ZEUS-Note 96-094
Sept. 1996

ZEUS NETWORK SWITCH CONNECTIONS PROJECT

(simulation study)

S.G.Basiladze', A.V.Kalagov', V.M.Rybnikov'
C.Youngman"
'- Moscow State University, NPI, NEAlab
"- DESY, ZEUS, Hamburg

Abstract

An investigation of the use of Network Switch Connection technology for the ongoing upgrade of the ZEUS detectors event building and TLT farm based on the solutions developed by the computer and network communications industry is described in the Note.

Current GSLT-EVB-TLT interconnect conditions, the advantages and limitations of using Network Switch Connections are considered in the first part of the Note. The modular system in C++ for DAQ-simulation is described in the second part. The modelling results of the current EVB and original TLT and several possible EVB-TLT upgrade versions are discussed in the third part.

The contents of this paper corresponds to the situation that took place at the beginning of 1996. The development of simulation tools allowing benchmark results to be used predictively in evaluating the event building performance.

1 Introduction

The current ZEUS Event Builder (EVB) was designed in 1991-92. Its task is to combine component front-end data into complete events, these are passed to the experiments Third Level Trigger (TLT) farm. Additionally the EVB is responsible for distributing the Global Second Level Trigger (GSLT) decision.

The EVB design foresaw input rates of ca.100 1/s with an event size of ca.150 Kbytes [1]. The rates seen during the first three years of HERA operation are shown in Table 1.

Table 1

HERA and ZEUS parameters evolution

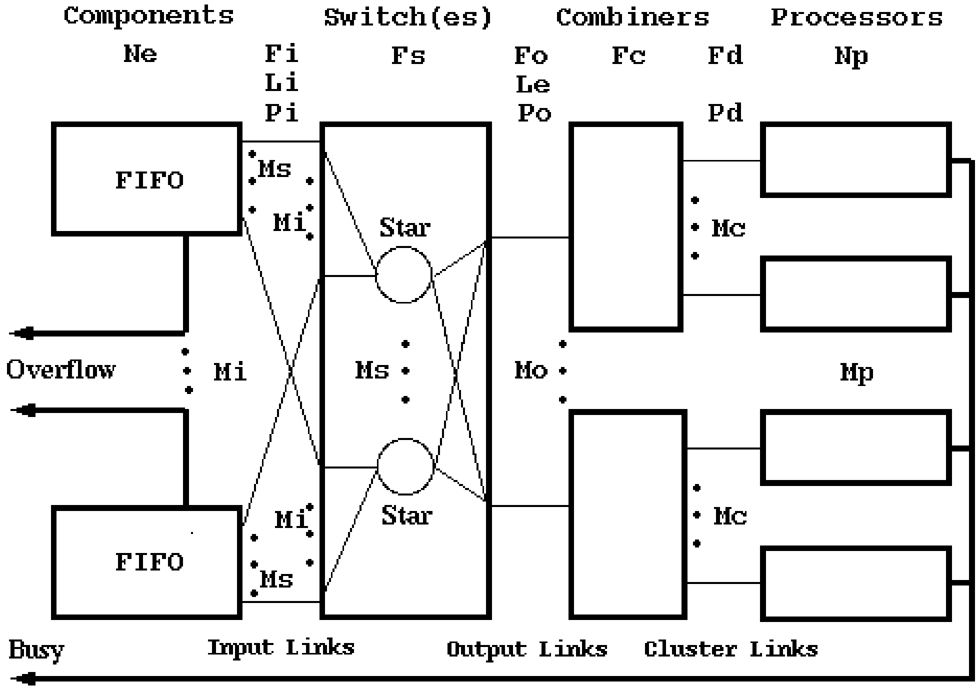
Year		Hera parameters			ZEUS parameters		
	Number of Bunches	Ie mA	Ip mA	Log10(L) (1/cm**2)*(1/s)	GFLT 1/s	GSLT 1/s	TLT 1/s
1992	9	2.0	1.5	27.6	20	10	5
1993	84	12.5	12.5	29.0	80	30	2
1994	180	30	40	30	150	15	3
1995	180	45	25	30.3	450	60	10
1996	180	65	35	-	700	90	15
Design	210	60	160	31	1000?	200?	30?

1.1 Current GSLT-EVB-TLT interconnect conditions

The general structure of a SLT-EVB-TLT interconnection, based on a switch, is shown in Fig.1.

The picture contains four layers of structural elements: Components, Switches, Combiners and Processors, and three layers of communication links: Input, Output and Cluster links. Each communication link may consist of several parallel connections (P) thereby increasing the total bandwidth; for instance, a byte line (8 conductors) has P=8 in comparison with a single serial link conductor with P=1.

The behavior of the entire system is determined by the parameters Ne through Np (see Fig.1). The overall throughput performance of the system is equal to Ne*Le. Conditions of operability can be defined by requiring that the throughput and bandwidth of each layer must not be less than this value. These inequalities, a)-f), are shown in Fig.1.



Definitions:

Ne - average rate of events
Le - length of event,
Li - length of component,
Mi - number of components,
Fi - input link bandwith,
Ms - number of stars,
Fs - star-switch bandwith,
Mo - number of switch outputs,
Fo - output link bandwith,
Mc - number of TLT-processors in the cluster,
Fc - cluster demultiplexor bandwith,
Fd - cluster link bandwith,
Mp - number of TLT-processors,
Np - average processor performance,

P_* - factor of line parallelism.

Operability conditions:

a) Fi > Ne*Li/Pi;
b) Fs > Ne*Le/Ms;
c) Fo > Ne*Le*Mc/Mp*Po;
d) Fc > Ne*Le*Mc/Mp;
e) Fd > Ne*Le/Mp*Pd;
f) Np > Ne/Mp;

Figure 1: The general structure of GSLT-EVB-TLT interconnection

In the EVB-TLT:	The maximum throughput rate (Ne) of each layer may be defined from inequalities a)-f);
Ne = 50 1/s,	
Le = 72 KB,	
Li = 16 KB (BCAL),	for the input links:
Mi = 20, Pi = 3,	Ne_max = Fi*Pi/Li = 110 ev/s, (1,a)
Fi = 0.6 MB/s,	for the switch:
Ms = 6,	Ne_max = Fs*Ms/Le = 2000 ev/s, (1,b)
Fs = 24 MB/s,	for the output links:
Mo = 6, Po = 8,	Ne_max = Fo*Mp*Po/Le*Mc=400 ev/s, (1,c)
Fo = 0.6 MB/s,	for the combiners:
Mc = 5, Pd=1/3,	Ne_max = Fc*Mp/Le*Mc = 410 ev/s, (1,d)
Fc = 5 MB/s,	for the cluster links:
Fd = 2 MB/s,	Ne_max = Fd*Mp*Pd/Le = 280 ev/s, (1,e)
Mp = 6*5=30,	for TLT-processors:
Np = 2.5 ev/s.	Ne_max = Np*Mp = 70 ev/s. (1,f)

As it can be seen the original GSLT-EVB-TLT [1-3] system performance has been adequate, but for the future improvements in HERA delivered luminosity the main ZEUS limitation factor now is the TLT-processors computing power.

An upgrade of the TLT computing power proposed by the TLT group is underway. In the case of TLT processors replacement (the same Mp):
Np = 7.5 ev/s, the rate of TLT-input events may be increased up to:
Ne = 230 1/s
and will be corresponded to the possible future GSLT rate (Table 1). But unfortunately, then current input (1,a), output (1,c) and cluster (1,e) links bandwidth will not be enough. The situation may become more nonacceptable if the length of component banks will be increased. For instance, if
Le=140 KB, then:

for the input links:	Ne_max = 55 ev/s,
for the output links:	Ne_max = 200 ev/s,
for the cluster links:	Ne_max = 140 ev/s.

It means that future HERA progress in the luminosity may make an upgrade of the GSLT-EVB-TLT system essential.

Previously, 1991-95, the experiment's TLT consisted of SGI workstations. They were connected to the EVB VME output buffers via the FNAL branch bus which is directly connected to the SGI's backplane.

The TLT upgrade cannot take the form of extending the branch bus and purchasing additional identical SGI machines. The reasons for this are the developments which have occurred since the initial design. Modern, high performance, SGI workstations no longer have a VME backplane and the FNAL branch bus concept has been matched or even exceeded by solutions developed by the computer and network communications industry. FDDI, ATM and HIPPI connections are examples of the link technologies currently available, see Table 2 [4-6].

Table 2

High Performance Interconnects

Parameters		Fubus+	SCI	HIPPI	FibChan	FDDI	ATM
Topology		Bus	Ring PtP	PtP	PtP	D-ring PtP	Star
Medium		Copper	Copper	Copper Fiber	Fiber	Fiber	Fiber Copper
Mode	bit	256 par	16 par	32 par	1 ser 1 ser	1 ser	1 ser
Bandwidth	MB/s	1800	1000	100	100	12.5	19.3 (125)
Distance	m	0.5	10	25	1000	2000	2000
Packet size	B	Free	256	1024		4500	48
Chip set		TI	Dolfin	Avail.	Avail.	NS, AMD	Trans. Corp.
Industrial support		Weak	None	Yes	IBM, HP	Strong	Strong
Node price	$			5000		1600	4000

PtP - Point-to-Point
par - parallel
ser - serial

Network Switch devices are currently popular when discussing the point to point connectivity previously provided by the branch bus, see Table 3 [5-8].

Table 3

Network Switches

Parameters		FDDI links	ATM links	
		DEC_GS	Gen.Datacom.	Hughes NS
Switching throughput,	MB/s	450	800	312
Maximum ports		34	32	16
Buffer size / port,	kB		64	512
Other links		ATM	Ethernet	-
Price	k$	50	125	90

In this paper we present an investigation of the use of Network Switch Connection (NSC) technology w.r.t. the ongoing upgrade of the ZEUS detectors event building and TLT farm.

2 Some remarks about the functional requirements of interconnection layers

The most efficient solution to a real design will be achieved when the loading level of the structural and communication links should be nearly 100 percent.

If only one of the layers is the main bottle-neck for the data flow then its throughput or bandwidth (1/F) will define the overall system dead time. The dead time leads (with increasing data input rate) to an increase in the input data loss, when data is rejected, or in the data flow blocking rate when data flow control is implemented, This is shown in Fig.2 where a simple shaper unit (Tout=1/F) is applied to random pulse interval generator. Flow control is provided by NANDing the shaper output with the generator. As can be seen from the picture: Effi=Output_Pulse_Number/Input_Pulse_Number is near to 100 percent when the ratio P_in/T_out is 2-4 for random and is close to 1 for regular input periods.

2.1 Components

The input data flow regularization can be carried out through FIFO memories on the GSLT component outputs. The simulation results of FIFO overflow losses (due to limited depth), the latency and queue occupancy values may be taken from Fig.3 as functions of the ratio of the (regular) read period - T_read to the (average) write period - T_write. As can be seen FIFO depths of >4 are sufficient to regulate random input stream with small losses.

The FIFO nearly full occupancy signals, from all the components, should be OR'd (see Fig.1) thus blocking further event input. This data flow control prevents FIFO overflow and possible data corruption.

The ratio of the component data length to the input link bandwidth is the time needed for transferring component data to the EVB switch input:

$$\mathrm{Ti} = \mathrm{Li}/\mathrm{Fi}. \tag{2}$$

Real component data lengths may differ substantially, therefore the component with the maximum data length may limit the rate of data transfer; from the operability condition a) it is required that:

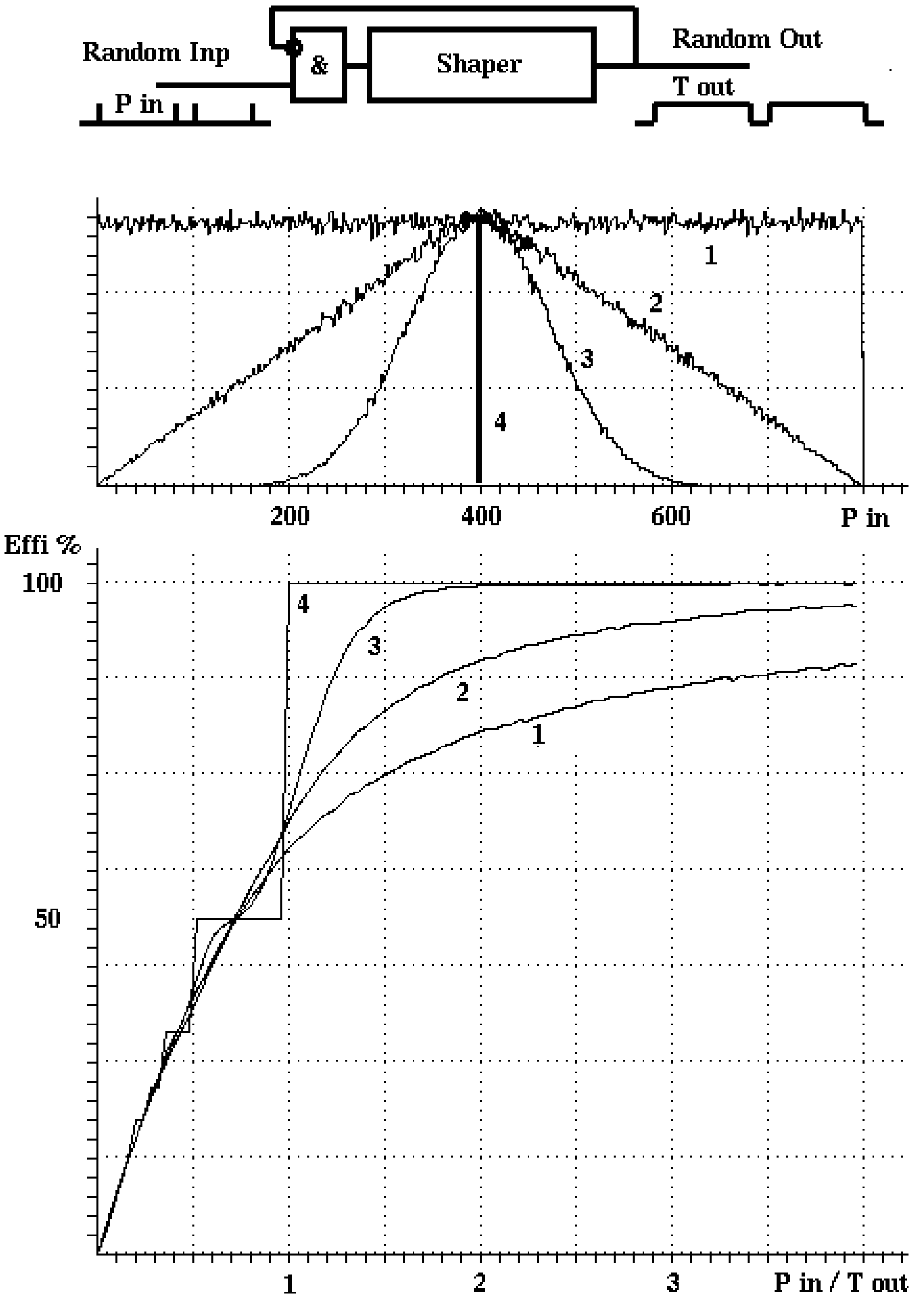


Figure 2: The efiiciency of Shaper for the random input/output flow

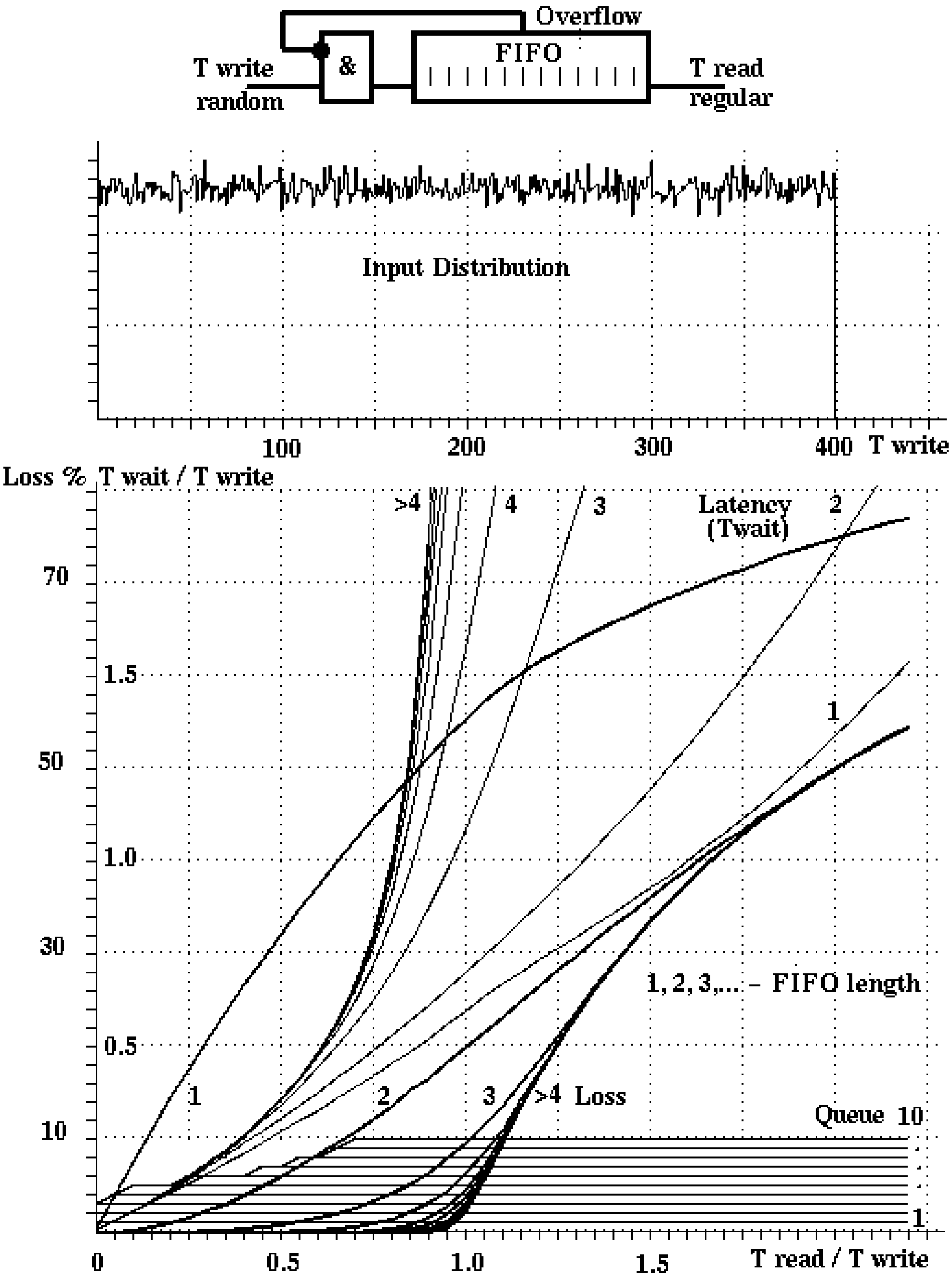


Figure 3: Queue, Latency and Loss distributions for FIFO

$$1/(\mathrm{Ti})\mathrm{max} > \mathrm{Ne}. \tag{3}$$

If this condition is not fulfilled, at the communications link layer, then the data of the component with maximum length should be divided into several parts (demultiplexed) in such a way that for every part this condition would be fulfilled. Conversely, components with small data lengths may be merged (multiplexed) in the common data flow, under the condition:

$$1/[\mathrm{SUM}\ (\mathrm{Ti})\mathrm{min}] > \mathrm{Ne}. \tag{4}$$

2.2 Switch

The Event Builder switch section has non-symmetrical functions (it combines dispersed event data from the components at its outputs). If the switch section is based on the Bus interconnection (VME, for instance) then Ms=1. In the case of Barrel shifter or Crossbar based switches Ms is equal to the number of processor clusters; each cluster is connected to one combiner and may contain more than 1 processor. The advantageous of the Crossbar is that its switches operate independently (asynchronous) whereas the Barrel shifter switches operate concurrently (synchronously).

In the custom designed EVB, when Ms>>1, a number of parallel output lines Po may be used for each processor cluster. The output link bandwidth, Fo, in this case may be less in Po times, because the EVB is able to combine the event data from several components in parallel. Additionally some parallel input lines may be used for one component data transfer. The original ZEUS EVB, see Fig.4, contains 8 parallel output links (Po) to each TLT cluster (T800) and 3 input links (Pi).

Every switch output (if they are autonomous) should have the throughput defined by inequality b).

2.3 Combiners

The data component combiners have two buffer memories: a components input buffer and an event output buffer. The combiner operability condition d) coincides with condition b), if Ms=Mo so far as Mo=Mp/Mc. Then the required number of combiners is:

$$\mathrm{Mo} > \mathrm{Ne}^{*}\mathrm{Le}/\mathrm{Fc}. \tag{5}$$

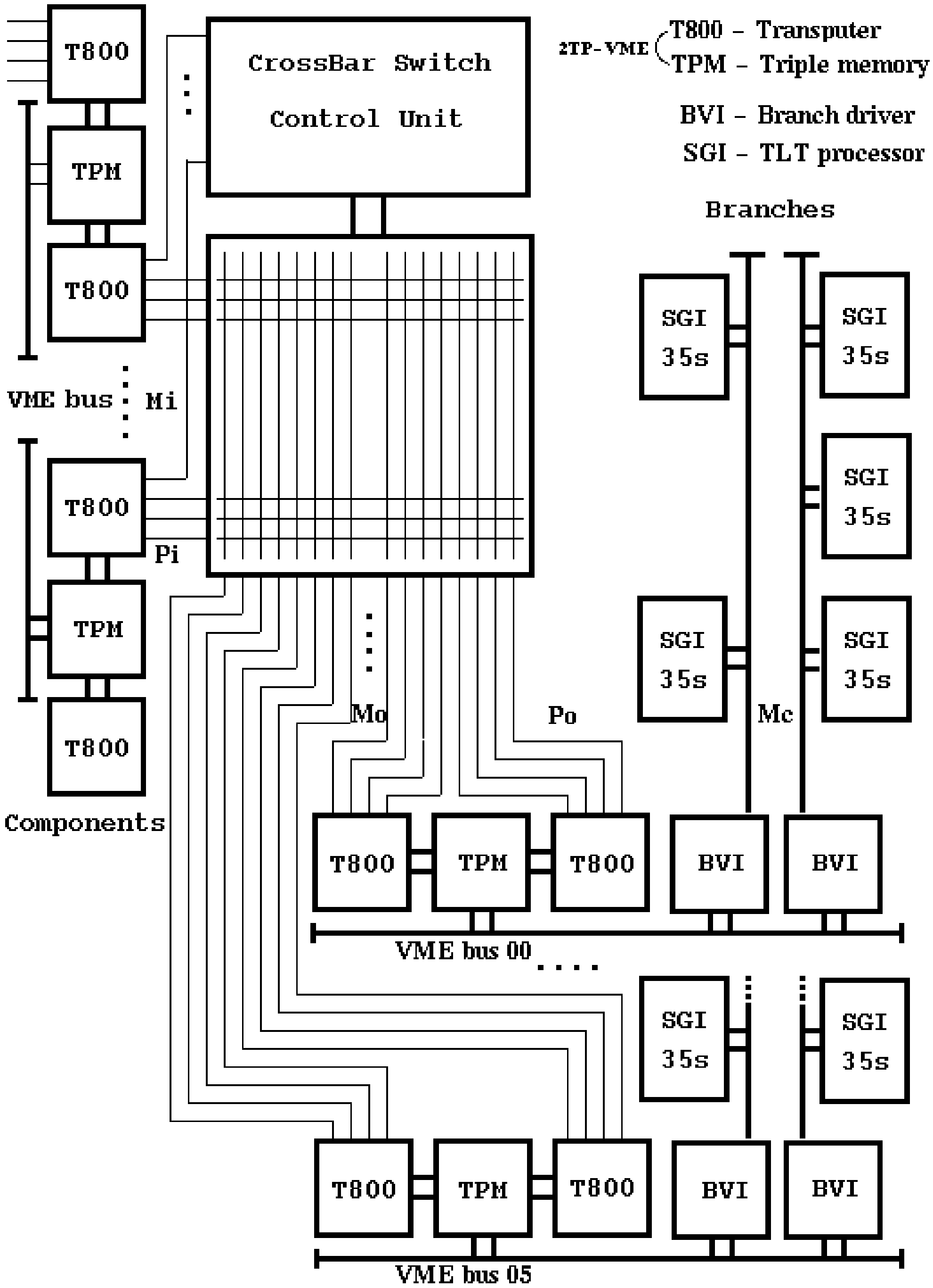


Figure 4: ZEUS EventBuilder and original TLT hardware

2.4 TLT processors

The number of TLT processors follows directly from the condition f):

$$Mp > Ne/Np, \tag{6}$$

where the corresponding number of processors in the each cluster (for 1 combiner) is:

$$Mc = Mp/Mo. \tag{7}$$

If the number of processors are less than (7) or if the components have insufficient smoothing (regularization) FIFO buffers then feedback flow control is necessary, see Fig.1, to block input.

The event distribution control among the processors may be based on the routine algorithms: priority, round robin or random distribution. The result will not depend on the distribution mechanism if the latter provides a processor loading level near to the maximum.

2.5 Data flow control

Two main operations are the basis of data flow control in the EVB:
1) a definite mechanism of free TLT-processor (address) specification for the new event;
2) a mechanism of data flow blocking on the system input if the all units in one of its layers are fully loaded or near to overflow.

TLT-processor address specification may be carried out
- at the moment of the new event arrives on the system input, then the address of a free processor is written into each component destination target address location;
- at the moment when the combiner completes assembling an event, then the processor address is written as the event destination point (in this case the component destination target address is specified as a combiner number using a round-robin distribution mechanism.)

The first specification mechanism is necessary if the processors assemble the event from components themselves, or when the number of combiners is equal to the number of TLT-processors (Mo=Mp and Mc=1). This mechanism leads to the pauses in the TLT-processors operation, the duration of which is equal to the event building time. As a result the TLT-processor efficiency decrease to:

$$Ep = Fo \ / \ (Fo + Le^{*}Np). \tag{8}$$

The second specification mechanism has no such drawback, but FIFO-memory for assembled events is required in the combiners and the combiner bandwidth performance should be higher than the processor attached:

$$Fc > Np*Mc. \tag{9}$$

The data flow blocking mechanism (the synchronization) may be realized by means of:
- "global" feedbacks from "all busy" units in each layer to the system input (all the processors Busy signal as shown in Fig.1);
- "local" feedbacks to the upper stream data flow units (is shown in Fig.1 as Overflow signals from the FIFOs).

The first synchronization mechanism is preferable if only one of the layer (for instance, processors) is the main limitation factor in the system. The choice of the synchronization mechanism has no influence on the system throughout rate.

3 Network Switch Connections in ZEUS

In the case of using 100 Mbit/s (12 MB/s) LAN interface for input/output links with new TLT processors and if:

Le = 140000 B,	for the input links:	Ne_max = 375 ev/s,
Li = 32000 B,	for the output links:	Ne_max = 514 ev/s,
	for the cluster links:	Ne_max = 514 ev/s.

i.e. such interface as FDDI has facilities adequate to EVB/TLT with more than 10 times higher Ne*Le. Note that this is a comparison of bandwidth and does not include control overheads.

Table 4

Event Distribution Investigations [4,9,10]

Parameters		**ZEUS switch**	**NA48**	**FNAL**	**LHC**	**(SSC)**	**KEK**
Ne - rate of SLT events	**1/s**	**200**	**300**	**10000**	**25000**	**1000**	
Le - length of event	**kB**	**100**			**75**	**1000**	
Mi - number of SLT components		**20**		**64**	**16384**	**400**	**1000**
Li - length of components data	**kB**	**20**				**2.5**	
Fi - EVB input links bandwidth	**MB/s**	**5**	**100**	**20**	**19.3**	**125**	**100**
Ms - number of switch stars		**10**		**1**	**16**	**3**	**4**
Fs - star--switch bandwidth	**MB/s**	**12**		**1000**	**12500**	**100000**	**125**
Fo - EVB output links bandwidth	**MB/s**	**5**	**100**	**20**	**19.3**	**125**	
Mp - number of TLT processors		**30**		**64**	**2048**	**1000**	**500**
Np - processsor performance	**ev/s**	**7.5**	**-**	**-**	**-**	**-**	**-**
Input links		**FDDI**	**HIPPI**	**(ATM)**	**ATM**	**FibChan**	
Switch(es)		**DEC_GS**	**NetwSC**	**FNAL**	**Alcatel**	**ANCOR**	**KEK**
Output links		**FDDI**	**HIPPI**	**(ATM)**	**ATM**	**FibChan**	

3.1 Advantages of using Network Switch

The advantages to the experiment of connecting TLT machines to a Network Switch are clear in that it will be cheaper than specific VME-bus interfaces and will allow all workstations to act as TLT nodes provided with a switch connection controller such as FDDI. The current limitation of only SGI type machines participating in the TLT will also be removed. The SG Challenge-S, purchased by Toronto group for the TLT, provide computing power (130 ms/ev) equivalent to three SGI 35s (400 ms/ev) used in the original TLT. Future machines will presumably provide even greater performance.

Replacing the internal databus of the TLT will lead to the possibility of replacing the experiments Event Builder. As a second phase it should be possible to connect the components sub-system crates directly to the Network Switch. This requires that the current EVB functionality of "building" events be moved to the TLT. The dataflow mechanism currently provided by the EVB will be transparently replaced by the Network Switch. This second phase will reduce the experiments need to maintain the EVB system.

We also assume that machines connected via FDDI will be able to communicate with those connected via, say, ATM. This is just a question of the driver software on the host machines and line adapter cards at the Network Switch. The ability of using TCP/IP between workstations connected via a Network Switch protects our investment since machines coming with faster connections, than FDDI which is currently available, such as ATM will be able to coexist with existing machines.

3.2 Limitations and restrictions

A number of boundary conditions exist for replacing the TLT bus by a Network Switch solution:

- FDDI is currently available and a number of controllers have already been purchased;
- a transition phase may be needed where "old" SGI 35's machines will have to coexist with Network Switch connected machines (non Network Switch mixed FDDI connections were tested by the TLT group).

Whatever happens the transition phase must always leave the experiment with a working data acquisition system.

3.3 Hardware required for implementing a Network Switch solution

The ZEUS experiments front end electronics use VME as a standard bus and power supply. Interconnection of VME crates is normally carried out with transputer modules such as the 2TP_VME module (see Fig.4). The principle requirement, when providing a smooth transition and an effective end solution, will require providing VME modules with an interface to FDDI.

The number of VME to FDDI connections required is ca.6 for the EVB-TLT interface upgrade, 1 for the IBMLINK and ca.20 for the second phase extension to the components. Whereby some of the EVB-TLT interfaces will be freed by the second phase implementation.

The commercial FDDI board solution is more preferable in the current situation and means that software design will be a large part of the NSC project.

3.4 What ZEUS was doing in terms of FDDI

Some activities at ZEUS have involved FDDI. They are briefly described below.

3.4.1 IBMLINK replacement

A replacement of the IBMLINK using FDDI is described in ZEUS Note 94-96. The FDDI-VME board tested was a NAT FDDI29, this board came with TCP/IP support on the driving OS9 CPU. An IBMLINK setup could have the form shown in the Fig.5. A feature of NAT board is that it requires controlling from a host CPU.

3.4.2 TLT improvement

The TLT group was developing the 6-th VME crate system. This system emulated the original 1 to 6 VME crate system in use in the experiment at least in terms of the EVB input and IBMLINK output. This is shown in Fig.6, the FDDI-VME boards were used from RADSTONE.

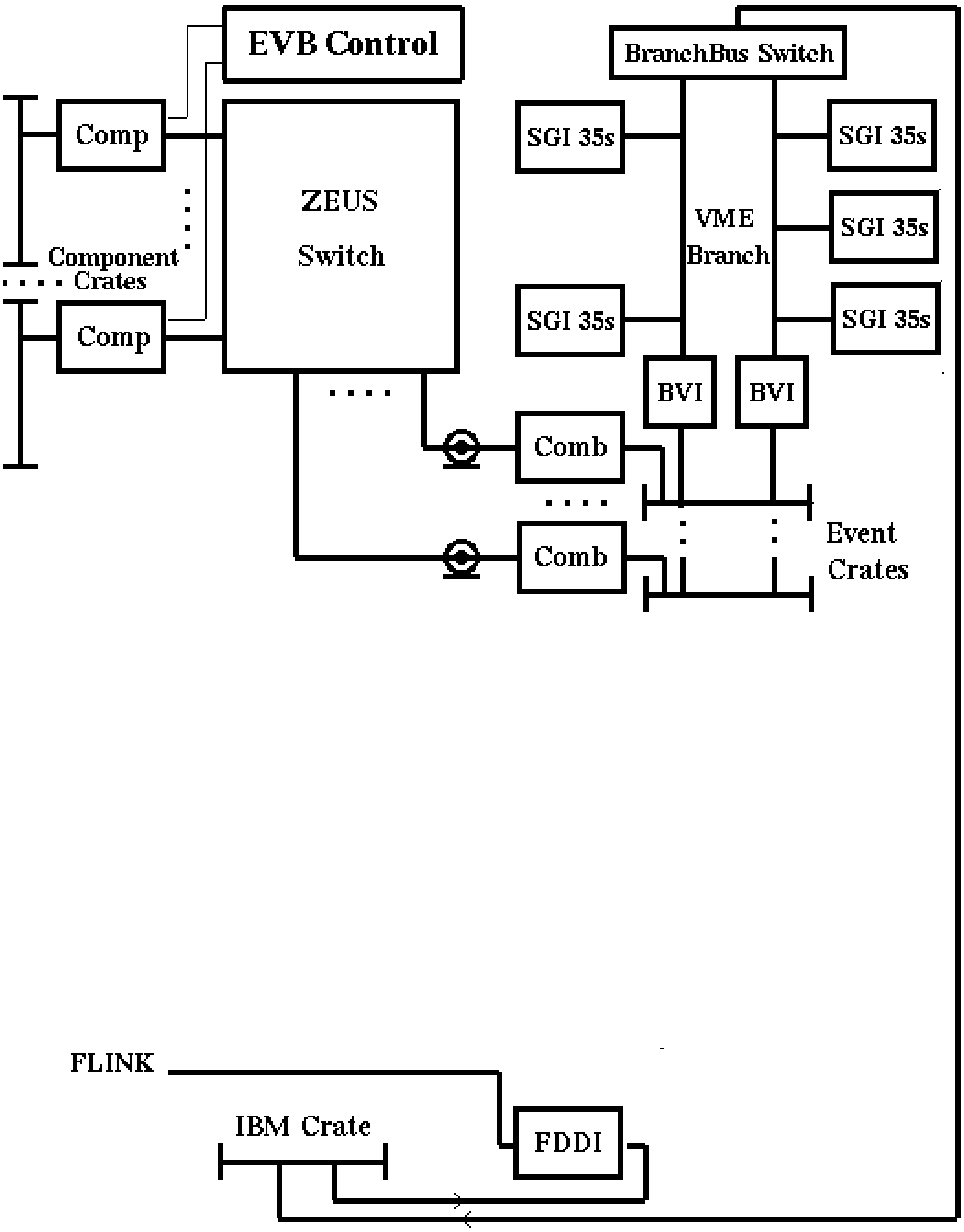


Figure 5: Phase 0,a: Applying FDDI in EVB-TLT IBM-link

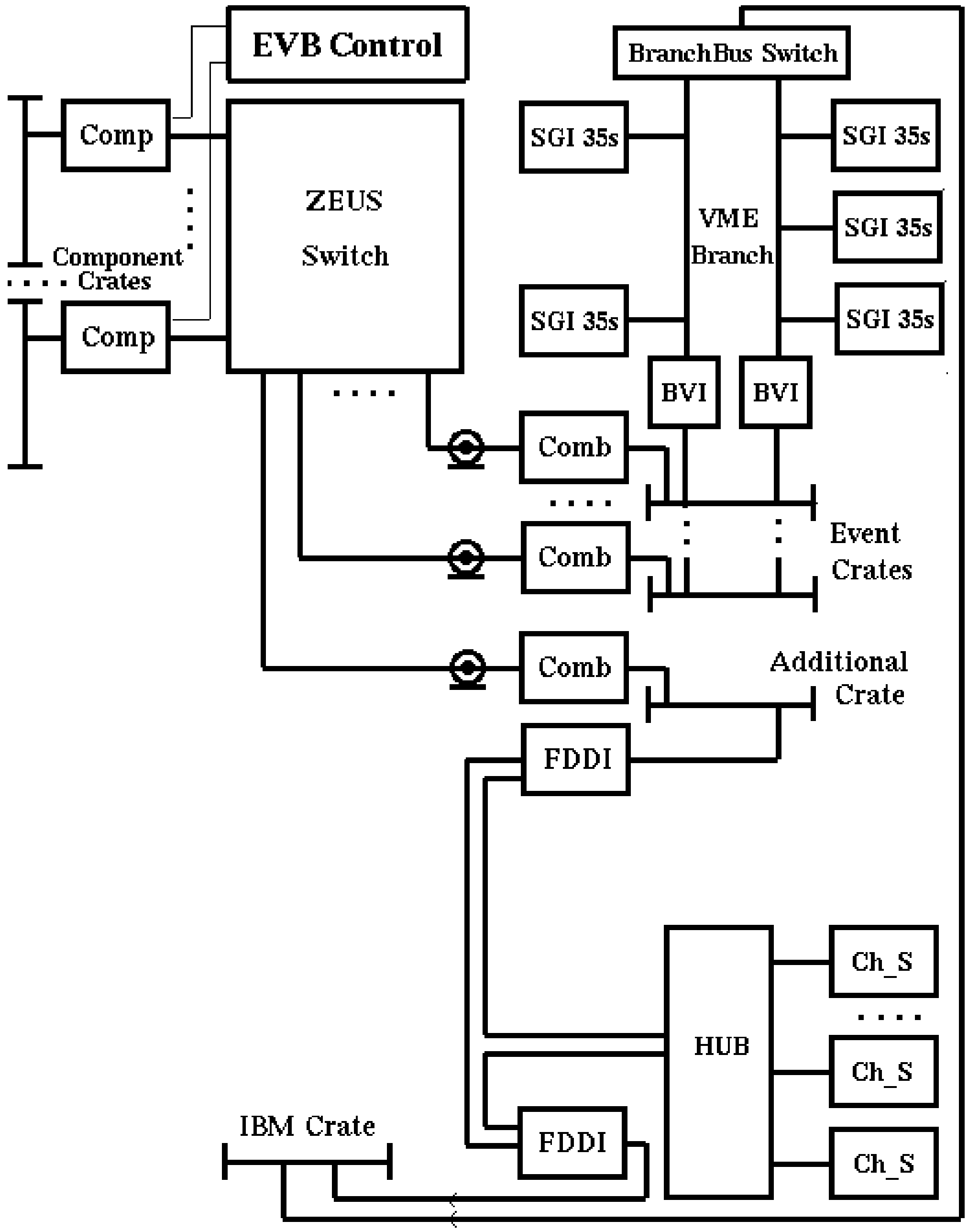


Figure 6: Phase 0,b: Using FDDI in the 6-th branch for TLT improvement

4 The modular system in C++ for DAQ-simulation

A number of programs such as Verilog, VHDL, MODSIM are available for computer simulation of NSC hardware. A common deficiency of such programs, due to their universality, is low emulation rate. Therefore C++ was used [11] for the custom oriented simulation programs design. Although the programming in a general purpose language like C++ is less straightforward, the run-time efficiency is much higher.

The described system is a set of programs realizing units for performance simulation of the nuclear electronics front-end, data aquisition and data transfer modules. The list of designed units is represented in the Table 5.

Table 5

Unit setup for simulation

Type	Function		Remark
101-2	Beam Event Random Generator		Multichannel
111	Pulse Shaper		Threshold and Delay control
121	Pipe Line Delay		Zero dead time
131	OR/AND Logic		Separate transition delay
132	Arbiter		Priority, round-robin, random
141	Master-Slave Interface		Addressed Slave
145	Multiplexor	M >- 1	Arbitered(M), addressed(1)
146	Mix-Demix	M >-< M	Common Link
147	Star Commutator	M >-< K	Arbitered(M), addressed(1)
148	Cross-Bar Switch	M >-< K	Arbitered(M), addressed(K)
151	FIFO memory for data blocks		Slave-Master write
161	Combiner of Events		Multicomponent
171	RAM memory for data blocks		Slave write-read
181	Token Network Adapter		Message fragmentation
191	Oscilloscope		Multichannel

The enumerated units may connect one to another for creating systems of required configuration. The units of types 141-181 carry out Master-Slave Handshake protocol for timing. The unified handshake lines are:
RQ - ReQuest for Slave, AK - AcKnowledge from Slave,
IR - InterRupt from Master, GR - GRant for Master.

The basic handshake protocols are given in the Table 6.

Table 6

Inter Unit Handshake protocols

<table>
<tr><th>Mode</th><th>Function</th><th>Protocol</th></tr>
<tr><td>Slave Write</td><td>=|W|</td><td>IRu -> GRu -> WMu-> WMd-> IRd -> GRd</td></tr>
<tr><td>Slave Read</td><td>=|R |</td><td>IRu -> GRu -> RMu -> RMd -> GRd -> IRd</td></tr>
<tr><td>Master Write</td><td>|W|o=</td><td>RQu -> AKu -> WSu -> WSd -> RQd -> AKd</td></tr>
<tr><td>Master Read</td><td>|R |o=</td><td>RQu -> AKu -> RSu -> RSd -> AKd -> RQd</td></tr>
<tr><td>Slave-Master</td><td>=W|o=</td><td>IRu -> RQu -> AKu -> GRu -> WMu-> WSu->
WMd -> IRd -> WSd -> RQd -> AKd -> GRd</td></tr>
<tr><td>Slave-Master</td><td>=R |o=</td><td>IRu -> RQu -> AKu -> GRu -> RSu -> RMu->
RSd -> AKd -> RMd -> GRd -> IRd -> RQd</td></tr>
</table>

<table>
<tr><td>= - Handshake</td><td>| - Interface border</td></tr>
<tr><td>o - Master</td><td>W,R- Write, Read</td></tr>
<tr><td>WM- Write data from Master,</td><td>RM - Read data for Master,</td></tr>
<tr><td>WS - Write data for Slave,</td><td>RS - Read data from Slave,</td></tr>
<tr><td>S*u - Signal transition-Up,</td><td>S*d - Signal transition-Down</td></tr>
</table>

Handshake lines are used also as command lines; besides all the units have "write" or "read" data lines. The command and data formats are shown in the Table 7.

Table 7

Unit Command and Data formats

<table>
<tr><td>Bits:</td><td>31</td><td>30 - 16</td><td>15</td><td>14 - 08</td><td>07 - 00</td></tr>
<tr><td>COMMAND:</td><td>[E:</td><td>Event Number</td><td>|W:</td><td>Destination Global :</td><td>Local]</td></tr>
<tr><td>DATA:</td><td>[*:</td><td>Tempo</td><td>|* :</td><td>Source Global :</td><td>Local]</td></tr>
<tr><td>TOKEN:</td><td>[*:</td><td>00...</td><td>|1 :</td><td>00...</td><td>]</td></tr>
</table>

E = Error   W = Write

For message routing the addresses of sources and destinations are subdivided into two - global and local parts. The Event number on the control line is a label of the message. As it can be seen from the Table 7 there is no real data on the data lines. Only the length of the message is represented as the information:

$$\text{Message length} = (\text{Message duration on data line}) * \text{TEMPO}, \quad (9)$$

the TEMPO value is asserted on the data line.

All the units have unified set of Control and Status Registers (CSR). As an example, the Interface Unit CSR set is listed below:

IDEN	of Unit	- Loc.Addr. 0
IDEN	of SuperVisor	- Loc.Addr. 1
NEXT	tick request	- Loc.Addr. 2
TR-O	TRansition of RQ	- Loc.Addr. 3
TR-1	TRansition of GR	- Loc.Addr. 4
TR-2	TRansition of DO	- Loc.Addr. 5
	1: Write	
	2: Read	
GADR	Global ADdRess iden	- Loc.Addr. 6
TOUS	TimeOUt Status	- Loc.Addr. 7
TOUT	TimeOUT value	- Loc.Addr. 8
LOSS	without AK-GR	- Loc.Addr. 9
MODE	unit state MODE	- Loc.Addr.10
	0: IDLE	
	1: IR-I arrived	
	2: RQ-I set if AK-I off	
	3: GR-I set if AK-I on (handshake)	
	4: DATA passing	
	5: IR-I removed	
	6: RQ-I remains on (if state = 2)	
	7: AK-I arrived (RQ-I reset)	
IR-P	IR Previous state	- Loc.Addr.11
AK-P	AK Previous state	- Loc.Addr.12
WM-P	WM Previous state	- Loc.Addr.13
RS-P	RS Previous state	- Loc.Addr.14
MX-P	MultipleXor Pointer	- Loc.Addr.15

The following parameters can be varied through CSR: the threshold, delay, duration, address iden and transmission tempo, timeouts of access or acknowledge waiting. The indication of the operation modes and the internal states of units are performed by means of these registers as well as diagnostics. As an example, the Interface Unit Losses messages are the following:

from:	-	up to:			
01	-	99	- if IR-I(a)u	and	MODE!=0 or
			- if DATA>0	when	MODE!=3 or 4
100	-	9900	- if IR-Iu	and	MODE!=0
10000	-	990000	- if AK-Iu	and	MODE!=2
1000000	-	99000000	- if IR-Id	and	DATA>0
100000000	-	9900000000	- if AK-Id	and	DATA>0

The main internal and external unit signals can be printed at any desirable time interval. The each unit can be selected for diagnostic printing through its unique iden in CSR (see an example above).

The short description of designed units is given below.

4.0.3 Units and their functions

Random Generator. This Unit imitates the physical events and their components random generation from the accelerator or after a trigger system. An optional number of the output channels are available for this unit (see Fig.7). Each channel has a RQ-AK handshake signals pair; the RQ signal can be asserted in the different ways:
- regular period,
- random distribution.

There are two versions of the Generator: 101 and 102. The first (more fast) version may generate "follows from flat" random distributions (see Fig.2,a):
- flat,
- triangular,
- Gaussian;

the left or right edges of 2 latter distributions may be cut off. The second version may generate the distributions of any form in accordance with the Preset Table. The signal period, distribution peak width and its position may be varied in the both versions.

Each channel can operate independently from others or synchroniously with the first channel.

When AK is received the data-out (DO) signal is asserted, its amplitude may be varied by means of the special function SetAmpl().

The series of signals generation is possible if the common input strobe signal is asserted. In other way one signal is generated when strobe signal transition-Up appears. Each channel has Busy input for blocking data

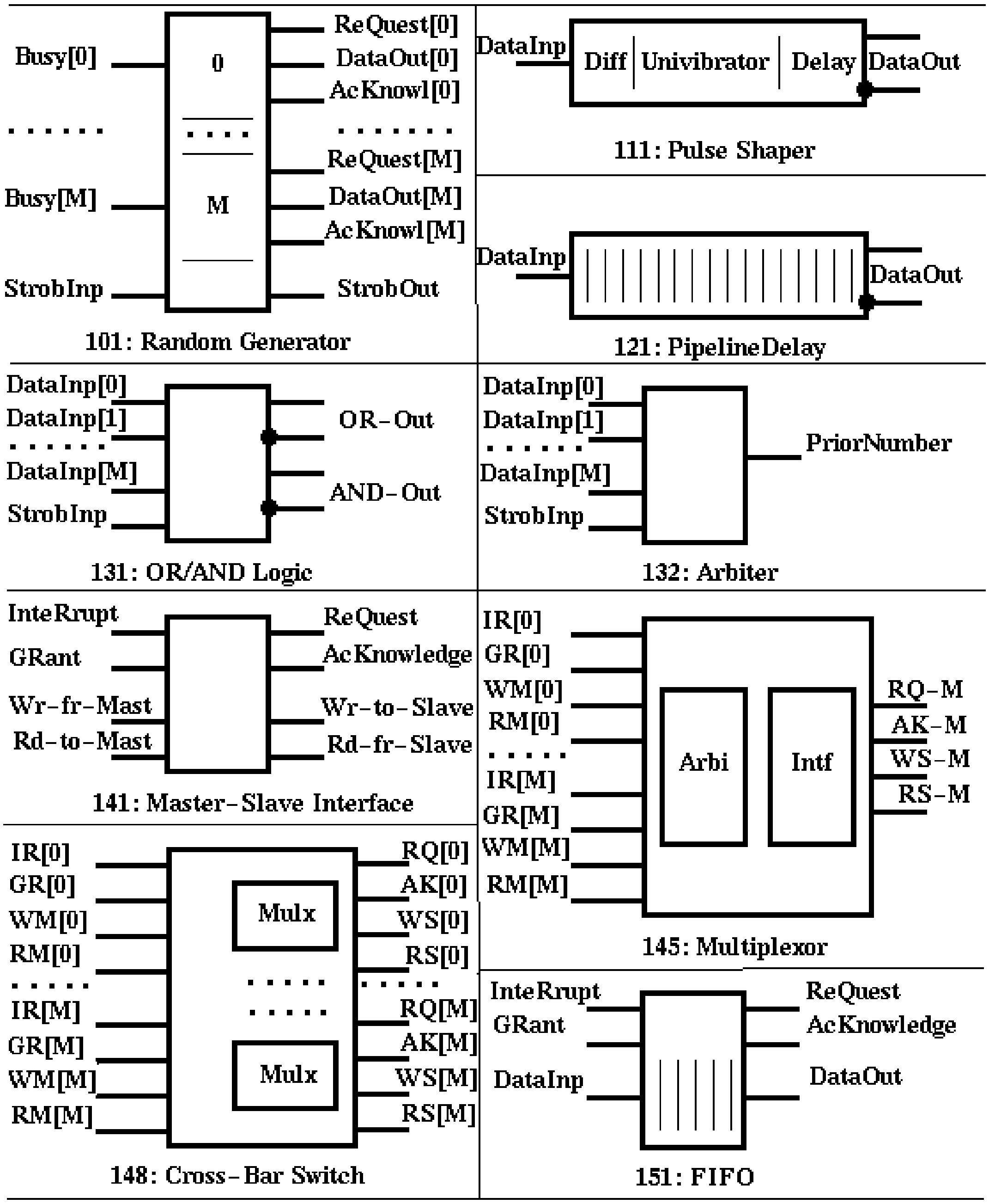


Figure 7: Tme main SIMU modules block-diagrams

output signals. Each AK signal, if asserted, delays generation in all the channels (ORed Wait).

As an example, the DESY ZEUS component latencies (SLTs decision delays) version 102 approximations for one of the real Run-94 are shown in Fig.8 (the curves No 00-07). When the last of SLT arrives GSLT decision is generated (see the curve 08); GSLT periods distribution approximation is represented as the curve No 09. The curve No 10 is the approximation of the TLT-processors decision time distribution. The approximations of ZEUS component data lengths distributions are shown in the Fig.9. The average data length for whole event is equal to 72 KB.

Shaper. This unit changes state if the input signal goes over a threshold level. The threshold level, output signal delay and duration values are stored in the CSR. The shaper dead time is equal to the sum of delay and duration values.

Pipe Line. This unit is a program controlled zero dead time delay for analog and digital signals.

Logic unit. Boolean logic unit implements OR/NOR and AND/NAND functions for optional number of input signals. The delays of transition-Up and transition-Down of output signals are controlled separately.

Arbiter. An optional number of the input channels is available for this unit (see Fig.7). If the input strobe-signal is asserted and the global destination address is equal to the Arbiter address iden (stored in CSR) or this address iden is negative then the Arbiter selects one of the input signals using the algorithm:
- logical or position priority, or
- round-robin, or
- random choice.

The selected input number is asserted on the output.

Interface unit. This unit provides IR-GR handshake from the Slave side and RQ-AK handshake from the Master side. The unit has no internal data buffer. If the global destination address iden is equal to the Interface address iden (stored in CSR) or if this address iden is negative (no address selection) then the Unit implements cooperate operation of Master and Slave sides accordingly 2 last protocols in the Table 2.

If AK signal was asserted before IR appearance it delays the beginning of the transaction cycle (no RQ assertion).

Multiplexor. This unit has an optional number of inputs and includes the following subunits: Arbiter, Commutator (controlled by Arbiter) and

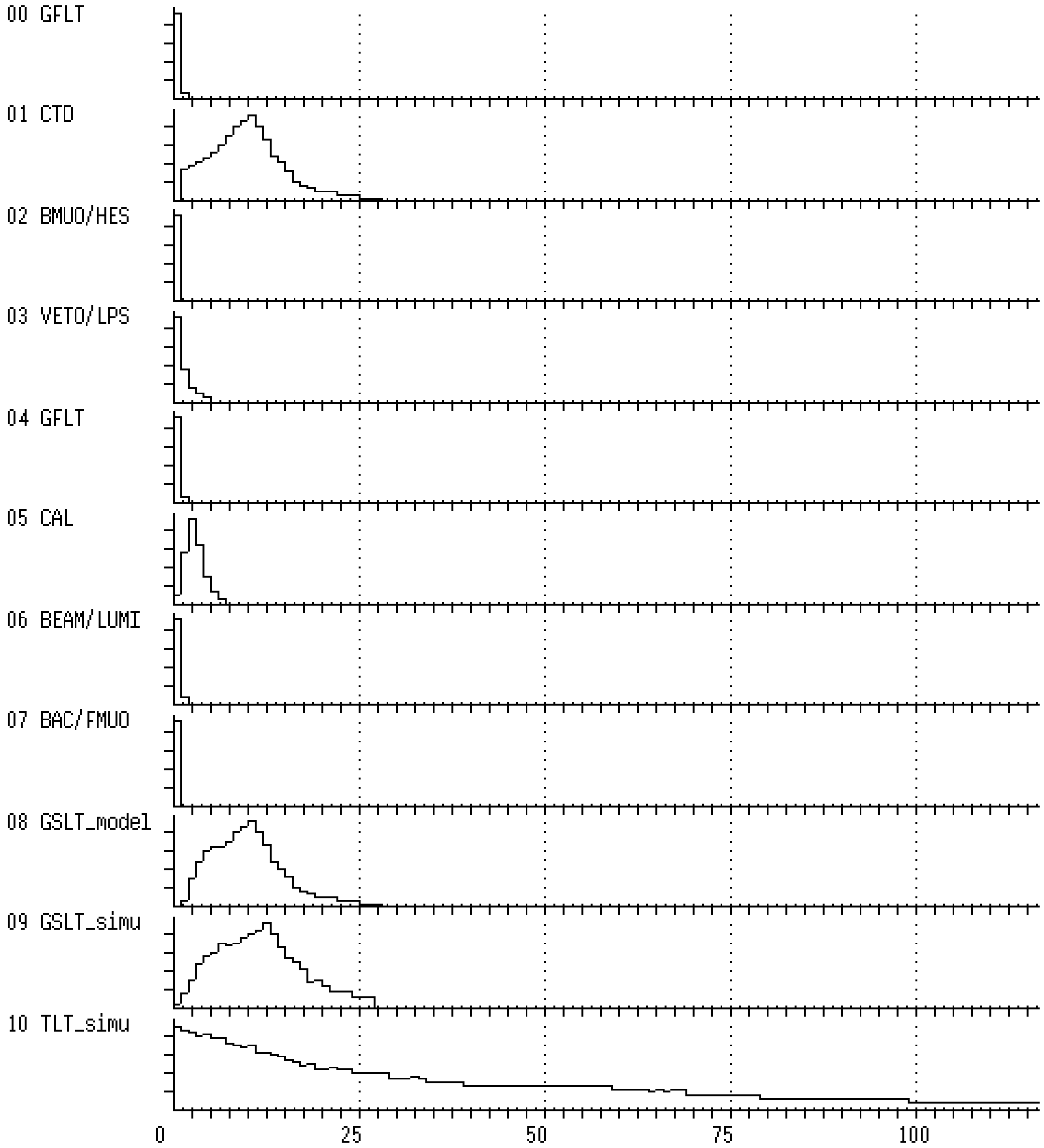


Figure 8: Component Latences, GSLT (ms) and TLT (ms/10) distributions (as generated)

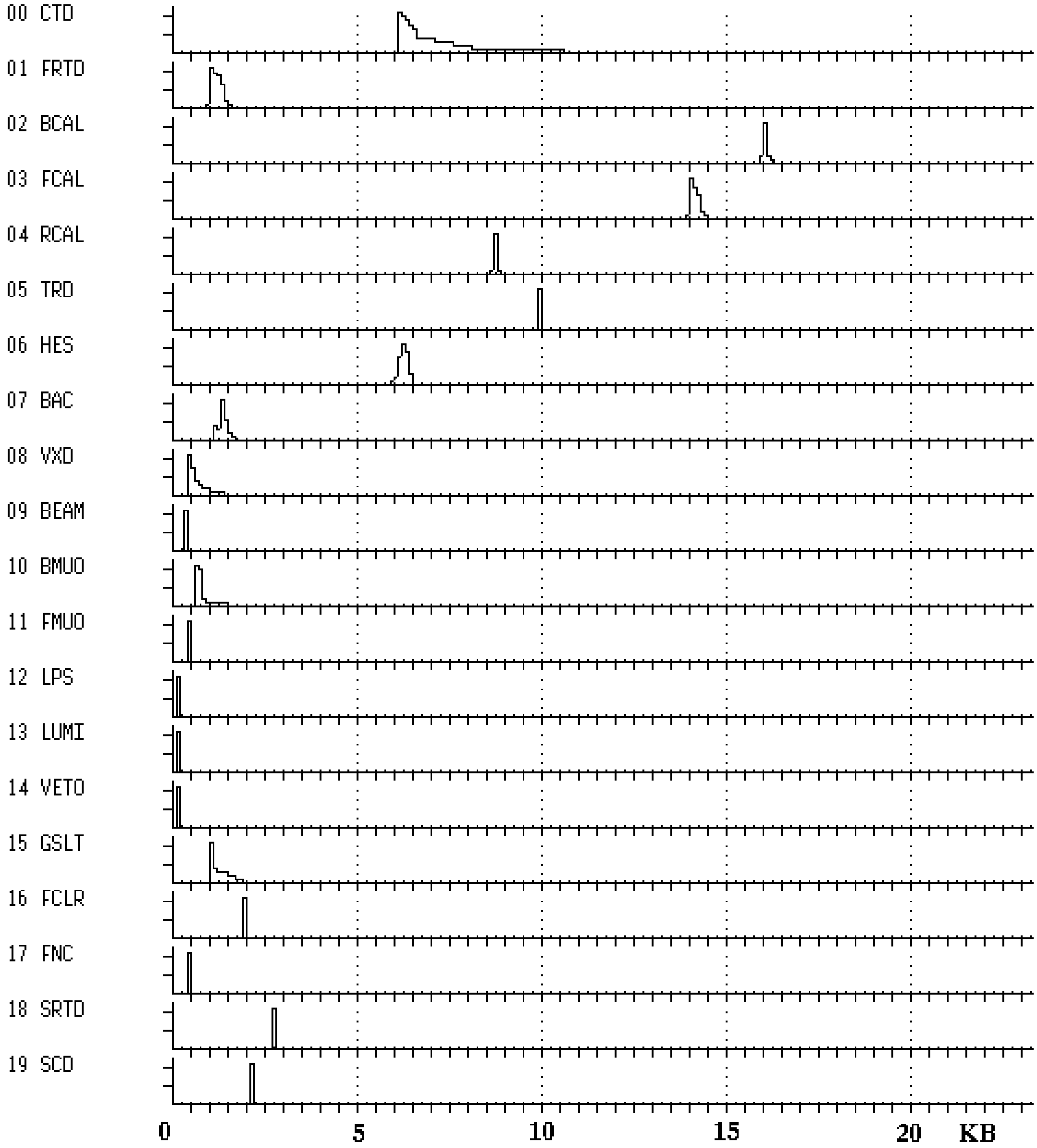


Figure 9: Component bank length Distributions (as generated)

Interface Unit. The multiplexor transfers either the messages that global destination address has coincidence with the address iden stored in CSR or all the messages if the address iden is negative.

Mix-Demix unit. This addressable unit has an optional number of the inputs and the same number of outputs; moreover every input signal is transferred on the output with the same number. In the rest Mix-Demix operation is similar to Multiplexor one.

Star Commutator. An optional number of the input and output channels are available. The input commutator is controlled by Arbiter as in the 2 previous units; the output commutator - by the local destination address.

Cross-Bar Switch. The number of its inputs (M) and outputs (K) are optional, but unlike Star it can transfer K messages simultaneously due to its K internal Multiplexors (see Fig.7).

FIFO-memory. This unit is intended for operations with blocks of data according to First Input - First Output order. Total buffer size is defined in CSR. Data writing to the buffer is carried out on the Slave side. If during receiving data the IR signal falls to zero then the block writing is cancelled. Written block transmission is carried out on the Master side. If during data transmission AK signal falls to zero the transmission will be repeated till the successful completion. The first transmission takes place after the first block is received; the receiving of additional blocks and their transmission are independent. In the case of buffer overflow the GR signal remains asserted as long as a block is transmitted on the Master side.

Transmission Tempo on the Master side can be set in CSR. If it is higher than received one the output block duration will be increased, in the opposite case - decreased.

Combiner. The unit performs the unification of component data into one whole event (each component gives a block of data). The number of combined events is only limited by the total buffer size (set in CSR). The event transmission starts on its combining and proceeds along with other event components receiving . In the rest Combiner operation is similar to FIFO (see Fig.7).

RAM-memory. The unit is double side Slave; one side is used for writing and the other - for reading blocks of data. It is able to store data blocks of different sizes (the sum of blocks length is limited only). The blocks can be read out (with memory clearing) in a random way according to the whole destination address. In the rest RAM-memory operation is similar to FIFO.

Token Network adapter. This unit consist of two parts - Master and Slave, each part contains the FIFO unit. When the Master receives a Token it starts to transmit data block stored in its FIFO to the Network ring. Masters are allowed to hold the Token for the time defined in the CSR.

The Slave side compares global destination address with ones stored in its address table. In the case of coincidence input data block is stored in Slave's FIFO.

All the blocks are divided into fragments before transmitting to the Ring (fragment size is CSR defined). Both sides can operate in 2 modes: with or without assembling the fragments in original data blocks.

Oscilloscope. The unit displays an optional number of timing diagrams on the workstation screen. The diagram size is usually limited by the size of the screen. Therefore the image can be compressed through the special algorithm that keeps the largest signal transition inside compression interval. The image can be expanded also for detailed observation. The compression and expansion factors are stored in CSR.

Data writing is possible only in the time limits defined in the CSR and only after Strobe-signal transition-Up, therefore the Oscilloscope together with the Logic Unit can be used as a Logic Analyzer. Stored timing diagrams are scaled according to the screen size.

4.0.4 Event driving program

The described Unit-programs can be directly used in the Main-program and can be called as the subroutines in the every Tick of discrete time. In this case inter-units links and system control functions are defined in the Main-program.

If in the modelling system the number of simultaneously operating units is negligible in comparison with the total number of units it means that the majority of Unit-program calls will be "useless" (no Unit-state changes during Unit-program executing). In such a situation there is an advantage to use Event Driving program that collects the Units-requests for action and then distributes them at the definite time Ticks.

A version of such Event Driving program has been designed. In this version the described Unit-programs are placed into the special "container". The latter provides request-message passing for actions and action command receiving. The Unit interface to container is implemented by means of "NEXT" and "TR-i" CSR registers (see an example above).

In this case the modelling system control functions should be described as separate Control Unit. All the links between Units are defined in the separate list. The Event Driving Main program supports Event queue operation only.

The comparative tests have revealed that Event Driving program provides no gain in simulation time for our modelling systems, because they are relatively simple ones and their units are in action for the most time.

5 Network switch connections simulation

5.1 Original GSLT-EVB-TLT simulation

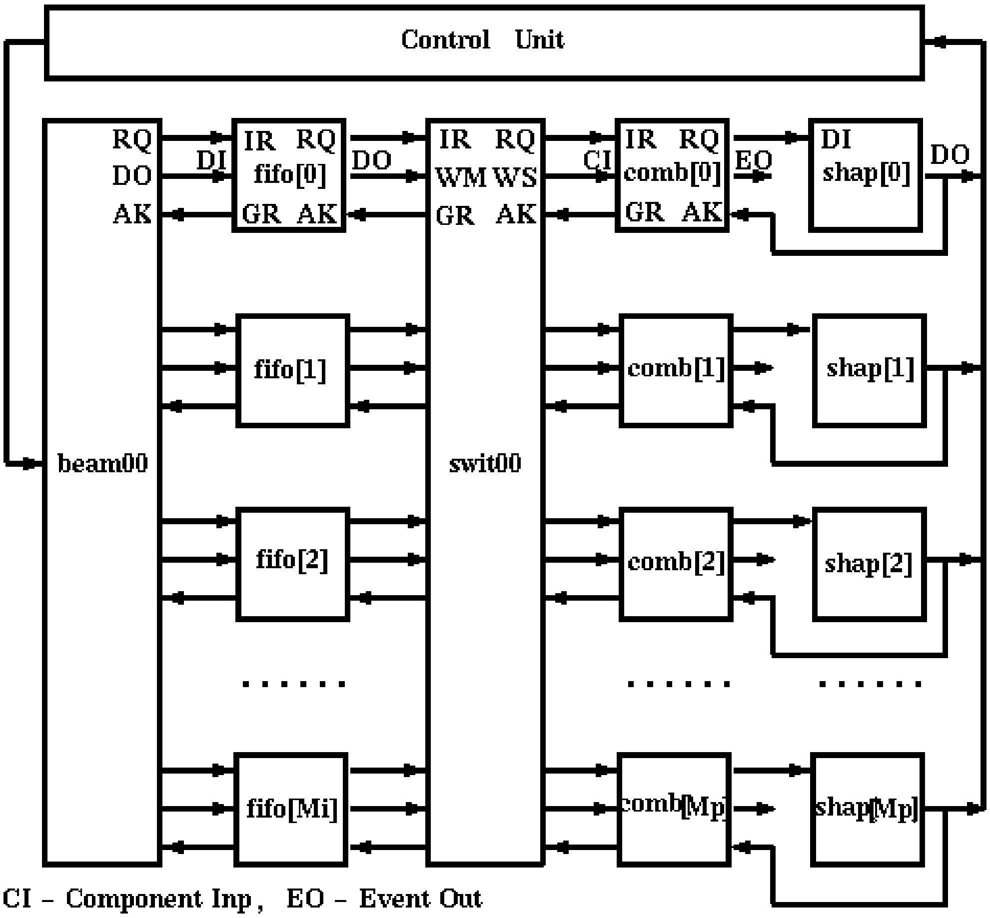
Switch based EVB logical model The simplest logic model of Event Builder based on Cross-Bar Switch is shown in Fig.10. It contains:
- 20-channel (components) Event Generator and smoothing FIFO buffers;
- Cross-Bar Switch with 20 inputs and 30 outputs;
- 30 Combiners for event unification;
- 30 Shapers as the Third Level Trigger processors;
- Control Unit for the data flow control.

In ZEUS Event Builder all the components send requests simultaneously to the Control Unit to obtain the Combiner number for the current event. Waiting for the acknowledges takes about 1 ms in average and causes the "dead time loop of data control". On the one hand, this time is much less than component transfering time (9 ms) via input links and these control loops may be carried out in parallel for several events. On the other hand, the main limitation factor in the original GSLT-EVB-TLT is a shortage of processing power. Therefore, this dead time has no essential influence on throughput rate and has not been taken into account in the models. The results presented here are from the models which cannot produce a data corruption either.

As each combiner prepares the events for one processor only and as they run in serial, the processors address specification is produced by Control Unit for the components (on the Generator output immediately). If all the processors are busy the Control Unit asserts Busy signal on the Event Generator input. During the suspending time a number of possible (may be generated) events is counted. These losses give a nonefficiency value of EVB (in the percentage expression this value is equal to the relative dead time).

The timing diagram of such model simulation is shown in Fig.11. For making the essential restrictions more easy-to-grasp the average beam period is taken small enough (about 1 ms) and TLT-decision time is chosen as a constant (equal to average value).

The diagrams 00-03 show FIFO operation for component 02 (ch=02) that has maximum data length (see Fig.9). The Combiner operation for the second channel is given in the diagrams 04-07. Diagram 08 illustrates component 02 smoothing FIFO filling up. The diagram 09 shows time in-



Te = 17 ms / ev, Mi = 20, Tp = 400 ms / ev, Mp = 30,

c00 = 7.2 KB, c01 = 1.4 KB c02 = 16.3 KB, c03 = 14.2 KB

c04 = 8.7 KB, c05 = 10.0 KB, c06 = 6.1 KB, c07 = 1.3 KB,

c08 = 0.5 KB, c09 = 0.4 KB, c10 = 1.3 KB, c11 = 0.5 KB,

c16 = 2.0 KB, c17 = 0.5 KB, c18 = 2.8 KB, c19 = 2.2 KB.

Tick = 100 mcs

Figure 10: Switch based EVB logical Model

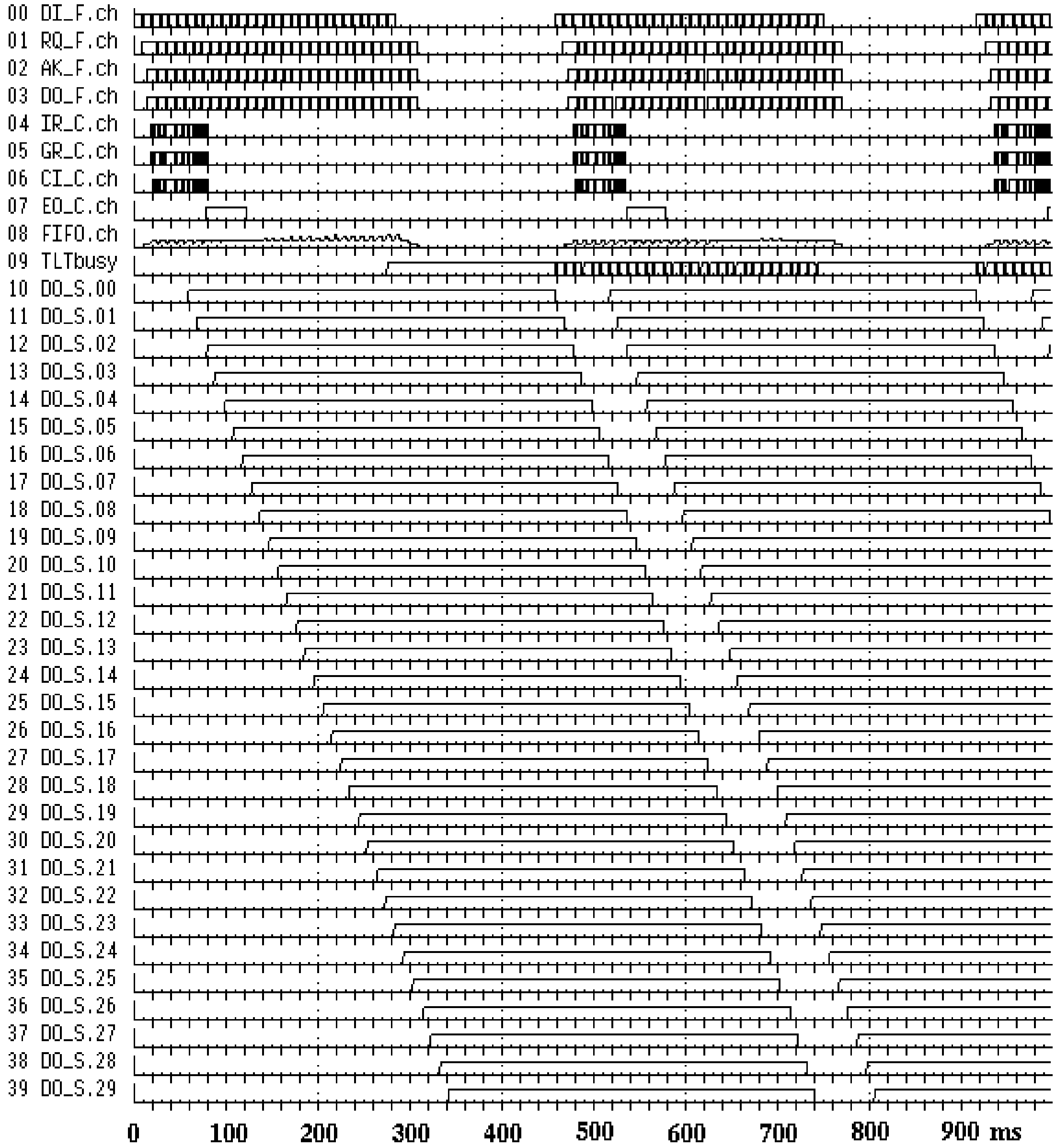


Figure 11: Switch based EVB logical Model timing diagrams

tervals when all TLT-processors are busy. The diagrams 10-39 represents TLT-processors operation.

In accordance with "before event building" processor specification mechanism there are 15-20% pauses in TLT processors time of event selection.

The following obvious conclusions can be made of Fig.11:

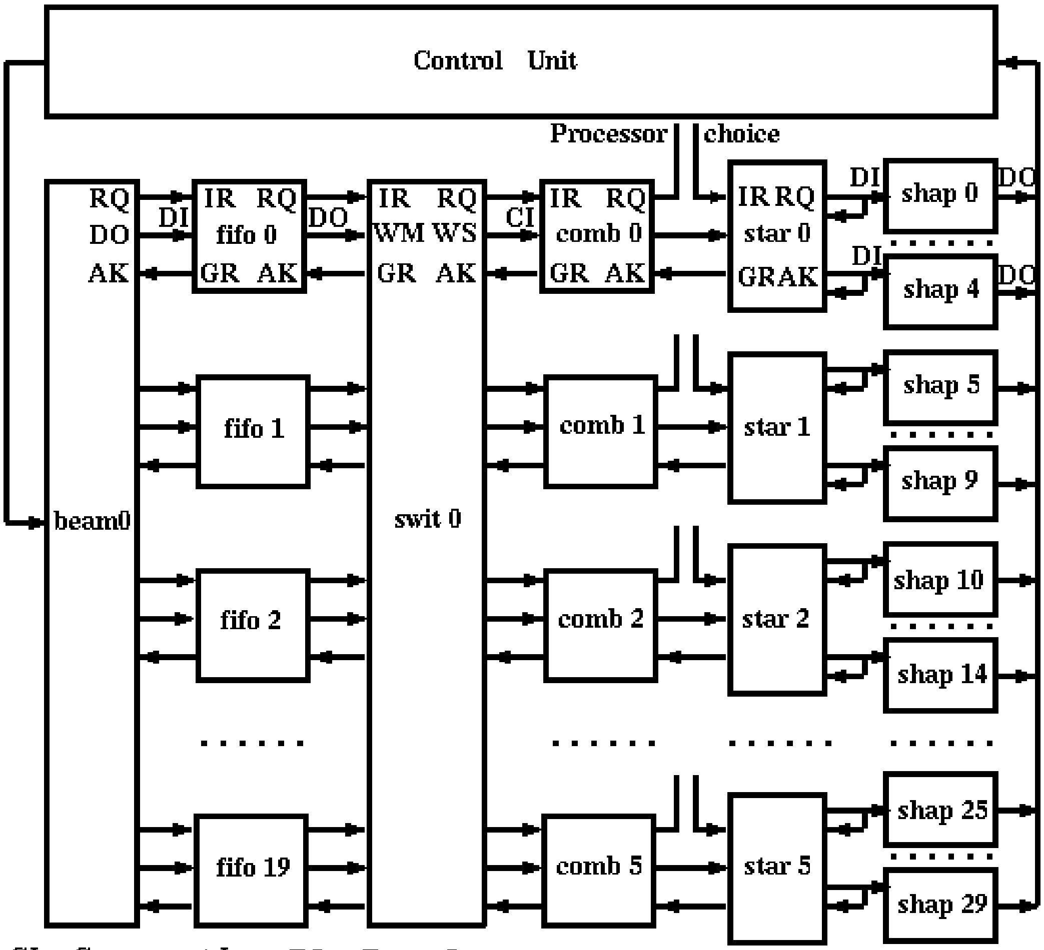
- the Third Level Trigger total processing power is not enough for such ultra-high beam intensity (the TLT-busy intervals between events taking in the 00-03 and 09 diagrams);
- the maximum data length component (02) limits the total rate of event taking within TLT-processors ability intervals;
- the combiners are evidently underloaded, it means that their number can be reduced.

ZEUS EVB-TLT logical model The next logical model corresponding to the original ZEUS EVB-TLT is shown in Fig.12. The main parameters are as in the previous example. But the number of combiners is reduced to 6 ones and processors are united in 6 groups (5 processors in the each group) by means of Star-commutators. The events are distributed between combiners in comply with round-robin order. When the event is assembled it is waiting for the next processor vacation in its combiner cluster.

The results of modelling this EVB-TLT version are represented by timing diagrams in Fig.13. It is easy to see that in this structure the Processors and Combiners performance corresponds one to another. As it follows from diagram 08 the necessary FIFO size is 3-5 banks of component 02. As before the processor power is not enough (in comparison with other units) but TLT processors operate without void intervals, it increases their productivity on ca. 20% (7).

ZEUS EVB-TLT full model As the original EVB-TLT throughout rate is limited by the shortage of processors power only, therefore the simple logical model should be enough for the general description of EVB-TLT behaviour. Nevertheless the full model of the EVB-TLT was created; it corresponds to the EVB-TLT block-diagram in Fig.4. The realization of EVB-TLT main parts by means of simulation Units is shown in the Fig.14.

The component buffer (see upper T800 in Fig.4 - transputer and its memory in the component 2TP-VME module) is represented by MEMOry Unit. The transmission transputer (lower T800) and its links are represented by three FIFO Units. The SWITch has 60 inputs channels



CI – Component Inp, EO – Event Out

Te = 14 ms / ev, Mi = 20, Tp = 400 ms / ev, Mp = 6*5 = 30

c00 = 7.2 KB, c01 = 1.4 KB c02 = 16.3 KB, c03 = 14.2 KB

c04 = 8.7 KB, c05 = 10.0 KB, c06 = 6.1 KB, c07 = 1.3 KB,

c08 = 0.5 KB, c09 = 0.4 KB, c10 = 1.3 KB, c11 = 0.5 KB,

c16 = 2.0 KB, c17 = 0.5 KB, c18 = 2.8 KB, c19 = 2.2 KB.

Tick = 100 mcs

Figure 12: ZEUS EVB-TLT logical Model

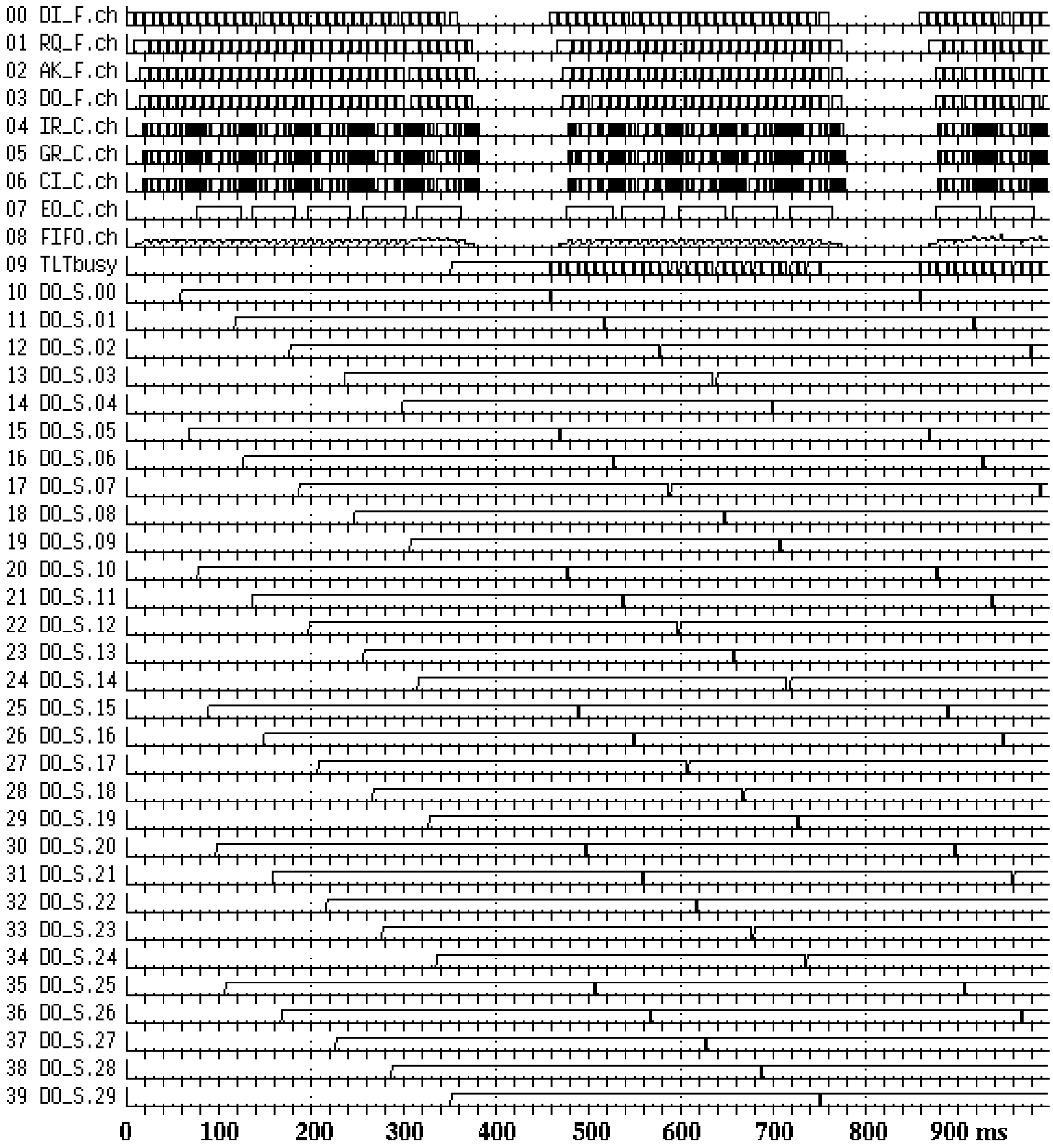


Figure 13: ZEUS EVB-TLT logical Model timing diagrams

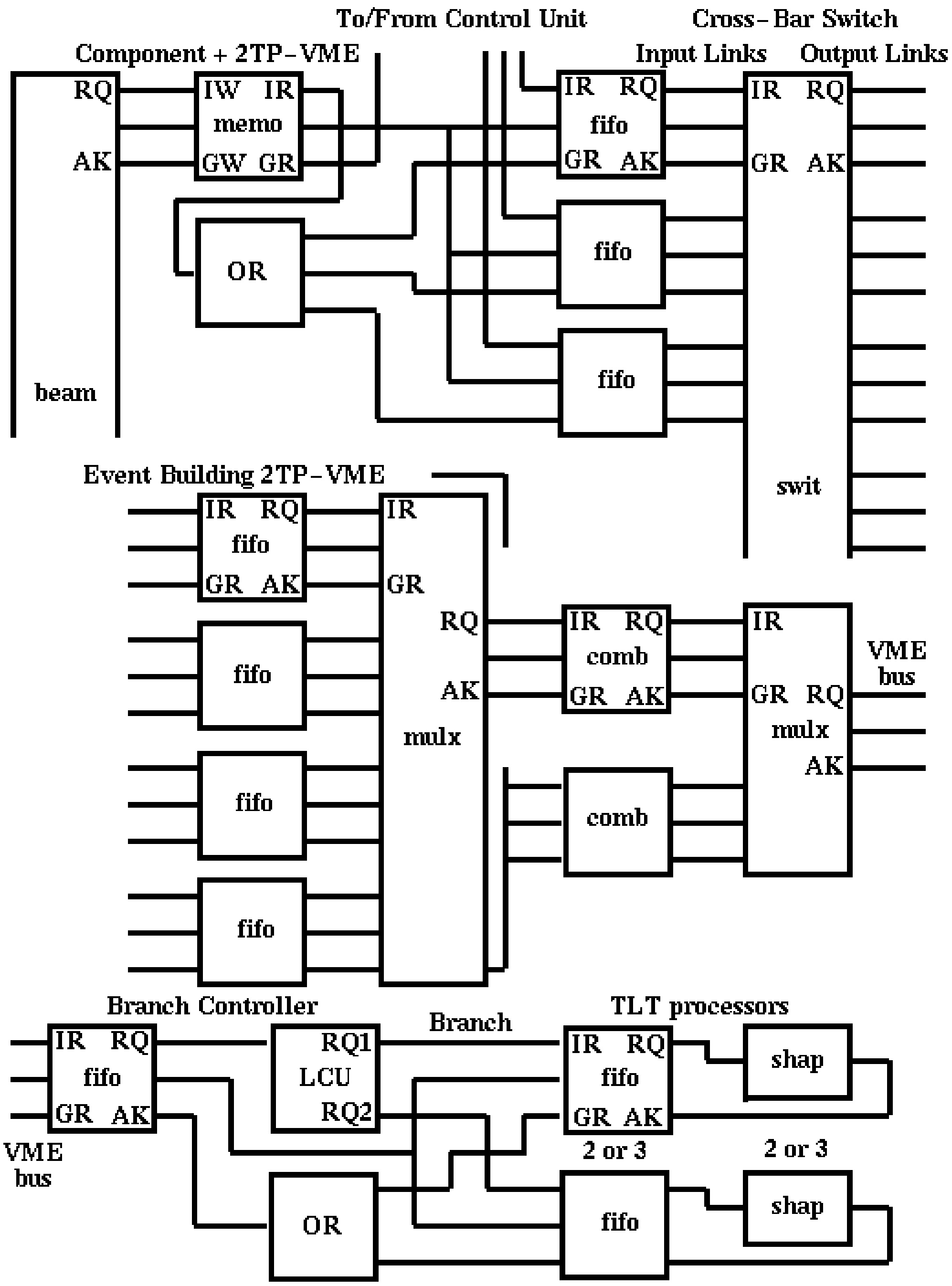


Figure 14: The original EVB-TLT simulation elements

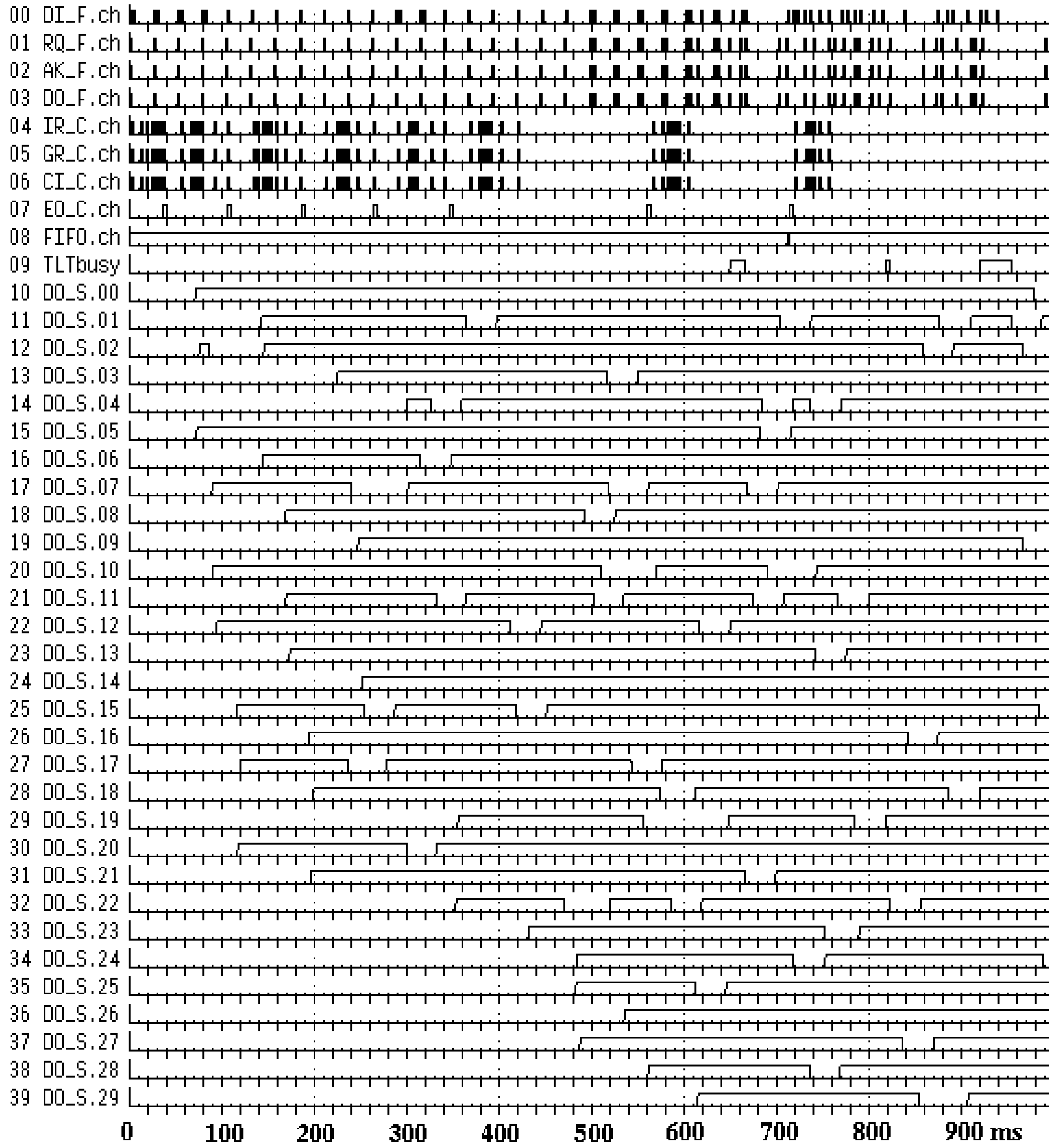


Figure 15: The original EVB-TLT timing diagram

and 48 outputs channels. The event building 2TP-VME module is collected from 4*2 FIFO Units with 2 Multiplexors, 2 Combiners and 1 Multiplexor for VME output. Branch controller (BVI in Fig.4) is the FIFO Unit and at last the TLT processors are the combination of FIFO plus Shaper Units. For switching RQ signals from one unit to the several ones the special Local Control Units (LCU) were designed. For ORing GR-signals the Logical NAND/NOR Unit was used.

The transmission rate defines by the TEMPO value in MEMO, FIFO and COMB Units. For the input and output transputer links the rate was preset on the 600 KB/s level. For the VME-bus the 15 MB/s rate was chosen and for Branch - 2 MB/s. Apart from holding the rate values the MEMO and FIFO Units represent the real EVB buffers in the corresponding nodes.

The timing diagrams of EVB-TLT model operation are shown in Fig.15. In the contrast to the other diagrams this one represents the real distribution of TLT-decision intervals. The data flow control is realized in the same manner as in the real EVB - through input MEMO Unit overflow signals. The overflows create the void intervals (diagrams 00-03) between data receiving; in the very beginning, as it can be seen from Fig.15, the data taking rate is higher (the processors are free) and then the rate drops down to a steady state.

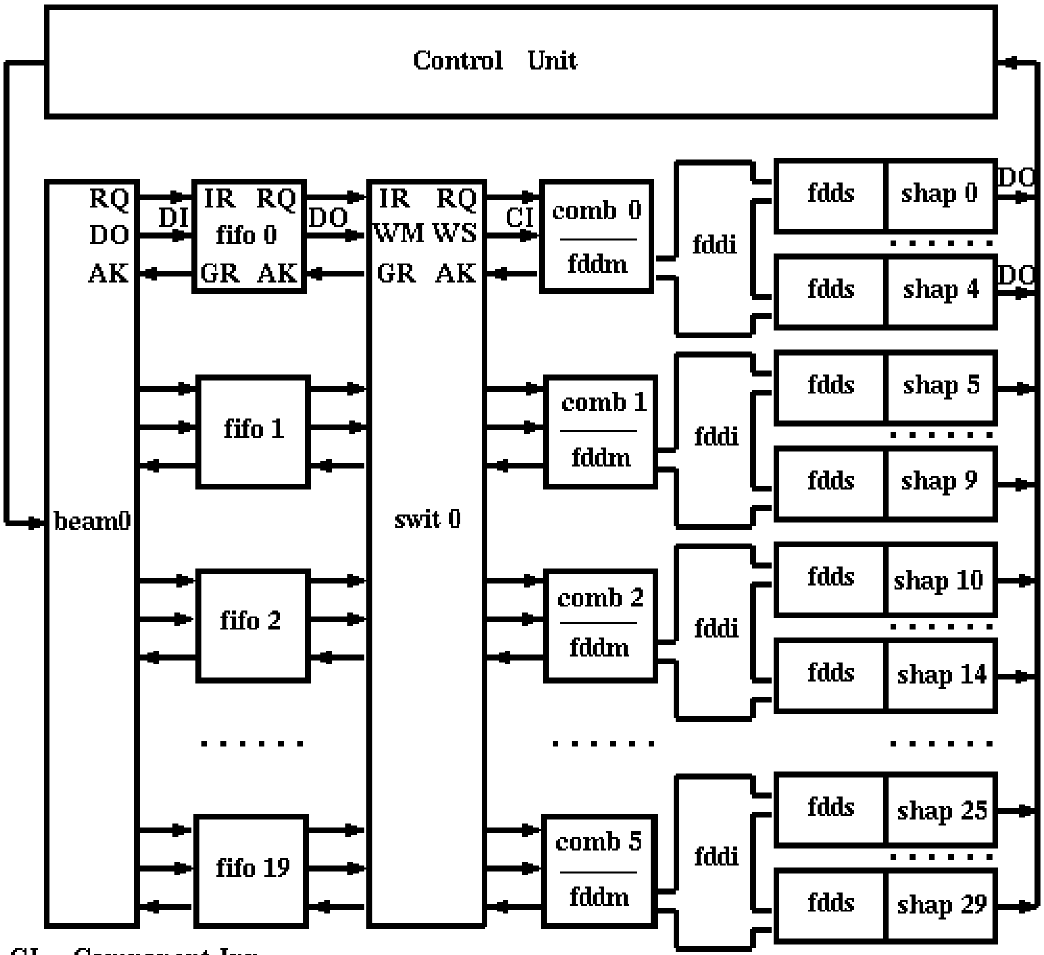
5.2 NSC simulation

EVB-FDDI-TLT logical model The model shown in Fig.16 corresponds to "Phase 1" of the NSC project. It consists of the ZEUS EVB-TLT logical model but FDDI rings with Master (FDDM) and Slave (FDDS) interface Units are used instead of the Star demultiplexors.

The simulation timing diagrams are shown in Fig.17. As above, component 02 FIFO operation is shown on the diagrams 00-03 and COMBiner operation in the second cluster is shown on the diagrams 04-07. Diagram 08 illustrates component 02 smoothing FIFO filling up and diagram 09 - all TLT-processors Busy intervals. The diagrams 10-39 represents TLT-processors operation (they are obviously underloaded).

ZEUS NSC model The model shown in Fig.18 corresponds to "Phase 2" of NSC project. It has

- 20 Token Network adapters for components (one adapter per one component);



CI – Component Inp

Te = 11 ms / ev, Mi = 20, Tp = 130 ms / ev, Mp = 6*5 = 30

c00 = 7.2 KB, c01 = 1.4 KB c02 = 16.3 KB, c03 = 14.2 KB

c04 = 8.7 KB, c05 = 10.0 KB, c06 = 6.1 KB, c07 = 1.3 KB,

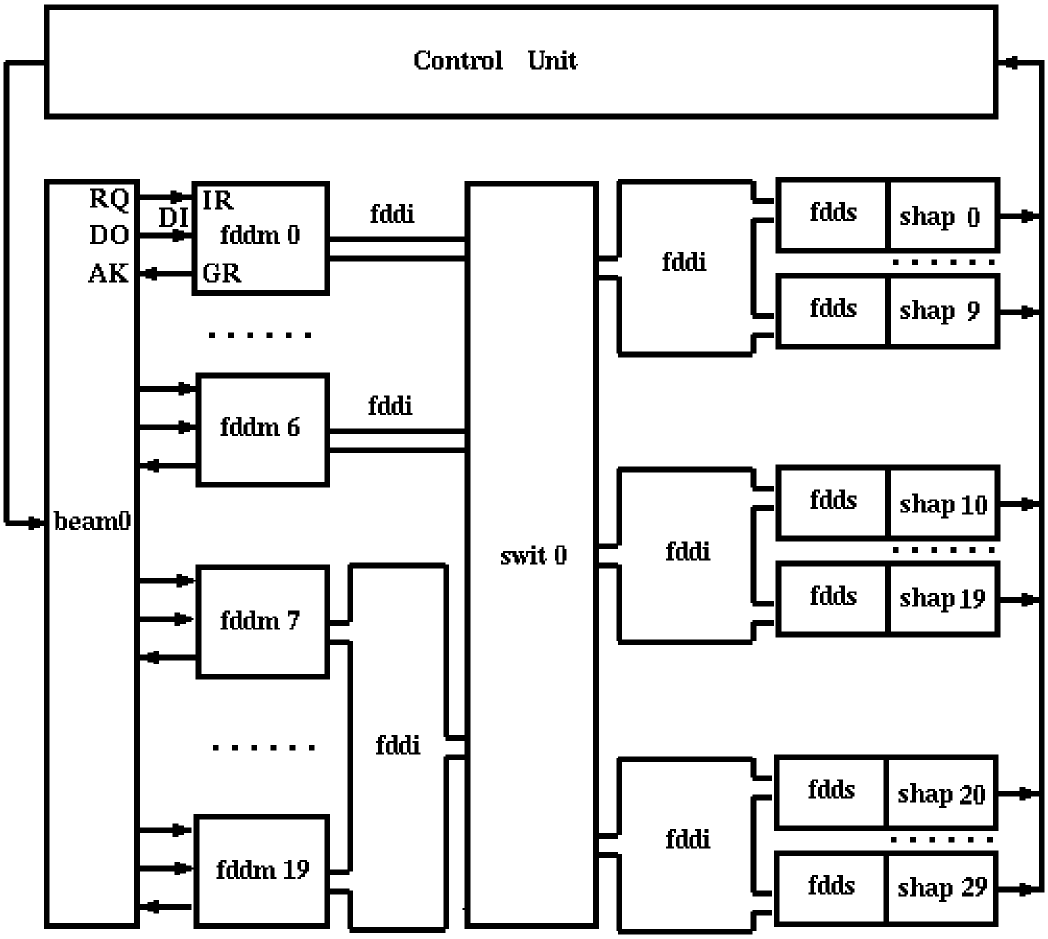
c08 = 0.5 KB, c09 = 0.4 KB, c10 = 1.3 KB, c11 = 0.5 KB,

c16 = 2.0 KB, c17 = 0.5 KB, c18 = 2.8 KB, c19 = 2.2 KB.

Tick = 10 mcs

Figure 16: ZEUS EVB-FDDI-TLT logical Model

Figure 17: ZEUS EVB-FDDI-TLT logical Model timing diagrams



Te = 6 ms / ev, Mi = 20, Tp = 130 ms / ev, Mp = 3*10 = 30

c00 = 7.2 KB,	c01 = 1.4 KB	c02 = 16.3 KB,	c03 = 14.2 KB
c04 = 8.7 KB,	c05 = 10.0 KB,	c06 = 6.1 KB,	c07 = 1.3 KB,
c08 = 0.5 KB,	c09 = 0.4 KB,	c10 = 1.3 KB,	c11 = 0.5 KB,
c16 = 2.0 KB,	c17 = 0.5 KB,	c18 = 2.8 KB,	c19 = 2.2 KB.

Tick = 10 mcs

Figure 18: ZEUS COMP-FDDI-TLT Model

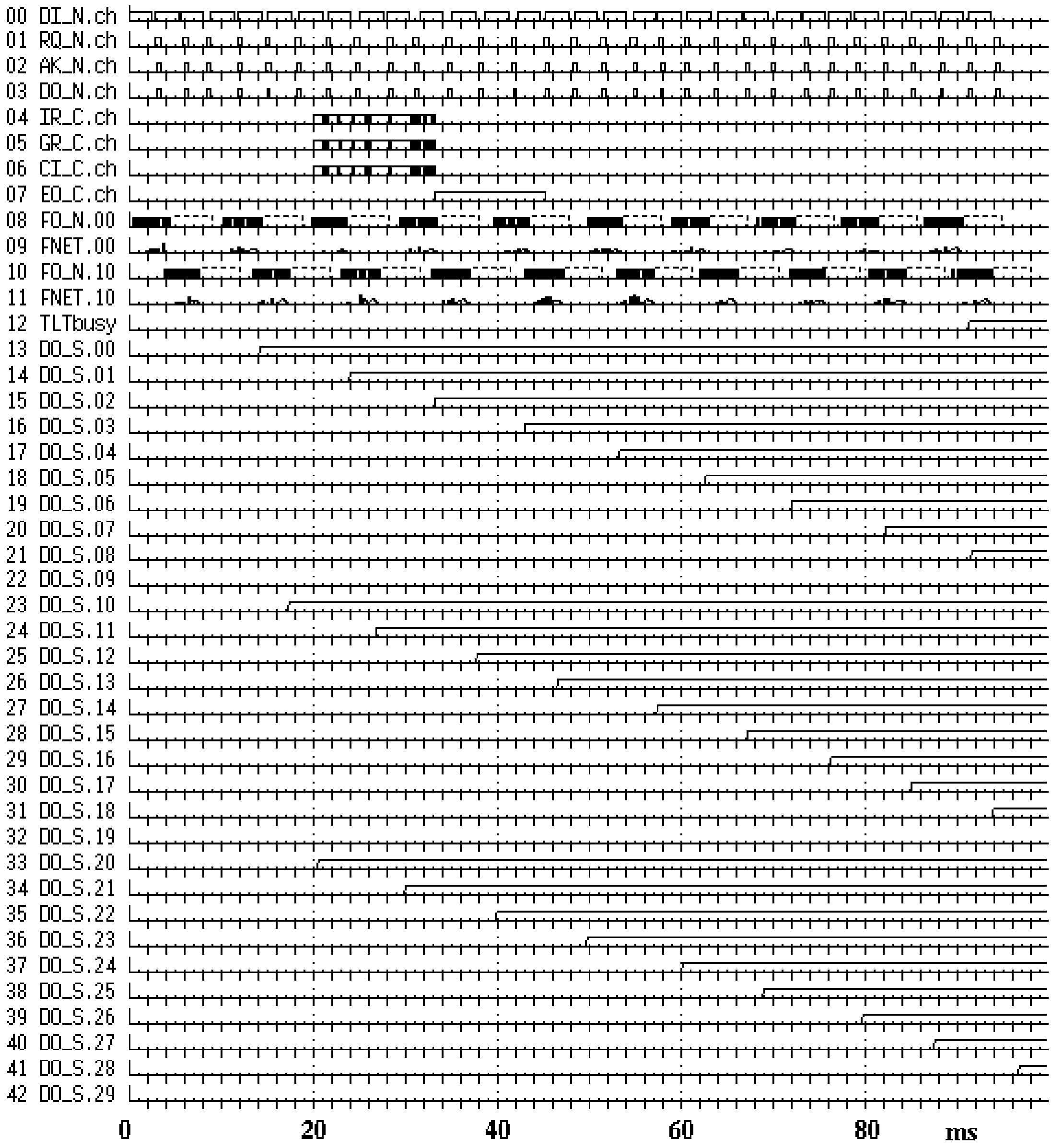


Figure 19: ZEUS COMP-FDDI-TLT Model timing diagrams

- GIGAswitch that consists of Cross-Bar Switch and 11 FDDI line cards;
- 30 TLT-processors (with FDDI cards), they are connected to GIGAswitch by means of three FDDI rings.

For reducing the number of used FDDI line cards in the GIGAswitch the components with small length data block are combined by one local FDDI ring (connected to the GIGAswitch via a common line card).

The model simulation timing diagrams are represented in the Fig.19 (time scale is reduced). The diagrams 00-01 show the component data banks storing in a Token Network adapter. The diagrams 02-05 show the component data transferring via FDDI ring (02) to a line card of GIGAswitch (04-05). The diagrams 08-11 represent switch output rings operation (2 rings) and Network adapters smoothing FIFOs filling up. The diagrams 13-42 show the operation of processors. The diagrams 08 and 10 are shown for 2 cases: for FDDI rate 12 MB/s (solid lines) and for real FDDI modules rate - 5 MB/s (dashed lines). These diagrams revealed the fact that FDDI is underloaded in the both cases.

It was considered that the processors are building the events themselves; it takes a part of their time.

5.3 The discussion of the simulation results

The Fig.20 represents in graphical form the results of EVB-TLT simulations. It shows the efficiency of the described EVB models as a function of the average event period (Te) for the real distributions of GSLT, TLT-decision time (Fig.8) and component bank lengths (Fig.9).

As the first outcome of the models investigation it may be mentioned that the simple logical models "1" and "2" give approximately the same results as "3" - more complex one (the error is less than 10 percents).

The main part of the efficiency curve may be represented as a straight line. The efficiency saturation point for "2", "3" and "5" models corresponds to

$$Te = 1 \ / \ (\ Mp*Np \ ) \qquad (10)$$

- is defined by the number and computing power of processors; other EVB layers are "transparent" for data flow.
Due to the empty intervals between TLT-processor operation (Fig.11 in comparison with Fig.13) the logical model "1" has the saturation Te ca.15% less than logical model "2". In the main parts of the curves there is

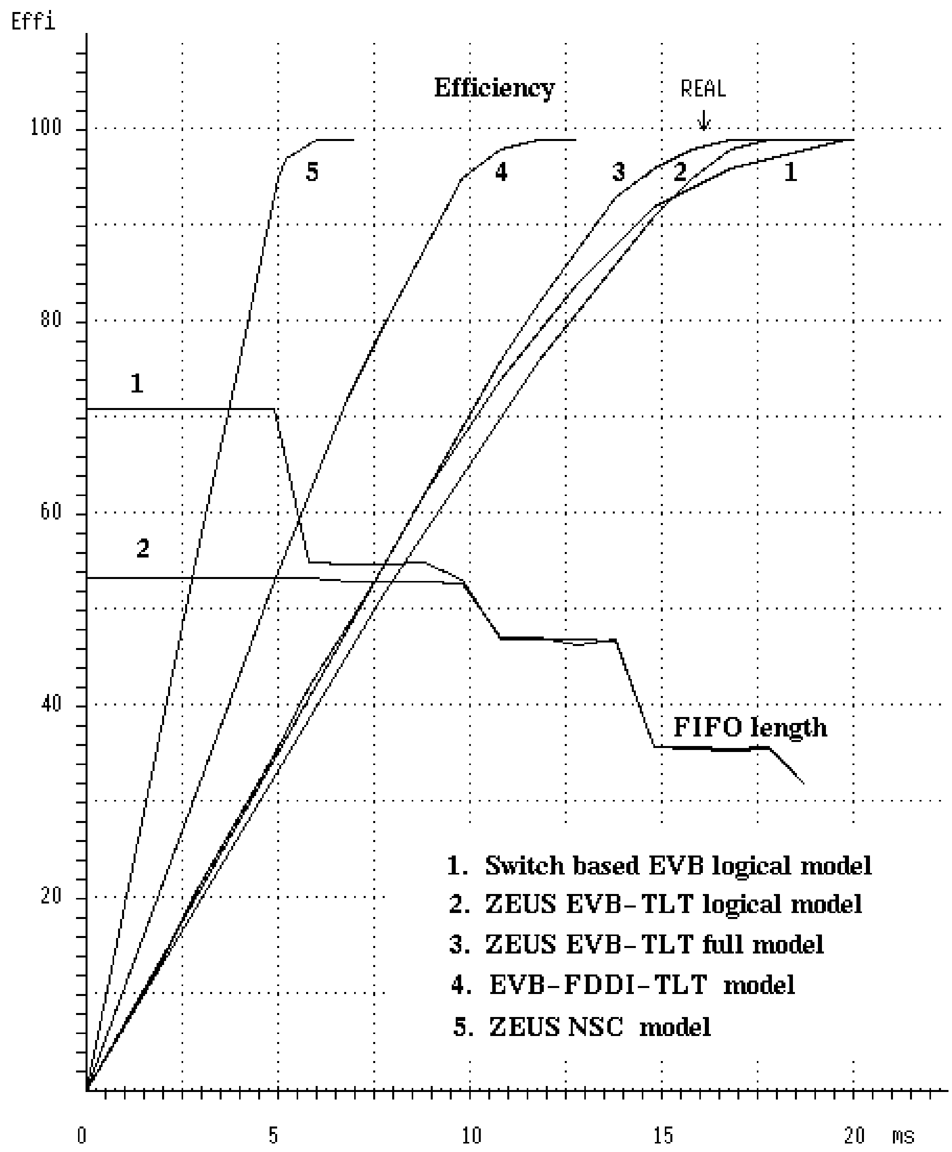


Figure 20: The EVB-TLT EFFIciency as a function of average Event period

no such difference of these models because, in contrast to Fig.13, there are many random "empties" between TLT-operations in both models with a random TLT-decision time (see, for example, Fig.15 - TLT-processor finished fast but Combiner is not ready yet).

For the model "4" the component 02 (BCAL) data length (or CrossBar switch input and output links bandwidth) is the limitation factor (the processors are underloaded - see Fig.17).

For the model "5" the efficiency curves are similar both for 12 MB/s and for 5 MB/s FDDI throughput rate.

The graphs of smoothing FIFO filling up are shown for "1" and "2" models in Fig.20 also (for 02-channel). In correspondence with Fig.3 the necessary length of smoothing FIFO is about 4-5 of component 02 banks length.

The substantial difference is in the parameter of latency for the logical and full models. When efficiency is less than 100% the number of building events (distributed along the system) for the logical model is equal 5-6 but for the EVB-existent model it is higher - 24-25, because the latter has many temporary fifo-buffers along the data flow.

From the model "5" results it may be concluded that ZEUS EVB throughout rate may be increased in ca.2.6 times if 30 new Ch-S (INDY) processors will be used and the Network Switch Connections are implemented using FDDI and a DEC GIGAswitch. If a separate processor will be used in the each FDDI ring for event building (as a combiner) the productivity may additionally be ca.15% higher.

6 Conclusion

The GFLT, GSLT and TLT trigger systems were conceived to sustain the HERA design luminosity with output rates of 1000, 100 and 5-10 Hz respectively. Whereby the effectiveness of the FC system is unknown.

Additional improvements to HERA, after '95, can be expected. Currently the proton beam emittance is significantly smaller than anticipated. Thus increasing demands on the trigger and readout system can be expected.

The experiment must, therefore, have a good understanding of the current system and its limitations and an understanding of potential upgrades. The significant price performance improvements in TLT farm machines and the availability of commercial network switches provide one such upgrade.

The estimations carried out on Page 4 show that in all cases where the Network is connected downstream of the current EVB; the total throughput is limited by the EVB links.

The MSU-NPI project is investigating the possibility of using a network switch with the DAQ environment. A full implementation of the new event builder within the experiment can be carried out as well.

The authors would like to thank I.Manjavidze, J.Vermeulen and L.Wiggers for the usefull discussions and K.Tokoshuku, M.Kasemann, F.Clebana, U.Behrens and S.Polenz for the information of ZEUS GFLT-GSLT-TLT parameters.

7 Literature

1. U.Behrens, L.Hagge, W.O.Vogel, The Eventbuilder of the ZEUS Experiment, DESY-Note 93-008, February 1993.

2. M.Crombie, D.Bandyopadhyay, S.Bhadra, F.Clebana, R.S.Orr, Third level trigger interfacing, ZEUS-Note 90-112, Oct. 1990.

3. ZEUS Contributions to the Real-Time 93 Conference, Vancouver, Canada, DESY-Note 93-091, July 1993.

4. Proceedings of "Computing in High energy Physics'92' Conference, CERN Report 92-07, December 1992, pp.57, 184, 188, 192, 196, 381, 681.

5. K.Tolly, D.Newman, FDDI bridges and routers, Data Communications, August 1993, p.66.

6. J.T.Johnson, ATM networking gear: welcome to the real world, Data Communications, October 1993, p.66.

7. R.J.Souza et al., The GIGAswitch system: a high performance packet switching platform, Digital technical Journal, V.6 No 1, 1994.

8. GigaRouter system description, Netstasr, Rev.1, November 1993.

9. Proceedings of International Data Acquisition Conference, Fermilab, October 26-28, 1994, pp.19-23, 33-40.

10. RD-31 Status report'95, CERN LHCC Note 95-14, March 1995.

11. J.C.Vermeulen, Simulation of data-acquisition and trigger systems in C++, EAST Note 93-22, October 1993.