+\ell^-$ process is one of the Flavor Changing Neutral Current (FCNC) processes. In the Standard Model, FCNC processes are forbidden at tree level. However, higher order diagrams, such as penguin diagrams and box diagrams, induce FCNC within the Standard Model (see Fig. 2.6). Such loop diagrams are expected to be sensitive to a new physics. Since heavy particles beyond the Standard Model could contribute to the additional loop diagrams, various parameters, such as branching ratio, may well be deviated from the expected value by the Standard Model.

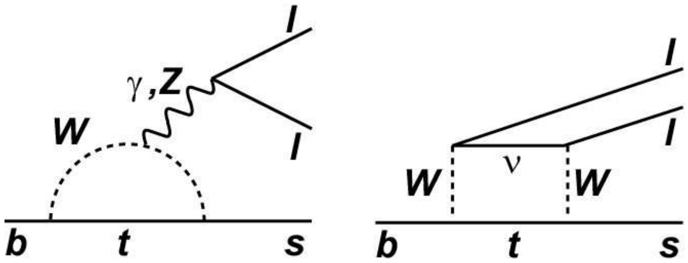


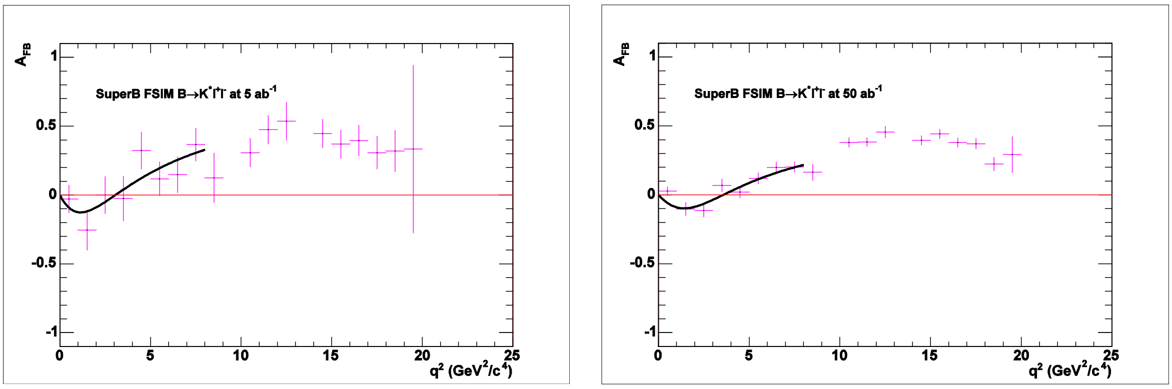
Figure 2.6: The Feynman diagrams of $b \to s\ell^+\ell^-$ process

The one target of the SuperKEKB for this process is the forward-backward asymmetry. The forward-backward asymmetry in $B \to K^*\ell^+\ell^-$, defined as

$$\overline{A}_{FB}(q^2) = \frac{N(q^2; \theta_{B\ell^+} > \theta_{B\ell^-}) - N(q^2; \theta_{B\ell^+} < \theta_{B\ell^-})}{N(q^2; \theta_{B\ell^+} > \theta_{B\ell^-}) + N(q^2; \theta_{B\ell^+} < \theta_{B\ell^-})}, \tag{2.4}$$

is an ideal quantity to disentangle the Wilson coefficients C_9 and C_{10} together with the sign of C_7, where q is the dilepton mass. Within the Standard Model, there is a zero crossing point of forward-backward asymmetry in the low dilepton invariant mass region, while the crossing point may disappear in some SUSY scenarios. Another important new physics effect can be searched for by using the $B \to K^*\ell^+\ell^-$ or $B \to X_s\ell^+\ell^-$ forward-backward asymmetry; $SN(2)$ single down-type quarks and tree-level Z flavor-changing-neutral-current.

Figure 2.7 shows the expected $\overline{A}_{FB}$ at 5 ab^{-1} and 50 ab^{-1} as a function of q^2. It can be seen that the crossing pattern of the forward-backward asymmetry will be already visible at 5 ab^{-1} and will be clearly observed at 50 ab^{-1}.



(a) 5 ab^{-1}

(b) 50 ab^{-1}

Figure 2.7: Forward-backward asymmetry in $B \to K^*\ell^+\ell^-$ at 5 ab^{-1} (a) and 50 ab^{-1} (b).

$b \to sq\overline{q}$

The recently observed disagreement between the value of the angle ϕ_1 measured in the penguin process $B \to \phi K_S^0$ and the precisely measured value in $B \to J/\psi K_S^0$ suggests the existence of a new *CP* phase in the penguin process $b \to sq\overline{q}$. The $B \to \phi K_S^0$ decay, which is dominated by the $b \to ss\overline{s}$ transition, is an especially unambiguous and sensitive probe of new *CP*-violating phase from physics beyond the Standard Model.

Figure 2.8 shows an example of a fit to events in a MC pseudo-experiment for the $B \to \phi K_S^0$ and $J/\psi K_s^0$ decays at 5 ab^{-1}. The large deviation can be observed with a single decay channel $B \to \phi K_S^0$ at the SuperKEKB.

2.3.2 Accelerator

Figure 2.9 shows the conceptual view of an accelerator for the SuperKEKB [10]. The design luminosity of the SuperKEKB is 1 - 5 × 10^{35} cm^{-2}s^{-1}. The SuperKEKB collider will be constructed by re-using most of the components of KEKB accelerator, in particular the ring magnets and klystrons used to supply RF power to the cavities. But, there are many components that need to be modified or newly developed; RF system, vacuum system, feedback system and so on. A crossing angle of the current KEKB accelerator is 30 mrad in order to keep the beam separated. The crab crossing scheme, which effectively creates a head-on collision, will be used at SuperKEKB. Beam energy exchange, with electrons injected to the LER instead of the HER and positrons injected to the HER, is also under consideration to reduce the effect of the photo-electron cloud and to minimize the injection time. The machine parameters of SuperKEKB is shown in Table 2.1. Figure 2.10 shows the plan of the KEKB accelerator upgrade.

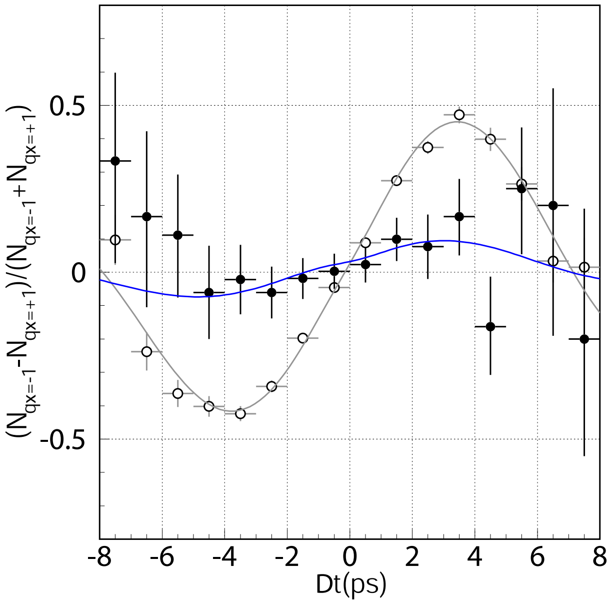


Figure 2.8: Raw asymmetries for $B \to \phi K_S^0$ (close circle) and $B \to J/\psi K_s$ (open circle) at 5 ab^{-1}. Input values are $S_{\phi K_S^0} = +0.24$ and $A_{\phi K_S^0} = +0.07$.

Table 2.1: Machine parameters of SuperKEKB

Parameters	LER / HER	Unit
Beam energy	3.5 (e$^-$) / 8.0 (e$^+$)	GeV
Beam current	9.4 / 4.1	A
Particles/bunch	1.18×10^{11} / 5.13×10^{10}	
Number of bunch	5018	
Horizontal β at IP	0.2	m
Vertical β at IP	0.003	m
Horizontal emittance	24	nm
Crossing angle	0 (crab)	mrad
Luminosity	5	$\times 10^{35}$cm^{-2}s^{-1}

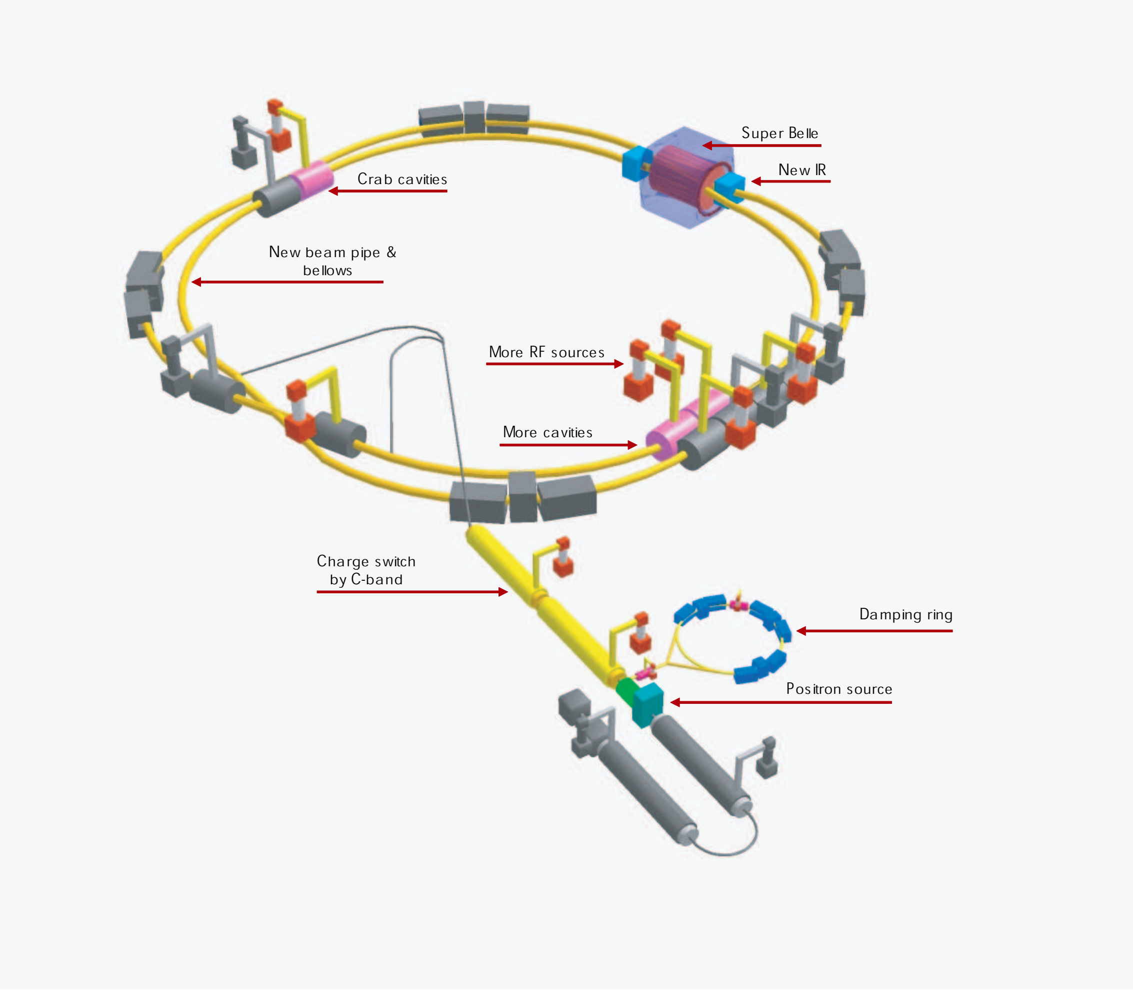


Figure 2.9: Conceptual view of an accelerator for SuperKEKB.

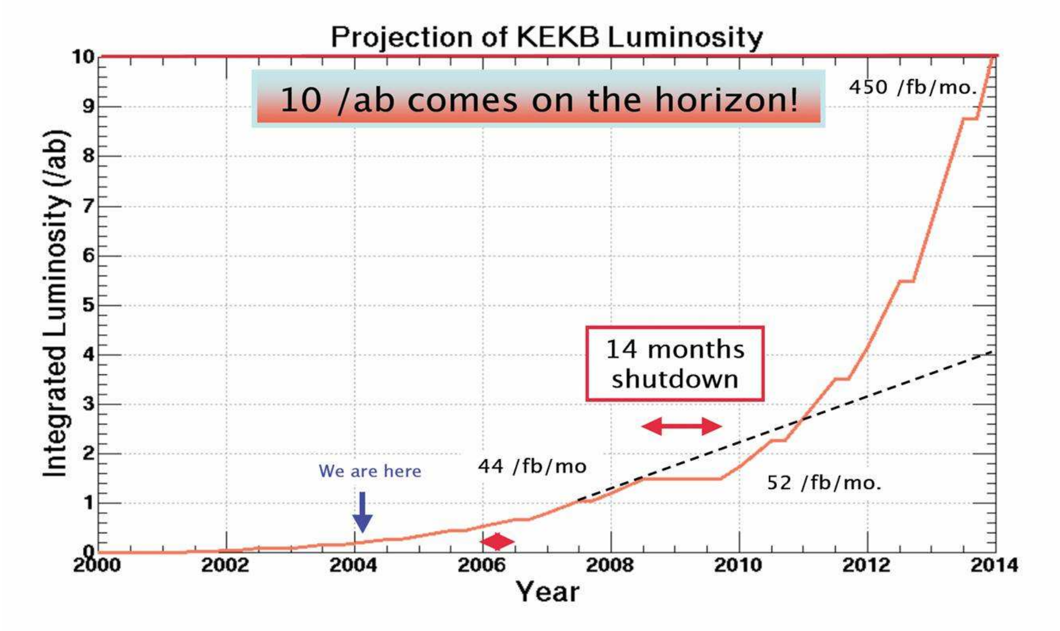


Figure 2.10: KEKB accelerator upgrade plan. The horizontal axis and the vertical one are integrated luminosity and year, respectively.

2.3.3 Detector

Figure 2.11 shows the conceptual view of a detector for SuperKEKB. The upgraded Belle detector for SuperKEKB consists of vertex detector, central tracker, particle identification detector, calorimeter, and K_L and muon detector. In order to maintain and to evolute the current performance in the higher background environment, each detector is upgraded gradually. We introduce the detail of DAQ system for SuperKEKB in Section 3.2. The detail of another detectors can be found elsewhere [10].

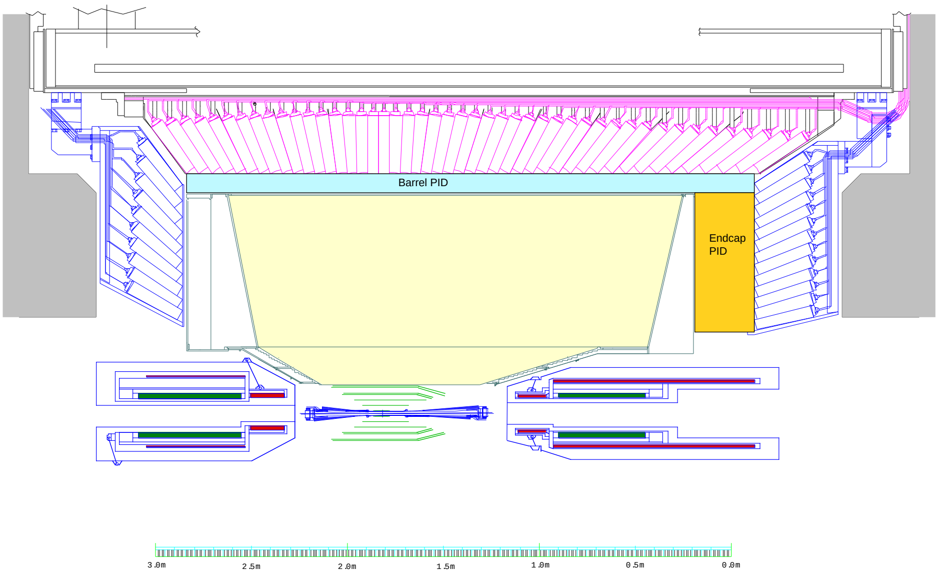


Figure 2.11: Side view of a detector for SuperKEKB.

Chapter 3

Data Acquisition System for SuperKEKB

In this chapter, we describe the data acquisition (DAQ) system of Belle and SuperKEKB. The current Belle DAQ system is given in Section 3.1. In Section 3.2, the problems of Belle DAQ system toward SuperKEKB and the overview of the DAQ system for SuperKEKB except the event building farm is described. The event building farm for SuperKEKB is written in Section 3.3.

3.1 Current DAQ system of Belle

The Belle DAQ system deals with the data flow from the analog signals by the individual sub-detectors of Belle to their digitized data to save in mass storage for offline analysis. Figure 3.1 shows the global data flow scheme of the Belle experiment. We can divide the Belle DAQ system into three parts: a front-end readout part, an event building part and a mass storage part. In the front-end readout part, detector outputs are digitized and the digitized data are sent to event building part. Then the event building part works to construct one event data from the data of individual sub-detectors. In the mass storage part, the built data are recorded to a tape library.

3.1.1 Front-end Readout – Q-to-T and multi hit TDC

The front-end readout part in Belle DAQ system takes care of the analog signals from individual sub-detectors in the Belle experiment and digitizes them [11]. The analog signals from sub-detectors except SVD are digitized by a unified readout system based on the Q-to-T conversion with FASTBUS TDCs [12, 13]. The principle of a Q-to-T technique is shown in Fig.3.2. A signal from a detector is connected to a capacitor. When the signal reaches the peak, the capacitor holds and start discharging. The output pulse width indicates the pulse hight of the original signal, and the leading edge (down edge) indicates the timing of the original signal. So we can measure both the timing and the height with one channel of TDC. The Q-to-T technique is effective to reduce the number of readout channels. For timing digitization, we use multi hit TDC LeCroy 1877S, which has 96 channels and holds up to 16 hits per channel.

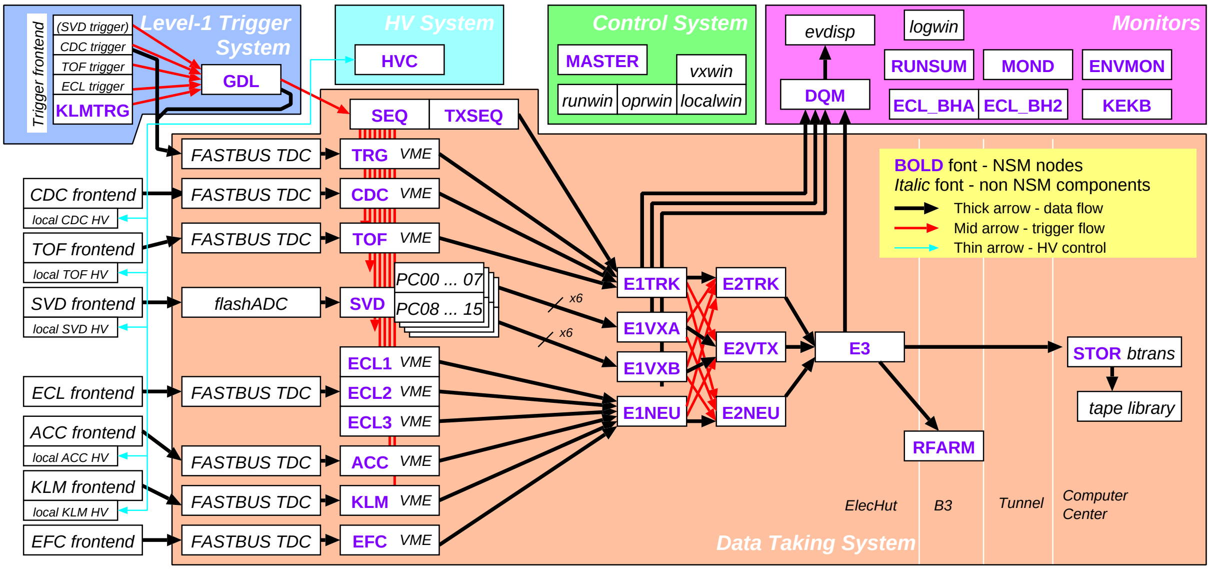


Figure 3.1: Overview of Data flow scheme in the Belle experiment. The current Belle DAQ comprises the front-end part (*FASTBUS TDC*, VME of each sub-detector), the event building part (E1xxx, E2xxx, E3), and the mass storage part (*tape library*)

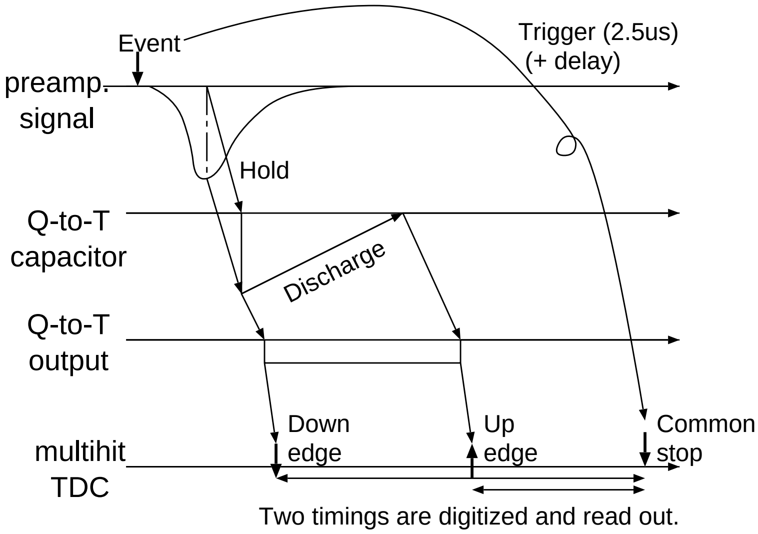


Figure 3.2: Principle of the Q-to-T technique

Digitized data are sent to a VME crate by a FASTBUS Processor Interface (FPI) and transfered to the event building farm through a 100base-TX network (see Fig. 3.3). One FASTBUS crate has one FPI, and the subsystem controller on VME requests all FPIs to collect data from TDC modules. The data from SVD are processed with a PC-based readout system and sent to the event building farm via the network [14].

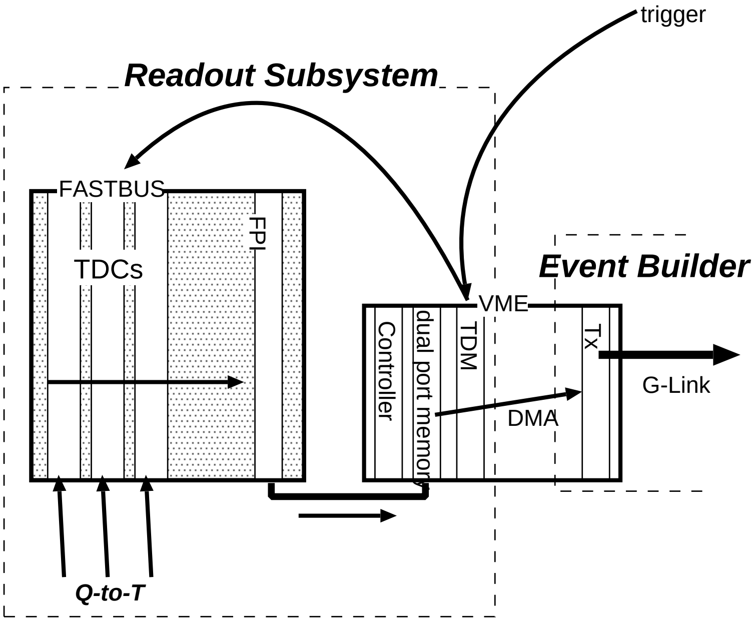


Figure 3.3: A schematic view of the unified TDC readout system. The data from the Q-to-T system is read out by multi hit TDC and sent to the event builder.

The current Belle readout system is not pipelined and, it has a readout dead time since we use the gate and delay method. The Figure 3.4 shows the distributions of readout time for some of detectors during when the data cannot be record. The readout time for SVD is almost fixed 30 μsec. The time for other detectors consists of two components. One is the constant latency caused by the readout overhead which is around 30 μsec. The other is the component which is proportional to data size and makes the tail component in the distribution. The relation between the total dead time fraction and the level-1 trigger rate is shown in Fig. 3.5. At a typical trigger rate of 400 Hz, the fraction of the dead time is around 2 %, which is reasonably small for the data acquision of the Belle experiment.

3.1.2 Event Building – Switchless event building farm

The event building part proceeds background reduction and form event data. We use a "*switchless*" system in which we connect all PCs in the point-to-point mode [15], to avoid any network congestion. The event building farm consists of three layers of PC servers (see Fig. 3.1). The first layer servers receive the data from the front-end readout part, perform a partial event building and carries out a software trigger (Level 2.5 trigger) processing using the partially built event data. The trigger signal is sent to the second layer servers to reject the event data. The second layer servers decide to send the event data to the third

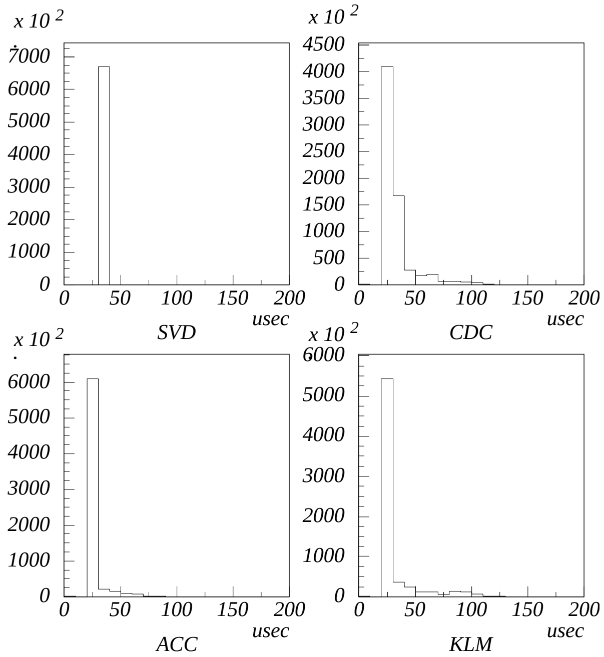


Figure 3.4: The distributions of the readout time for SVD, CDC, ECL, and KLM

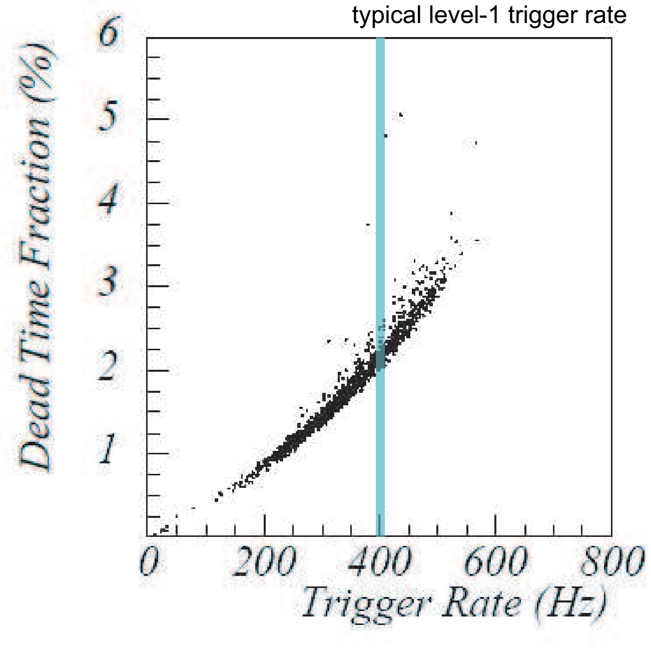


Figure 3.5: The dead time fraction as a function of the level-1 trigger rate. The typical trigger rate is about 400 Hz at current Belle condition.

layer server or discard them. The third layer server performs final event data construction and the online event selection [18]. The selected event is sent to the storage system. The current *switchless* event building farm is working well at the typical Belle trigger rate of 400 Hz(see Fig. 3.5). All PCs except third layer server of the current event building farm are equipped with four Intel PentiumⅢ CPU operating 700 MHz. The third layer server is equipped with two Intel Xeon CPU operating 3.06 GHz.

3.1.3 Mass Storage – High speed tape library

Sony Petasite tape library system with Sony DTF2 drivers is used as the mass storage system. We use SPARC workstations as the storage servers in order to use the tape library control software for Solaris operating system. We confirmed that the DTF2 drive provides the designed write speed of 24 MB/sec.

3.2 Overview of DAQ System for SuperKEKB

In this section, we describe the problems of the Belle DAQ system and the DAQ system for SuperKEKB. We show the requirement of the DAQ system for SuperKEKB in Subsection 3.2.1. We show the problems of the current Belle DAQ system in Subsection 3.2.2. The strategy of the DAQ system for SuperKEKB is written in Subsection 3.2.3. The readout system for SuperKEKB is described in Subsection 3.2.4. We show the overview and the design of event building farm for SuperKEKB in next Section.

3.2.1 Requirements to DAQ for SuperKEKB

The requirement of the DAQ system for the SuperKEKB experiment is much tighter than that of the Belle experiment. Table 3.1 shows the comparison of the design parameters of the DAQ systems between Belle and SuperKEKB.

Table 3.1: The list of DAQ design parameters of Belle and SuperKEKB.

	Belle	SuperKEKB
Luminosity(cm^{-2}s^{-1})	1.4×10^{34}	5×10^{35}
Physics trigger rate	140 Hz	1-5 kHz
Maximum trigger rate	500 Hz	10-30 kHz
Event size at L1	40 kB/event	200-300 kB/event
Data flow rate at L1	20 MB/sec	$>$ 2 GB/sec
Data flow rate at storage	10 MB/sec	$<$ 250 MB/sec

At the SuperKEKB experiment, the level-1 trigger rate is expected to be 10-30 kHz which is twenty times and more higher than the Belle experiment. The event size is expected to increase to 200-300 kB/event from current event size of 40 kB/event at the Belle, since 1) the number of channels will be increased and 2) we plan to sample the waveform of output signals for some sub-detectors. Due to the higher luminosity and the inflated event size, the data flow rates at the level-1 trigger is more than 2 GB/sec which is factor hundred and more higher than that of the Belle experiment. We need a novel system to handle such a high data flow rate. A system clock of the readout timing is chosen to be 42.33 MHz, one twelfth of the SuperKEKB RF clock of 508 MHz. These parameters depend on the luminosity increase of SuperKEKB.

3.2.2 Problems of the Current DAQ System

Although the current Belle DAQ system works well for the current Belle condition, there are several problems that prevent us from scaling up the system to be used in the - 30 times higher luminosity of 5 × $10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. For example, a linear extrapolation of the dead time fraction of current Belle system gives the dead time of more than 50 % (see Fig. 3.5). We think there are three limitation to use the Belle DAQ system at the SuperKEKB experiment.

The first limit comes from the front-end electronics and their readout. In the current FASTBUS-TDC based system using Q-to-T technique, it takes about 30 μsec in total to read out the TDC data of one event. This is too long when we need to handle 10 - 30 kHz trigger rate. Hence, we need a new deadtime-less readout system, which has pipeline buffer instead of gate and delay method.

The second limit is in the event building farm. The current event building farm is based on PCs connected one another via TCP/IP network without any large-scale network switch. The current event building farm is not scalable to the 30 times larger luminosity even if the PC performance is improved. Actually, we assume that the performance of the event building farm will saturate at a trigger rate of 600 - 700 Hz, which corresponds to a luminosity of about 2 × $10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ because of a lack of CPU power of the current event building farm. A parallel data processing is required to solve the problem.

The third limit is in the data storage. We expect the data storage rate will be 250 MB/s in the SuperKEKB. The current maximum data storage rate is 24 MB/s using high speed tape device.

The design strategy to overcome these limitations is discussed in the following subsection.

3.2.3 Strategy of DAQ for SuperKEKB

We need to develop a new DAQ system to satisfy the requirements listed in the subsection 3.2.1 for the SuperKEKB experiment. We still plan to keep the design concept of the current Belle event building farm as much as possible. The DAQ system consists of readout, event building farm and the storage parts. Each parts have difficulties toward the SuperKEKB experiment as described in the previous section. The new DAQ system must have the following features:

- Pipelined readout,
- Separated data streams, and
- Parallel data record devices.

The design strategies for the requirements are as follows:

1. Employ the pipeline based readout electronics to keep data taking during trigger decision.

2. Use of common readout platform as possible to handle the pipelined readout electronics.

3. Build up events in multi stages to manage the large number of readout modules and the large data size.

4. Adopt unit-style event building module to construct a scalable system to work with the luminosity increase.

5. Record the data onto disk directly.

The first and second strategies are to reduce the dead time of the readout and to ease the maintenance for the readout part. The third and forth strategies are to have DAQ processing scalability and to disperse the data flow in the event building farm part. The fifth strategy is to overcome the requirement of the data storage rate at the storage part.

Figure 3.6 show the schematic drawing of DAQ system. In the following subsections, we discuss the readout system and the event building farm system.

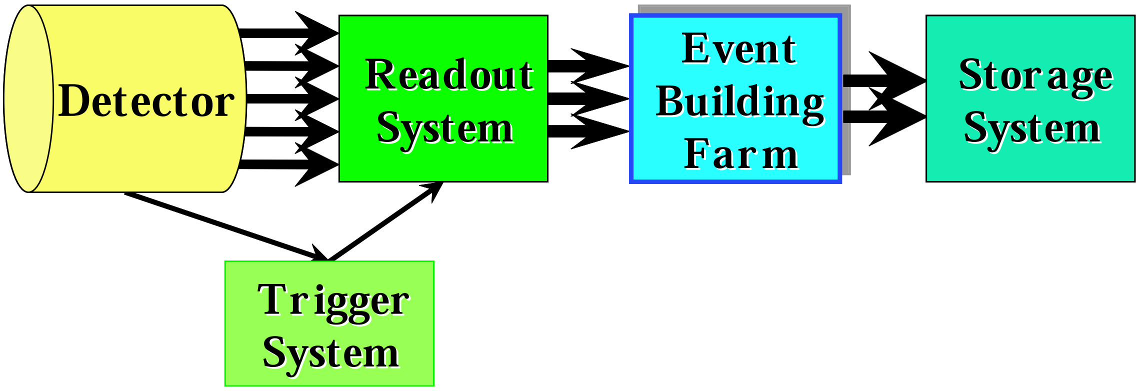


Figure 3.6: The schematic drawing of DAQ system for SuperKEKB.

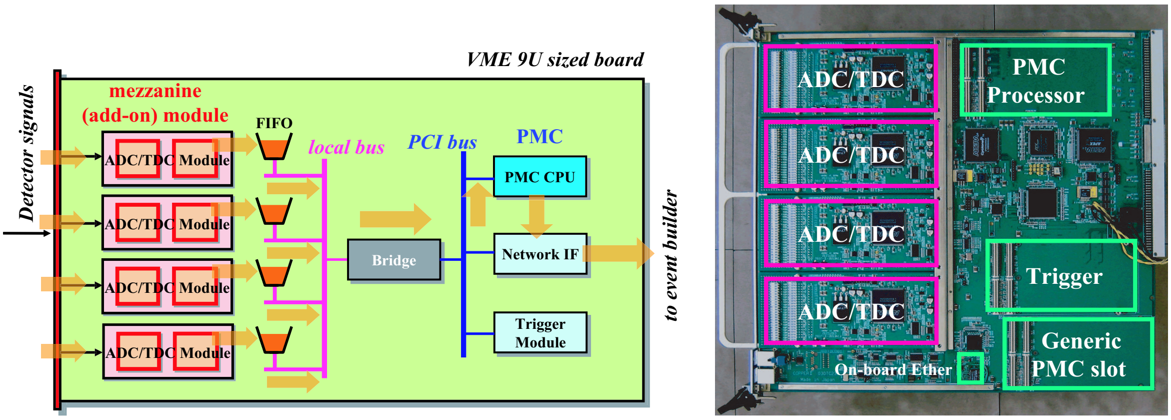
3.2.4 Common Readout System

To take care of the event rate of 30 kHz with the event size of 300 KB, a pipelined readout system is essential to handle the high trigger rate of 30 kHz with a low dead time. The readout system consists of a set of modularized common readout platform called a common pipeline platform for electronics readout (COPPER) [17]. The COPPER board is a VME 9U board to mount digitizer modules, a trigger timing receiver, a CPU module used for the on-board data processing.

Figure 3.7 shows a schematic drawing and a photograph of COPPER. One COPPER board is equipped with four slot for digitizer modules, four readout FIFOs for event buffering, and three PCI mezzanine card (PMC) slot. The CPU module on the COPPER board is a commercially available PMC module. It can be easily upgraded to use an up-to-date

CPU to increase the processing power. Many commercial PMC products are available; Ethernet cards, Gigabit Ethernet cards, memory modules and so on.

The digitizer modules are equipped with a L1 pipeline FIFO so as to record the digitized signal without readout dead time (see Appendix B). The L1 trigger signal is distributed to every COPPER module. The COPPER boards are received the L1 trigger signal by the trigger timing receiver from *Trigger Timing Distribution* system(see Appendix B).



(a) The schematic drawing

(b) The photograph

Figure 3.7: The schematic drawing and the photograph of COPPER.

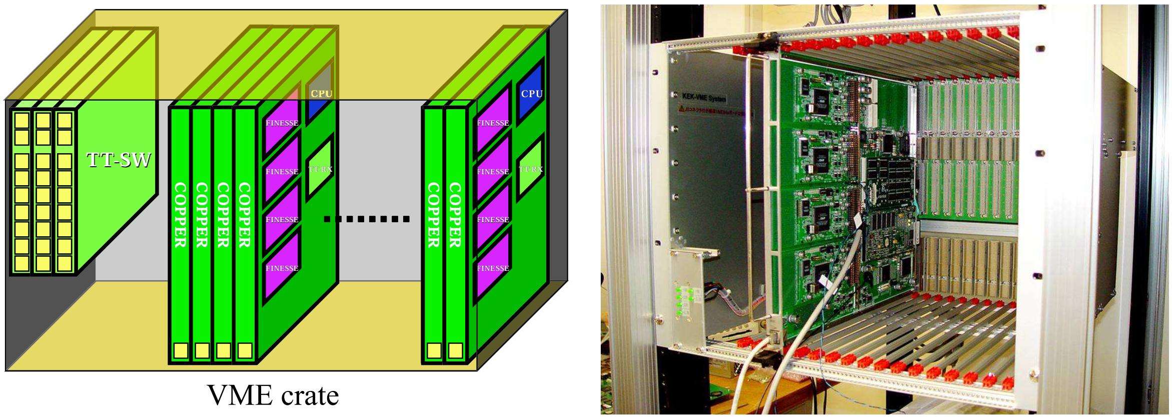
At SuperKEKB, the number of the COPPER boards is considered to be an order of 1,000. The total number of channels of central drift chamber for SuperKEKB is expected to be about 15,000 channels. Thus the number of the COPPER boards for central drift chamber is about 150 if the number of channels of one COPPER is assumed to be about 100. Figure 3.8 shows the picture of the COPPER with VME crate. One VME crate has 16 9U slot for the COPPER board and four 6U slot for trigger modules. The digitized data by digitizer modules are sent to the event building farm through the Ethernet of the COPPER board.

3.3 Desgin of Event Building Farm for SuperKEKB

In this section, we describe the event building farm for SuperKEKB and the software architecture for event building.

3.3.1 Multi-stage Event Building Farm

To perform the event building and data reduction with the $>$ 1000 readout modules with the trigger rate of - 30 kHz, we design a multi-stage event building and multiple unit structure. Figure 3.9 shows the global design of the event building farm for SuperKEKB. The designed event building system consists of three stages; Stage1, Stage2, and Stage3. At Stage1 (readout stage), event fragments are gathered from the readout modules to



(a) The schematic drawing

(b) The photograph

Figure 3.8: The VME crate with the COPPER board.

manage the large number of readout modules. At Stage2 (distribution stage), to have the scalability, the gathered event fragments are sent to one of the event building units, which are located in parallel at Stage3. At Stage3 (full building stage), the event fragments from Stage2 are built to form an event and the built event is sent to one of the level-3 farm units.

The purposes and the functionalities of each stages are summarized below.

Stage1 (Readout Stage)

Stage1 indicates the readout networks, which is to collect the event fragments from readout modules. For easy maintenance and cost and space reduction, the number of readout PCs have to be as small as possible. We employ small network switches in this stage. An overview of a part of Stage1 is shown in Fig. 3.10. Digitized signals from the front-end electronics of each sub-detector are first fed into the readout modules. The triggered data are then sent to the readout PCs via a network switch. The network switch is connected with 20 readout modules, this number corresponds to one VME create, by 100Base-TX. Each of readout modules and the network switch are connected by 100Base-TX, and the switch and the readout PC are connected by 1000Base-T.

The readout PCs perform partial event building for data ransferred from the VME create. A data reduction of the partially built event is also performed in the readout PCs. It is important for the readout PCs to read out from the readout modules in such high rate of 30 kHz at this stage

The details of Stage1 are described in Chapter 4.

Stage2 (Distribution Stage)

Stage2 indicates a network matrix for data transmission (distribution network matrix).

The readout PCs send the collected data from readout modules to one of *event-building units*. To distribute the collected data to *event building unit*, distribution network matrix

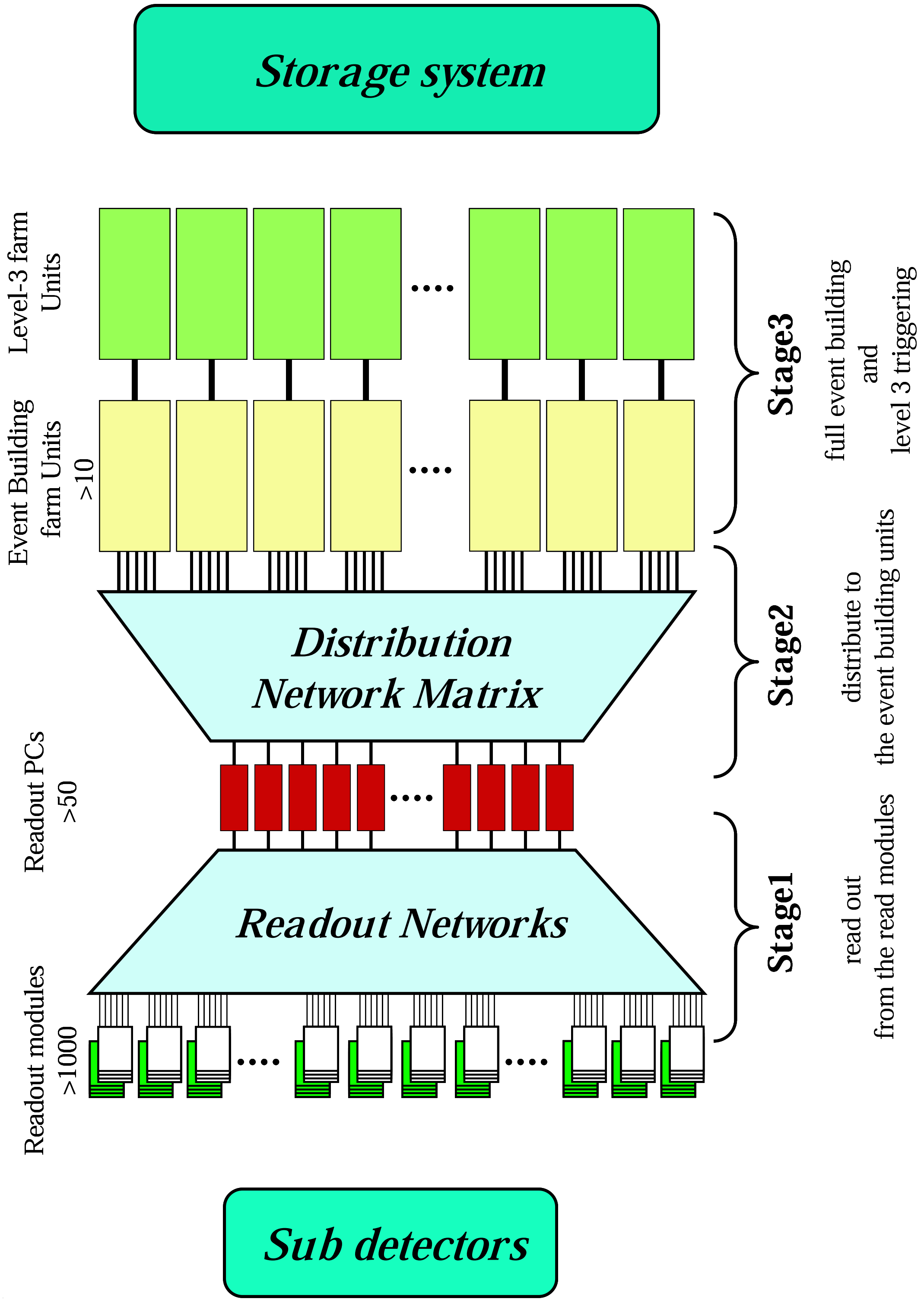

Storage system
Level-3 farm Units
Event Building farm Units >10
Distribution Network Matrix
Readout PCs >50
Readout Networks
Readout modules >1000
Sub detectors
Stage3
full event building and level 3 triggering
Stage2
distribute to the event building units
Stage1
read out from the read modules

Figure 3.9: Global design of the event building farm for SuperKEKB

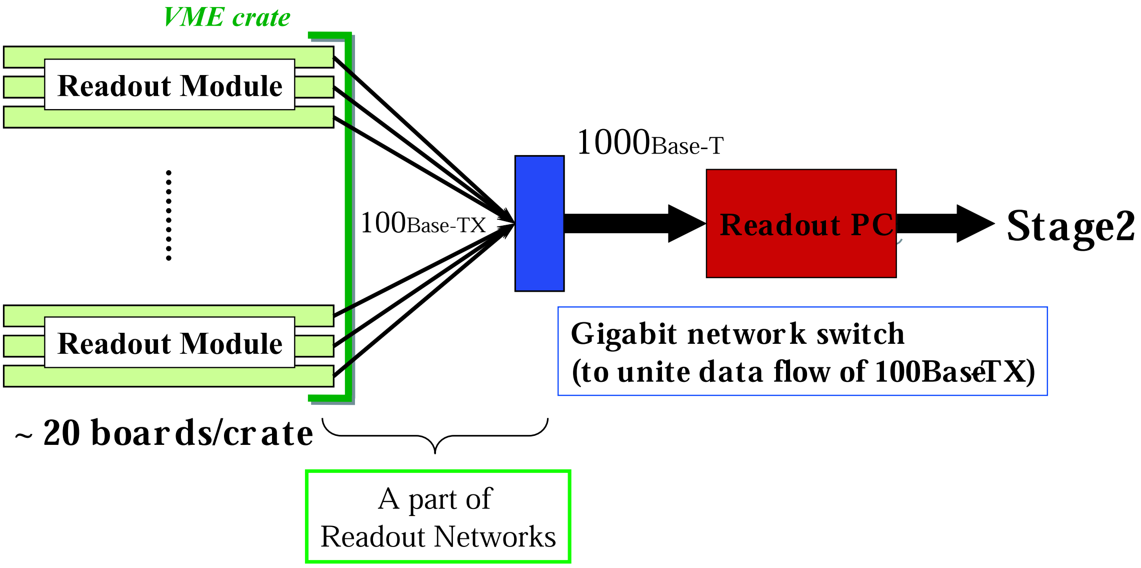


Figure 3.10: The schematic drawing of a part of Stage1. The black arrow is a data flow. Stage1 consists of the readout modules (COPPER boards), a gigabit network switch and a readout PC.

is constructed between the readout PCs and the *event building units*. Figure 3.11 shows the overview of Stage2 and the connections between them. We plan to employ about 10 *event-building units* at the beginning of SuperKEKB experiment. We can easily add more event building units to deal the increase of luminosity at SuperKEKB. The event fragments from sub-detectors are sent to one of the event building units. We can disperse the CPU usage of one *event building unit*, thus, the system can be scalable to handle the luminosity increase. Each of the readout PCs and event building units are connected by 1000Base-T via a network switch.

The details of Stage2 are described in Chapter 5.

Stage3 (Full Event Building Stage)

Stage3 indicates the multiple unit array, which consists of the event building farm unit and level-3 trigger farm unit. Each event building unit in Stage3 has almost the same structure as the one used in the Belle DAQ system consisting of three layers of PC arrays. Figure 3.12 shows the overview of Stage3. In the first layer-PCs of an event building unit, all of the event fragments from one sub-detector are gathered, and a software trigger (*level-2.5* trigger) processing is performed. In current Belle experiment, a fast track trigger using CDC information combine with the hardware trigger information is running as the *level-2.5* trigger. Our software framework [16] is designed to work for run equally in both online and offline environment so that the software trigger for online processing can be developed easily using offline PCs. When one of the first-layer PCs fires an event rejection signal from level-2.5 trigger, it is sent to all second-layer PCs so that the whole event data is discarded in the second layer.

The final event building is performed in the third-layer PC in the event building unit.

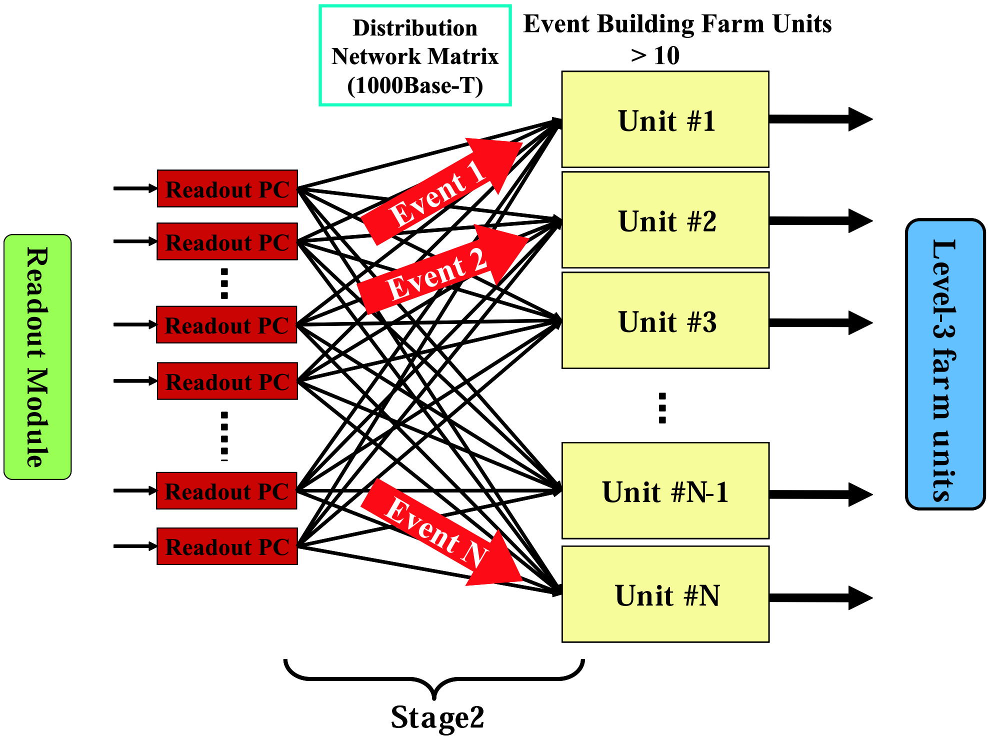


Figure 3.11: The overview of Stage2. Stage2 consists of the readout PCs, the distribution network matrix and the layer-1 servers in the event building units.

A fully built event is then sent to a level-3 trigger farm, which is directly connected to the output from the third-layer PCs. A full event reconstruction is performed in the level-3 trigger farm unit and then a sophisticated event selection is performed. The selected events are finally sent to the data storage system.

We set up a proto-type event building farm with the design concept discussed in this section. We study the performance of Stage1 and Stage2 in the new DAQ system for the SuperKEKB experiment using the prototype event building farm. The results are discussed in the following two chapters.

3.3.2 Software Architecture

We develop the software to receive the event fragments from the many connections and perform the event building. The software structure in the readout PC is schematically drawn in Fig.3.13. An ellipse, a circle and an arrow indicate a Linux process, a shared memory and a data flow, respectively.

The data transfer is based on the TCP/IP (Transmission Control Protocol/Internet Protocol) to guarantee the reachability to the next PC and the ordering of the data fragments. The data from each sender PC are received at a TCP socket through network interface card (NIC) by one *receiver process* and stored in a shared memory buffer. The data are not merged at this moment. The *event build process* collects the data fragments from all shared memory buffers and builds the data record. The built data record is stored

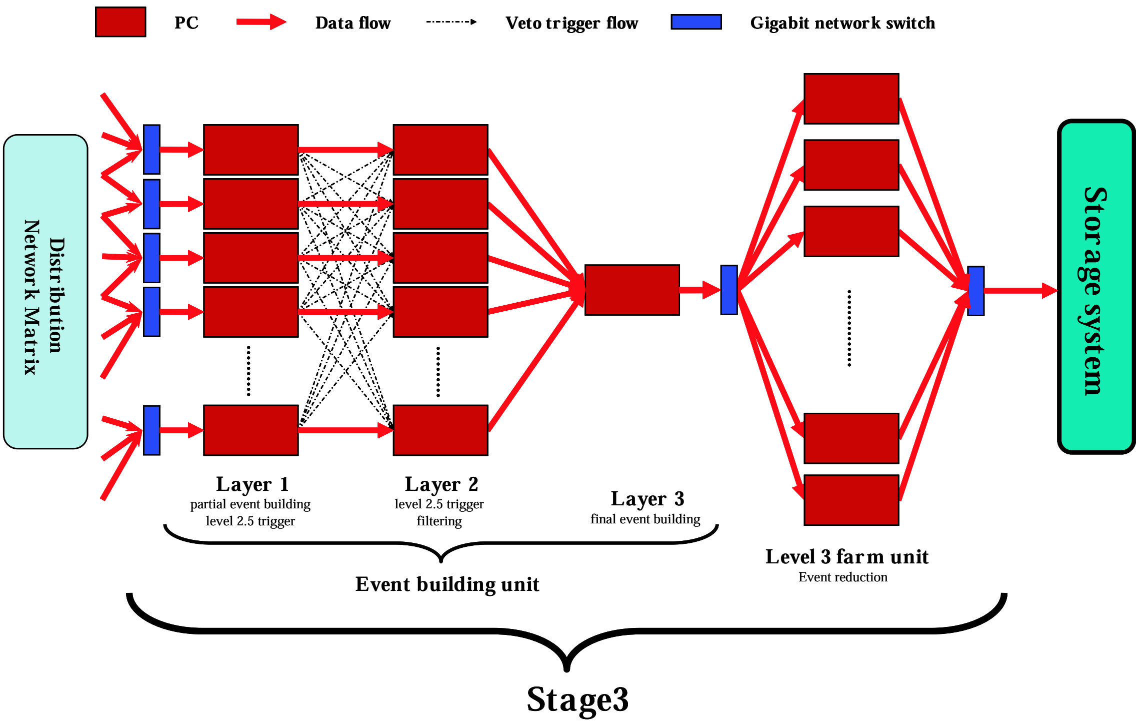


Figure 3.12: The schematic drawing of one event building unit and one level-3 farm units

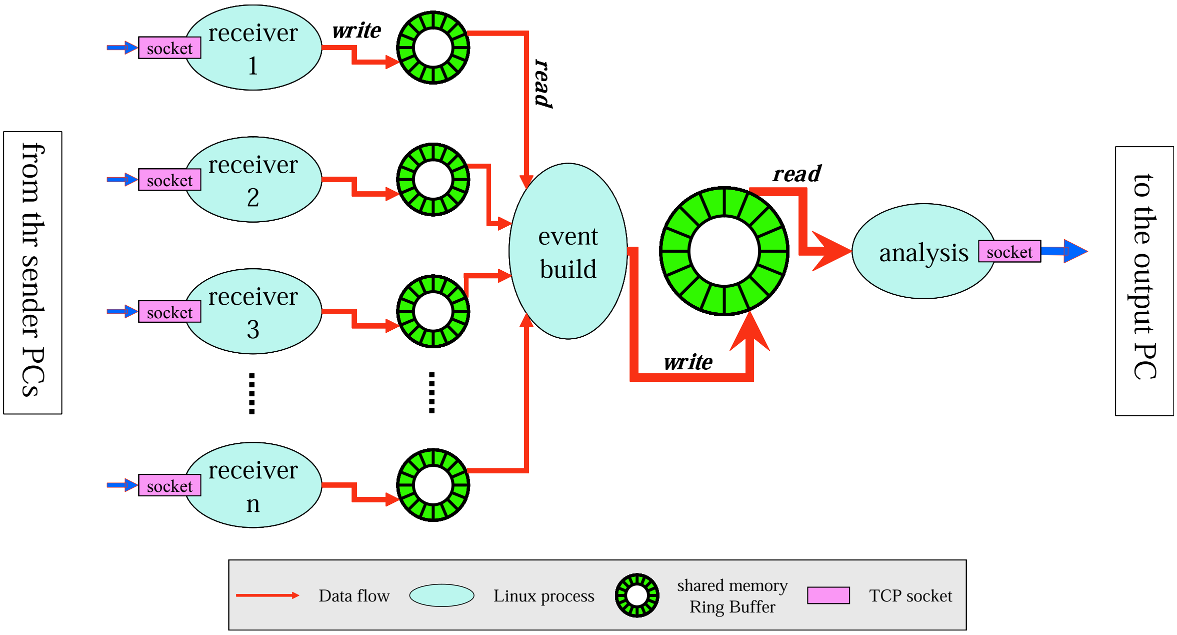


Figure 3.13: The software architecture for event building in the readout PC in case that n-th sender PC are connected.

in another shared memory. The built data is read by the *analysis process* and is sent to the output PC. In real situation, we can process the built data record to reduce the data size. The number of *receiver processes* equals to the number of sender PCs. When the number of the sender PCs is "n", the total number of processes is "$n + 2$"; n *receiver processes*, one *event build process* and one *analysis process.*

Ring Buffers of the Shared Memory

The shared memory buffer forms a ring buffer, which is divided into fixed size segments. This ring buffer can hold data up to 1000 event so that the *receiver process* can receive data from sender PCs as much as possible. We use SystemV semaphore to synchronize the received data fragments. The number of the stored event fragments and the number of empty segments in the ring buffer are recorded in the semaphore.

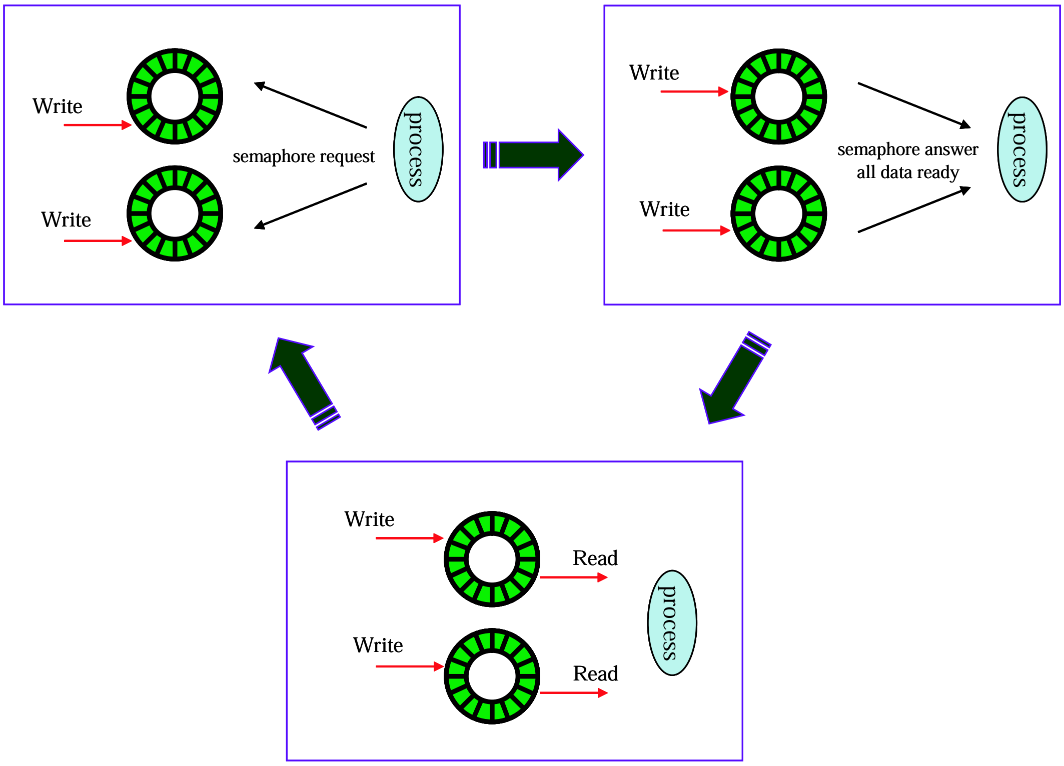


Figure 3.14: The schematic drawing of how the ring buffer works when the number of ring buffers is two.

Figure 3.14 shows how the ring buffer works. A Linux process asks the semaphore if data with a specific event number is available or not. When all event fragments belonging to the same event number are ready in the ring buffer, the Linux process accepts the reply from the semaphore and read the data from all shared memory buffers.

By changing the number of connections and the depth of the sheard memory, we can use this software at every Stage of the event building farm.

Chapter 4

Performance Study of Stage1 Event Building

To avoid the network congestion, all PCs of the event building farm are connected each other via point-to-point connections in the current Belle system. However the number of readout modules is considered to increase to more than 1000. To ease the maintenance and to reduce the cost and space, we need to keep the number of the readout PCs to be as small as possible. We then employ the small network switches in Stage1 of the event building farm for SuperKEKB. By using the network switch, we can reduce the number of network interface cards and the number of readout PCs, which gather the event fragments and performs the partial event building. We have to make sure that the network switch does not restrict the network flow and cause the network congestion. To decide the configuration of Stage1, we study the performance, which one readout PC collects the event fragments from many readout modules assuming the maximum number of 20. This chapter describe the performance studies of Stage1.

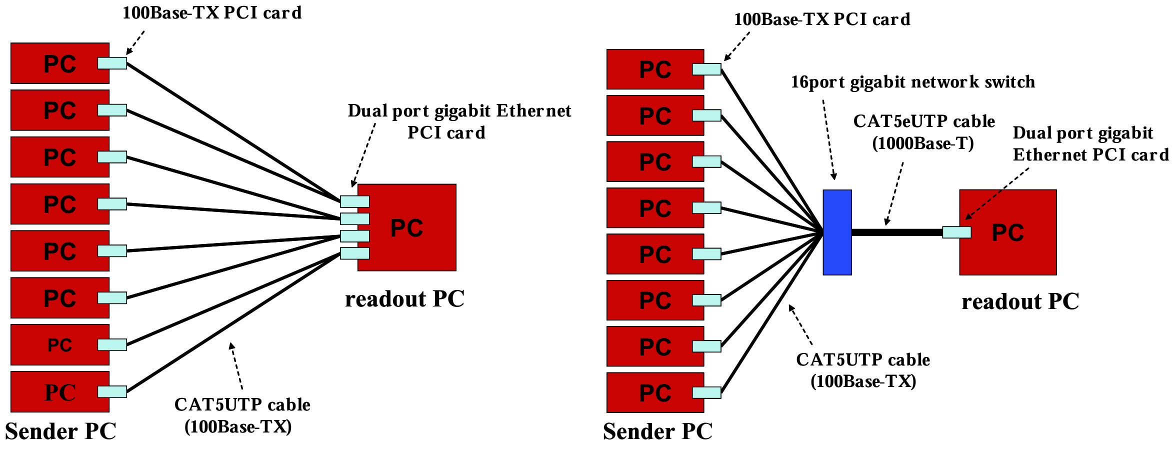
4.1 Basic Network Performance Study

In order to compare the readout network using the network switches (the network switch mode) with the one using the point-to-point connections (the point-to-point mode) and prove that the network switch is not the bottleneck of data transmission, we measure the event rate and the data transfer rate varying a event size. We also vary the number of PCs from one to eleven, which create the pseudo data and send it to one readout PC. We use PCs to simulate the readout modules described in the Subsection 3.2.4 and call them "sender PCs." In the network switch mode, each of the sender PCs and the readout PC is connected via the network switch. In the point-to-point mode, each of the sender PCs is connected directly with the readout PC and the network switch is not used. The readout PC does not build the received data so as to make no CPU usage except the data receiving process in the readout PC.

4.1.1 Setup for Basic Network Performance Study

Test configuration of the point-to-point mode is shown in Fig.4.1(a). Each of sender PCs and the readout PC are connected directly via 100Base-TX with a CAT5 UTP cable. Test configuration of the network switch mode is shown in Fig.4.1(b). The sender PCs and the network switch are connected via 100Base-TX. The network switch and the readout PC are connected via 1000Base-T with a CAT5e UTP cable. The network switch is used to combine the 100Base-TX data flow into the flow of 1000Base-T. The switch we used is FXG-16TX, which is produced by PLANEX. It has 16 gigabit ports and 272 kB packet buffer memory. The transfer mode of this switch is the store&forward mode. The switch is also equipped with flow control, which employs IEEE802.3x in full duplex mode.

We use PCs, in which Red-Hat9 with 2.4.20-8smp kernel are installed, instead of a readout modules. The readout PC is equipped with two Intel Xeon CPUs operating at 2.46 GHz and Red-Hat9 with 2.4.20-8smp kernel is also installed. The CPUs of the readout PC are operated with Hyper Threading Technology enabled. Figure 4.2 shows the picture of the network switch and the readout PC.


100Base-TX PCI card
PC
PC
PC
PC
PC
PC
PC
PC
Dual port gigabit Ethernet PCI card
PC
readout PC
CAT5UTP cable (100Base-TX)
Sender PC
100Base-TX PCI card
16port gigabit network switch
CAT5eUTP cable (1000Base-T)
Dual port gigabit Ethernet PCI card
PC
PC
PC
PC
PC
PC
PC
PC
PC
readout PC
CAT5UTP cable (100Base-TX)
Sender PC

(a) point-to-point mode

(b) network switch mode

Figure 4.1: The schematic drawings of the test configuration to study basic network performance.

4.1.2 Comparison of Network Switch Mode with Point-to-point Mode

We measure the transfer rate of sender PC varying the event size per event. The result of the comparison of the network switch mode with the point-to-point mode is shown in Fig.4.3, when the number of sender PCs is eight.

We observe that the transfer rate of the both modes reaches close to the maximum transfer rate of 100Base-TX. When the event size is less than a few hundred bytes, the transfer rate of the network switch mode is 10 % higher than that of the point-to-point

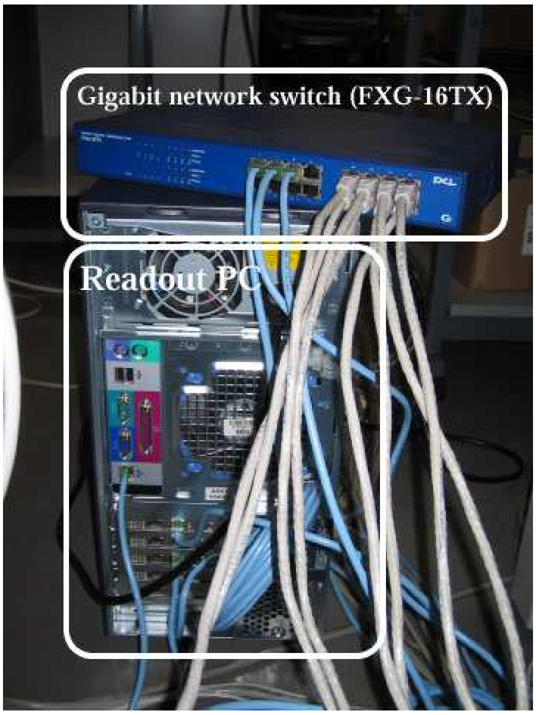


Figure 4.2: The picture of the network switch (upper) and the readout PC (lower).

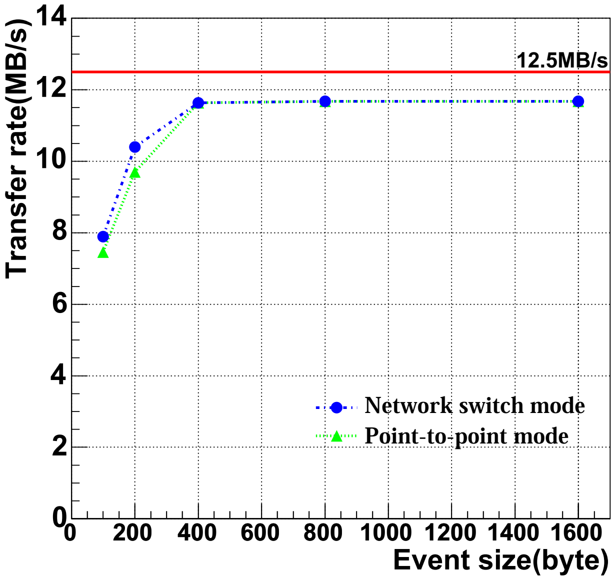
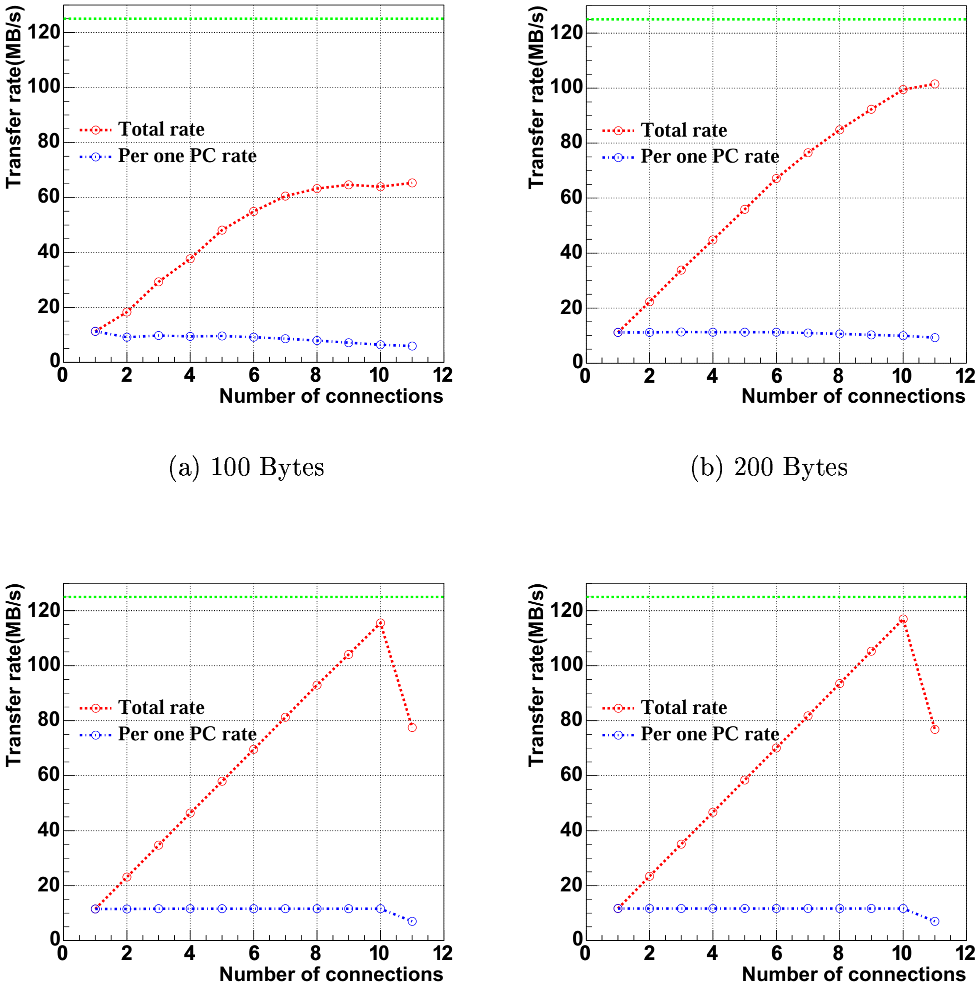


Figure 4.3: Comparison of the transfer rate of the network switch mode (blue) and that of the point-to-point mode (green) when the number of sender PCs is eight. The vertical axis shows the transfer rate and the horizontal axis shows the event size per sender PC. The horizontal solid line (red) is the maximum transfer rate of 100Base-TX (12.5 MB/s).

mode since the network switch has the packet buffer of 272 kB. We find that the network congestion does not occur in the network switch mode in the configuration of the number of sender PCs to be eight.

4.1.3 Number of Connections vs. Throughput

We measure the transfer rate varying the number of sender PCs from one to eleven and varying the event size from 100 to 800 Bytes. The results of the transfer rate measurement are shown in Fig.4.4.



(a) 100 Bytes

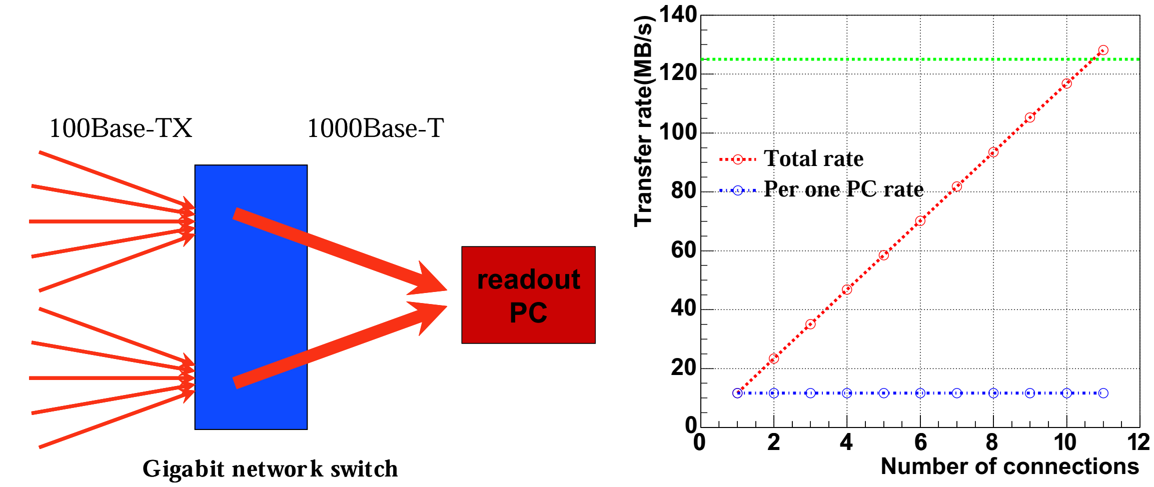
(b) 200 Bytes

(c) 400 Bytes

(d) 800 Bytes

Figure 4.4: The transfer rate as a function of the number of connections. The dashed (red) lines and the dotted (blue) lines show the total transfer rate and the transfer rate per sender PC, respectively. The green lines show the maximum transfer rate of 1000Base-T.

For the case of the event size of 100 and 200 Bytes, the total transfer rate is saturated at around eight or ten connections, respectively. The transfer rate per one sender PC slightly decreases. When the event size is 400 or 800 Bytes, the transfer rate per one sender PC reaches close to the limit of 100Base-TX), and the total transfer rate also reaches close to the limit of 1000Base-T at ten connections. In case of eleven connections, the total transfer rate and the single transfer rate drop to about 80 MB/s and about 7 MB/s in the 400 and 800 Bytes cases, respectively. This drop in data transfer rates indicates that the buffer memory of the network switch almost fulls since the connection between the network switch and the readout PC is over 1000Base-T (125 MB/s).



(a) 2 way

(b) 800 Bytes in 2 way

Figure 4.5: (a)The schematic drawing of data flow in network switch using two ports for output. (b)The transfer rate as a function of the number of connections. The dashed (red) line and the dotted (blue) line shows the total transfer rate and the transfer rate per Sender PC, respectively. The green line shows the maximum transfer rate of 1000Base-T (125 MB/s).

In order to solve the transfer rate drop at eleven connections, we use two ports to output the data (see Fig 4.5(a)). Figure 4.5(b) shows the result of the measurement using two ports for output. At the eleven connections, the total transfer rate exceeds the 1000Base-T capability and the single transfer rate does not drop in 100Base-TX since a load of data flow on one port of the network switch is dispersed. These two output data flows are independent of each other. By using the two ports for output, we can use full performance of the 1000Base-T even the number of connections is more than ten.

4.2 Performance of Stage1

In Stage1, the readout PC collects the digitized event fragments by the readout modules and performs the partial event building as described in Subsection 3.3.1.

Here, we summarize the requirements for Stage1.

- ◇ The readout PC should collect the event fragments from as many readout modules as possible to reduce the number of the readout PC. The maximum number of readout modules we consider is 20 modules.
- ◇ The readout PC should perform the partial event building in the high trigger rate of SuperKEKB. The typical trigger rate and the maximum trigger rate are considered to be 10 and 30 kHz, respectively.

4.2.1 Test Setup

To investigate the performance of Stage1, we set up the test bench as shown in Fig.4.6.

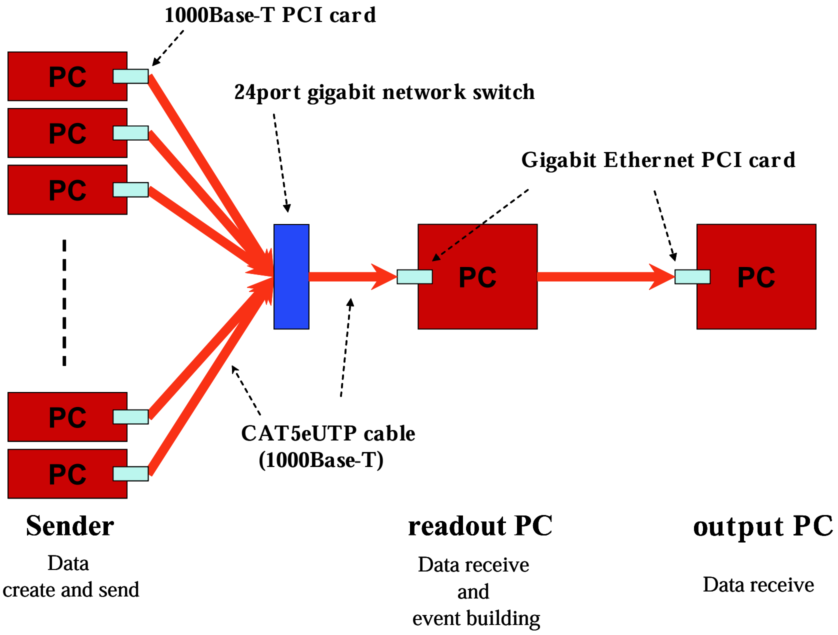


Figure 4.6: The test bench setup for Stage1. The red arrows show the data flow.

The test bench consists of sender PCs, which can simulate the readout modules, the readout PC and the output PC. A role of each PC is as follows as;

- **The sender PC** – emulate the readout module, generate a pseudo data and send them to the readout PC.
- **The readout PC** – receive the pseudo data, build up one event and send the built data to the output PC.
- **The output PC** – receive the built data.

The pseudo data, which are generated by the sender PCs, are collected and built up into a single event by the readout PC. The readout PC sends the built data to the output PC. Each of the sender PCs and the readout PC is connected with the 1000Base-T Ethernet via a gigabit network switch. We use 24 port gigabit network switch, which is FMG-24K provided by PLANEX in this study. The transfer mode of the switch is the store&forward mode. The packet buffer memory in the network switch is 1 MB. In the real situation, the connection between the readout modules and the readout PC is 100Base-TX. If a throughput from the sender PCs to the readout PC is larger than 12.5 MB/s (the limit by 100Base-TX) in this study, the throughput should be considered to saturate at 12.5 MB/s in the real situation. The readout PC and the output PC are directly connected with the 1000Base-T Ethernet using point-to-point mode. Each of all PCs is a SMP server equipped with two Intel Xeon CPUs operating 3.06 GHz with Hyper Threading enabled and Red-Hat9 with 2.4.20-8smp kernel is installed. We use the event building software as shown in Fig.3.13.

4.2.2 Results of Performance Measurement of Stage1

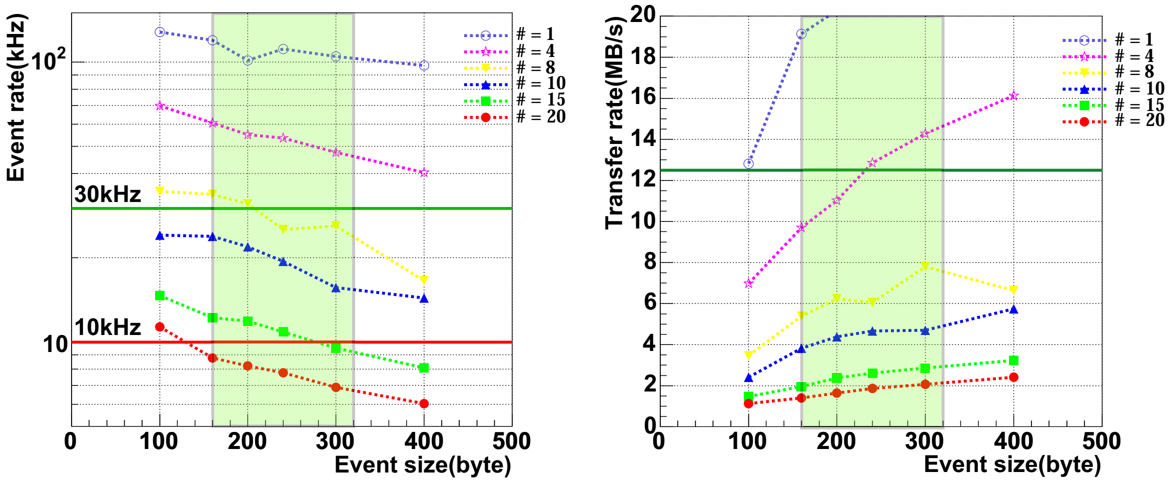
We study the performance of Stage1 varying the event size from 100 Bytes to 400 Bytes and the number of the sender PCs from one to twenty. The measured performances are the event rate and the transfer rate of the sender PCs. We define the event rate as the number of average events, which the sender PC has sent in every one second. Each sender PC sends the data to the readout PC as much as possible. So, the event rate indicates the upper limit of an acceptable trigger rate in Stage1.

Figure 4.7 shows the results of Stage1 performances as a function of the event size. We also define the typical event size corresponding to the one from the drift chamber with 10 % occupancy. One channel of the drift chamber corresponds to 16 bytes. The number of channels of one readout module is assumed to be 100. One event size from one readout module is 1600 Bytes. When an occupancy of the drift chamber is 10 %, the typical event size becomes 160 Bytes. As shown in Fig 4.7(a), the event rate decreases by increasing the number of the sender PCs and reaches the typical trigger rate of 10 kHz when the number of the sender PCs is less than 15. The event rate also decreases by increasing the event size. As shown in Fig.4.7(b), if the number of the sender PCs is four, the transfer rate is to be over 12.5 MB/s at the event size of 240 Bytes. In the real situation, we use 100Base-T in the connection between each of the sender PCs and the readout PC, and this transfer rate saturates at 12.5 MB/s. The transfer rate increases according to the increase of the event size as shown in Fig.4.7(b).

Search for Bottleneck

Figure 4.8 shows the event rate as a function of the number of the sender PCs at the typical event size of 160 Bytes. With the typical event size of 160 Bytes, the event rate reaches about 30 kHz when the number of sender PCs is less than eight. We find that the event rate decreases gradually as the number of the sender PCs increases. When the number of sender is 20, the event rate is about 8 kHz, which is not reached the typical trigger rate.

Figure 4.9 shows the transfer rate per one sender PC and total transfer rate. As shown



(a) Event rate of sender

(b) Transfer rate of sender

Figure 4.7: The result of the performance study of Stage1 for variable number of sender PCs. The horizontal axis shows the event size. The vertical axis shows the event rate (a) and the transfer rate (b), respectively. (a): The red horizontal line is the typical trigger rate. The green horizontal line is the maximum trigger rate of 30 kHz. (b): The horizontal green line is the limit of 100Base-TX. The light green area shows typical event size area where the occupancy of the drift chamber is assumed 10 % - 20 %

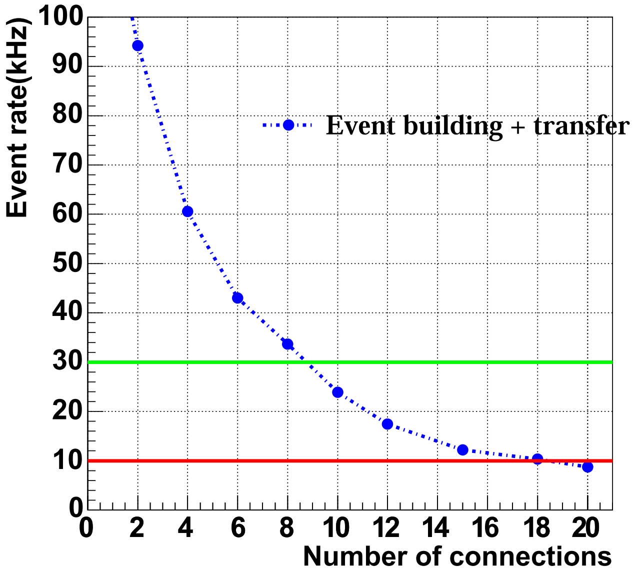
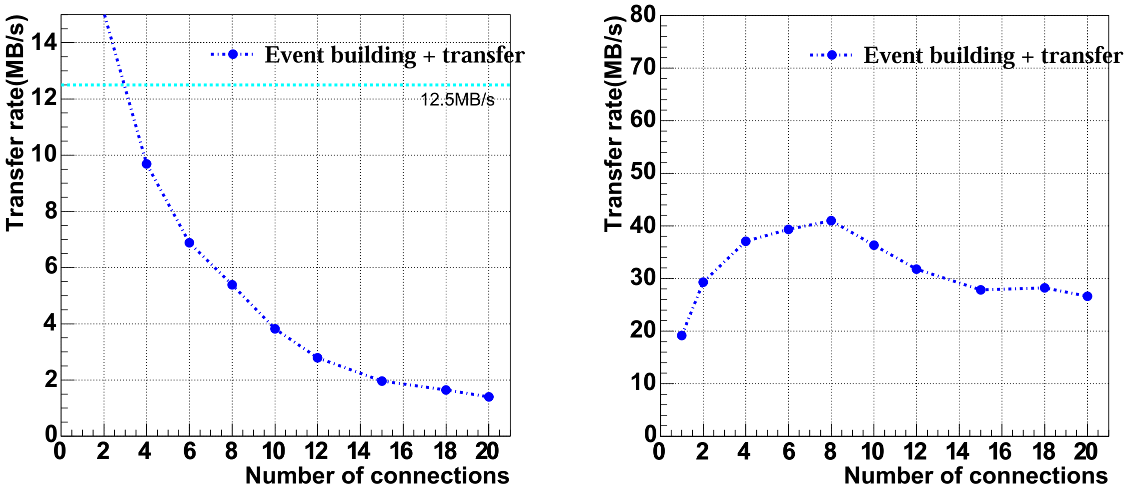


Figure 4.8: Event rate as a function of the number of sender PCs with the event size of 160 Bytes. The green and red horizontal lines show 30 kHz and 10 kHz trigger rates, respectively.



(a) Transfer rate per one sender

(b) Total transfer rate

Figure 4.9: The transfer rate per one sender (a) and the total transfer rate (b) as a function of the number of connections. The light blue dotted line (a) is a limit of 100Base-TX.

in Fig 4.9(a), the transfer rate do not reach the maximum transfer rate of the 100Base-TX. As shown in Fig 4.9(b), the total transfer rate do not use the whole throughput of the 1000Base-T. Therefore, Fig 4.9 indicates that the bottleneck is the software of the readout PC. We think that the *event building process* does not work well (see Fig.3.13). The next paragraph shows the study of the bottleneck.

Active Process Ratio

In the high trigger rate operation of 10 - 30 kHz, it could be considered that each process cannot use enough CPU time because of too many processes. To find out how each process uses the CPU, we monitor the status of each processes.The process status is described by R or S, where R means that the process is running or runnable in run queue and S means that the process is in wait queue by call of `interruptible_sleep_on()`. We define the active process ratio as the percentages of R processes to R + S processes. By examining the active process ratio, we can find out which of *event build process* or *receiver process* is the bottleneck. If the active process ratio decreases as the number of the sender PCs increase, the bottle neck is in *event build process.* If the active process ratio does not vary, the bottle neck is in the *receiver process.* The active process ratio is shown in Fig.4.10.

We observe the active process ratio decreases as the number of connections increase. If the number of connections is larger than 15, more than half of the processes is in S status, which mean the *event build process* cannot use the sufficient CPU resource. Thus, event building by *event build process* takes time, then *receiver process* is blocked since it cannot write the received data into the shared memory buffers until the buffers are cleared.

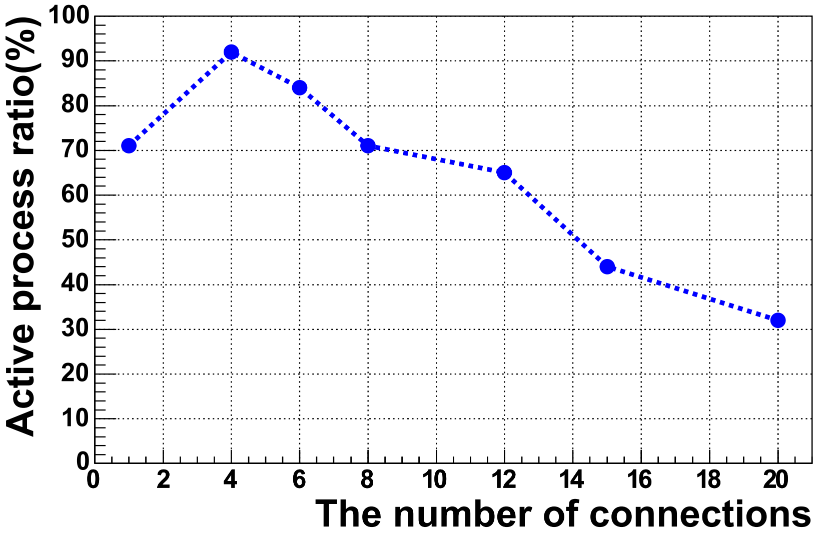


Figure 4.10: The active process ratio as a function of the number of connection.

Modification of Event Building Software

To resolve the bottle neck, we modify the software to reduce the number of processes. By reducing the number of processes, *event build process* can use more CPU resource. Figure 4.11 shows the schematic drawing of modification step.

One *receiver process* has one TCP socket in the previous method, the one *receiver process* has multiple TCP sockets in the modified method. The *receiver process* reads the event fragments from multiple TCP sockets using `select()` system call. We reduce the number of processes by getting the multiple TCP sockets together in one process. We measure the event rate varying the number of the TCP sockets per one *receiver process* from one to ten.

Table 4.1: The comparison of the number of processes in the readout PC.

Total number of process	# of connections per one receiver process	Event rate (kHz)	Improvement (%)
22	1	8.8	
12	2	12.1	37.5
7	4	13.2	50.0
6	5	13.4	52.2
4	10	9.5	8.0

In case of the sender PCs being 20, Table 4.1 shows the result of measurement using modified software. When the number of the collected TCP socket is four or five, the event rate is improved by about 50% and reaches about 13 kHz, which is over the typical trigger rate of 10 kHz. The reduction of the number of *receiver process* is effective to make *event build process* to use sufficient CPU time. As for the further software improvement,

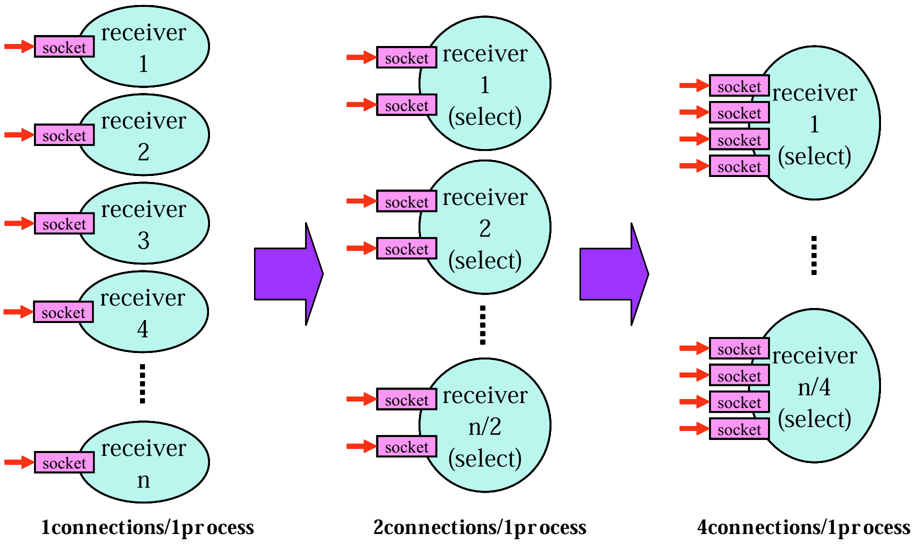


Figure 4.11: The schematic drawing of the receiver process of the readout PC using select function. One receiver process receives data from multiple TCP sockets. The number of the sockets are one (left), two (middle) and four (right).

reduction of the number of shared memory buffers is effective to carry out the faster event building of the *event build process*.

4.3 Summary of Stage1

To study the gigabit network switch, we compare the network switch mode with the point-to-point mode. We observe that the network congestion does not occur in the network switch.

We set up the test system of Stage1 and investigate the event rate. The event rate of Stage1 is measured to achieve 33.7 kHz when the number of the sender PCs is eight, and 13.4 kHz by using modified event building software when the number of sender PCs is 20 with the typical event size. For 30 kHz operation, we find that further software improvement and faster PC with multiple CPUs are required when the number of sender PCs is 20.

Chapter 5

Performance Study of Stage2 Event Building

At present, we use the *switchless* event building farm, which use no network switch. The current event building farm is working well in the typical trigger rate of 400 Hz as shown in Fig.3.5. The typical trigger rate of the SuperKEKB experiment, however, is considered to be more than 10 kHz. The current event building farm is not expected to work in such the high trigger rate. Thus, we have to design a new event building farm which tolerate the high trigger rate. Since we expect the luminosity will gradually increase after the beginning of the SuperKEKB experiment, the event building farm have to be scalable as the luminosity increase. To meet these requirements and to handle the large number of the readout PCs, we employ multiple unit structure. With this structure, we can make a large system by scaling up from current system easily. Another merit of employing the structure is that we can keep the concept of current event building farm, which works well. To distribute the event fragments to multiple units, the network matrix for data distribution (the distribution network matrix) is formed between the readout PCs and the multiple units (see Fig.3.11). We study whether the distribution network matrix has the scalability and work under the high trigger rate of 10-30 kHz. This chapter describes the study of the distribution network matrix of Stage2.

5.1 Requirements to Stage2

In Stage2, the readout PCs send the collected data to the multiple units in turn. The sent data are built to an event by PC servers of Layer 1 in Stage3 as shown in Section 3.3.1.

Here, we list the requirements for Stage2.

- ◇ The distribution network matrix between the readout PCs and the multiple units should work under the high trigger rate of 10 - 30 kHz. We define 10 kHz and 30 kHz as the typical trigger rate and the maximum trigger rate, respectively.
- ◇ To catch up the luminosity increase, the system should have scalability by adding the event building units.

5.2 Test Setup

To measure the tolerable event rate and the transfer rate of the distribution network matrix, we set up the test bench as shown in Fig.5.1. Figure 5.2 shows the picture of PC servers we used for the performance study.

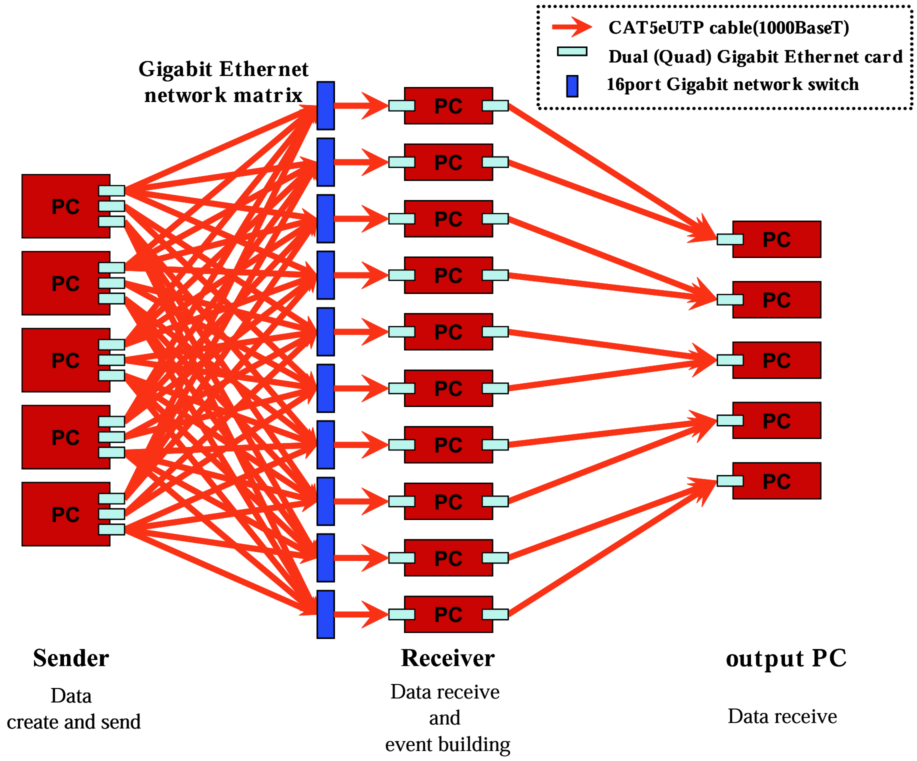

CAT5eUTP cable(1000BaseT)
Dual (Quad) Gigabit Ethernet card
16port Gigabit network switch
Gigabit Ethernet network matrix
PC
Sender
Data create and send
Receiver
Data receive and event building
output PC
Data receive

Figure 5.1: The test configuration and data flow for Stage2

The test bench comprises five sender PCs, ten receiver PCs and five output PCs. A role of each PC is as follows as;

- **The sender PC** – emulate the readout PC, generate a pseudo data and send them to the receiver PC.
- **The receiver PC** – emulate one of the layer-1 server of event building unit in Stage3, receive the pseudo data, build up one event and send the built data to the output PC.
- **The output PC** – receive the built data and emulate one of the layer-2 server of event building unit in Stage3.

Each sender PC has ten gigabit ports for the data transmission, equipped with two quad gigabit network interface card (NIC) and one dual gigabit NIC. Each receiver PC has a 16 ports gigabit network switch (FXG-16TX) provided by PLANEX. The total number of the network switches is ten. Each of the sender PCs and each of the receiver PCs are connected with 1000Base-T via the gigabit network switch. Each of the receiver PCs and each of

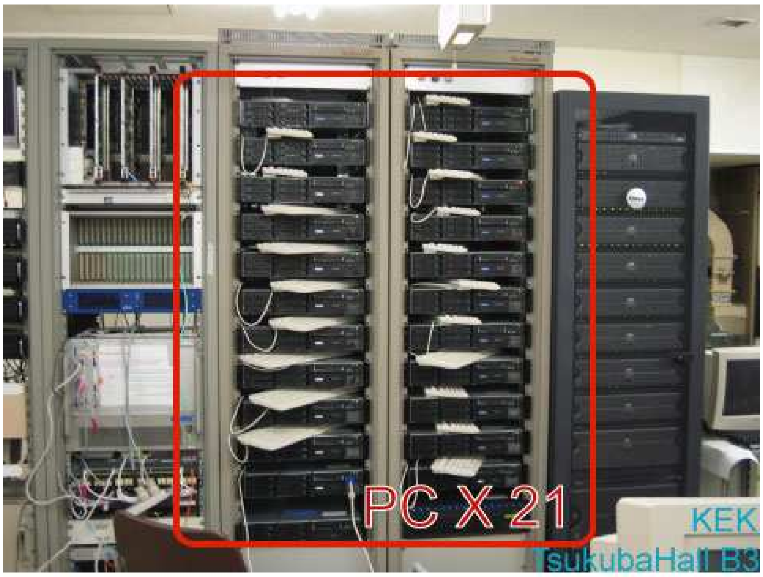


Figure 5.2: The picture of the PCs which we use for the Stage2 study in KEK Tsukuba Hall B3.

the output PCs are connected with 1000Base-T directly. Between the sender PCs and the receiver PCs, 5 × 10 network matrix is constructed as the distribution network matrix. All PCs are SMP servers equipped with two Intel Xeon CPUs operating 3.06 GHz and RedHat9 with kernel 2.4.20-8smp is installed. With Hyper Threading Technology enabled, each of the server PCs has four virtual CPUs, where two virtual CPUs corresponds to one physical CPU. The summarized parameters of the PCs are listed in Table 5.1.

The sender PCs create pseudo data and send them to the receiver PCs, which is selected in turn. For example, the sender PCs create the *n*-th event data and send them to *m*-th receiver PC. Next time, the sender PCs create the *n+1*-th event data and send them to *m+1*-th receiver PC. If *m+1* is larger than ten, which is the number of the receiver PCs in this test configuration, the sender PC send the *n+1*-th event data to the *1*-st receiver PC. The receiver PCs perform the partial event building with the data, which are received from the sender PCs. The built data are sent to an output PC by the receiver PC.

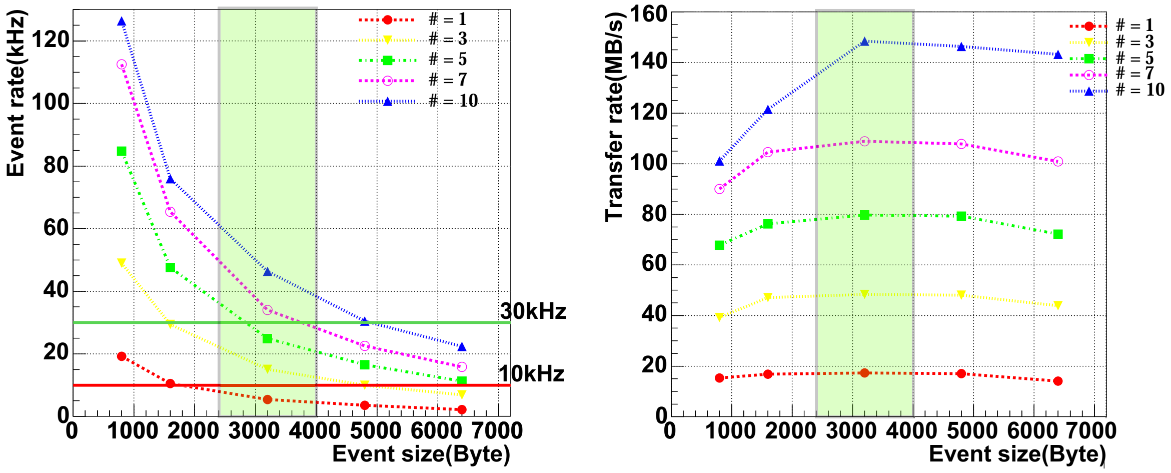
We use the event building software as shown in Fig.3.13.

Table 5.1: Parameters of the PCs used for the performance study.

	Equipment type
CPU	Intel Xeon 3.06 GHz Dual
Memory	1 GB
NIC	Intel PRO/1000 MT Quad(Dual) Port Server Adapter
NIC driver	Intel 82546(EB) controller
PCIbus	PCI-X (64bit/100MHz)
OS	RedHat9 with kernel 2.4.20-8smp

5.3 Results of Performance Measurement of Stage2

We measure the event rate and the transfer rate of the sender PC varying the event size from 800 to 6400 Bytes and the number of the receiver PC from one to ten when the number of the sender PCs is five. We define the event rate as the average number of the events that five sender PCs can send in one second. Each sender PC sends the data to the receiver PCs as much as possible. We also define the typical event size of Stage2 which corresponds to the data size from the drift chamber with 10 % occupancy. If the number of the readout modules connected to one readout PC is assumed to be 20 at Stage1, the typical event size from one readout PC is 3200 Bytes.



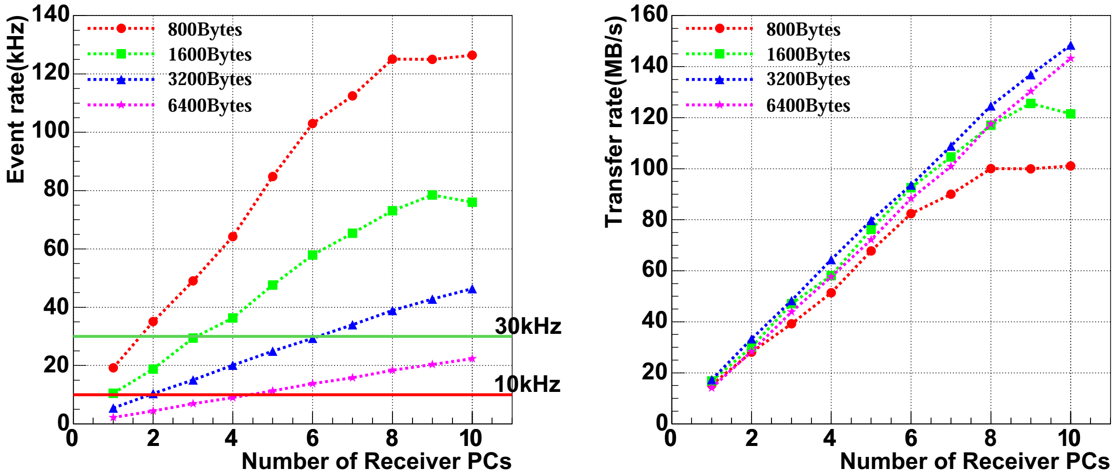
(a) Event rate

(b) Transfer rate

Figure 5.3: The event rate (a) and the transfer rate (b) of the sender PCs as a function of the event size varying the receiver PCs from one to ten when the number of the sender PCs is five. (a): The green and red horizontal lines show 30 kHz and 10 kHz trigger rate, respectively.

Figure 5.3 shows the result of the measurement as a function of the event size. Here the number of the sender PCs is five. As shown in Fig.5.3, the event rate and the transfer rate of the sender PC increase by increasing the number of the receiver PCs. When the number of the receiver PCs become ten, the event rate gets over the maximum trigger rate of 30 kHz with the typical event size of 3200 Bytes. The transfer rate is about 150 MB/s with the typical event size.

Figure 5.4 shows the result of the measurement as a function of the number of the receiver PCs. As shown in Fig.5.4, the event rate and the transfer rate are saturated when the event size is 800 or 1600 Bytes. The saturated transfer rate is 100 (120) MB/s with 800 (1600) Bytes. As shown in Fig.5.4(a), when the event size is less than 3200 Bytes, the event rate reaches the maximum trigger rate of 30 kHz. On the other hand, when the event size is 6400 Bytes, which corresponds to the event size of the drift chamber with 20



(a) Event rate

(b) Transfer rate

Figure 5.4: The event rate (a) and the transfer rate (b) of the sender PC as a function of the number of the receiver PC when the number of the sender PCs is five. (a): The green and red horizontal lines show the 30 kHz and 10 kHz trigger rate, respectively.

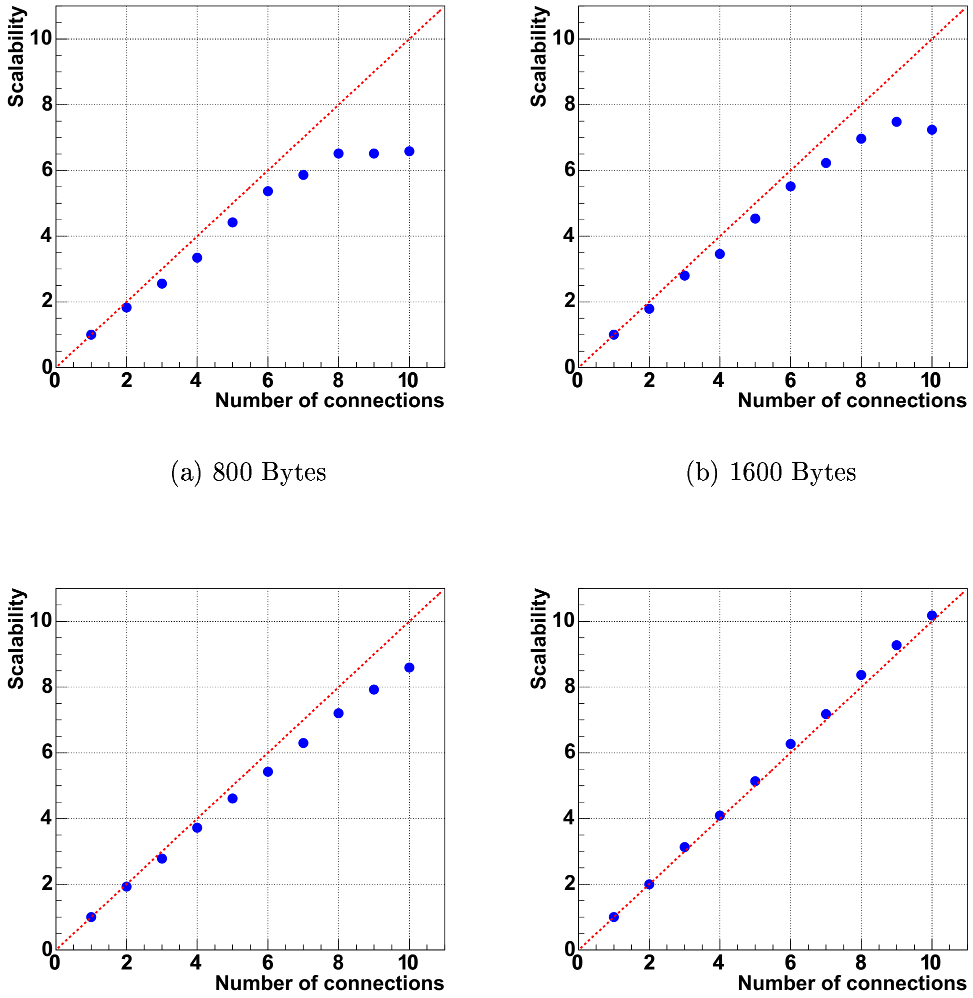
% occupancy, the event rate does not reach 30 kHz, which is the maximum trigger rate.

5.4 Scalability

In order to study the scalability of Stage2, we calculate scalability using the measured event rate, where we define the scalability as

$$Scalability = \frac{EventRate(\#ofReceiverPC = 1 - 10)}{EventRate(\#ofReceiverPC = 1)}. \tag{5.1}$$

Ideally the scalability is on a line whose inclination of the straight line is one. If the scalability is saturated, the tolerable event rate is not expected to increase by increasing the number of the event building units. Figure 5.5 shows the scalability as a function of the number of the receiver PCs. We vary the event size from 800 to 6400 Bytes and the number of the sender PCs is five. As shown in Fig.5.5(a) and 5.5(b), the scalability is saturated around 6.5 and 7.5 with the event size of 800 Bytes and 1600 Bytes, respectively. As shown in Fig.5.5(c), the scalability is below the straight line and is about 8.5 if the number of the receiver PCs is ten with the event size of 3200 Bytes. When the number of the receiver PCs is ten with the event size of 6400 Bytes, the scalability is on the straight line and is about 10 as shown in Fig.5.5(d). At all event sizes, all network throughput are not saturated as shown in Fig.5.3.



(a) 800 Bytes

(b) 1600 Bytes

(c) 3200 Bytes

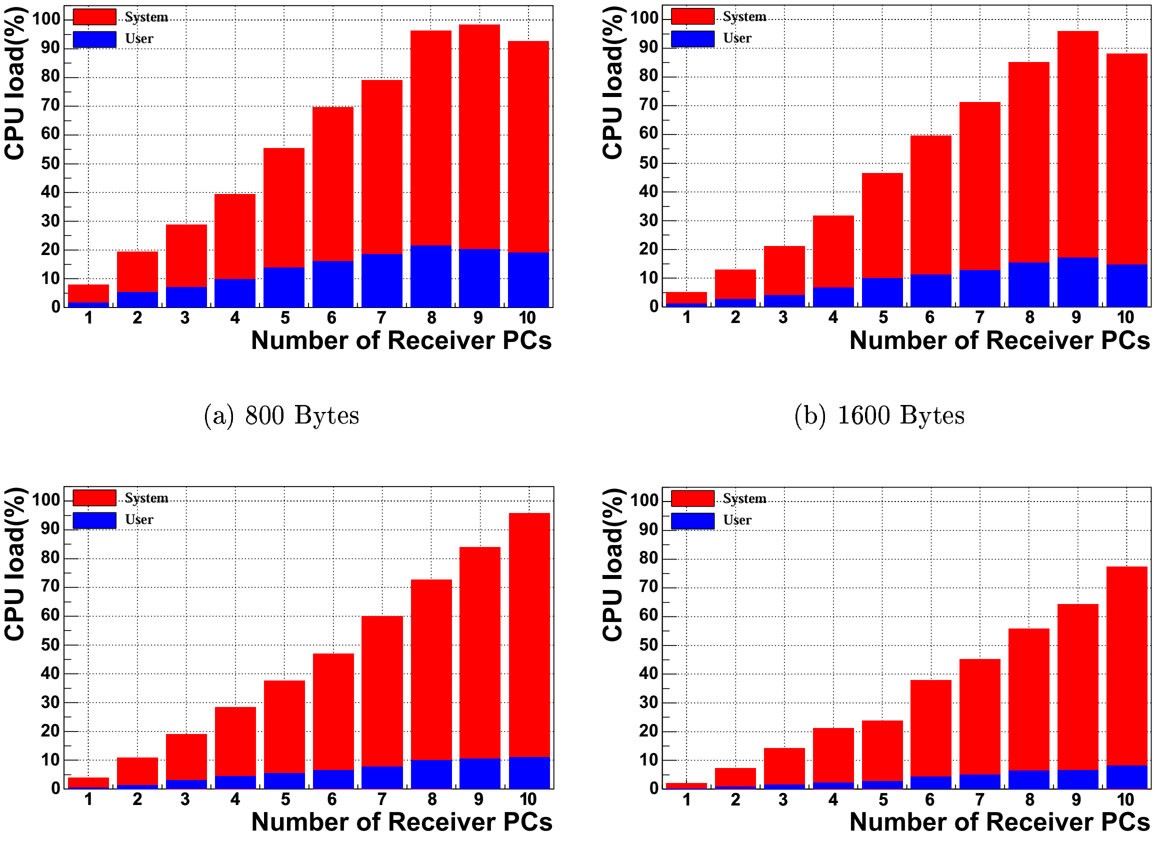
(d) 6400 Bytes

Figure 5.5: The scalability as a function of the number of receiver PC when the number of the sender PC is five. The red dotted line is a line whose inclination of the straight line is one.

Search for Saturated Reason

To investigate why the scalability is saturated, we measure the CPU load of the sender PC varying the number of the receiver PCs from one to ten. We monitor the CPU which the data sending process in the sender PC works. Figure 5.6 shows the measured CPU load varying the event size from 800 to 6400 Bytes when the number of the sender PCs is five. As shown in Fig.5.6, the CPU load of the sender PC increase by increase of the number of the receiver PCs. When the event size is 800 or 1600 Bytes, the CPU load is about 100% at ten or nine connections. The CPU load contains system load of about 80 % and user load of about 20 %. The system load is larger than the user load. The high system load indicates that the CPU of the sender PC is busy operating the NIC driver of the maximum of 10 ports and TCP/IP networking in a high rate as shown in Fig.5.4(a).

As shown in Fig.5.6(c), we find that the event rate of event size of 3200 Bytes reaches close to the limit since the CPU load reaches close to the limitation of 100 %. If the number of the receiver PCs is assumed to be more than eleven, the event rate will be saturated up to about 50 kHz, which is above the typical trigger rate.



(a) 800 Bytes

(b) 1600 Bytes

(c) 3200 Bytes

(d) 6400 Bytes

Figure 5.6: The CPU load of the sender PC. The horizontal axis and the vertical one shows the number of the receiver PCs and the CPU load, respectively.

We find that the saturated reason of the scalability is the very high CPU load. But we also find that scalability increases steady until the CPU load of the sender PC becomes 100%.

Upper limits of Event Rate in This Configuration

We also measure the event rate of the sender PCs for the case of the number of the sender PCs is one. If the number of the sender PCs is five, the *event building process* have to read the received data from five shard memory buffers. If the number of the sender PCs is one, the event building is performed faster than that in the number of the sender PCs is five since the *event building process* read the received data from only one shard memory buffer. Thus, we can know the upper limits of the event rate of distribution network matrix. Figure 5.7 shows the schematic drawing of the setup, when the number of the sender PCs is one.

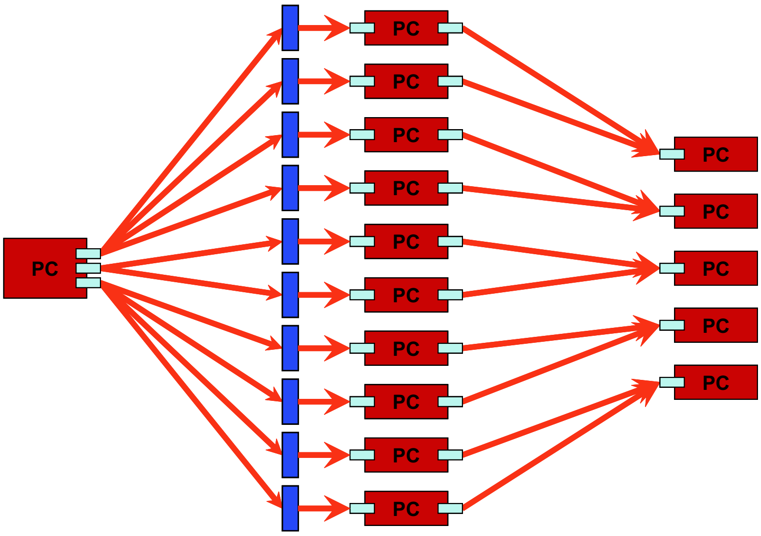
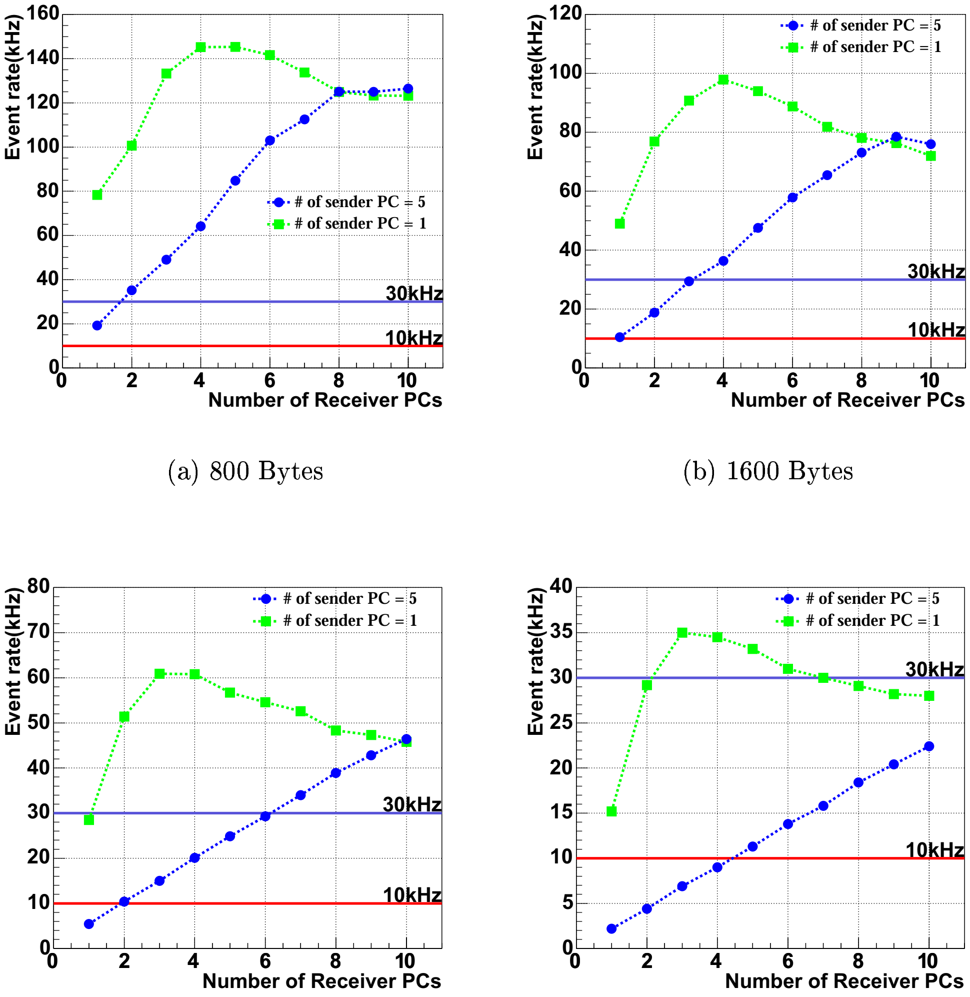


Figure 5.7: The setup of Stage2 and data flow when the number of the sender PCs is one.

Figure 5.8 shows the result of the measurement. We find that the event rate increases until the number of receiver PCs is four or three, and then decreases gradually up to ten receiver PCs or is saturated with all event sizes if the number of the sender PCs is one. At 800 and 1600 Bytes, we observe that the saturated event rates are the same point with both of configurations which the number of the sender is one or five. This indicates that the event rate with one sender PC determine the upper limit. At the event size of 6400 Bytes, the event rate of five sender PCs increase up to the event rate of one sender PC as shown in Fig.5.8(d). We observe that the event rates already reaches the upper limits when the event sizes are 800 Bytes, 1600 Bytes or 3200 Bytes.

We also measure the CPU load of the sender PC with the event size of 3200 Bytes. Figure 5.9 shows the CPU load of the sender PC varying the number of the receiver PCs from one to ten. When the number of the receiver PCs is four, the CPU load already becomes about 100%.



(a) 800 Bytes

(b) 1600 Bytes

(c) 3200 Bytes

(d) 6400 Bytes

Figure 5.8: The event rate as a function of the number of receiver PCs. The green line and the blues line show when the number of sender PC is one and five, respectively. The blue and red horizontal lines show 30 kHz and 10 kHz trigger rate, respectively.

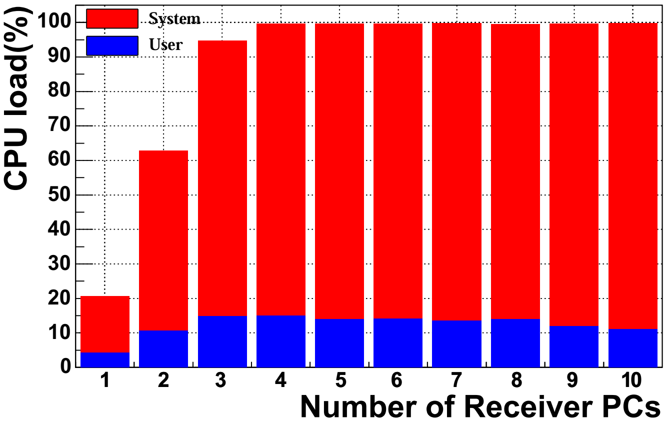


Figure 5.9: The CPU load of the sender PC with the event size of 3200 Bytes. The number of the sender PC is one.

5.5 Comparison with Another Configuration

To study another configuration using one network switch, we set up the test bench as shown in Fig.5.10. Each of the sender PCs and each of the receiver PCs are connected with 1000Base-T via one 24port gigabit network switch. Transfer mode of the switch is the store&forward mode. The packet buffer memory of the switch is 1 MB.

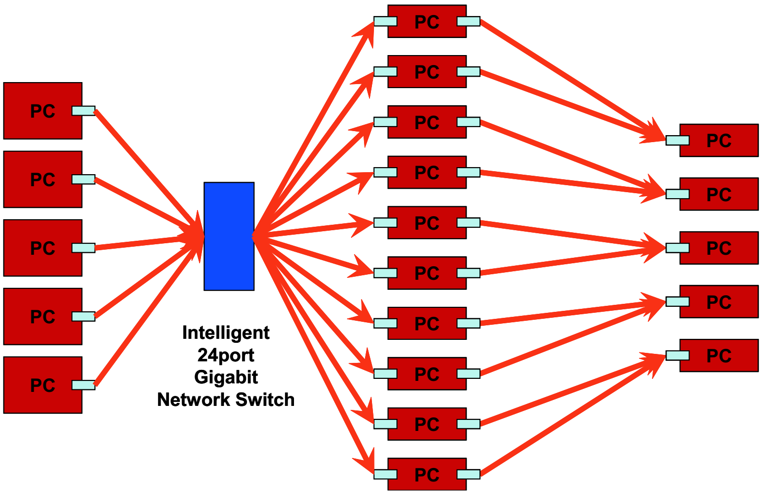
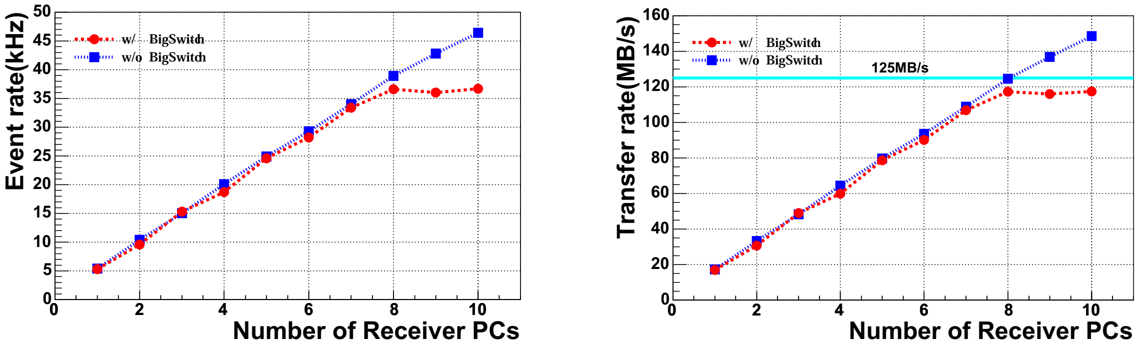


Figure 5.10: The setup of Stage2 and data flow using one network switch

The measurement of the event rate and the transfer rate are carried out at the event size of 3200 Bytes varying the number of the receiver PCs from one to ten. The result of the measurement is shown in Fig.5.11. If the number of the receiver PCs is less than seven, the result is similar to that in the previous configuration. If the number of the receiver PCs is over eight, the network between the sender PCs and the network switch is saturated as shown in Fig.5.11(b). The event rate is higher than the maximum trigger rate of 30 kHz at six receiver PCs.



(a) Event rate

(b) Transfer rate

Figure 5.11: The event rate and the transfer rate of sender as function of event size.

We measure the CPU load of the sender PCs with this configuration. Figure 5.12 shows the CPU load of the sender PC with one network switch. If we compare Fig.5.6(c) with Fig.5.12, the system load of the configuration with one network switch is smaller than that of the configuration with ten network switches (see Fig.5.1). Reason is as follows. The number of network port for each sender PC is only one in the configuration with one network switch. On the other hand, the number of ports of the sender PC is ten in the configuration using ten network switches. It is easier for the CPU of the sender PC to handle one port than handing ten ports. But the throughput of network between the sender PCs and the network switch is limited less than 125 MB/s. In addition, it is possible to occur the network congestion in the network switch as the number of the sender PCs and the receiver PCs increase.

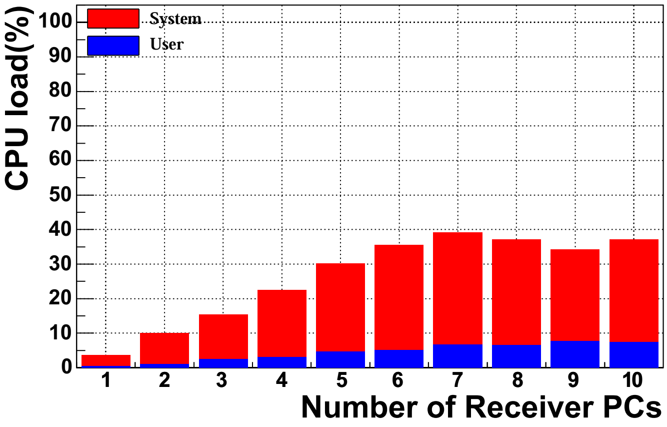


Figure 5.12: CPU load of the sender PC using one network switch

5.6 Modification of Distribution Network Matrix

In Section 5.3 - 5.5, we study two kinds of configuration;

1. Each receiver PC has a network switch. The total number of the network switches is ten. It is necessary for the sender PC to have ten network ports.

2. All receiver PCs have a network switch. The connection between the sender PCs and the receiver PCs is connected via the one network switch. The number of ports of the sender PC is one.

The first configuration has the problem that the upper limits of the event rate decrease since the CPU of the sender PC is hard to handle ten ports under the high trigger rate. The second configuration has the problem that the transfer rate between the sender PC and the receiver PC is to be limited 125 MB/s by the 1000 Base-T connection.

To avoid these problems, we modify the distribution network matrix as shown in Fig.5.13. Three or four receiver PCs share one gigabit network switch to reduce the number of ports of the sender PC. In this configuration, each sender PC has three ports. Thus the transfer rate between the sender PC and the receiver PC does not have the limitation of 125 MB/s. When the event size is 3200 Bytes and 6400 Bytes, we measure the event rate and the load of CPU varying the number of the receiver PCs with this configuration.

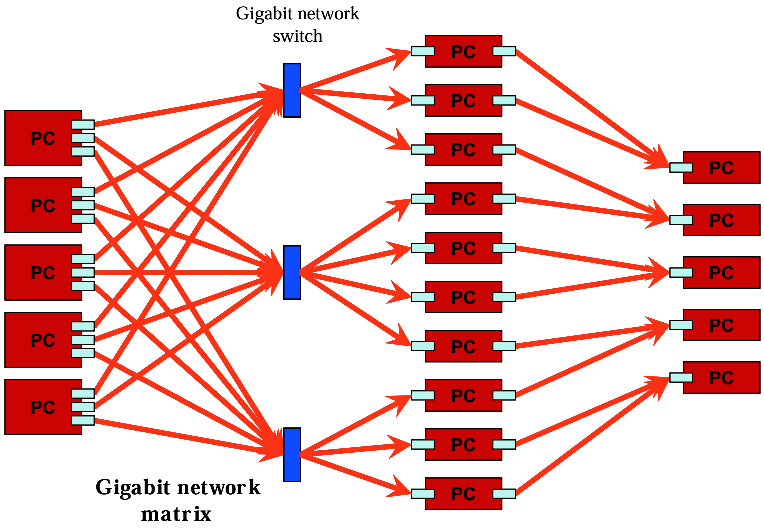
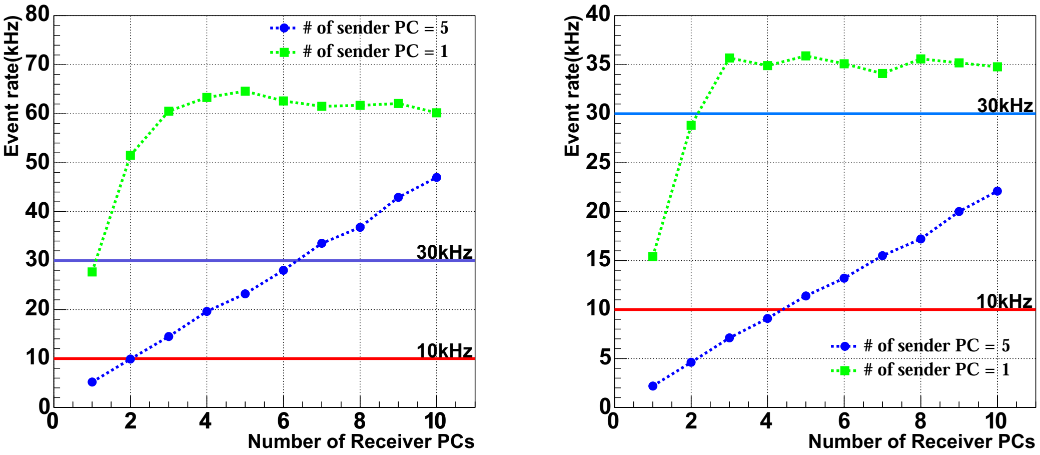


Figure 5.13: The setup of Stage2 and data flow modified the distribution network matrix

Figure 5.14 shows the measured event rate. If we compare Fig.5.8 with Fig.5.14, we find that the event rate of one sender PC does not decrease. When the event size is 6400 Bytes and the number of the receiver PCs is ten, the event rate of one sender PC is over 30 kHz, which is the maximum trigger rate. Since the event rate does not decrease with the event size of 6400 Bytes, we can expect to reach 30 kHz by increasing the number of the receiver PCs when the number of the sender PCs is five. The transfer rate of five sender PCs reaches 150.3 (141.3) MB/s with the event size of 3200 (6400) Bytes.



(a) The event size of 3200 Bytes

(b) The event size of 6400 Bytes

Figure 5.14: The event rate of the sender PC as a function of the number of the receiver PCs with the improved distribution network matrix to be compared with Fig.5.8. The red and blue horizontal lines are 30 kHz and 10 kHz trigger rate.

Figure 5.15 shows the CPU load of the sender PC. As shown in Fig.5.15, we observer that the CPU load is smaller than the configuration using ten ports since the number of ports of the sender PC reduce (see Fig.5.6(c)).

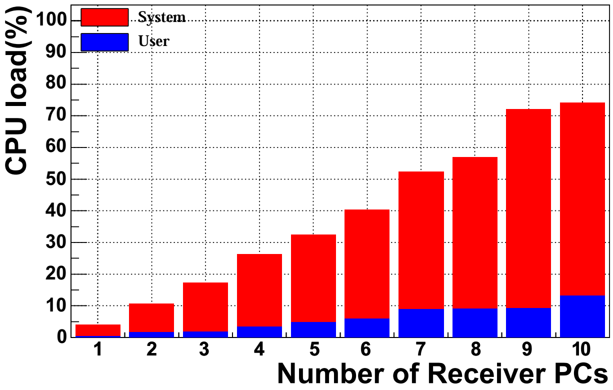


Figure 5.15: CPU load of sender PC using the modified network matrix

We also calculate the scalability of this configuration. At the event size of 3200 Bytes, the scalability is increased to be 9.0 from 8.5 since the CPU load becomes smaller than that of the first configuration.

For further reducing the CPU load, we consider that there are two solutions. The first solution is to increase Maximum Transfer Unit (MTU). The second one is using Layer-2

Frames to transmit data [19]. But, second solution needs to develop a new protocol to transmit data instead of TCP/IP technology. We think that first solution is effective for reducing the CPU load of the sender PC.

5.7 Summary of Stage2

We set up the 5 × 10 network matrix as the distribution network matrix for the data transmission. We study the performances of Stage2 by measuring the event rate, the transfer rate and scalability. With the typical event size of 3200 Bytes, the event rate and the transfer rate are 47 kHz and 150.3 MB/s, respectively, when the number of the sender PCs is five and that of the receiver PCs is ten. We observe that the distribution network matrix in Stage2 has scalability up to 9.0 which is satisfies the maximum trigger rate of 30 kHz by increasing the receiver PCs when the number of the sender PCs is five with the typical event size of 3200 Bytes. We also find that the bottleneck of the upper limits of the distribution network matrix is due to the high CPU load of the sender PC. By reducing the number of ports of the sender PC, we observe that the CPU load of the sender PC can be smaller than the previous configuration.

Chapter 6

Conclusions

6.1 Design Specification

In Chapter 4 and 5, we have investigated the performances of Stage1 (readout stage) and Stage2 (distribution stage). In this section, we consider the design specification of Stage1 and Stage2 based on the performance studies.

We assume the event building system with about 1000 readout modules (COPPER board) and 12 event building units, which consists of eight PCs of layer-1 servers. The schematic drawing of design specification is shown in Fig.6.1.

Stage1 (readout stage)

The 16 readout modules and one readout PC are connected through a gigabit network switch (16x1). The one switch has 16 input ports and one output port. If the transfer rate between the switch and the readout PC is over 125 MB/s of limits of 1000Base-T, we can easily add the output data flow. The outputs from 1000 COPPER boards are distributed to 64 readout PCs, each sending event fragments of a few hundred Bytes. In the performance study of Stage1 described in Section 4.2, the tolerable event rate reaches over 10 kHz when the number of the COPPER boards and the readout PCs is 16 and one, respectively. We expect that the maximum trigger rate operation of 30 kHz can be overcome by further improvement in software and hardware.

Stage2 (distribution stage)

The one readout PC has at least three ports for output. Each port is connected to one of gigabit network switches (8x4) in sets of three. One gigabit network switch has eight input ports and four output ports. The 64 readout PCs are concentrated into eight PCs of layer-1 PC servers of event building units. The one PC of layer-1 servers receive the data size of 8 x about 3 kB = 24 kB. In the performance study of Stage2 as shown in section 4.2, the upper limits of the event rate with the data size of 3200 Bytes exceeds 30 kHz. Thus the distribution network matrix is capable over 30 kHz in this configuration.

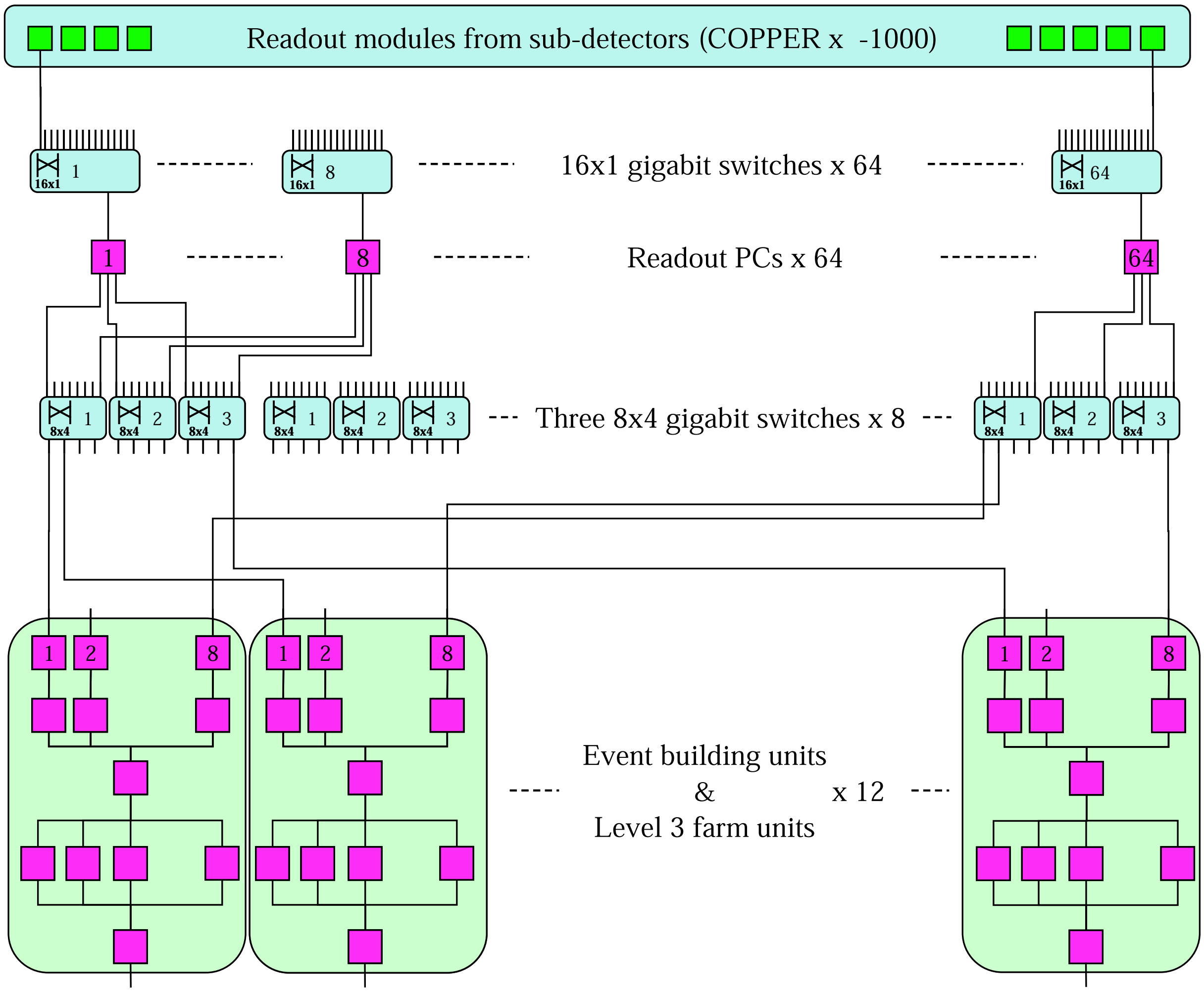

Readout modules from sub-detectors (COPPER x -1000)
16x1 1
16x1 8
16x1 gigabit switches x 64
16x1 64
1
8
Readout PCs x 64
64
8x4 1
8x4 2
8x4 3
Three 8x4 gigabit switches x 8
1
2
8
Event building units
&
Level 3 farm units
x 12

Figure 6.1: The design specification of the event building farm for SuperKEKB. The Stage1 (readout part) consists of about 1000 COPPER, 64 gigabit network switch (only number 1, 8 and 64 are shown) and 64 readout PC (only number 1, 8 and 64 are shown). The Stage2 (distribution part) consists of 8 gigabit network switch in sets of three, and 12 units.

6.2 Conclusions

We plan an upgrade of the KEKB accelerator, SuperKEKB, which will have 10 - 30 times larger luminosity than that of current KEKB. In SuperKEKB, the maximum level-1 trigger rate will be 10 - 30 kHz and the number of readout modules will be more than 1000. We need a new event building farm which meets the requirements above for the DAQ system of the SuperKEKB experiment.

We have designed the event building farm for SuperKEKB based on the current Belle's switchless event building farm. The new event building farm is designed to builds up events in multi-stages to manage the large number of readout modules. The multi-stage consists of three stages; the readout stage (Stage1), the distribution stage (Stage2) and the full event building stage (Stage3). We employ the unit structure in Stage3 to keep up with the luminosity increase by installing additional units. To determine the design specification and validate our design, we set up the proto-type event building farm and study the performance of Stage1 and Stage2.

We measure the event rate of Stage1. When the number of readout modules is eight, the event rate and the transfer rate is measured to achieve 33.7 kHz which satisfy the requirement of 30 kHz, with the typical event size of 160 Bytes. When the number of readout modules is 20, the even rate is measured 8.8 kHz. We observe that the bottleneck of data transmission of Stage1 is in the large number of processes in the readout PC. By modifying the event building software, the event rate are improved by about 50 % and achieved 13.4 kHz.

We measure the event rate, the transfer rate and the scalability of Stage2. We set up the 5 × 10 distribution network matrix. With the typical event size of 3200 Bytes, the event rate and the transfer rate per one readout PC are measured to be 46.4 kHz and 148.5 MB/s, respectively, which satisfy the requirements of 30 kHz and 100 MB/s with the event size of 3200 Bytes. We prove that Stage2 has the linear scalability on the number of the readout PCs with the typical event size. We also observe that the bottleneck of the saturation of the scalability is in the CPU load of the readout PC. To avoid the bottleneck, we prove that the reduction of the number of ports of the readout PC is effective.

We expect that Stage3 should not be a problem. We consider the design specification based on the performance studies of Stage1 and Stage2, and confirm that the designed event building farm system satisfies the requirements towards the expected hight luminosity of SuperKEKB. We conclude that the new DAQ system discussed in this thesis is the strong solution for the high luminosity experiment at SuperKEKB.

Acknowledgment

First of all, I would like to express my thank to my supervisor Prof. H. Aihara (Univ. of Tokyo) for giving me a chance to work on research in Belle experiment. I would like to acknowledge Dr. T. Higuchi of KEK, who encouraged me and discussed with me and answered my stupid questions.

I am deeply thankful to M. Iwasaki, H. Kakuno, T. Abe for their advise. I would like to express my appreciations to R.Itoh (KEK), M. Nakao (KEK), S. Y. Suzuki (KEK), Y. Nagasaka (HIT) and all members of Belle DAQ group for their help. I would like to thank all member of the Belle Collaboration.

I also thank all the member in our lab, A. Kusaka, K. Tanabe, R. Ishida, Y. Nakayama, Y. Nakahama, T. Tomura, T. Nakadaira and H. Kawai for their advise, discussions and joyful drinking.

Appendix A

Cabibbo-Kobayashi-Maskawa Matrix

In 1973, M. Kobayashi and T. Maskawa proposed a theory of quark mixing, which can introduce *CP* asymmetry within the framework of the Standard Model [4]. They demonstrated that quark mixing matrix with a measurable complex phase introduces *CP* violation into interactions. In the Standard Model, the quark-W boson interaction part of the Lagrangian is written as

$$L_{qW} = \frac{g}{\sqrt{2}}\{\overline{u}_L \gamma_\mu W_\mu^+ V d_L + h.c.\} \tag{A.1}$$

where g is the weak coupling constant, $u_L(d_L)$ represents the left-handed component of u-type (d-type) quarks, and V is a quark-mixing matrix. If all the elements of the quark mixing matrix V are real, the amplitude for a certain interaction and that for the CP conjugate interaction are the same. In order to violate *CP*, V should have at least one complex phase as its parameter.

In general, N dimensional unitary matrix has N^2 parameters with $N(N-1)/2$ real rotation angles and $N(N+1)/2$ phases. Since we can redefine phases of quark fields except one relative phase, $(2N-1)$ phases are absorbed and $(N-1)^2$ physical parameters are left. Among them, $N(N-1)/2$ are real angles and $(N-1)(N-2)/2$ are phases. The presence of phases means some of the elements must be complex and this leads to *CP* violating transitions. For the case of $N = 2$, two quark-lepton generations, there is 1 rotation angle (the Cabibbo angle) and no phase. This means *CP* must be conserved in the model with four quarks. For three generations, $N = 3$, there are three rotation angles and one phase so that *CP* can be violated. The quark mixing matrix for six-quark model can be written as

$$V \equiv \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{A.2}$$

$$\simeq \begin{pmatrix} 1-\frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} \tag{A.3}$$

Eq A.2 is called cabibbo-kobayashi-maskawa (CKM) matrix. The parameterization of CKM, eq A.3, is suggested by Wolfenstein [5], has four parameters λ, A, ρ and η. Experimentally, the parameters A and λ can be determined from tree-level decays and are rather well known [6]:

$$A = 0.84 \pm 0.04, \qquad \lambda = 0.2196 \pm 0.0023 \tag{A.4}$$

while ρ and η are not determined precisely, since their determination requires the measurement of V_{ub} and V_{td} which are of λ^3.

The unitary of CKM matrix leads to some constraints on its elements. For example, the B meson system is related to the following equation:

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0 \tag{A.5}$$

This equation gives a triangle in the complex plane as shown in Fig. A.1. The ϕ_1, ϕ_2, and ϕ_3 indicate the interior angles of the triangle.

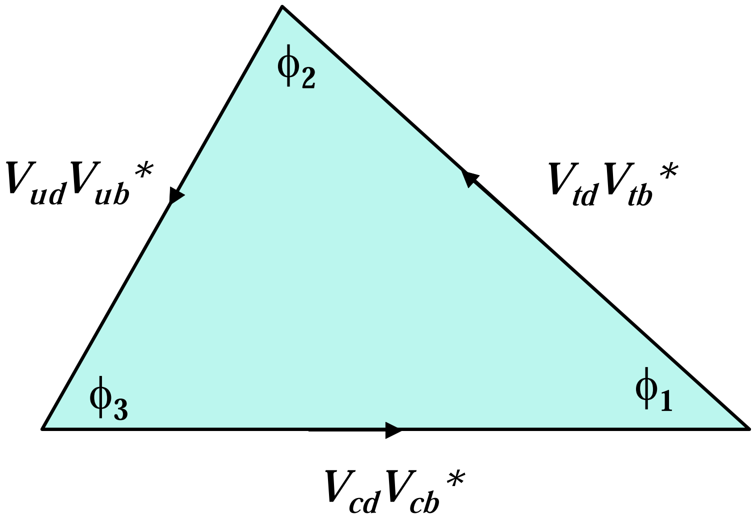


Figure A.1: The unitary triangle of CKM matrix in the B meson system

The triangle related to the B meson system is called *"Unitarity Triangle"*. Measuring the elements in the CKM matrix is equivalent to determine the three sides and the three angles of the Unitarity Triangle.

Appendix B

FINESSE and TTD

B.1 FINESSE

The digitizer modules are called FINESSE modules and various types of the FINESSE modules can be implemented according to the requirement of each detector. Figure B.1 shows the schematic drawing of the FINESSE module. The FINESSE modules are equipped with L1 pipeline FIFO so as to record the digitized signal without readout dead time. The FINESSE modules receives detector signals and digitizes them with an accuracy of up to 32 bits at the timing of system clock. The digitized signals are continuously stored in the L1 pipeline FIFO. The pipeline consists of four groups of FIFO buffers, which work as a ring buffer. Once the FINESSE modules receive an L1 trigger signal from the trigger timing receiver (TT-RX), the write pointer to a FIFO buffer is switched to the next buffer to freeze the contents of the digitized data in the buffer. The data in the pipeline buffer are then transfered to a readout FIFO on the COPPER board. The data in the readout FIFO are read out by direct memory access technique via a local-to-PCI bus bridge.

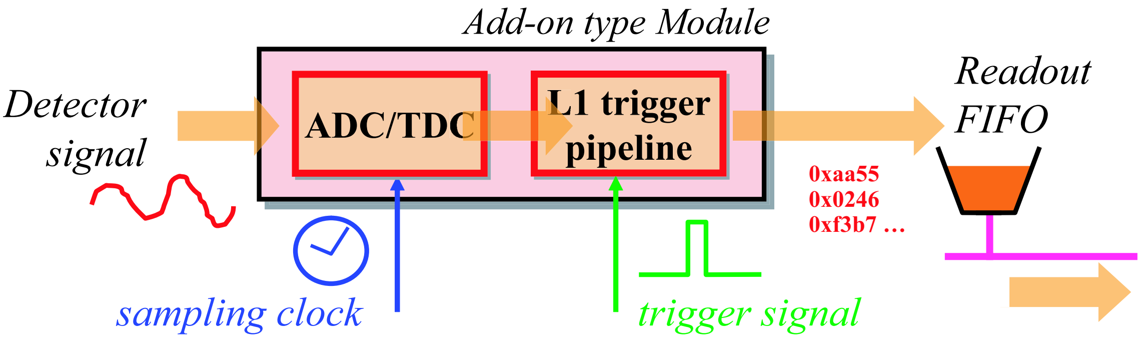


Figure B.1: The schematic drawing of FINESSE

B.2 TTD

The L1 trigger signal is distributed to every FINESSE module on the COPPER board and the *busy* response from each of them is collected by the *Trigger Timing Distribution*

(TTD) system, to synchronize the readout timing and order of the event fragment, which are sent to event building farm. Figure B.2 shows the overview of TTD system. In order to coordinate more than a thousand COPPER, we design a four-step cascaded distribution system using one-to-eight Trigger Timing Switch (TT-SW) modules. The trigger signal is received by a master Trigger Timing Input/Output (TT-IO) module, and eventually distributed to the Trigger Timing Receiver (TT-RX). The trigger signal is provided as an LVDS (Low Voltage Differential Signaling) signal; other signals, including the event tag and abort flag to downstream and the busy response and error flags to upstream, are encoded in a 10-bit serial-bus line over enhanced category-5 (CAT5e) shielded twist pair (STP) cable. The serial-bus is synchronized with a system clock. We define the format of the four pairs of the STP cable as shown in Fig. B.3. The TT-SW and TT-IO modules are VME 6U modules. The TT-RX module is a PMC daughter card to be attached onto the COPPER board.

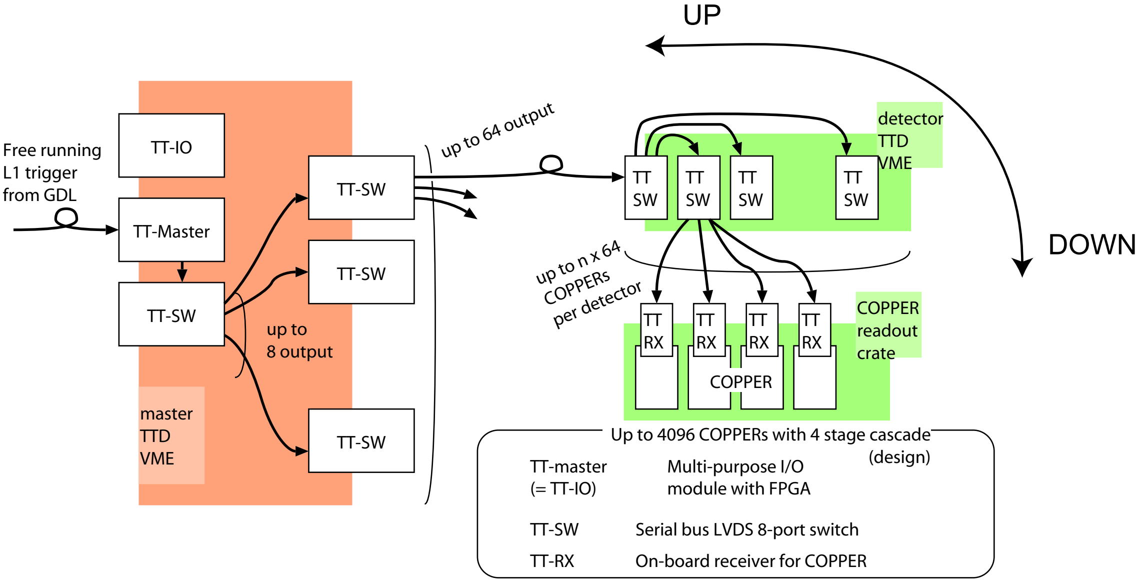
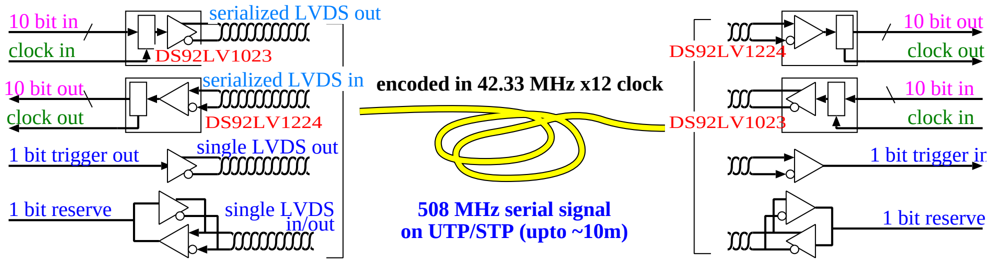


Figure B.2: Overview of the TTD system.



pair 1: upstream → downstream
- 4 bit event tag
- 2 bit event type
- 1 bit L1.5 abort
- 1 bit revolution signal
- 2 bit for error detection
- Self-decoded system clock (42.33 MHz)

pair 2: downstream → upstream
- 1 bit readout busy
- 1 bit user-I/O error
- 1 bit sync error
- 3 bit busy/error source
- 2 bit user defined signal
- 2 bit for error detection

pair 3: upstream → downstream 1 bit dedicated LVDS for the trigger timing
pair 4: upstream ↔ downstream 1 bit bidirectional reserved LVDS line

Figure B.3: Format of the STP cable with two sets of serial-bus and two LVDS lines.

Bibliography

[1] K. Abe *et al.* [BELLE Collaboration], Phys. Rev. Lett. **87**, 091802 (2001) [arXiv:hep-ex/0107061].

[2] B. Aubert *et al.* [BABAR Collaboration], Phys. Rev. Lett. **87**, 091801 (2001) [arXiv:hep-ex/0107013].

[3] J. H. Christenson, Cronin, Fitch and Turlay, Phys. Rev. **13** 138 (1964).

[4] M. Kobayashi and T. Maskawa, *Prog. Theor. Phys.* **49** 652 (1973).

[5] L. Wolfenstein, Phys. Rev. Lett. **105** 1945 (1983).

[6] K. Hagiwara *et al* (Particle Data Group), Phys. Rev. D **66** 010001 (2002), Available: http://pdg.lbl.gov/.

[7] Belle Collaboration, "Belle Technical Design Report," KEK Report 95-1 1995.

[8] KEKB accelerator group, KEK B-Factory Design Report, KEK Report 95-7 (1995).

[9] Belle SVD group, Technical Design Report of Belle SVD2

[10] K. Abe *et al.* (editted by S. Hashimoto, M. Hazumi, J. Haba, J. W. Flanagan and Y. Ohnishi), "Letter Of Intent for KEK Super B Factory", hep-ex/0406071, Available: http://belle.kek.jp/superb/loi/.

[11] S. Y. Suzuki, "Development of the Belle DAQ system," KEK Report 2001-10 (2001).

[12] M. Nakao, *Nucl. Instr. Meth.,* vol. **A379**,pp. 539-541, 1996.

[13] M. Nakao *et al.*, *IEEE Trans. Nucl. Sci.*, vol. **47**, pp. 56–60, 2000.

[14] H. Ishino *et al.*, "Data Acquisition System of the Belle Silicon Vertex Detector (SVD) Upgrade," in Proc. of conference (IEEE 2003).

[15] S.Y. Suzuki *et al.*, *Nucl. Instrum. Methods*, vol. **A494**, pp. 535–540, 2002;

[16] R. Itoh, "BASF - BELLE Analysis Framework," at Comptuing in High-energy Physics (CHEP97).

[17] T. Higuchi *et al*, hep-ex/0305088(2003); Y. Igarashi *et al*, physics/0305137(2003).

[18] R.Itoh *et al*, “Experience with Real Time Event Reconstruction Farm for Belle Experiment,” in Proc. of conference (CHEP 2004).

[19] V.Brigljevic *et al*, “The CMS Event Builder,” in Proc. of conference (CHEP 2003).