Master's thesis
February 1, 2000

The study of the Network based Event Builder

Hiroshi Yumura

Department of Physics, Kyoto university

Abstract

On 2th June 1999, BELLE detector observed first Beam interactions. The aim of the BELLE experiment is to measure angles of the unitarity triangle and to see the Kobayashi-Masukawa theory which is the origin of CP violation.

The BELLE Event Builder completed on Junuary 1999, and works stablily in efforts of Yamagata-san and members of KEK online group. The role of Event Builder is to collect event records from sub systems and reconstructs full event records with 15MB/s data rate.

In this summer, KEK-B Accelerator acheived a luminosity $10^{32} cm^{-2} s^{-1}$, and we will run at $10^{34} cm^{-2} s^{-1}$ (bunch crossing rate: 500Hz) in the future. In order to endure such high rates, BELLE Event Builder adopts Barrel-Shifter type, but the performance is not enough yet. Now that Evnet Builder works stablily at $10^{32} cm^{-2} s^{-1}$, we should prepre for coming $10^{34} cm^{-2} s^{-1}$ beam.

In this paper, I report the states of Event Builder and my study about Network based type which is one of prominent future plans. To establish the way to construct the event builder using commercial network devices enables us to construct cheap, flexible and reliable DAQ system. Though there are many points to study, I can see the possibility of the network based Event Builder which can have enough performances.

Contents

List of Figures

List of Tables

Chapter 1

introducion

Now, the current universe is composed entirely of matter, but it is known that it has been evolved from the matter-antimatter Big-Bang through exceedingly small asymmetry between matter and anti-matter. We could see so-called "CP violation" effects only in nuetral K meson system, but the Standard Model predicted that we can observe larger CP violation in B meson system. The origin of CP violation in the Standard Model is delived from the Kobayashi-Masukawa theory, so we can check the Kobayashi-Masukawa theory by measuring B mesons and determing the value of Kobayashi-Masukawa Matrix.

The Standard Model predicted that CP violation in B meson system is larger thanK meson, but the typical branching ratio of the decay which is usable for the obsevation of CP violation is rather small($10^{-4} \sim 10^{-6}$). In order to generate B mesons in large quantities, an asymmetric e^+e^- collider KEKB was planed, which will run with full bunch (bunch crossing rate: 500MHz) and accomplish the luminosity $\sim 10^{34} cm^{-2}s^{-1}$ and generate 10^8 B mesons a year.

To catch B mesons effectively in the luminosity $10^{34} cm^{-2}s^{-1}$ (1000 times higher than Tristan), we need the excellent triger system and the powerful DAQ system. Since the triger rate and data size are expected as $\sim 500Hz$ (physics event $\sim 100Hz$) and 30KByte/event respectively, the required performances for DAQ system is almost 15MByte/s.

In BELLE experiment, the Event Builder, which constructs full event records, consists of multiple event builder nodes (EB-N) and operates them in parallel, because 15MByte/s data rate exceeds the processing speed available in single node. The processing speed in one EB-N is a few MByte/s, so the 15MByte/s bandwidth is acheived with more than 6 EB-N. 12 sub detector node (IB-N) and EB-N are connected with barrel sifter type switching network and its control module.

When BELLE Event Builder was designed, the switching network which satisfied the requirement of BELLE DAQ system was only barrel sifter types which was developed at Fermilab. Nowadays, the wider computing network spread, the charper and the more powerful Ethernet devices become. It was known that it is possible that we can construct the event builder using Ethernet in principle which is useful, cheap and available. Since there are still many problems to construct the event builder by ethernet, we need to begin studies.

The BELLE experiment began on June 1999, and KEKB acheived the luminosity $10^{32} cm^{-2}s^{-1}$ this summer and aim $10^{33} cm^{-2}s^{-1}$ in autumn 1999. Since the enviroment of DAQ system is expected to be harder and header, we have to check the performance and problems of Event Builder again before the liminorsity is achieved $10^{34} cm^{-2} s^{-1}$.

In this papar, the present status of Event builder and its problems and studies of the ethernet based event builder are reported. Also, the possiblity of the ethernet based event builder is described.

Chapter 2

BELLE experiment overview

The aim of the BELLE experiment is to measure CP violation in B meson system. In this chapter, the process of the B meson causing CP violation and how the BELLE detector will catch it are shown.

2.1 Scientific Motivation

In Standard Model, quark mixing in weak interaction is shown by Kobayashi-Masukawa Model and CP violation is the consquence of a imaginary part of KM matrix.

2.1.1 KM matrix

The relation of quark mass eigenstate and weak eigenstate is shown by unitary matrix, and it has a imaginary phase and called KM matrix when the quark consists of 3 generations.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{2.1}$$

$$V = \begin{pmatrix} V^{ud} & V^{us} & V^{ub} \\ V^{cd} & V^{cs} & V^{cb} \\ V^{td} & V^{ts} & V^{tb} \end{pmatrix} \tag{2.2}$$

$$= \begin{pmatrix} c_{12} & s_{12}c_{13} & s_{13}e^{-\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{\delta_{13}} & -c_{12}c_{23} - s_{12}c_{23}s_{13}e^{\delta_{13}} & c_{23}c_{13} \end{pmatrix} \tag{2.3}$$

where $s_{ij} = \sin\theta_{ij}$, $c_{ij} = \cos\theta_{ij}$ and δ are free parameters of unitary matrix and i,j $\in$ 1,2,3 are generation labels.

Applying the approximate parameterizaion in (2.3), $s_{12} = \lambda \simeq 0.22$, $s_{23} = A\lambda^2$ and $s_{13}e^{-i\delta} = A\lambda^3(\rho - i\eta)$, this is the Wolfenstein Parameterization which is often used in B physiscs.

$$V \simeq \begin{pmatrix} 1-\lambda^2/2 & \lambda & \lambda^3 A(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & \lambda^2 A \\ \lambda^3 A(1-\rho-i\eta) & -\lambda^2 A & 1 \end{pmatrix} + O(\lambda^4) \tag{2.4}$$

2.1.2 Unitarity triangle

The relation of Unitarity of KM matrix shown as

$$V_{ij}^* V_{ik}^* \quad = \quad \delta_{jk} \tag{2.5}$$

can be represented as a triangle in imaginary plane.

In case of nuetral B meson($j = b, k = d$), this triangle is usable because each lengths of sides is same order. Using wolfenstein parameterization,

$$V_{tb}^* V_{td}^*, \quad V_{cb}^* V_{cd}^*, \quad V_{ub}^* V_{us}^* \quad \sim \quad \lambda^3 \tag{2.6}$$

the triangle on imaginary plane, called Unitarity triangle shown in Figure 2.1, has areas ,and three angles are defined as,

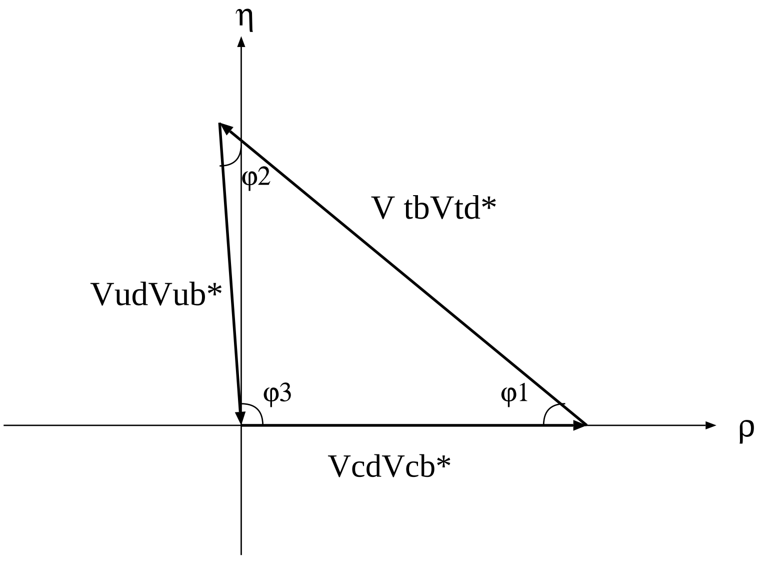


Figure 2.1: Unitarity triangle

$$\phi_1 \equiv arg\Big(\frac{V_{cb}^* V_{cd}}{V_{tb}^* V_{td}}\Big)$$

$$\phi_2 \equiv arg\Big(\frac{V_{ub}^* V_{ud}}{V_{tb}^* V_{td}}\Big)$$

$$\phi_1 \equiv arg\Big(\frac{V_{cb}^* V_{cd}}{V_{ub}^* V_{ud}}\Big)$$

In oreder to test the KM model as a origin of CP violation, it's very effective to measure the length of sides and angles of the Unitarity triangle. If angels are small and the triangle has no area, there would be a new physics beyond the Standard model.

2.2 CP violation in B Meson system

2.2.1 Direct CP Violation

CP symmetry is violated when there is a complex coupling constant. We can take a whole phase factor arbitrarily, but when we cannot make all of the coupling constants real, CP violation occurs.

Direct CP violation can be observerd as the difference of decay probability between one mode and its CP transfered mode only when there are more than two independent amplitude contribution to the probability. Consider the mode $X \to Y$ which has two independent amplitude.

$$\begin{aligned} Amp(X \to Y) &= A_1 e^{i\delta_1} e^{i\phi_1} + A_2 e^{i\delta_2} e^{i\phi_2} \\ &= A_1 e^{i\delta_1} e^{i\phi_1} \left[1 + \frac{A_2}{A_1} e^{i(\delta_2 - \delta_1)} e^{i(\phi_2 - \phi_1)} \right], \end{aligned} \tag{2.7}$$

where $\delta_{1,2}$ are the strong phase which are independent of weak phase $\phi_{1,2}$. The decay probability is proportional to square of the absolute value of the amplitude.

$$|Amp(X \to Y)|^2 = A_1^2 \left[1 + \frac{A_2^2}{A_1^2} + 2\frac{A_2}{A_1} \cos(\delta_2 - \delta_1 + \phi_2 - \phi_1) \right]$$

For the CP transformed mode, weak phase changes its sign and strong phase does not.

$$|Amp(X^{CP} \to Y^{CP})|^2 = A_1^2 \left[1 + \frac{A_2^2}{A_1^2} + 2\frac{A_2}{A_1} \cos(\delta_2 - \delta_1 - \phi_2 + \phi_1) \right].$$

So, the asymmetry arises,

$$Asym \propto |Amp(X \to Y)|^2 - |Amp(X^{CP} \to Y^{CP})|^2 = -4A_2 A_1 \sin(\delta_2 - \delta_1) \sin(\phi_2 - \phi_1).$$

Thus, to observe the violation, the mode has to have two independent amplitude, and both the weak phase and strong phase must be different between the two amplitude ($\delta_2 \neq \delta_1,\ \ \phi_2 \neq \phi_1$).

2.2.2 Indirect CP Violation

When B^0 and $\bar{B}^0$ mesons decay into the same CP eigenstate f, the CP violation aasymmtry A_f is given by,

$$A_f = \frac{R(B^0 \to f) - R(\bar{B^0} \to f)}{R(B^0 \to f) + R(\bar{B^0} \to f)} = sin(2\phi_{CP})\ sin(\Delta m \Delta t) \tag{2.8}$$

where Δm is the mass defference between two B^0's and Δt is the difference of proper time for B^0 and $\bar{B^0}$ decays. $R(B^0 \to f)$ is the decay rate of $B^0 \to f$.

The asymmetry is produced by the interference of two amplitudes; a direct decay $B^0 \to f$ and a decay through the $B\bar{B}$ 0scillation $B^0 \to \bar{B}^0 \to f$ which is called as the indirect CP violation. Experimentally, the measurement of Δt is essential for the observation of the asymmetry. We can measure Δt using the Lorentz boost of the $\Upsilon(4S)$ state produced by the asymmetry of beam energies as described later. The angle ϕ_{CP} is the phase difference between two amplitude, which is derectly related to the internal angle of the unitarity triangle.

In BELLE experiment, we try to measure angles of the unitarity triangle by following decays to CP eigenstate.

$$\phi_1 : B^0 \to J/\Psi K_s^0$$
$$\phi_2 : B^0 \to \pi^+\pi^-$$
$$\phi_3 : B^0 \to D^0 K$$

2.2.3 $B^0 \to J/\Psi K_s^0$

The decay mode $b \to ccs$ is used to measure the angle ϕ_1. The time dependent CP asymmetry is written as follows;

$$Asym[J/\Psi K_s^0] = sin(2\phi_1)\ sin(\Delta m \Delta t)$$

There're two large contrubutions for this decay; one is tree diagram (Figure 2.2) and the other is pengin diagram (Figure 2.3).

The contribution in weak phase from the penguin diagram is written as the sum of three terms, $(V_{ub}^* V_{us} + V_{cb}^* V_{cs} + V_{tb}^* V_{ts})$, which can be expressed as $\sim V_{cb}^* V_{cs}$, using the unitarity relation $V_{ib}^* V_{is} = 0$. So the weak phase penguin diagram is approximately same as $V_{cb}^* V_{cs}$, which is same with that of tree diagram.

There're two large contrubutions for this decay, but weak phases are almost same and direct CP violation hardly occures. Also, $J/\Psi \to l^+l^-$ and $K_s^0 \to \pi^+\pi^-$ is very clean. This is why the decay mode $B^0 \to J/\Psi K_s^0$ is called as "Gold Plate Mode".

2.2.4 $B^0 \to \pi^+\pi^-$

The decay mode $b \to u\bar{u}d$ is used to measure the angle ϕ_2. The time dependent CP asymmetry is, if the contribution is only from tree diagram (Figure 2.4),

$$Asym[\pi^+\pi^-] = sin(2\phi_2)\ sin(\Delta m \Delta t)$$

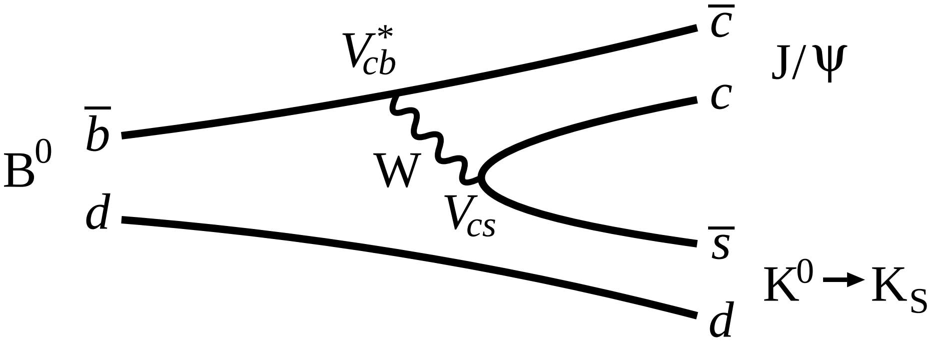


Figure 2.2: The tree diagram of $b \rightarrow ccs$

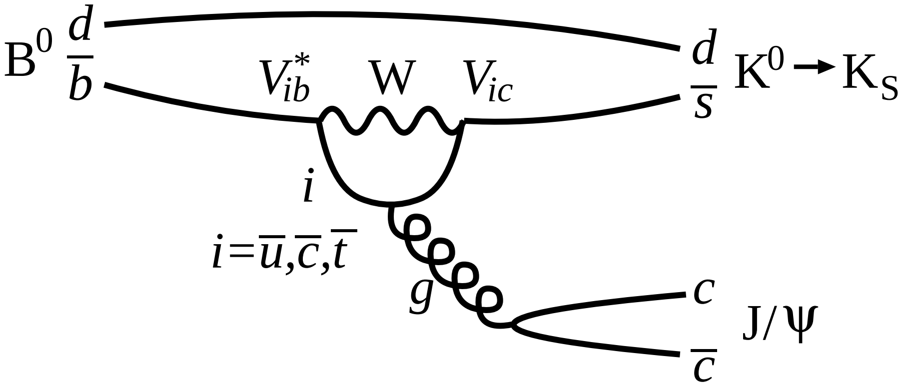


Figure 2.3: The penguin diagram of $b \rightarrow ccs$

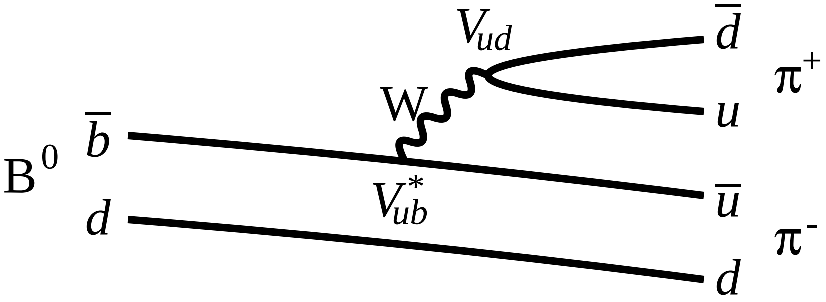


Figure 2.4: The penguin diagram of $b \to u\bar{u}s$

However, unlike $B^0 \to J/\Psi K_s^0$, the contribution from the penguin diagram (Figure 2.5) is not negligible, because the weak phase $V_{id}V_{ib}^*$ are different from that of tree diagram. So the direct CP violation may occur.

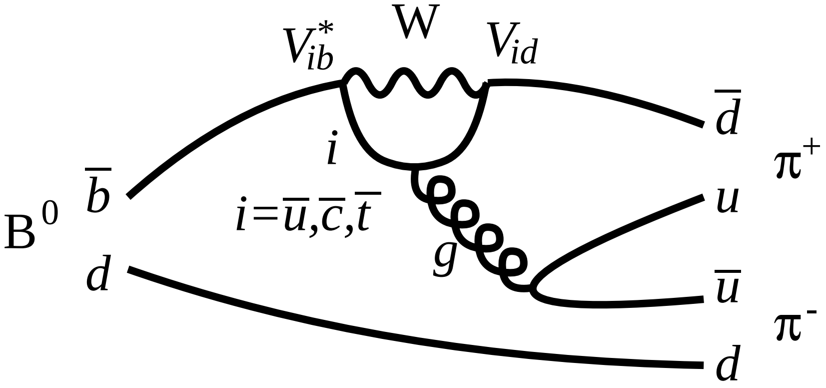


Figure 2.5: The penguin diagram of $b \to u\bar{u}s$

It is possible, in principle, to disentangle the effects of the tree and penguin contributions for the final states $\pi\pi$, using isospin relations. This method requires to measure the decay rate of $B^0 \to \pi^0\pi^0$, $B^0 \to \pi^+\pi^-$, $B^+ \to \pi^+\pi^0$.

2.3 BELLE experiment

The goal of BELLE is to measure an asymmetry in proper time distribution when one of $B\bar{B}$ pair decays into a CP eigenstate. So, it is important to know which B meson decayed into a aimed CP eigenstate. At first, when one of $B\bar{B}$ pair decays, we try to catch high momentum lepton or Kaon and thier charge in order to know which B meson decayed. It is called tagging. Secondly, we try to catch the decay of the other B meson. The proper time distribution is measured by the difference of decays of B meson pair (see Figure 2.6).

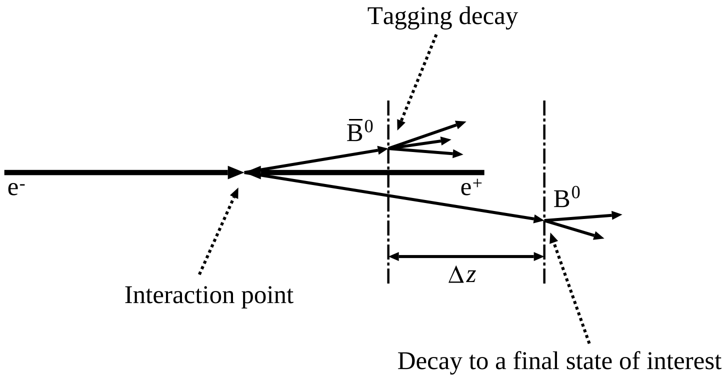


Figure 2.6: Tagging and proper time measurement

In order to acheive these requirements, the BELLE Detector (Figure 2.7) must have following characteristics.

- accurate vertex detection around interaction point;
- B meson reconstruction and tagging;
  - charged particle tracking and momentum measurement
  - good Particle Identification, especially charged π, K, lepton
  - γ ray measurment for π^0 reconstruction

Treating rare decays of B meson, each detector is required to have good resolutions and efficiencies.

2.3.1 KEKB Accerelator

The KEKB Accerelator (Figure 2.8) is an e^+e^- asymmetric collider at $8\times$ 3.5GeV (center-of-mass energy 10.58GeV), which generates $\Upsilon(4S)$ that decays into $B\bar{B}$ dominantly. The Lorentz boost factor($\beta\gamma$) is 0.45 which enables us more precise measurment of the time evolution of B mesons.

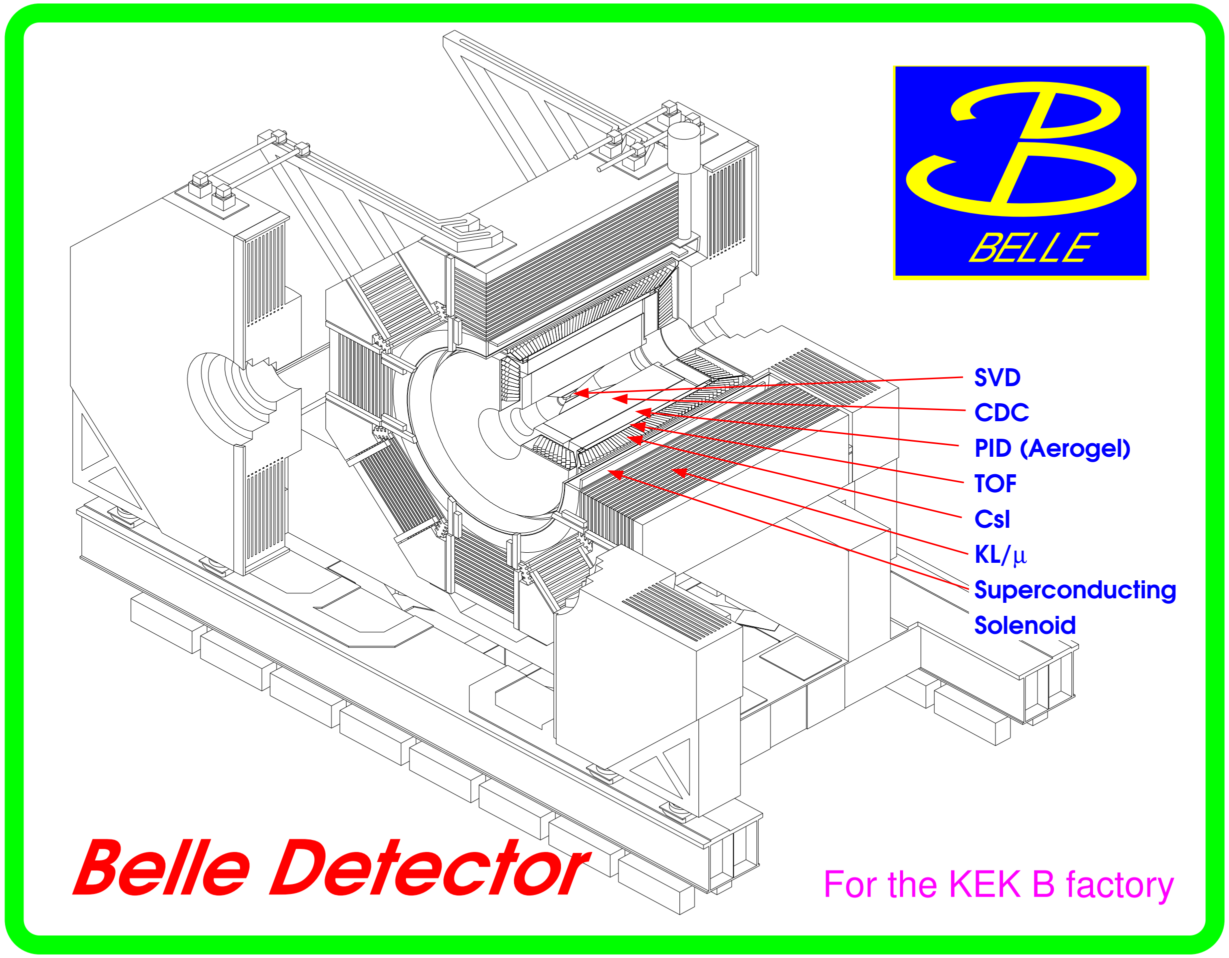


Figure 2.7: BELLE detector

To acheive the luminorsity $10^{34} cm^{-2} s^{-1}$, 5000 bunches will be accumulated in each ring whose circumstance is 3 Km. The beam crossing angle at the interaction region is $\pm 11 mr$, which allows us to avoid using the bending magnet to separate e^+e^- beam. The parameters of KEKB is summarized in Table 2.1

1.without wigglers 2.with wigglers

Ring	LER	HER	
Energy (E)	3.5	8.0	GeV
Circumference (C)	3016.26		m
Luminosity (L)	1×10^{34}		$\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
Crossing angle (θ_x)	± 11		mrad
Tune shifts (ξ_x/ξ_y)	0.039/0.052		
Beta function at IP (β_x^*/β_y^*)	0.33/0.01		m
Beam current (I)	2.6	1.1	A
Natural bunch length (σ_z)	0.4		cm
Energy spread (σ_E)	7.1×10^{-4}	6.7×10^{-4}	
Bunch spacing (s_B)	0.59		m
Particles/bunch	3.3×10^{10}	1.4×10^{10}	
Emittance ($\varepsilon_x/\varepsilon_y$)	$1.8 \times 10^{-8}/3.6 \times 10^{-10}$		m
Synchrotron tune (ν_s)	$0.01 \sim 0.02$		
Betatron tune (ν_x/ν_y)	45.52/45.08	46.52/46.08	
Momentum compaction factor (α_p)	$1 \times 10^{-4} \sim 2 \times 10^{-4}$		
Energy loss/turn (U_0)	$0.81^1/1.5^2$	3.5	MeV
RF voltage (V_c)	$5 \sim 10$	$10 \sim 20$	MV
RF frequency (f_{RF})	508.887		MHz
Harmonic number (h)	5120		
Longitudinal damping time (τ_ϵ)	$43^1/23^2$	23	ms
Total beam power (P_b)	$2.7^1/4.5^2$	4.0	MW
Radiation power (P_{SR})	$2.1^1/4.0^2$	3.8	MW
HOM power (P_{HOM})	0.57	0.15	MW
Bending radius (ρ)	16.3	104.5	m
Length of bending magnet (l_b)	0.915	5.86	m

Table 2.1: Parameters of KEKB

2.3.2 SVD(Sillicon Vertex Detector)

SVD is placed around the interaciton point to measure the decay vertices of the two B mesons. The measurment of proper time distributions means the measurment of the distance between the decay vertices along the beam direction. It is shown by

$$\Delta t \simeq \Delta z / c\beta\gamma = (z' - z)/c\beta\gamma \tag{2.9}$$

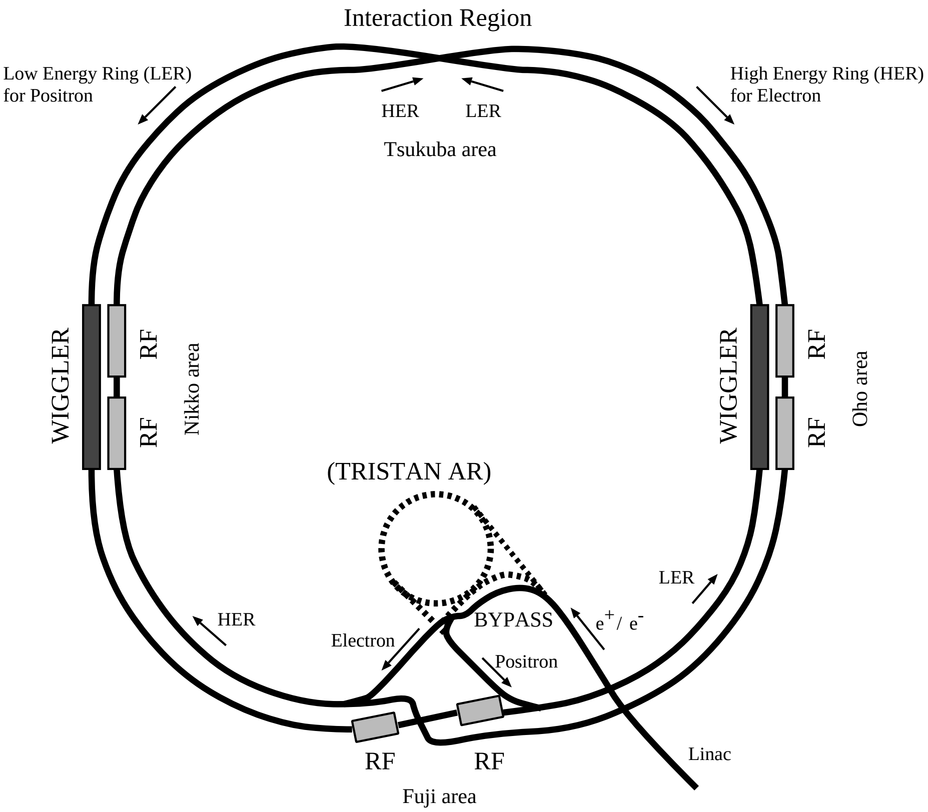
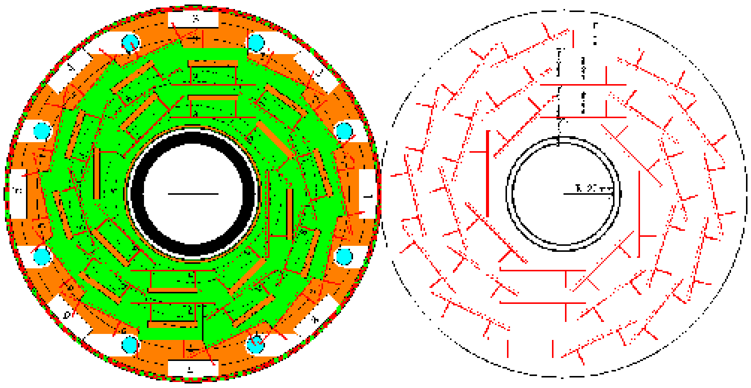


Figure 2.8: Configuration of the KEKB accelerator system

where $\beta\gamma$ is the Lorentz boost factor due to the asymmetry beam energies($\beta\gamma = 0.42$ at KEKB), Δt and Δz are the difference of time and distance of B mesons respectively. Considering the life time of B meson ($\tau = 10^{-12}s$), Δz is less than $1mm$. So SVD has 3 layer and will achieve less than $100\mu m$ in Δz resolution.



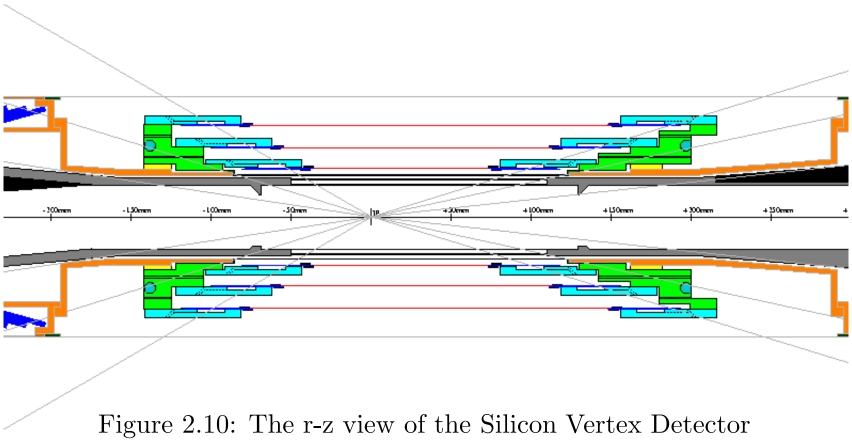


Figure 2.10: The r-z view of the Silicon Vertex Detector

2.3.3 CDC(Central Drift Chamber)

CDC measured the momentum of charged particles. CDC is covered with super conducting Solenoid which generates 1.5T in z direction. The energy loss of the charged particle through the gas of CDC, called dE/dx, is also measured and used to identify the kind of charged particle which momentum is less than 0.8 GeV/c.

CDC consisnts of three types layer, axial, stereo and cathode. axial wires are streched pallarel to z direction and stereo wires are streched to have 40∼70m rad from axial wires, so not only $r - \phi$ distance, but z direction tracking are available.

The transverse momentum distribution of charged particles from B decays peaks around 200 MeV/c, so to have good momentum resolution around 200 MeV/c is important to get good mass resolution when B meson is reconstructed.

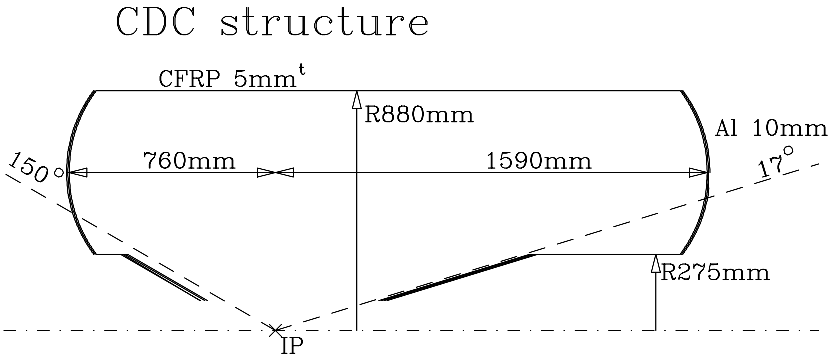


Figure 2.11: Structure of Central drift Chamber

2.3.4 ACC(Aerogel Cerenkov Counters)

A threshold Cerenkov counter system based on siloca aerogel radiator is used for K/π separation (1.2GeV/c ∼ 4GeV/c) in BELLE. The aerogel approach is a well establish technology, and The development of the technique for producing low-refractive-index aerogel with excellent hydrophobic propeties enable us to use up to 4Gev/c.

2.3.5 TOF/TSC

TOF (Time of Flight counters) uses plastic scintillation counters and measures the flight time. For 1.2m flight path, a system of counters with 100ns is effective for particles momenta below 1.2GeV/c, which encompasses 90in $\Upsilon(4S)$ decays.

TSC (Trigger Scintillation counters) provides a clean signal for trigger.

TOF/TSC system consists of 128 barrel TOF counters and 64 barrel TSC counters.

2.3.6 ECL(Electromagnetic Calorimetry)

The main purpose of the electromagnetic calorimeter is the detection of photons fron B meson decay with high efficiency and good resolution. Since the one third of end-products of B meson decay is π^0, detections of low energy gamma, especially below 500MeV/c, is important. Also good resolutions for

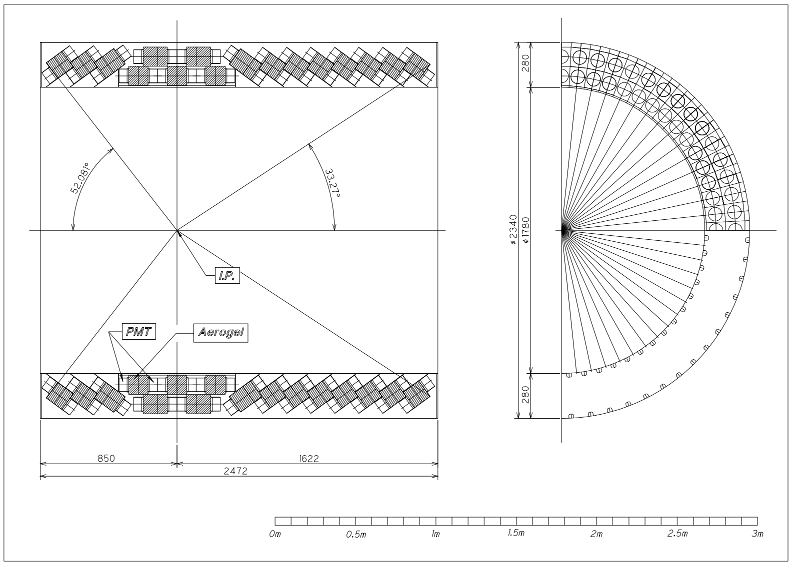


Figure 2.12: The Aerogel Cherenkov Counter

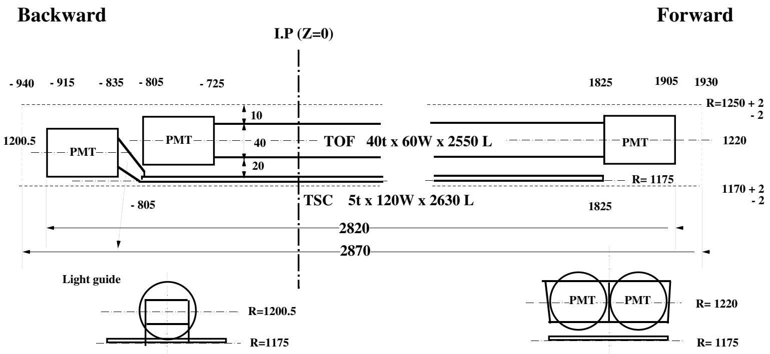


Figure 2.13: Time of Flight system

high energy gamma are needed, because important two body decay such as $BK^*\gamma$ and $B\pi^0\pi^0$ generate gamma which energy is more than 2GeV. The Electoron identification uses the shape of energy deposite in ECL, so ECL is important for hadron rejection, too.

2.3.7 KLM(K_L and Muon detector)

KLM detector is placed outside the superconducting solenoid and consists of an octagonal barrel and two endcaps that are composed of a sandwich of fourteen iron absorber plates and fifteen gaps.

The decaymode using μ is $B^0 \to J/\Psi$ and $J/\Psi \to \mu^+\mu^-$ is very important for measurment of ϕ_1, because it's easy to reconstruct. Also the cleanness of the Muon identification is important for the flavor tagging and semileptonic decays.

2.4 Trigger system

In BELLE experiment, we will run at quite high luminosity like $10^{34}cm^{-2}s^{-1}$, with almost continuous beam (bunch crossing rate :500MHz). In such a sitiation, trigger must owe a big role and resposibility. The requirement of the trigger system aree listed as follows:

1. High efficiency for all ecents fo physics interest; especially $\simeq$ 100% efficiency for $B\bar{B}$ events
2. Redundant, independent triggers for important physics events; this will help to monitor the trigger efficiency for those modes.
3. Powerful reduction of background trigger rate to a tolerable level.
4. Facility of trigger rate monitoring, calibration, self-diagnostics.

BELLE triger system is composed of detector su-trigger system and global decision logic. Figure ?? shows a schematic view of the trigger system. GDL gathers informations from those sub-trigger system so as to judge whether a physics event occurred or not. Especially owing to TSC trigger, the event timing is decided, and final trigger is put out 2.2μ sec after the event occurred.

The final trigger starts the DAQ cycle. It is received by the sequence controller and distributed to each DAQ subsystems. Receiveing the final trigger, DAQ subsystem starts digitizing the event data, and finishes within 200μ sec. The data will be collected and stored by DAQ system. While the DAQ subsystems are busy for digitizing, the next trigger is blocked. The dead time will be around 10% when trigger rate is 500Hz. Too high trigger rate will choke the DAQ system. So the trigger sustem must reduce the trigger rate which the DAQ system can work with safely.

In the full-bucket operation of KEK-B, the trigger rate for physics processes wiil be $\sim$ 100Hz. Details are described in Table 2.2.

Meanwhile, the total background trigger rate is estimated to be $\sim$ 100Hz which is due to cosmic ray background $\sim$20Hz, beam background $\sim$50Hz and so on. Since background estimation is not so exact, the trigger system must flexibly handle the case of higher background rate than expected.

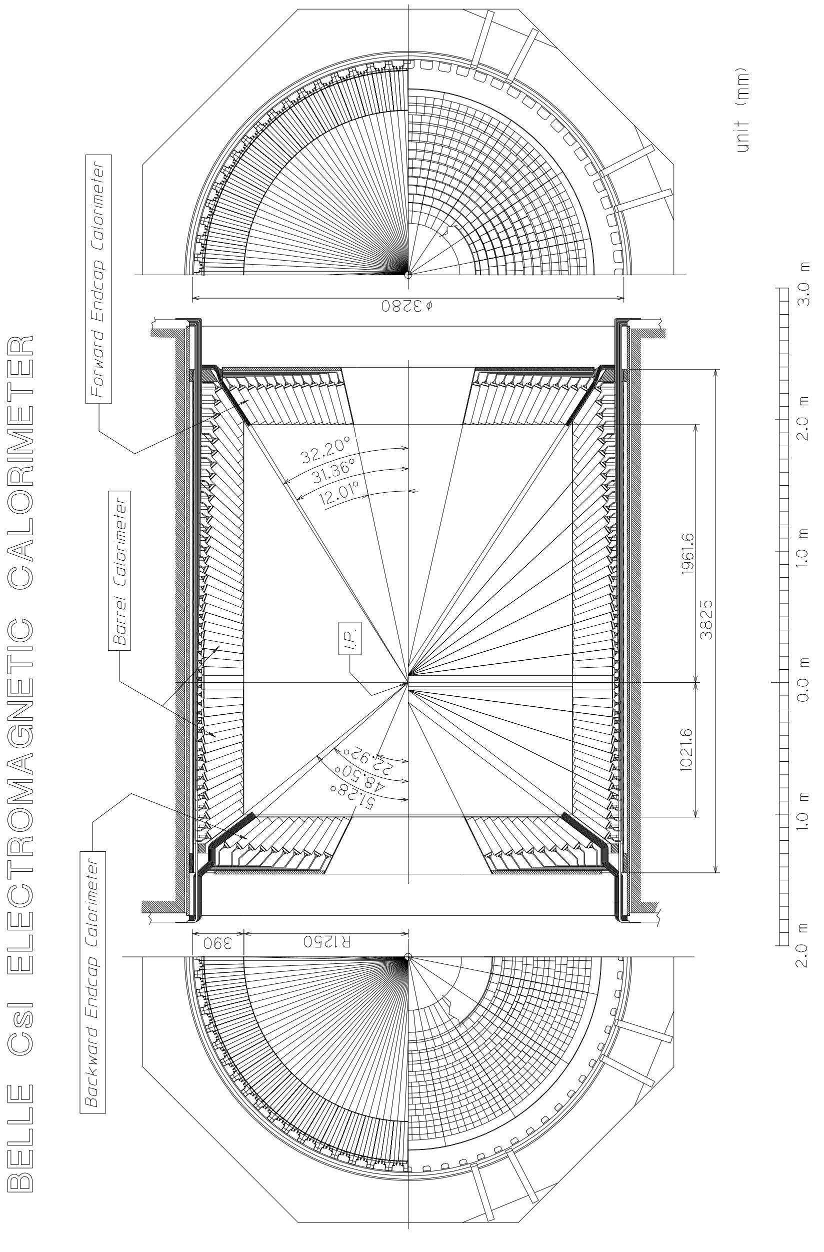


Figure 2.14: The Electro-Magnetic Calorimeter

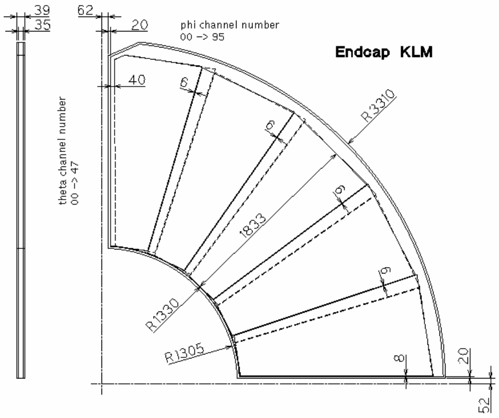


Figure 2.15: Front view of a quadrant the endcap KLM

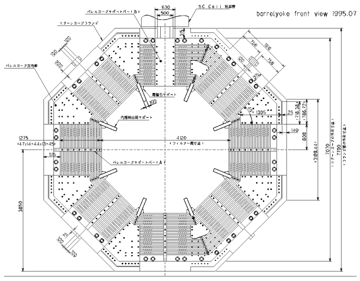


Figure 2.16: Front view of a barrel the endcap KLM

Physics Process	Cross Section (nb)	Rate (Hz)
$\Upsilon \rightarrow BB$	1.2	12
Hadron production from continuum	2.8	28
$\mu^+\mu^- + \tau^+\tau^-$	1.6	16
BhaBha ($\theta^{lab} \geq 17°$)	44	4.4[1]
$\gamma\gamma(\theta^{lab} \geq 17°)$	2.4	0.24[1]
2γ processes ($\theta^{lab} \geq 17°; p_t \geq 0.2GeV$)	~15	~35[2]
Total	~67	~96

Table 2.2: the summary of multi point test

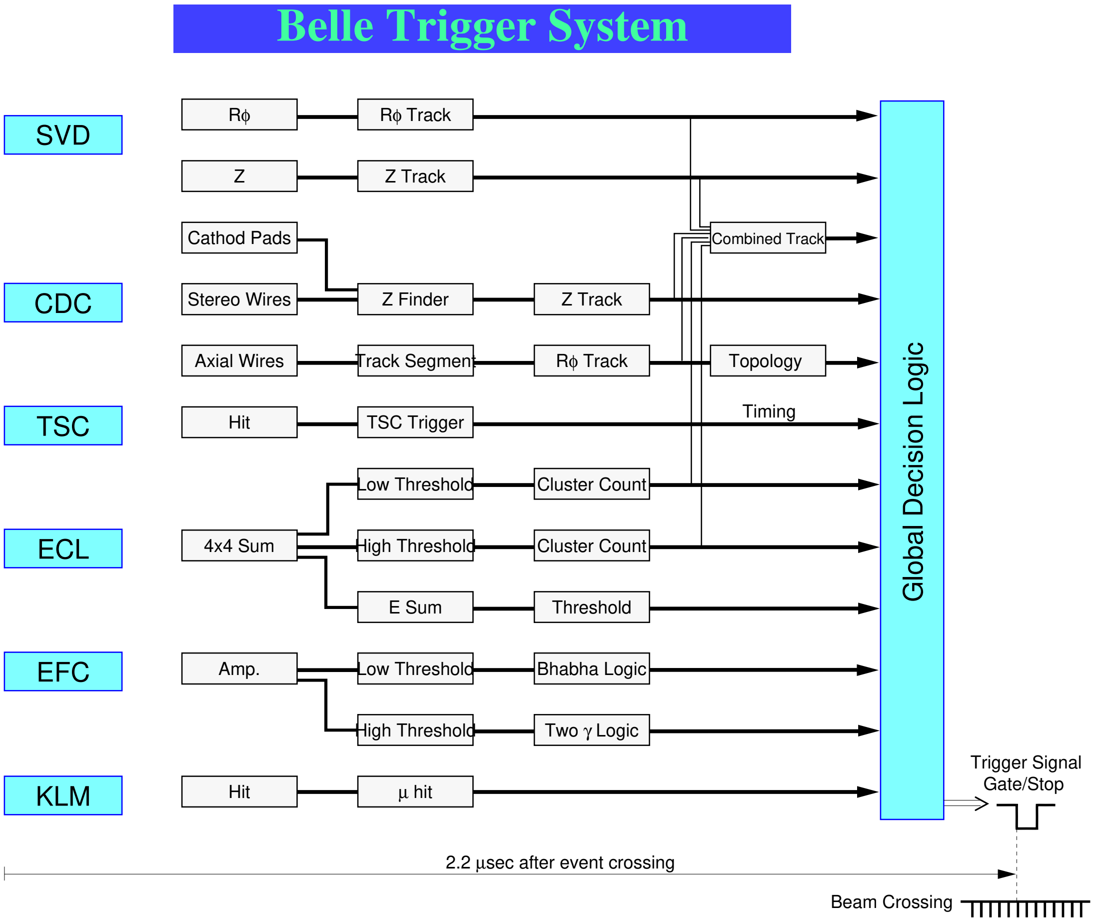


Figure 2.17: Schematic view of trigger system

2.5 Data acquisition system

The basic requirements to BELLE data acquisition system is,

1. Maximum trigger rate 500Hz,
2. The number of readout channel 143K,
3. Data size per event 30KB,
4. Maximum data transfer rate 15MB/sec.

The flow of DAQ system is shown in Fig 2.18.

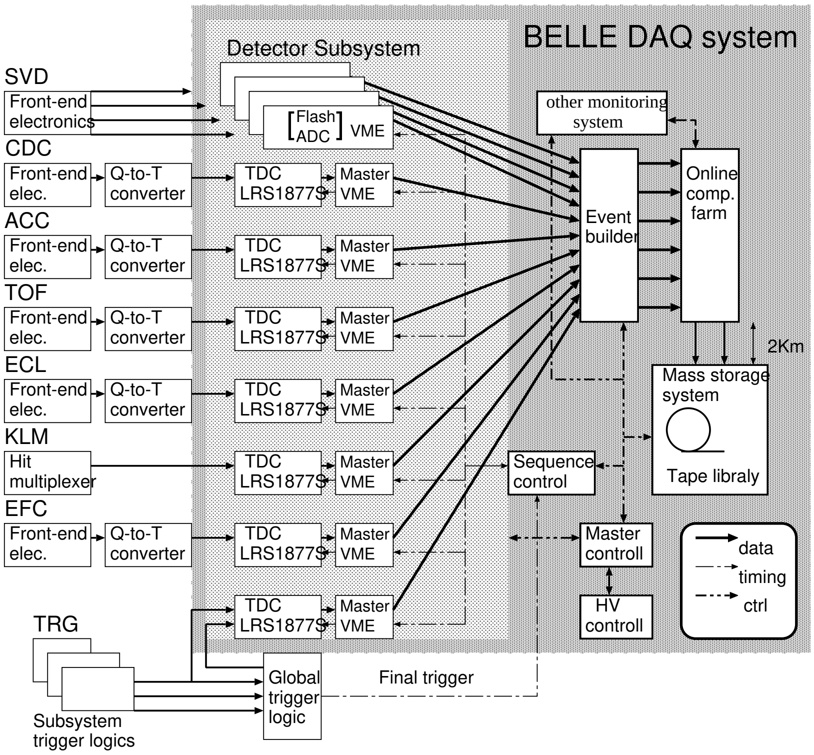


Figure 2.18: Global data flow

There are mainly four parts, detector subsystem, Event builder, online computer farm and Mass storage system that transfer the event data,and other control and monitoring systems, which are described on detail.

2.5.1 Detector Subsystem

The composition of detector subsystems is shown in Figure 2.19. There are four role for detector subsystems.

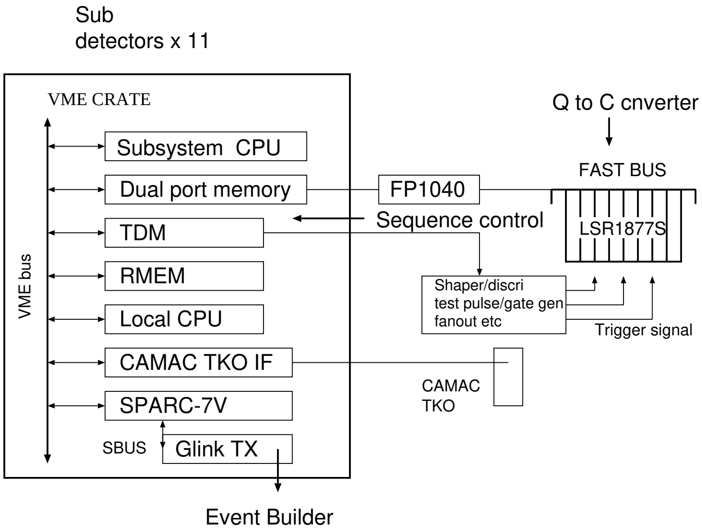


Figure 2.19: The architecture of subsystems

1. Read out the event data from FASTBUS
2. Pack the event data
3. Put the tag to the event data
4. Send the event data to event builder node

Receiving a final trigger, FASTBUS begins digitiizing the event data which is converted from analog data to time information by Q-to-T converters and delayed 2.2μs for waiting for trigger judgement. While digitizing for about 200μsec, the sequence controller blocks the next trigger. The digitized event data are transfered to Dualport memory which capacity is 128KByte. The subsystem CPU MVME177 collects and packs the event data which are scattered on Dualport memory. The event data are transfered to event builder interface, GLink-TX, after being put the event information.

In this scheme, the event data are digitized, read, and transfered to event builder.

2.5.2 Event Builder

The event data digitized are combined into a single event record by an event builder, which converts the "detector-by-detector" parallel stream to an "event-by-event" data river.

The event builder consists of 12 GLink-TX nodes (TX) and 6 GLink-RX nodes (RX) with a interconnecting switching network between them (see Figure 2.20). TX is composed of CPU board SPARC-7V and Sbus GLink-TX board, and RX is SPARC-7V and GLink-RX board. The event data colleted by subsystem CPU are read by SPARC-7V and transfered to GLink-TX. Glink-TX has a 128KByte FIFO and trasfered FIFO data to FIFO of GLink-RX by GLink via oprical fiber.

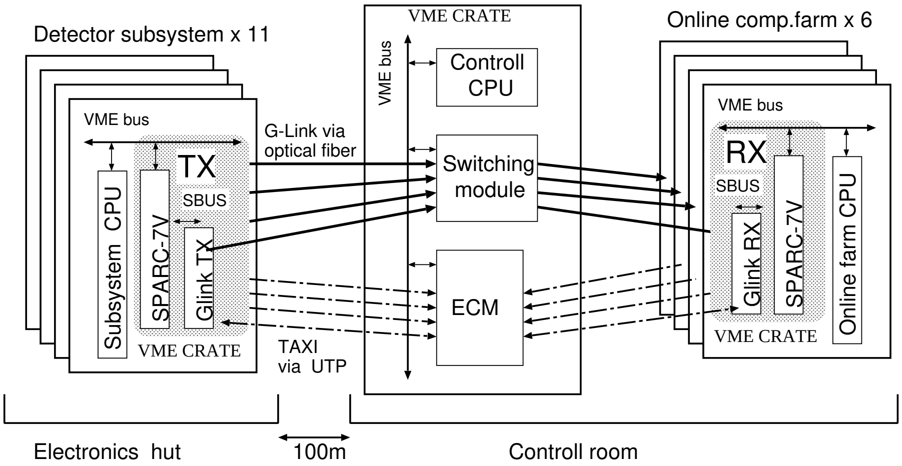


Figure 2.20: The architecture of event builder

The switching network is composed of ECM (External Controll Module), switch module and controll CPU board. ECM is conneted with all TX and RX by TAXI protocol, so the CPU can controll the switch and the data transfer timing. Using controll CPU, synchronized data transmission is performed.

The data compositions of each detector is shown in Table 2.3. In synchronized data transmission, the difference of data size may cause the fall of the total throughput, because data transmission timings are decided by the latest node.

subdetector	data size(KByte)	data size of one IB-N
SVD	16	4
CDC	9	9
ECL	4	4
Othes	1∼2	1∼2
Total	30	

Table 2.3: the result of 2×2 test

The detail of event builder is described in next chapter.

2.5.3 Online Computer Farm

The event builder output is transferred to an online computer farm called as "L3 tregger", where another level of event filtering is done after a fast event reconstruction. The online farm fillters unwanted events and reduces the number of events stored in mass storage system.

The requirement on the online computer farm is as follows:

1. a processing power at leats 5000 VAX quivalents;
2. a data rate from the event builder of more than 15MB/s; and
3. a transfer rate to the data server (mass storage system) of of more than 15MB/s.

Figure 2.21 shows the architecture of the online computer farm. The online computer farm consists of several VME crates. Each crate contains five Motorola MVME-188 CPU module which is equipped with four event processor chips that shared 64MB memory. The event data are fillterd by each CPU in parallel, and send to the mass storage system.

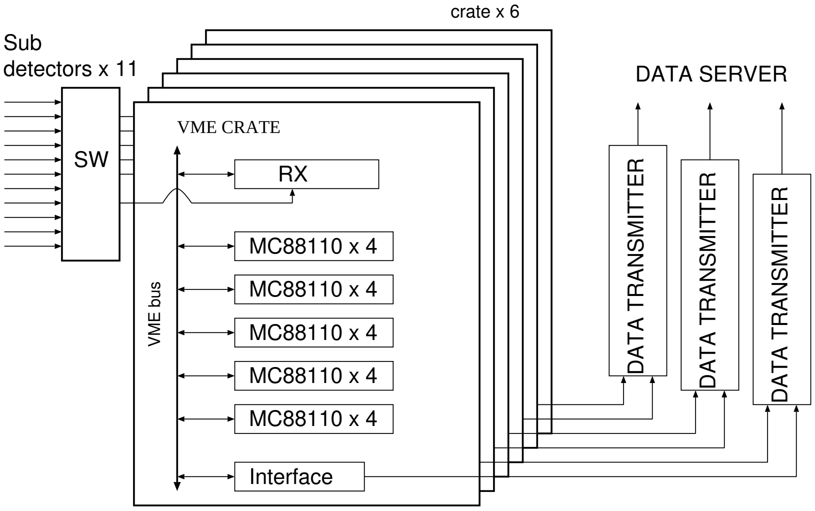


Figure 2.21: The architecture of the online computer farm

2.5.4 Mass storage system

The mass storage system is a cluster of magnetic tapes, which is capable to record 30TB/year. The data stored in the mass storate system is read out by offline computers and analized. In this sense, the mass storage system is the interface between DAQ sysytem and analysis(offline computers).

The mass storage system and offline computers are in ???, which is located for 2Km from Tsukuba are. The data are transfered using optical fiber link.

2.5.5 Other subsystem

The DAQ system has several other subsystem including: the event display, the data quality monitor, the monitor and alarm, the high-voltage controll, and a system to comminicate with the accelerator control system. These subsystems are connected with Network-Shared-Memory. It should be emphasized that these subsystems are to be unified into a single system, instead of having independent system for each subdetector. This is important in a large system such as the BELLE daq system, because this makes the operation easier and more reliable.

Chapter 3

BELLE Event Builder

In high energy physics experiment, many kinds of detectors are used in parallel. The event data taken by each detectors are collected and reconstructed to one event data by event builder.

In BELLE experiment, the barrel-shifter type event builder is running, which must handle a 15MByte/s data rate as described in Chapter 2. In this chapter, the construction of BELLE event builder and its performance are reported on detail.

3.1 BELLE event builder over view

The event builder consists of mainly three parts, 12 input buffer nodes (IB-N), 6 event builder node (EB-N) and the network devices which connects them. The data taken by each subdetectors are written in each IB-N, and transfered to one EB-N via the network. This is how the full event records are reconstructed in event building system.

Generally, the network connecting IB-N and EB-N can link one IB-N with one EB-N simultaneously, so IB-N must wait while one EB-N patrols other IB-N and collects data (see Figure 3.1). It means that the performance of the event builder is decided by the processing speed of one EB-N and the throughput of the network.

In BELLE, the data rate 15MByte/s exceeds the processing speed available in single EB-N, so multiple EB-N operating in parallel is needed. For operating EB-N in parallel, synchronized switching network is adopted. Using the switch and sincronized the data transmittion, it is possible that some IB-N link with some EB-N (see Figure 3.2) and several event data are reconstructed in parallel. Also, BELLE event builder has the scalablity by using switching network. It means that total processing speed can be expanded easily by increasing the nuber of EB-N and changing the network configuration.

To have the scalability is one of the most important point for event builder which is required high processing speed as BELLE.

BELLE event builder is composed of four parts (see Figure 2.20), Sbus G-Link module TX and RX, switching module and ECM, which are described on detail.

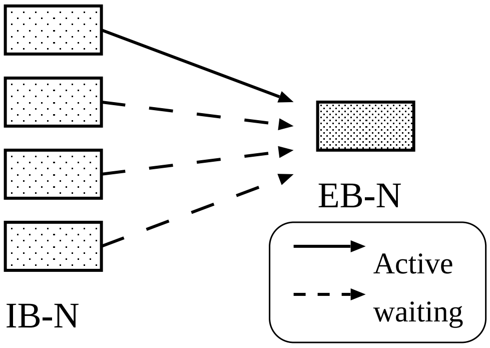


Figure 3.1: usual event builder : only one node is active simultaneously

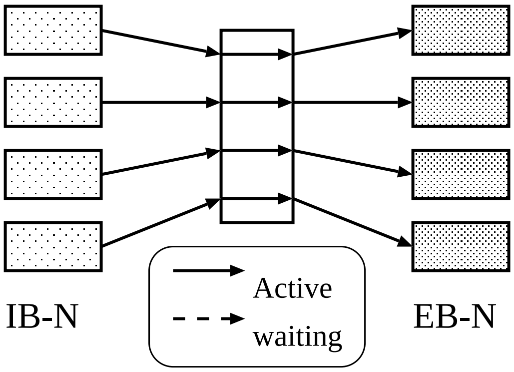


Figure 3.2: multiple nodes event builder : reconstruction is done in parallel

3.1.1 the switching system

BELLE event builder adopted the Barrel-Shifter type switching module which is one of the most simplest switching method. Event builder using Barrel-Shifter switch is developed in Fermilab. Figure 3.3 shows the operation of the 4x4 event builder using barrel shifter switch.

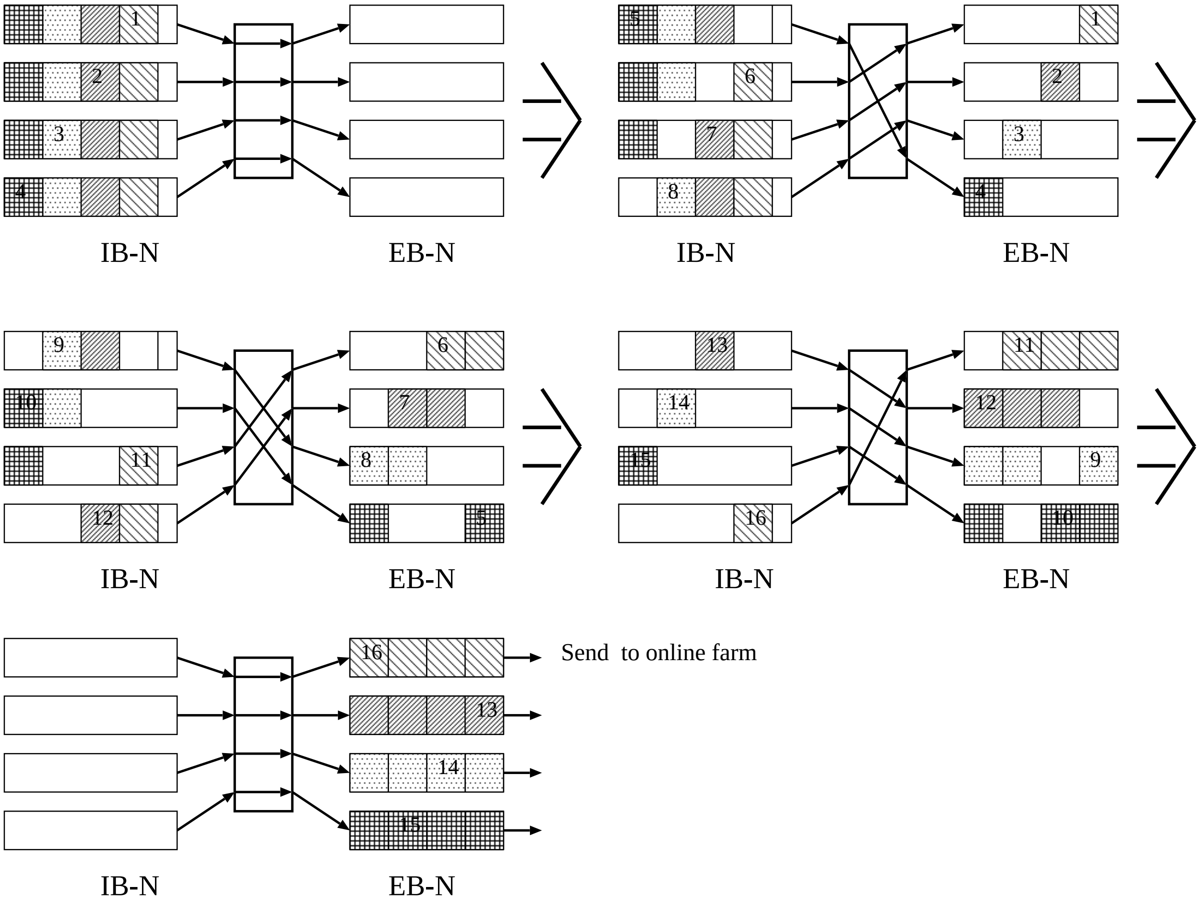


Figure 3.3: 4x4 barrel shifter event builder

Since the each data transmittion is synchronized, the switch configuration is changed after all data transmittion completed. It means that data transfer timing is deceided by the node which data size is large and the transfer time is the longest in all nodes. This may become a cause of the fall of the peformance if data size of each node is quite different.

This module is a VME module and one module has two 4x4 switches. The switching configuration is controlled by the controll CPU through VME bus.

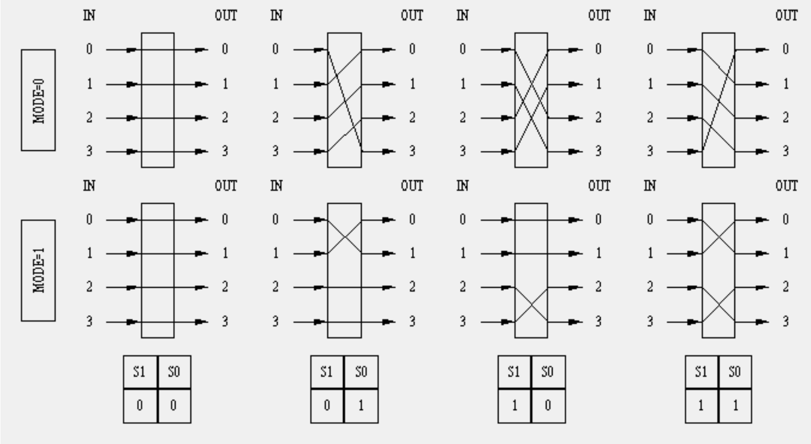


Figure 3.4: switching mode

3.1.2 Sbus G-Link module TX and RX

The G-Link chip produced by HP is used to transmit event data from IB-N to EB-N. The G-Link chip acheives the throughput 1.17Gbps. Figure 3.5 shows the block diagram of G-Link TX module. TX and RX module have a 128KByte FIFO and SBus interface. the data is written into TX FIFO through SBus, sent to RX FIFO through G-Link and read out from RX FIFO through SBus.

The TAXI chip is used for the controll line between TX, RX and ECM. It achieves the throughput 100Mbps, which is sufficient for ECM to controll TX and RX.

3.1.3 ECM(External Controll Module)

ECM controlls the synchronized data transmittion. ECM is a VME module and connected with all TX and RX through TAXI chip, so the controll CPU can controlls all data transfer timing through ECM. The controll timing of ECM is shown in Figure 3.6.

By sending Strobe to TX and RX and receiving Ack, synchronized data transmittion is performed.

3.2 performance

The details of the BELLE event builder is discribed in previous section, so now the performance of the BELLE event builder can be estimated.

Now the event builder has 6 EB-N, so 6 event data are reconstructed simultaneously. The data flow of the event builder is as follows:

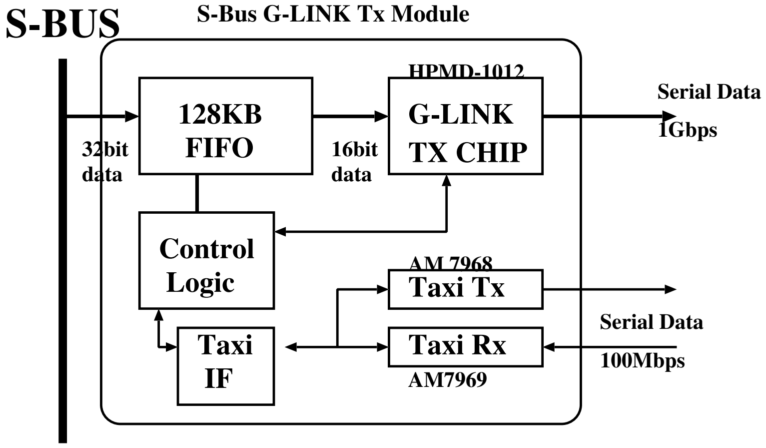


Figure 3.5: The block diagram of G-Link TX module

1. read out event data from subsystem CPU by mmap for SVD and DMA for others;
2. fill TX by Sbus mmap;
3. send from TX to RX by G-Link;
4. read out from RX by Sbus mmap,

which is shown in Figure 3.7. The throughput of event builder is estimated by the sum of the time which is taken by each steps. In 12×6 event builder, 6 event data are transfered in 12 times switching, so the time for trasmitting one event data is estimated by the sum of the the time for reading one event from subsystem CPU and switching two times.

3.2.1 data transfer subsystem to event builder

As noted before, event data are read out from subsystem CPU by mmap or DMA. The throughput from subdetector CPU to event builder is tested by H.Matsuo and in his master thesis [2]. Figure 3.8 and 3.9 shows the throughput from subdetector CPU to event builder and the time it takes. Since the data size of SVD and CDC is expected as 4KByte and 9Kbyte, the time for reading out event data is expected as 1.5ms and 0.8ms respectively.

3.2.2 Sbus mmap and Switching

The event data are written into G-Link TX FIFO and read out from RX FIFO by Sbus MMAP. Each speed is measured as 14MByte/s, and the time for another ECM process is measured as less than 50μsec which can be ignored.

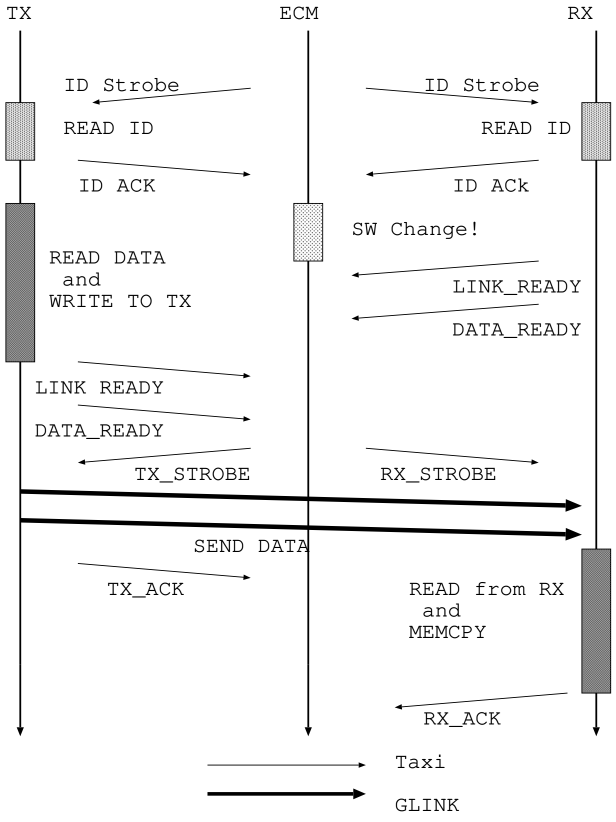


Figure 3.6: The ECM controol timing

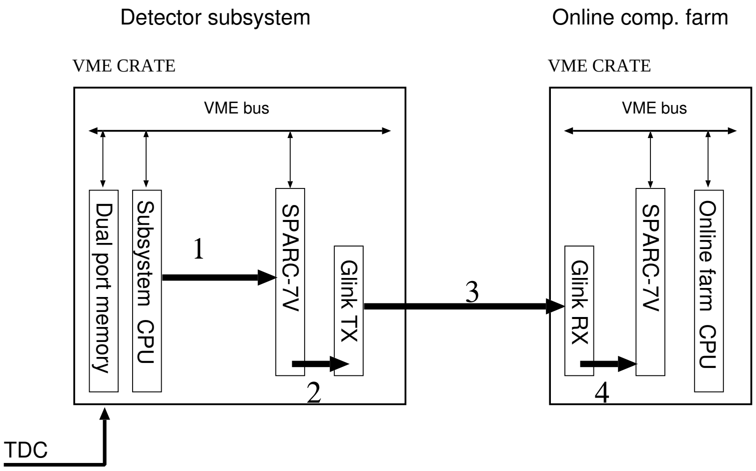


Figure 3.7: The data flow of event builder

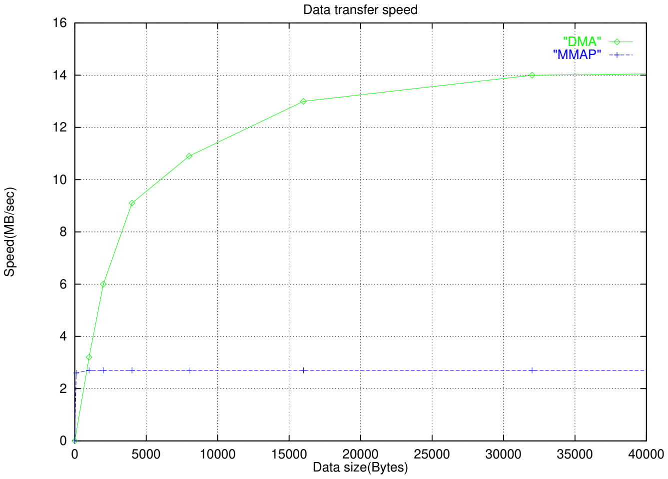


Figure 3.8: The data flow of event builder

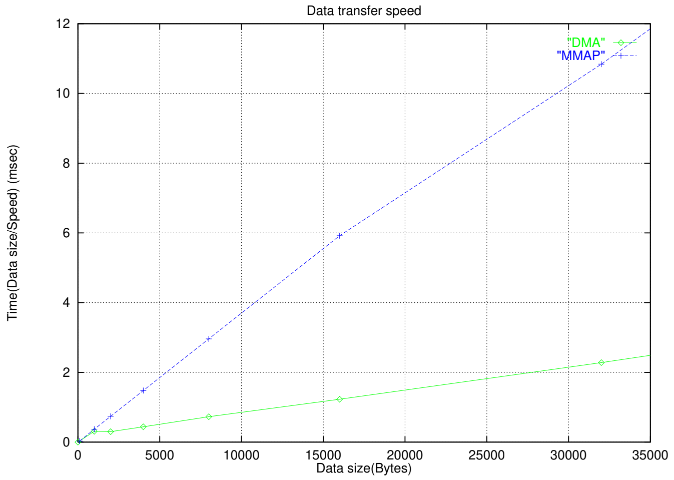


Figure 3.9: The data flow of event builder

3.2.3 summary

In event builder, the performance can be estimated by the total time of 3 Steps which ECM controlls (see Figure 3.10).

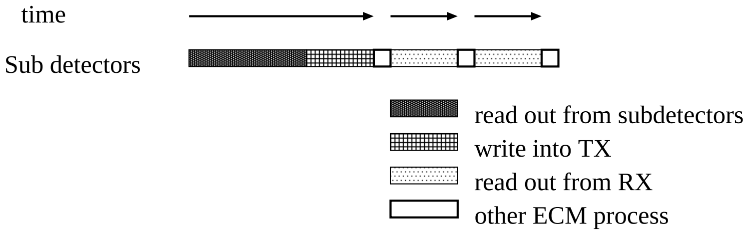


Figure 3.10: The data flow of event builder

The total time can be calcurated as,

$$(\frac{4Kbyte/event}{2.7MByte/s} + \frac{4Kbyte/event}{14MByte/s})(SVD) + \frac{9(CDC) + 4(SVD)(Kbyte/event)}{14MByte/s} = 2.6ms/event \quad (3.1)$$

which shows that the possible event rate of event builder is $1s/2.6ms = 385Hz$. The data transfer ability is expected by the data size which is shown in Figure 3.11. In the run which begans in Oct 1999,

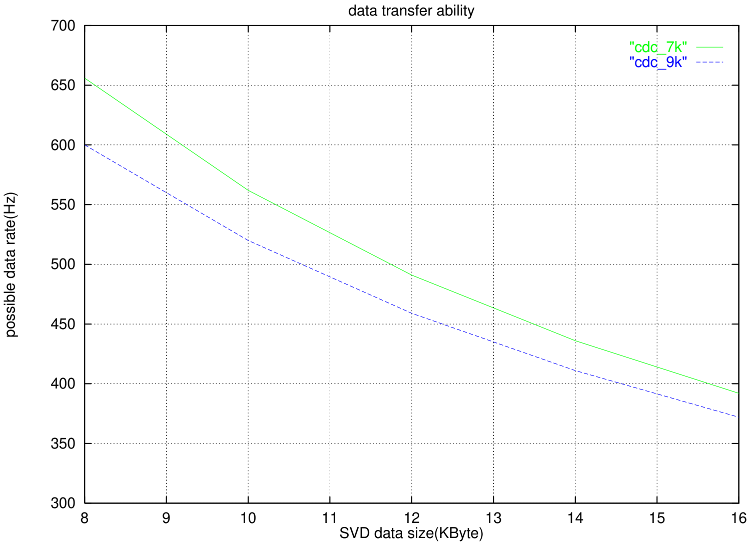


Figure 3.11: The data transfer ability

the avarage data size of SVD and CDC are 9KByte and 7KByte respectively, so the event builder can operate more than 500Hz.

Also it is suggested that the performance of event builder become 1.4 times faster by changing programs. So we will achieve the data processing rate 500Hz in present setup soon.

3.3 problem

3.3.1 Scalability

The important characteristics of BELLE event builder is to have a scalability. Now BELLE DAQ system consists of 12×6 system, which will be expanded to 12×8 system. The network is composed of barrel-shifter switch of 4×4 and 2×2.

Figure 3.12 shows the configuration of the 12×6 network, which is pretty complicated. It is possible,but rather difficult to expand the network to 12×8 setup.

3.3.2 GLINK

In 12×6 setup, the event data go through the switch 4 times. When we try to construct 12×8 or another setup, the event data go through the switch more than 5 times. It is reported that the data broken rate raises severely from 5 times, which can be suppressed by using Invert-mode of the switch which the switch reverses and outputs the signal as 0 to 1 and 1 to 0.

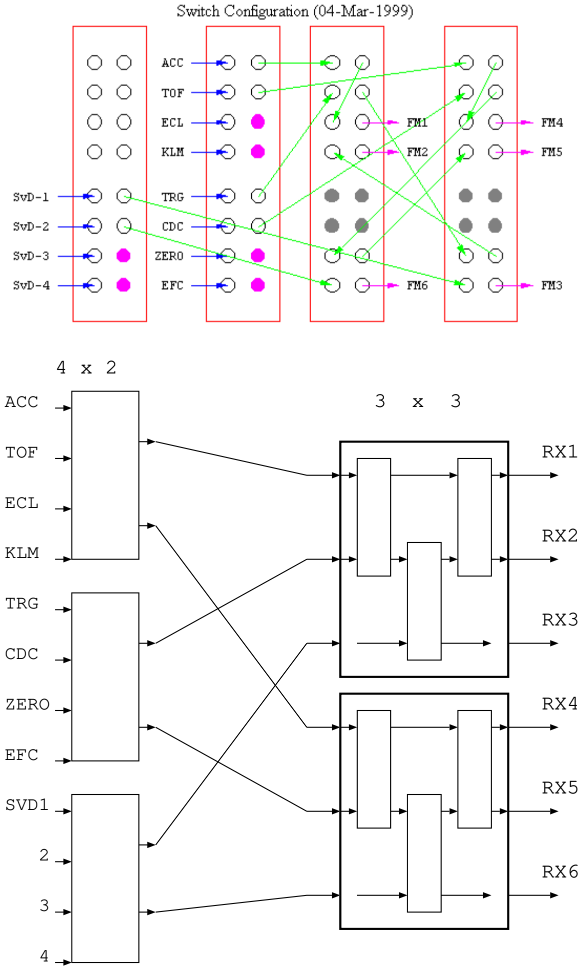


Figure 3.12: 12×6 switch configuration

Also the data broken problem is reported, which occurres when same continuous states as 0000000.. or 11111... are sent. This problem can be avoided by using exclude-or twice.

The program and the setup become more complicated. However, there is no way to solve these problem fundamentally. We should prepare for substitutes for G-Link.

3.3.3 performance

The performance of the event builder is described in previous section on detail. The total processing power linealy depends on the data size of subdetectors, especially SVD and CDC. The data size of SVD and CDC tend to be small by efforts of each sub detector group, but there are still some uneasinesses when the luminorsity increases.

The performance problem can be solved if SVD subsystem can use DMA or Sbus-DMA are introduced.

3.4 summary

In this chapter, The details of BELLE event builder are described.

The aim of BELLE event builder is to have the processing speed 15MByte/s, which will be satisfied. In order to use multi nodes in parallel and have the scalability, barrel-shifter type switch and G-Link are used. But, it is not easy to deal with them because they are specialized and ingeneral devives for event builder. So, there are some problems about the management and the scalability of the event builder.

Chapter 4

The network based event builder

In chapter 3, the details of BELLE event builder is discussed. Recently, the plan of upgrading the event builder begins to solve the problem. One of the most influential plan is the Ethernet based one, which is composed of conventional network devices.

In this chapter, problems in developing the network based event builder is discussed, and the goal of this paper is setted.

4.1 overview

When BELLE event builder was planed, the Ehternet devices which satisfied the requirement of BELLE DAQ system are very expensive. Recently, the price of the Ehternet devices is going to be cheaper and cheaper, and the performance becomes higher and higher. The Ethernet based event builder has a possiblity to correrspond from small experiment to large experiment which is planned in the future. So it is very important to study the characteristics and the problem of the Ethernet based event builder.

In this paper, we aim to contstuct the test system and know the characteristics of the Ethernet based event builder composed of 100Base-TX and TCP/IP, which is expected to be applied for the large experiment as BELLE.

4.1.1 setup

The Ethernet based event builder consists of 100Base-TX deveice, the switching hub and TCP/IP protocol which connect them.

There are some sender hosts and receiver hosts on the network. Sender hosts change the destination each time and try to send the data to all reveiver hosts equally. Figure 4.1 shows the setup of the Ethernet based event builder.

The Switcing Hub automatically controlls the connection and send the packet to distinations. The setup is exceedingly simple and reliable, so we have only to develop the programs which controll the data transmission.

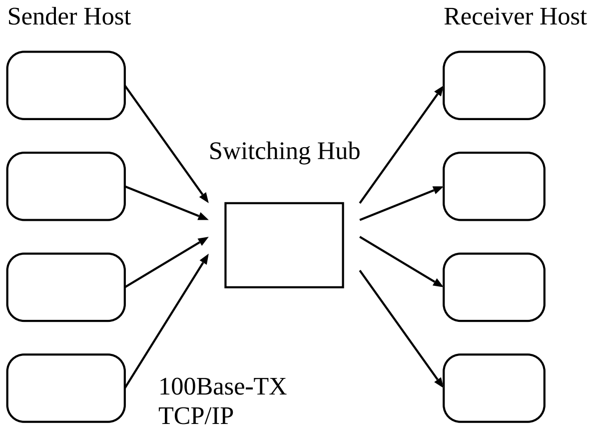


Figure 4.1: The setup of the ethernet based event builder

4.1.2 merits

Using conventional devices, The Ethernet based event builder is very useful and reliable. Merits of the Ethernet based event builder is shown as follows:

1. setup is exceedingly simple;
2. conventional network devices are reliable and available;
3. the cost is very cheap, and going to be cheaper;
4. performances are easily expanded by increasing node;
5. we can use PC for DAQ, we don't neet specialize device;
6. the capacity of the network is basically decided by its basic components, for example, 100Base-TX.

Making good use of these merits, we aim to construct the event builder easily, cheaper and flexibly by the requirement of the experiment.

4.2 problems in development

In developing the Ethernet based event builder, there are many problems to solve. In this section, the details of problems and ways out of them are discussed.

4.2.1 software

The development of the Ethernet based event builder results in the development of network programs if performances of devices are completely understood. Also, performances of devices are measured by using network programs. It is important to establish the way to test devices and develop the network program easier.

For the Ethernet based event builder, it is disiable to use same programs on all IB-N and EB-N respectively, because it makes the development and the maintenance easier. This time I use the benchmark tool DBS for performance measurment, and the library KONOE for whole network programs. To have a consistent policy is important in developing programs because it makes us easy to maintain and develop.

4.2.2 Selection of devices

In the Ethernet based event builder, the conventional devices are used. There are many devices in the world, and we have to choose one. It is expected that devices which performance are unsettled are the switching hub, so the selection of the switching hub is the most important.

The points in selecting devices are as follows;

1. to have enough performance;
2. to endure heavy load;
3. to have a stable performance.

Devices are deceided by checking these three points in each test, because there is no guarantees whether devices work correctly in next test.

Once devices are deceided, we have to know the basic performance of each devices. Their basic performances become benchmarks for many tests, and the influence of setup is estimated .

It is very important to know limitations and characteristics of each devices.

4.2.3 tuning

There are many parameters in Ethernet based event builder, and they fairly affect the total performance.

In order to get maximum performances, we have to examin all parameters one by one. The values of parameters is basically decided by 1×1 test.

4.2.4 influence of using multi nodes

When multi nodes are used, the throughput per one sender may be lower than 1×1 setup. If lower, it is important to understand the reason and expect how the throughput changes as the number of hosts increase. By knowing its mechanism, we can expect how many total throughputs are acheived when multi nodes ethernet based event builder is constructed.

As long as devices are working correctly, it is expected that there are some relations between total throughput and the number of senders and receivers.

4.3 goals

The aim of this paper is to construct the test system of the network based event builder, and know the characteristics. The goal is setted as follows, for construction;

- to establish the way to measure the basic performance of devices;
- to know the basic performances of each devices;

and for performance test of multi nodes ethernet based event builder,

- to understand the mechanism of multi nodes transmission;
- to know the way to expect the throughput and relations with the result of 1×1 test.

If these characteristics be aparent, I think of how to apply to BELLE experiment, and discribe the possiblity of the ethernet based event builder.

Chapter 5

Tools and component

The details of the ethernet based event builder are described in Chapter 4. This time, the ethernet based event builder are composed of 100Base-TX, TCP/IP and switching hub.

In this chapter, I introduce devices and tools I used for constructing the ethernet based event builder.

5.1 KONOE

KONOE (toolKit for Object oriented Network Distributed Online Enviroment) is the UNIX based toolkit developed by KONOE collabolation.

KONOE aims to supply class libralies and user enviroments which enable us to develope client-server model DAQ programs easier and faster. As the performance of netwok and computers advanced, the distributed and scalable DAQ system can be made by connecting machines with TCP/IP protcol. To develope or port softwares for such a DAQ system, the speciality of each platform and the experience were needed and it took long times and many troubles. KONOE releases us from trobles by using object oriented programs. KONOE has following characteristics.

1. to have enough performances
2. easy to operate
3. specified for network distributed enviroment
4. usable on all UNIX system

Using C++ where performances needed and dependences of machines exist, we can use same programs on different platforms with enough perfomances. Also JAVA and Tcl/Tk based GUI liblaries are prepared for users operating easier.

KONOE make us very easy to make network programs, which is already used for the program of BELLE high voltage control. This time, I use KONOE for whole network programs.

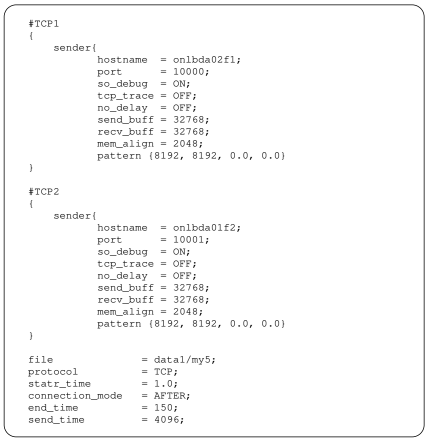
5.2 DBS

DBS (Distributed Benchmark System) is the powerful tool developed by Yukio Murayama NAIST [1] , which aimed to give network with multi-point configuration heavy load and measure performance index to see TCP specifications [3].

In network based event builder, the network is expected to be loaded heavily because sub detectors send all event data to their destination nodes via network. So DBS is avilable not only to see differences of the performance in changing configurations but to estimate the performance of devices which is very important to construct the event builder.

DBS has follwing features:

1. The DBS can treat multiple TCP connection spanning muptiple hosts simultaneously
2. The DBS has the capability of measuring the change of application level throughout at every data transmissions
3. The connection pattern and the traffic pattern are designated in script file
4. Perl and gnuplot based command "dbs_veiw" is prepared
5. use "xntp" for time synclonization of all hosts



```
#TCP1
{
    sender{
            hostname  = onlbda02f1;
            port      = 10000;
            so_debug  = ON;
            tcp_trace = OFF;
            no_delay  = OFF;
            send_buff = 32768;
            recv_buff = 32768;
            mem_align = 2048;
            pattern {8192, 8192, 0.0, 0.0}
}

#TCP2
{
    sender{
            hostname  = onlbda01f2;
            port      = 10001;
            so_debug  = ON;
            tcp_trace = OFF;
            no_delay  = OFF;
            send_buff = 32768;
            recv_buff = 32768;
            mem_align = 2048;
            pattern {8192, 8192, 0.0, 0.0}
}

file              = data1/my5;
protocol          = TCP;
statr_time        = 1.0;
connection_mode   = AFTER;
end_time          = 150;
send_time         = 4096;
```

Figure 5.1: DBS command file

---

[1] http://infonet.aist-nara.ac.jp/member/yukio-m/

5.3 TCP/IP

TCP/IP protocol is the most important technic used in internet which is widely spreading all over the world. TCP/IP representes for two different protocols, TCP(Transmission Control Protocol) and IP(Internet Protocol). IP is the protocol for controling routing between hosts, but it is not reliable because lost or cascade packets are not monitored. So TCP protocol is used to supply with reliable data trasnmission by using IP packet equipment which is not reliable.

For event builder, the reliance is the most important. If packets are lost or out of turn, a event data must be abandoned. So I adopt the TCP/IP and try to construct the reliable evnet builder.

5.4 Fast Ethernet(100BASE-TX)

Fast Ethernet was the method standardized 1995, which works on same cables with 10BASE-T and 10 times faster. The network based event builder supposes to use Fast Ethernet because the throughput 100Mbps(12.5MByte/s) per node is fast enough. But as the throughput is expected to go down if multi nodes are connected, I have to develope the way to expect the throughput at any setup.

In these tests, two Fast Ethernet devices are tested. The one is the combination of SPARC-5V and Fast Ethernet Card X1018A. In BELLE experiment, SPARC-7V is used in sub detector nodes, so it is expected that Fast Ethernet Card X1018A is used when the BELLE adopts the ethernet based event builder. The other is S-7/400Ui Fujitsu. The maximum throughput of X1018A is reported to 4.2MByte/s by Ryoji Inoue[2] because of the problem of SBus control DSP chip. Meanwhile, S-7/400 acheives almost 100Mbps(see chapter 6).

The basic performance of the Fast Ethernet device is also to be checked.

5.5 Switching Hub

The switching hub is defined as the network component which can receive the packet, store it, and send it to destination port.

The movement of switching hub looks like the postal serving system. The switch knows where everyone lives. At first, the switch receives mails. Secondly, the switch checks the address and deliver. If it is a broken mail, the switch send a message to a consignor.

The connections Port A to D, C to F and B to E are established simultaneously with no delay(see Figure5.2). The switching hub is suited to get stable throughputs when the number of port increases. The performance of the switch is decided by 4 basic factors, the number of port, the size of buff, the mechanism of packet transporting and backplane architectures.

5.5.1 mechanism of transmission

There are following three types in switching technics.

1. store and forward

---

[2] http://onlax2.kek.jp/ inoue/

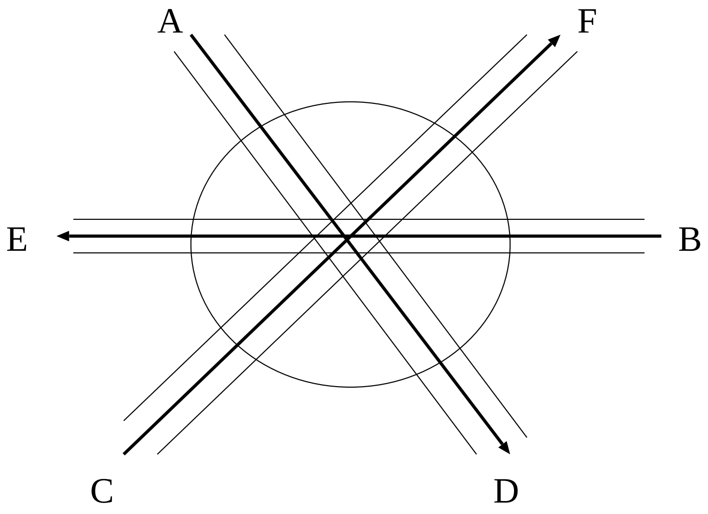


Figure 5.2: the movement of Switching Hub : Multipul data streams can go through the switch without interactions

2. cut and through
3. modefied cut and through

In store and forward method, the received packet is completely stored in switch buff and put out to destination port. The large packet may cause the fall of the throughput because the delay in the switch is equivalent to the packet size.

Meanwhile in cut and through method, the switch checks only the destination address. So even a broken packet is send to the desitination.

The modefied cut and through method adopts the advantage of two method. The switch check the first 64 Byte of the packet, by which almost all broken packet can be removed. This switch looks like store and forward type for short packet and cut and through type for long packet.

5.5.2 stability of transmitting

The switch is expected to devide the band of network evenly when a destination band is overloaded. The movement of the switch can be seen comparing Switch with Router using 10Base-T and 2×1 test shown Figure5.4.

The switch receives the packets and send it only to destination port. Meanwhile, the router send received packets to all connected ports, so there is always only one connection. Figure5.5 and Figure5.6 show the time evolution of the throughput. It can be seen that the destination node is completely shared with two sender by the switch and not by the router.

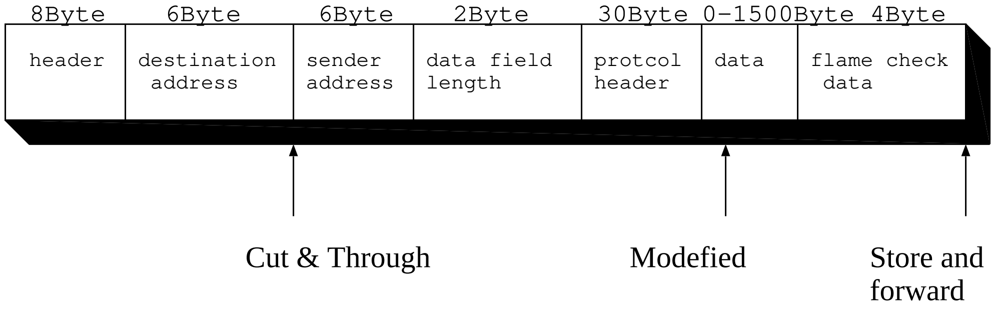
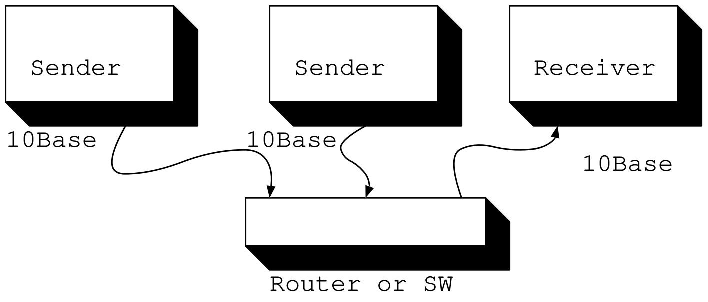


Figure 5.3: Ethernet flame and transmit starting point



Sender,Receiver: SPARC CPU-5V
Router : Allied telesis Centre Com820
SW : Corega FastSW4
OS : SunOS 5.5.1

Figure 5.4: 2x1 setup

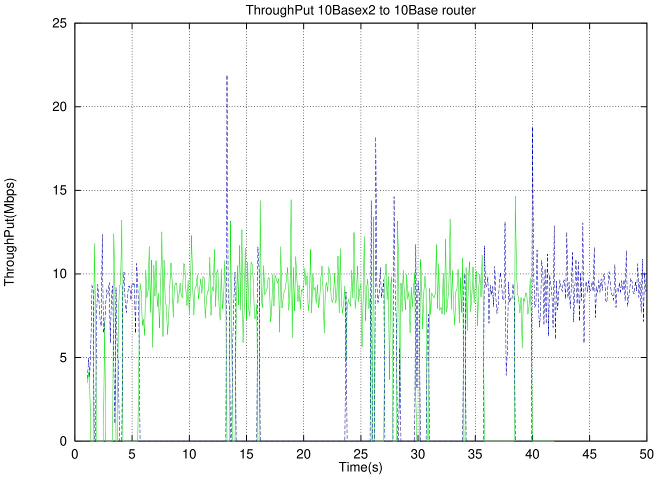


Figure 5.5: 2x1 throughput : router

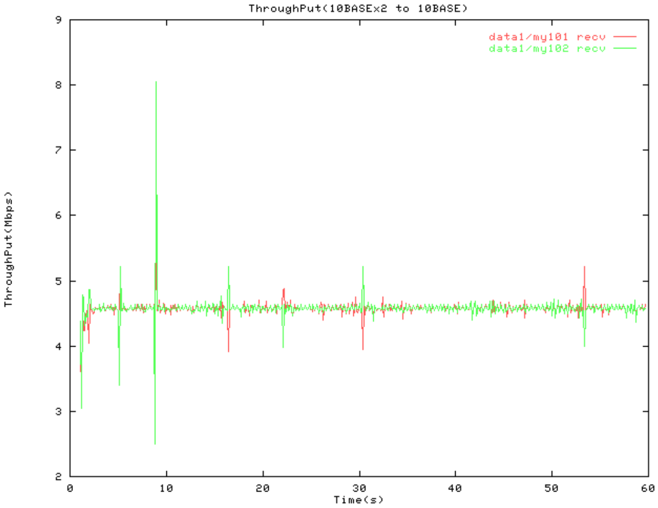


Figure 5.6: 2x1 throughput : sw

Chapter 6

Preparing for the study

The compornent of network is described in Chapter 5. In this chapter, the performance of Fast Ethernet and Switching Hub is tested by using mainly 1×1 setup. There are many parameters to decide and problems to solve. It is important to measure basic performances to appear the problem.

6.1 Parameters

TCP socket has many parameters and it seriously affects to throughput. Also data size which is sent once is expected to cause a fall of throughput because small packet size increase the number of system calls. The important parameters are checked and decided at 1x1 setup.

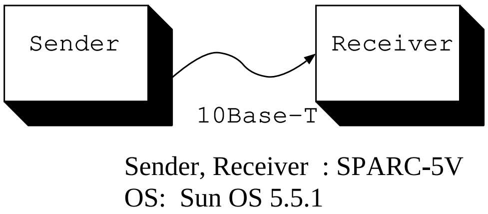


Figure 6.1: 1x1 setup

6.1.1 TCP socket parameter RCV_BUF

TCP socket has many parameters (see Figure 6.2) and the function "setsockopt()" is prepared in C libraly to set them.

RECV_BUF sets the size of buffers where TCP stores the recieving packets. Large buffer is expected better since small buffer causes the fall of throughputs by increasing the number of system calls.

```
SO_DEBUG        SO_REUSEADDR

SO_KEEPALIVE    SO_DONTROUTE

SO_LINGER       SO_BROADCAST

SO_OOBINLINE    SO_SNDBUF

SO_RCVBUF       SO_TYPE

SO_ERROR
```

Figure 6.2: socket parameters

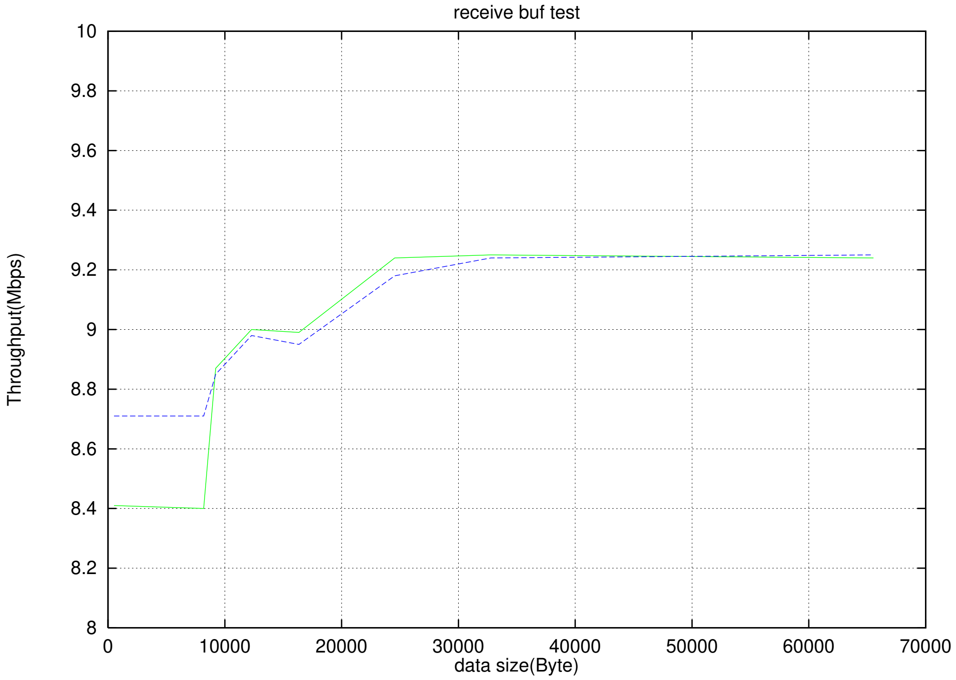


Figure 6.3: RCV_BUF test

Figure 6.3 shows the influence of the size of RCV_BUF on throughputs, Green and blue line show the packet size 8KByte and 32KByte respectively. It can be seen that throughputs are stable when RCV_BUF is large enough. So RCV_BUF is setted to 65535 Byte which is maxmum size allowed in system. Also SND_BUF is setted to 65535 Byte in order to decrese the number of system calls.

6.1.2 packet size

A packet size means the size of data sent or received by one system call. Large packet size is expected to raise throughputs because the overhead of system call is larger. When a packet size is raised to some extent, the network is saturated with continuous data flows and throughputs are limited.

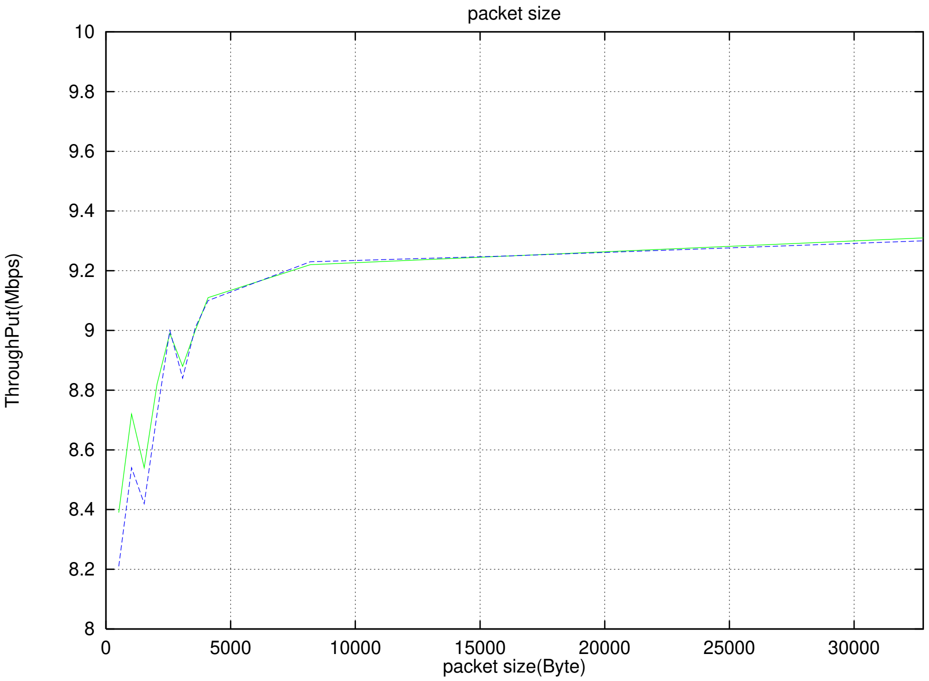


Figure 6.4: packet size test

The result of the test is shown in Figure 6.4. The x axis shows the packet size and blue and green line mean the size of RCV_BUF 8KByte(default) and 64KByte respectively. The network seems to be saturated when packet size is setted more than 10KByte.

The data size of one sub dectector node is less than 4KByte except CDC. So some event data should be packed into one packet and send for efficient transmitting.

6.2 Data Reconstruction

TCP is the powerful protocol which checks lost packets and orders, but there are one weak point. Sender hosts know the size of packet they send, but receiver hosts don't. Also, if packets are sent continuously, the receiver doesn't know where the boundary of the packet is. This is because packets are devided into small flames which size is less than about 1500 Byte (see Figure 5.3) on real network, and the receiver gets these small flames. For example, the sender sends 10KByte pakets, but anyone knows how many times does "receive" needed for the receiver to get 10KByte. 1 or 2 or 3 ? it depends on the condition of the network and hosts.

6.2.1 packet structure

Using TCP, users have to prepare for a method to know the packet size and the boundary. This time, this problem is solved by giving packets header and end mark. The structure of packet is shown in Figure 6.5.

This is the packet structure used in BELLE event builder now. The packet is made from three parts, packet header, event header and data space. The size of data space is 4KByte, which is the largest size sub detectors except CDC need for one event. Also one packet can have many event data at the same time. This supposes the scheme that one data space is filled with one real data, some event data is packed into one packet and the packet is sent to server hosts.

In this paper, how many data are packeted into one packet is disscused later. Usually, 6 event are packed and send.

6.2.2 packet reconstruction

All the packet informations are in the packet header. If the packet header is caught, the position of the packet end mark, the beginning of event data and the top of next packet can be identified. However, it is often the case that the packet header is not caught and lost. This is because the packet header is already read from TCP buffer when the previous event is read. Also, the packet end mark is often lost, because the received buffer size is smaller than the packet size. Considering these points, the packet reconstruction scheme is made.

Figure 6.6 shows the scheme. In this scheme, throughputs is expected to be lower because the sender host writes informations in the packet header and the event header every time, and the receiver host must use the function "memcpy" many times to reconstruct packets. The throughput must be estimated as the amount of event data transmitted. So the size of header and end marks must be subtructed from total data size the receiver host received. This is also the reason why throughputs seem decreasing.

The cause of the fall of throughputs must be studied because the fall caused this scheme may become a bottleneck.

6.3 influence of data reconstruction

Throughputs are expected to fall as described in previous section. In this section, the influence of data reconstruction is estimated.

The flow of the test is as follows.

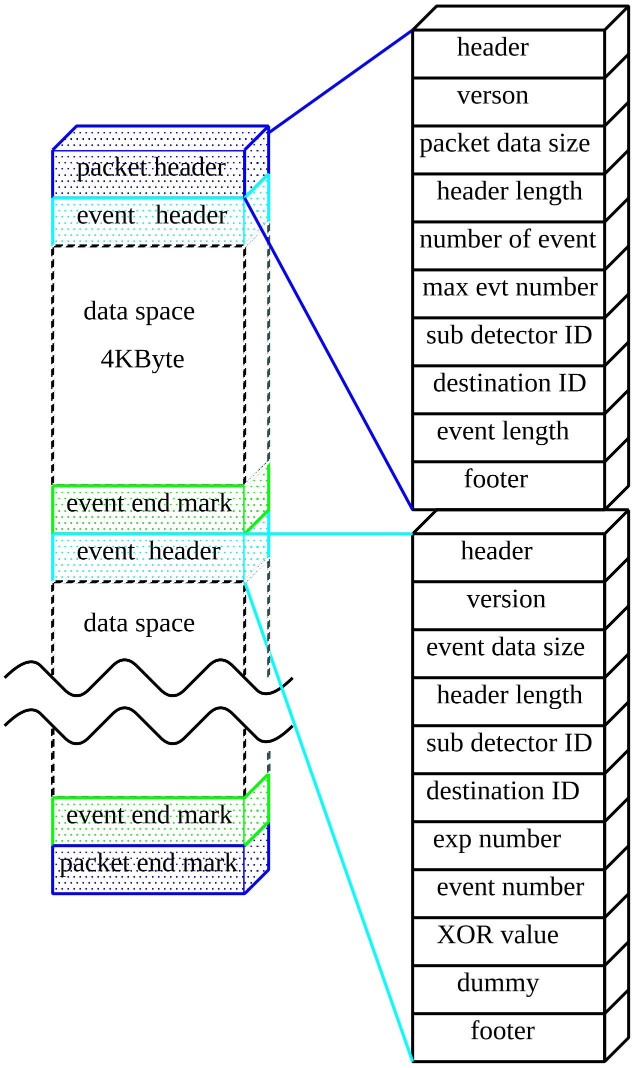


Figure 6.5: the structure of the packet

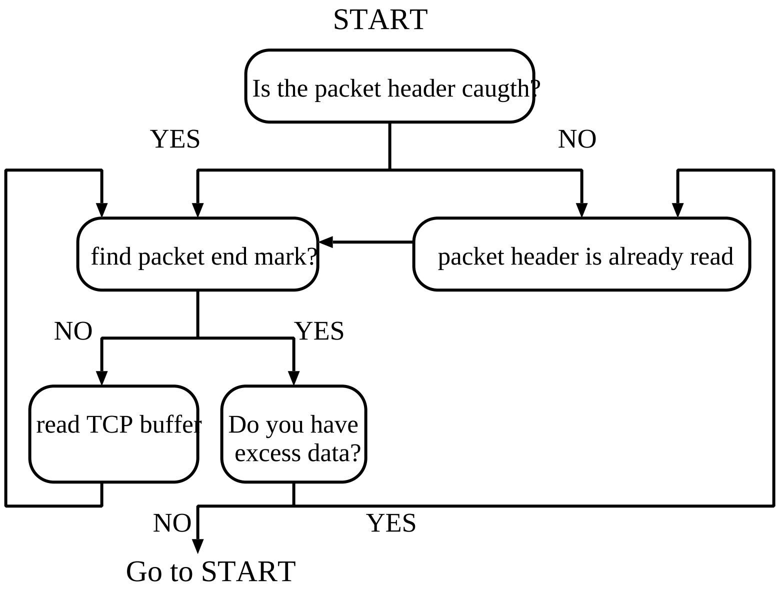
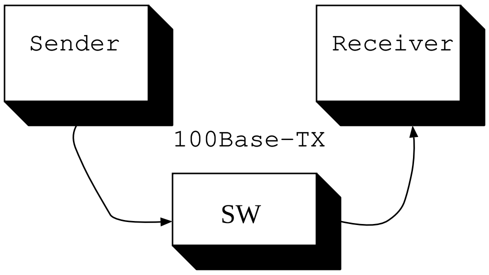


Figure 6.6: the scheme of packet reconstruction



Sender, Receiver : SPARC-5V or S-7/400Ui
SW: allied telesis FS-708E

Figure 6.7: 1×1 test

1. 1×1 test (Figure 6.7) and compare with normal transmittion
2. check which hosts limitted the throughput by 1×2 or 2×1 test

The result of this test is the basic parameter for estimating a multi point network.

6.3.1 1×1 test

The 1×1 test is performed using SPARC-5V and S-7/400Ui for hosts respectively. The result of the test is shown in Table 6.1.

	normal(Mbps)	reconstruction(Mbps)
SPARC-5V	38	25
S-7/400Ui	94	93

Table 6.1: the result of 1×1 test

As expected, thoughputs decrease in SPARC-5V test. CPU powers of the sender and the receiver in SPARC-5V test are resumed 30So It seems that throughput is limited to 25Mbps by CPU powers.

Meanwhile, S-7 doesn't decrease and the network are saturated by hosts.

6.3.2 1×4 test

1×4 setup is used to measure the maximun performance of the sender SPARC-5V. Sending the packet to four receivers by turns in 1×4 setup, the load of one receiver becomes one fourth comparing with 1×1 setup in which receivers limite the throughputs. 1×4 setup is used because there are only two Fast Ethernet cards for SPARC-5V. Also, these tests show that the switching hub devides packets to both nodes equally on Fast Ethernet.

The result is shown in Table 6.2.

	Sender(Mbps)	Receiver(Mbps)
SPARC-5V	35	8.7

Table 6.2: the result of 1×4 test

The throughput of sender seems to be quivalent to normal transmitting. It is apparent that the performance of the receiver host becomes a benchmark in building the multi point network.

6.3.3 the influence of data size on 100 Base-TX

In the test of 6.3.1, 6 event are send, which means that packet size is about 25KByte. Here, the number of evnet data are changed and the influence is checked.

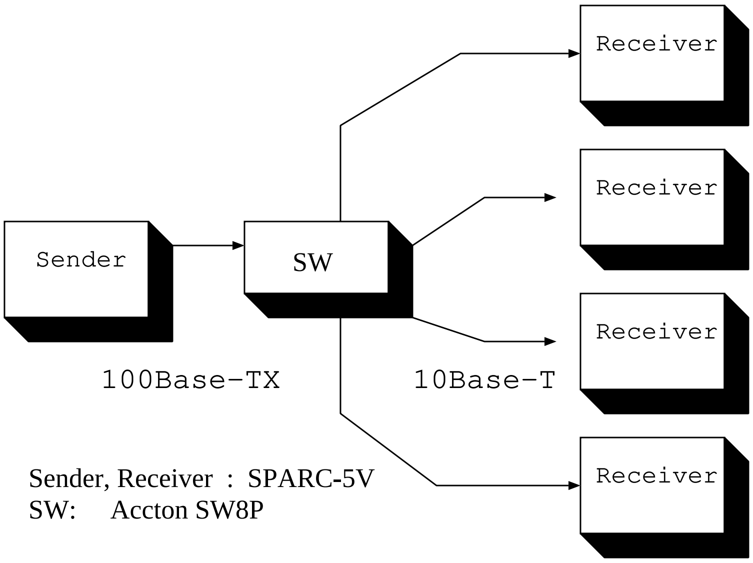


Figure 6.8: 1×4 test

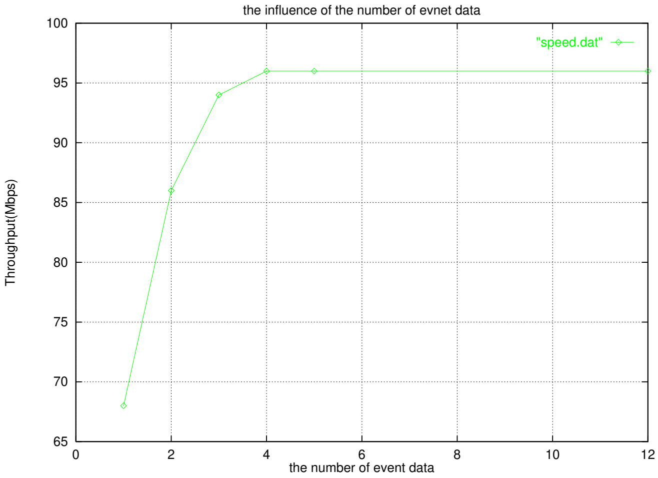


Figure 6.9: the influence of the number of event data on throughput

The result of the test is shown in Figure 6.9. It can be seen that the influence is large and we should use more than 3 event to avoid the fall of the throughput when we construct test systems.

Chapter 7

The study of multipoint network

The maximum throughputs of the sender and the receiver are measured in Chapter 5. It is important to keep the throughput as fast as the maximum when the number of hosts increases.

In this chapter, the multi point network is built and investigated. The multi point network is defied as some sender hosts and receiver hosts are connected each other, and sender hosts send packets to all receiver hosts. At first, 1×2 test is expanded, and other methods are discussed later.

7.1 The expansion of 1×2 type

One of the most simple method to construct the multi point network is the expansion of 1×2 type shown in Chapter 5. The sender sends packets to receiver hosts in order, and receiver hosts receive arrived packets intently. At first 2×2 setup is tested, and expanded to the advanced.

7.1.1 2×2 setup

2×2 setup is constructed by putting two 1×2 setup together. Setup is shown in Figure 7.1.

On the receiver, two processes run so that packets send by each hosts are received in parallel. The result is shown in Table 7.1. One line means one connection, and throughputs are measured by one connection.

	expected(Mbps)	measured(Mbps)
S-7/400Ui	46	25

Table 7.1: the result of 2×2 test

The throughput in 2x2 test is expected as 46Mbps beucause the maximum throughput of one receiver (92Mbps) is shared with two process. But the measured is 25Mbps. The reason of the drop of throughputs must be checked. Also, the sw of Accton made is not working collectly in this test (see Appendix).

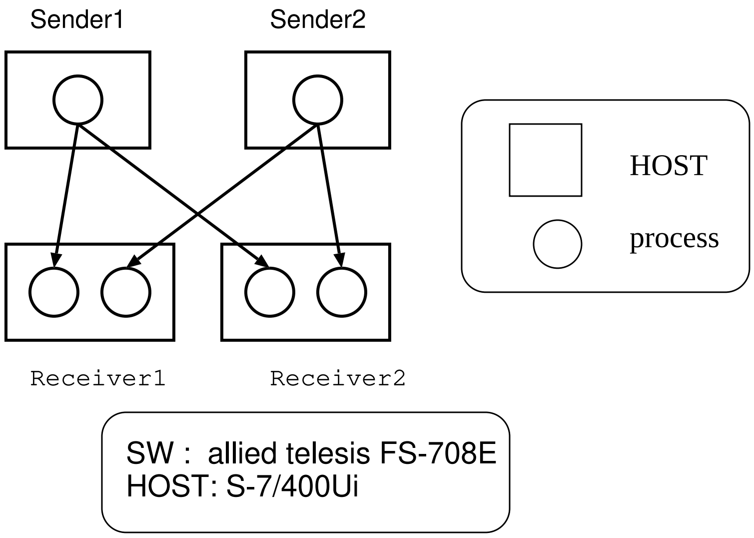


Figure 7.1: 2×2 test

7.1.2 Considerations

It is important to investigate the reason of the drop for proceeding to advanced setup.

In this test, it is expected that two sender hosts are synchronized because same CPU is used. In a ideal case, the destination of the sender changes synchronouly (see Figure 7.2) and throughputs are preserved.

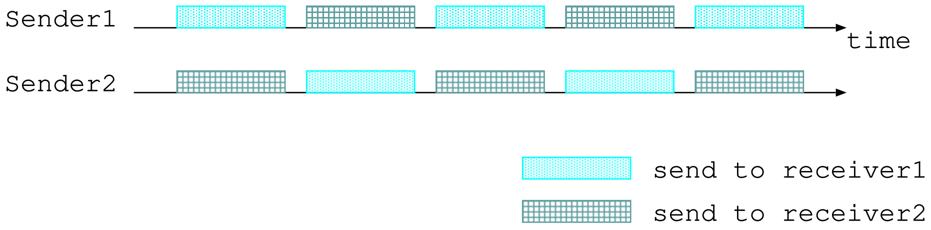


Figure 7.2: ideal case:synchronized transmittion at 2×2 setup

The result of the test shows that two hosts are not synchronized and it takes twice longer time than 1×2 test to send one packet. This can be explained by thinking the mix of packets. When two packets which destination is same arrive at the switch at the same time, two packets are completely mixed at ethernet flame level, so it seems to take long time for the sender host to send one packet (see Figure 7.3).

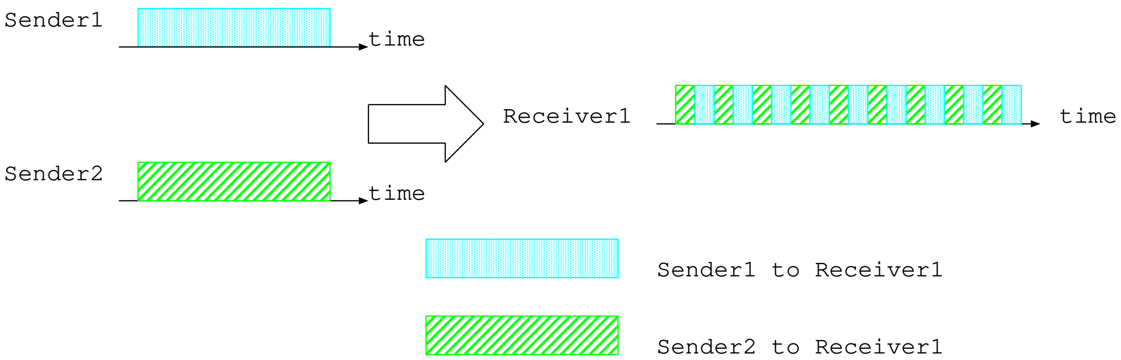


Figure 7.3: packets mixing

This can be proved by 1×1 + 1×2 test (Figure 7.4). The result of test shows that only Sender1 suffers the influence. This is because packets of Sender2 are mixed into packets of Sender2 by the switch.

	before connection(Mbps)	after connection(Mbps)
Sender 1	93	46
Sender 2	46	46

Table 7.2: the result of 1×1 + 1×2 test

Using this scheme, the throughput of network can be expected.

7.2 The test of expanded network

In previous section, how packet collisions happen the fall of the throughput are desicribed. In this section, the network is expanded and tested.

7.2.1 The expectation of the throughput

The throughput of one connection can be expected.

In 2×2 test, throughputs are a half of maximum throughput in 2×2 setup because one receiver has two connections and packets which two senders send are mixed. Here, maximum throughput in 2×2 setup can be caluclated as 46Mbps because the throughput of one sender is devided into two connections. So the expected throughput is defined as,

$$expected\ \ throughput\ \ of\ \ one\ \ connection = \frac{94}{N.S \times N.R} \tag{7.1}$$

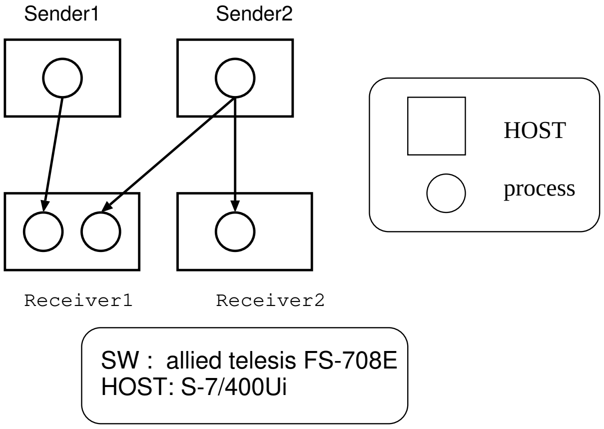


Figure 7.4: 1×1 + 1×2 test

where N.S is the number of senders and N.R is the number of receivers. (7.1) shows that the total throughput of all conneciotns is always 94Mbps which is maximum throughput of 100Base-TX.

(7.1) can be expressed as

$$expected\ \ throughput\ \ of\ \ one\ \ sender = \frac{94}{N.S} \tag{7.2}$$

(7.2) is also important because not only total throughputs but the throughput of one sender host must be investigated in constructing event builder. The throughput of one sender has no relations with the number of receivers.

7.2.2 The expanded network test

Table (7.3) shows the result of the test.

The measured throughputs are roughly agreeed with the expected. This result shows that packets are mixed at the switch and throughputs are fallen as expected. If the switch works normaly at 12×6, The total throughputs are 94Mbps(12MByte/s) and 35Mbps(4MByte/s) for S-7 and SPARC-5V respectively, which is inadquent for event builder.

What is the most important point in this scheme is that the scalablity which is one of the most important features of event builder is completely lost by packet mixing effect.

7.3 Synchronized data transmittion

The packet mixing occured when some sender hosts intend to send packets to same destination host. In order to avoid packet mixing and have the scalablity, data transmittion must be synchronized.

	expected(Mbps)	measured(Mbps)
1×2	47	47
2×2	23	25
3×2	16	17
4×2	11.5	13
5×2	9.5	10.5
2×3	16	17
3×3	10.5	12.5
4×3	8	9

Table 7.3: the result of the expanded test

In this section, synchronized data transmittions are evolved and tested. At first, the scheme how the transmittion is synchronized is described. Second, the basic performance is measured using 1×1 setup. Third, multi point network is investigated.

7.3.1 How to synchronize?

One of the most simple way to synchronize data transmittion is to use socket transmittions. The principle is as follows(see Figure 7.5),

1. the controll host sends start-flags to 1 and 4
2. 1 and 4 send finish-flags to the controll host after sending packets to receivers respectively
3. If the controll host receives finish-flag, send start-flags to2 and 3
4. 2 and 3 send finish-flags to notice the controll host finishing the transmittion

In this schems, it is expected that throughputs are dropped because the delay of flag transmittion arises and throughputs of one sender is not changed if the number of senders or receivers increases.

7.3.2 Measurment of basic performance at 1×1 setup

It is important to measure basic performance at 1×1 test because it's easy to see what is a bottleneck of the limit of throughputs. 1×1 setup is shown in Figure 7.6, and the result is in Figure 7.7.

In the case, only one setup is tested, we can achieve the throughput 94Mbps. However, when two setup is tested simultaneously, the throughput falls to 66Mbps. This is because the data size received once by the receiver is quite small in two setup case. It means that the data transfer efficiency of the switching hub falls by loads. In the experiment, the network and the switching hub are loaded heaviry, so we adopt 66 Mbps as standard performance.

The resumed CPU power in senders and receivers are 60% and 40% respectively. It is apparent that the load of servers and receivers doesn't causes the fall of throughputs because CPU are not saturated

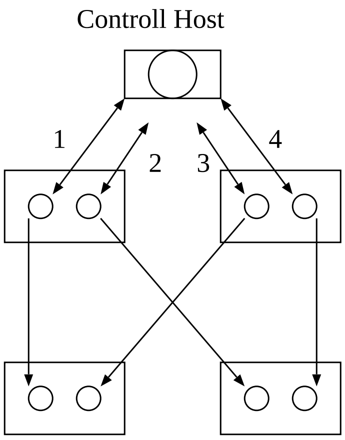


Figure 7.5: the synchronized data transmittion

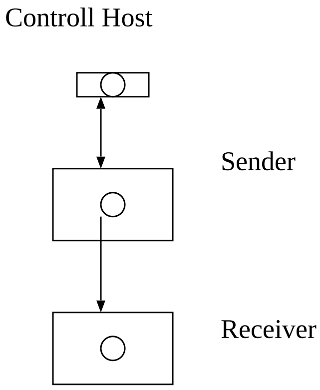


Figure 7.6: Synchronized 1×1 test

yet. When the number of senders increases, the load of the host increases proportionaly and may cause the fall of throughputs. Meanwhile, the load of the controll host is only 5% in 1×1 test, and the influence are estimated as rather small.

Even if 12 senders are used and the load causes the fall of throughputs, it can be solved by changing the performance of the switching hub and each hosts.

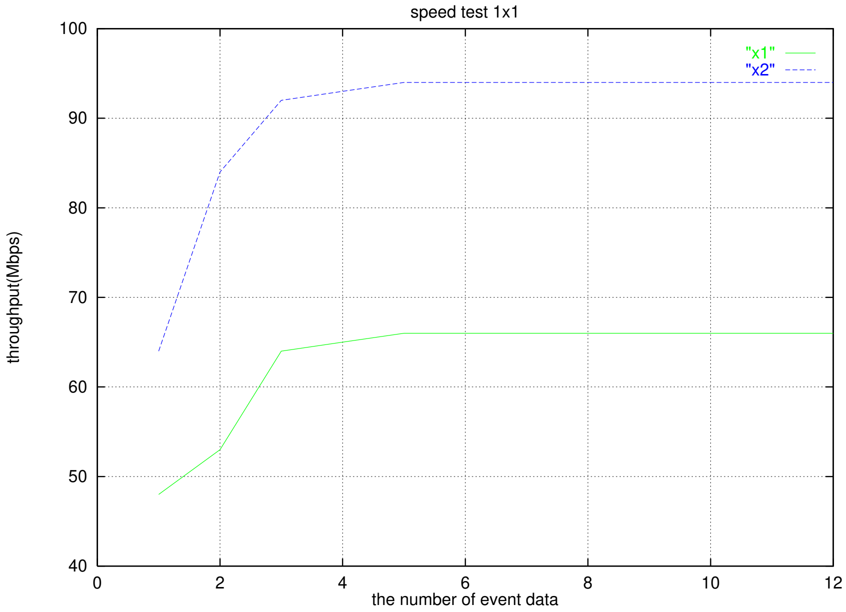


Figure 7.7: the influence of the number of event data on synchronized 1×1 test

7.3.3 The measurment of multi point setup

Now that the basic performances has been measured, multi point setup are tested. The result of multi point test is shown in Table(7.4).

The result of tests shows that throughputs are not fallen. 3×2 and 4×2 look slow, but this is because senders can send packets two of three times and two of four times at 3×2 and 4×2 setup (see Figure 7.8). Considering this point, throughputs of one sender are calculated as 69Mbps and 72Mbps respectively, which are almost maximum speed of synchronized transmittions. The influence of the signal host is not seen yet.

This test shows that the switching hub and synchronized transmittions are available for multi point network.

	throughput of one sender(Mbps)	one connection(Mbps)
1×1	66	66
2×2	62	31
3×2	46	23
2×3	60	20
3×3	63	21
4×2	36	18

Table 7.4: the result of synchronized multi point test

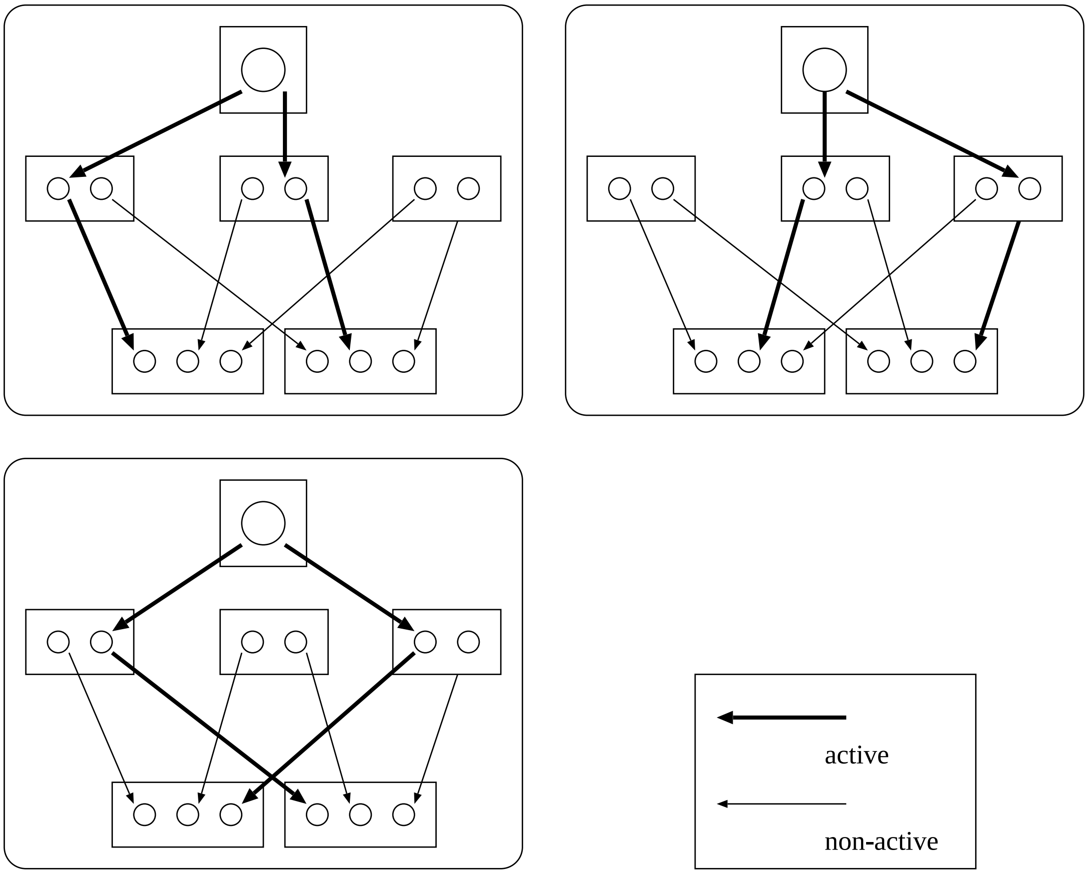


Figure 7.8: Timing table of 3×2 test: connections of one host are active twice, and non-active once

7.3.4 Expecting throughput

As described before, to expect throughputs are important. From now on, how to expect throughputs of synchronized setup are described.

When the number of senders and receivers are same or the number of senders is smaller than receivers, connections of senders are always active, and throughputs of one sender are maximum(66Mbps). The number of connections of one sender is equivalent to the number of receivers. So, throughputs are devided by the number of receivers. This can be shown as,

$$throughput\ \ of\ \ one\ \ sender = 66Mbps(constant)\ \ (if\ \ N.S \leq N.R) \tag{7.3}$$

$$expected\ \ throughput\ \ of\ \ one\ \ connection = \frac{66}{N.R} \tag{7.4}$$

If the number of senders is larger than receivers, as 4×2 and 3×2 case, connections of senders are not always active. The ratio of active and non-active is given by the ratio of the number of senders and receivers(see Figure 7.8). This is written as,

$$expected\ \ throughput\ \ of\ \ one\ \ sender = 66 \times \frac{N.R}{N.S}\ \ (if\ \ N.S > N.R) \tag{7.5}$$

$$expected\ \ throughput\ \ of\ \ one\ \ connection = \frac{66}{N.S} \tag{7.6}$$

where N.S and N.R are the number of senders and receivers. (7.5) shows that the throughput of one sender increses if the number of receiver increase. The throughput of 12×6 setup can be expected as 33Mbps(4Mbyte/s) per sender and 392Mbps(49MByte/s) by all connections, which can be expanded to 66Mbps(8Mbyte/s) and 792Mbps(99Mbyte/s) by using 12×12 setup.

7.3.5 Synchronized test using SPARC-5V

The event builder consists of SPARC boards, so the performance of SPARC-5V is needed. However, there are only two Fast Ethernet cards as described before, only the performance of 1×1 is measured. 10Base-T is used for the connection between the controll host and the sender. The result is shown in Table 7.5.

The measurment of S-7/400Ui shows that the throughput measured 1×1 becames a basic parameter, so it is expected that the measured value is available for multi point setup using SPARC as S-7 is. Using this assumption, the total throughput can be calculted as 150Mbps(19MByte/s) for 12×6 setup and 300Mbps(37.5MByte/s) for 12×12 setup.

7.4 Summary

In this chapter, multi point network is constructed in two ways, synchronized and non-synchronized, and performances are measured, which is summarized in Table 7.5.

	synchronized	non-synchronized	SPARC-5V
1×1	66	93	25
Throughput of one sender	$66 \times \frac{N.R}{N.S}$	$\frac{93}{N.S}$	$25 \times \frac{N.R}{N.S}$(expected)
Total throughput	66 × N.R	93	25 × N.R (expected)

Table 7.5: the summary of multi point test

Using the synchronized scheme, total throughputs of 12×6 setup are expected as 99Mbyte/s for S-7 and 19Mbyte/s for SPARC-5V, which satisfied the demand of the BELLE event builder. These setups are scalable, so throughputs can increase by the number of receivers.

The important points to construct the network based event builder are to keep packet size more than 3 event (∼13 KByte) and not to send same receivers simultaneously. If we keep these two points, we can construct the network based event builder easily.

Chapter 8

The possibility of the network based event builder

The performances of the network based event builder is expected in Chapter 7. In this chapter, problems which ocurr in applying for BELLE are discussed at first, and the possiblity is discussed next.

8.1 Application for BELLE experiment

In the test of multi nodes, the total throughput of 12×6 network based event builder using SPARC-5V is expected as 18.5MByte/s. This performance seems to be satisfied with the requirement of BELLE DAQ system 15MByte/s. In this section, I will describe what is needed when the network based event builder is introduced in BELLE.

8.1.1 subsystem

The required data rate for BELLE DAQ system is shown in Table 8.1 again. It can be seen that the data size of CDC is especially larger than other subdetectors. The throughput per one sender is allowed

Subdetector	data size(KByte)	data rate/subsystem(Mbps)
SVD	16	16
CDC	8	32
ECL	3	12
TRG	3	12
KLM	2	8
others	2	8

Table 8.1: the required data size of subdetectors

to 12.5Mbps in 12×6 setup. The required performance of subsystems of SVD and CDC exceeds, and ECL and TRG are in danger of saturation .

In order to be satisfied with the requirement of subdetectors except CDC, we have to use 12×12 setup if possible, or at least 12×8 setup, by which we can get throughputs more than 17Mbps per one sender.

For CDC subsystem, the easiest way to solve the problem is to change SPARC-7V to SPARC-50T. Using SPARC-50T, we can expand the throughput about twice as fast as the throughput of SPARC-5V[1]. This speed is enough for CDC subsystem.

Also, there is another plan, which is to devide CDC subsystem into two. The data rate of one CDC subsystem becomes a half, which SPARC-5V can deal with.

The basic problem for subsysytem is that the performance of SPARC-7V and FastEthernet card is poor. There is no better solutions than increasing performance by improving Drivers.

8.1.2 EB-N

The receiving ability for a SPARC-7V is measured as 25Mbps, which is satisfied with the required ability for one EB-N. But it may take some overhead for online farm to read event data, so it is not sure that SPARC-7V can be satisfied with the required data rate for one EB-N.

One of the most important merit of the network based event builder is that we can use PC for online farm and we don't need to use SPARC board and VME Crate which limite the performance 25Mbps. Now Linux based DAQ system begins to be discussed. Using PC, we can easily construct 12×12 setup and get enough performances.

8.1.3 timing matrix of event sending

The throughput 25Mbps means that the packet which size is 25KByte is sent to destinations at 125Hz. In order to achieve 500Hz, we need to send more than 4 event once. This time, I think of the case 6 event are packed and sent simultaneously. Packing 6event, the packet size increases to some extent. We can send the packet efficiently.

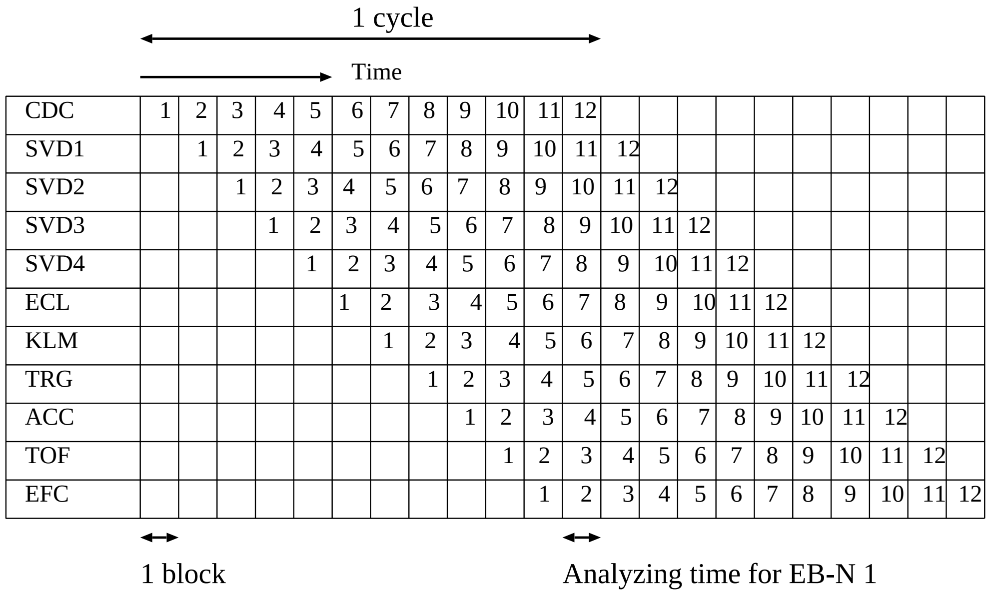
The example of timing matrix in 11×12 is shown in Figure 8.1. Each subsystems send packets after reading out 6 event. One block means 6 events, and one cycle 6×12=72 event. After reading out 6 event from subsystems, packets are sent to EB-N. Basically, DAQ system is synchronized by trigger system, so we don't need to worry about packet mixing. We have only to read and send data.

In Figure 8.1, the order of subsystem is decided by the data size. In EFC, the event data which should be sent to EB-N 1 must wait for 10 blocks. It means EFC subsystem must store 6×10=60 event. The data size of EFC is the smallest of all. If the data size is large, it is difficult to store many event data. This is why the order is decided by the data size.

In EB-N 1, the full event data are completed by receiving from EFC subsystem. After receiving from EFC, there are one vacant block for EB-N 1. This is because there are only 11 subsystem, and EB-N 1 can analyze the event data in this period. Even if it takes long time for EB-N to analyzing 6 reconstructed event data, it won't be serious problem because we can obtain the analyzing time only by increasing the number of EB-N.

---

[1] reported by E.Inoue http://onlax2.kek.jp/ inoue/



CDC	1	2	3	4	5	6	7	8	9	10	11	12										
SVD1		1	2	3	4	5	6	7	8	9	10	11	12									
SVD2			1	2	3	4	5	6	7	8	9	10	11	12								
SVD3				1	2	3	4	5	6	7	8	9	10	11	12							
SVD4					1	2	3	4	5	6	7	8	9	10	11	12						
ECL						1	2	3	4	5	6	7	8	9	10	11	12					
KLM							1	2	3	4	5	6	7	8	9	10	11	12				
TRG								1	2	3	4	5	6	7	8	9	10	11	12			
ACC									1	2	3	4	5	6	7	8	9	10	11	12		
TOF										1	2	3	4	5	6	7	8	9	10	11	12	
EFC											1	2	3	4	5	6	7	8	9	10	11	12

Figure 8.1: Timing matrix of 12×12 setup

8.1.4 summary

To introduce the network based evnet builder in BELLE, we should change DAQ system to more than 12×12 setup. If we can improve the Driver of FastEthernet card, we don't need to change setup, and we have only to change G-Link to FastEthernet card. However it is difficult both to extend present 12×6 setup and to improve the Driver, we should wait for introducing the network based event builder till online computer farm is changed to PC based one.

8.2 enlargement

The total throughput of the network based evnet builder depends on the number of hosts and their basic components. In this section, the enlargement of the network based evnet builder is discussed.

8.2.1 the number of hosts

It seems that the total performace increases till the total sending ability of IB-N by increasing the number of EB-N. But there is a limit in the number of socket connection which one host can open. When we want to construct 32×32 event builder for example, we can't connect IB-N and EB-N directly. we have to think of new scheme.

8.2.2 the influence of basic performance

In the test of chapter 7, the influence of basic performance can be seen. If 100Base-TX is replaced with 1000Base, the total throughput is expected to be 10 times faster. But we need to reconstruct packets in EB-N. Though the throughput is expected to be limitted by the performance of EB-N, we can get enough performance of ∼1000Mbyte/s by using 1000Base ethernet, excellent switching hub and dozens of EB-N.

Chapter 9

summary and Future Prospects

The goal for this paper is to construct and measure the performance of the test system, and establish the ways to construct the network based evnet builder. It is expected that the network based evnet builder wil be introduced to almost all the experiments with regardless of their scales, because it is powerful, cheap, reliable, flexible and easy to operate. So, we have to establish the ways to construct the network based evnet builder.

The study begins from the selection of the devices. Though there are a lot of conventional devices for Ethernet, we have to select one device by checking that it satisfies the required performance. The devices are reliable, but they have own features each other. To know their features is more important in constructing the network based evnet builder.

All the programs for the network based event builder is the network program. In order to make it easy, the library KONOE is introduced. KONOE is designed to construct DAQ programs, which gives full plays to develope programs of measurments and socket trasnmissions. Also, the benchmark tool DBS is used to test the performance of devices. Using KONOE and DBS, the development of programs becomes very easy.

One of the most important targets in this paper is to know the characteristics of the network based event builder. The test system is construted by using 2∼7 hosts. Though there are many troubles, I succeed to construct the test system which has scalabilities. The important points are to keep the packet size larger and not to send same receivers simultaneously. To have scalabilities means that total throughput can be expanded by the sum of the transmission ability of IB-N. By scalabilities, the flexiblity, which is one of the most important features, is established.

The network based event builder is expected to be introduced in BELLE, so the study for applying it for BELLE is done at the same time. For BELLE, the basic performance of SPARC board and Fast Ethernet card is not so enough that it is difficult to be used in present 12×6 setup. The network based event builder will be introduced when DAQ system is upgrade.

As noted before, the network based event builder has various merits. It is expected that to establish the ways to construct the network based event builder can be applied not only for BELLE, but for various high energy experiments and nuclear experiments which are planed in future.

Appendix A

switch test

Appendix B

data reconstruction

Bibliography

[1] Y.Ushiroda,"Development of the BELLE Central Trigger System Global Decision Logic", Master's Thesis, Dept. of Phys. Kyoto univ. (1998)

[2] H.Matsuo,"Development of the BELLE event builder", Master's Thesis, Dept. of Phys. Kyoto univ. (1997)

[3] Y.Murayama,"Design and Implementation of DBS: a performance evaluation system for multi-point TCP connections", Master's Thesis, Nara Institute of Science and Technology (1996)

[4] The BELLE Collaboration, "a Study of CP Violation in B Meson Decays Technical Design Report",KEKK Report 95-1 (1995)

[5] R.Breyer and S Riley,"Switched and Fast Ethernet,Second Edition", ZIFF DAQVIS PRESS (1996)

[6] The BELLE Collaboration, "Letter of Intent for A Study of CP Violation in B Meson Decays",KEKK Report 94-2 (1994)