7 December 1985

DATA RATES FOR EVENT BUILDERS AND PROCESSOR FARMS

by

Thomas Devlin
Department of Physics and Astronomy
Rutgers - The State University of New Jersey
P.O. Box 849, Piscataway, New Jersey 08854

I have found it educational to perform a very naive calculation of data rates through a matrix of event builders and a processor farm with the architecture currently popular amongst devotees of this subject. I have no doubt that sophisticated use of queuing algorithms such as RESQ will yield a much more precise picture of such a data pipeline. Nevertheless, there may be others out there who, like me, feel more comfortable with a simple numerical/algebraic picture of these multi-Gigabyte data rates.

Assume that various subsystems of a large detector assemble data blocks locally with a typical block size of 30 kiloBytes (kB), and that each can send that block to an event-building matrix of dual-port memories over a high bandwidth 32-bit bus. Fig. 1 shows a schematic view of the structure. A resource controller for the matrix ensures that one horizontal row of the

matrix is dedicated to a particular event as long as needed. A processor node in the "farm" is assigned at the same time, and a path established to it from the assigned row in the matrix. Block writes into the memory modules on the event building row can occur independently on the various vertical buses from the detector, and horizontal block output to the farm can occur any time a block input is finished and the appropriate bus is free. This complicates the resource management problem, but it makes more efficient use of the facility.

Calculation of rates is fairly simple. Let N_1 and B_1 be the number and bandwidth of vertical buses in from the detector, and N_2 and B_2 the number and bandwidth of horizontal buses out to the farm. Although RESQ may fine-tune this number, to first order $N_1B_1 = N_2B_2$, i.e. total output = total input. There seems no reason to adopt an inferior technology for either input or output buses, so we can assume $B_1 = B_2 = B$ and $N_1 = N_2 = N$.

Try some numbers here. If block transfers can occur at rates of one word per 100 nsec, then B = 40 MB/sec. N = 32 is a familiar number, and it yields a total bandwidth NB = 1,280 MB/sec. If, as claimed, a typical SSC event will have an average size S = 1 MB of raw data, then this choice of numbers could handle R_o = 640 events/sec out of the level-2 trigger when operating at 50% duty factor (D = 0.5). ($R_o = NBD/S$) This falls within the spectrum of estimates I have heard for this stage of the data acquisition system.

The data flow through the farm is not nearly as well defined. Most discussions assume a tiered structure of successively more refined analysis and restrictive cuts. This could all occur in software within one node, or it

could be implemented in the hardware architecture of the farm (Fig. 2). I have not confronted the data flow between tiers for the latter case, but, assuming that it is tractable, my-analysis applies to both situations. In the latter case, it suggests how processors should be allocated among the several tiers.

For definiteness, let's assume a 5-tiered analysis structure with cuts in the number of events at each of the first four tiers. The final tier is assumed to be full reconstruction in which all events are passed on to the data logging device. Let the event rate into the first tier be R_a, the average analysis time (in Vax11/780 seconds = Vsec) per event t_a, and the fraction passed on to the next stage f_a. In a similar fashion, we define $R_b \ldots R_e$, $t_b \ldots t_e$ and $f_b \ldots f_e$. Of course, $R_a = R_o = NBD/S$, $f_e = 1$. Relationships are straightforward: $R_b = f_a R_a$, $R_c = f_b R_b = f_a f_b R_o$, etc. The total analysis capacity required, in Vsec per second is

$$T = R_o(t_a + f_a t_b + f_a f_b t_c + f_a f_b f_c t_d + f_a f_b f_c f_d t_e)$$

If each node in the farm is one Vax equivalent, then this is also the number of processors in the farm. The individual terms in the sum indicate the allocation of resources amongst tiers.

We now have to make some assumptions about the analysis strategy. The final rate written to tape should be about 1 event/sec. Thus, we have the

constraint $f_a f_b f_c f_d = 1/640$. An assumption as good as any other at this point is that $f_a = f_b = f_c = f_d = 1/5$. We also have a popular estimate, $t_e = 1000$ Vsec. The rest depends on the analysis and cut strategy in the several tiers. Let's suppose that the first tier merely refines the Level-2 trigger and takes about $t_a = 0.1$ Vsec. One possible assumption is that each tier takes an order of magnitude longer than the previous tier: $t_b = 1$ Vsec, $t_c = 10$ Vsec and $t_d = 100$ Vsec. This provides sufficient definiteness to compute the results in Table 1. The numbers used in this example assume the farm operating at 100% duty cycle, whereas the data bus/event builder system was assumed to run at 50% duty cycle. Fluctuations would require some additional processor power. Purely statistical fluctuations might be accomodated by a 2 std. dev. increase in the number of processors. If one node equals one Vax, then $2\sqrt{1984} = 89$ extra nodes. Real fluctuations are more likely to come from luminosity, detector or trigger changes, and result in a change in the rate at all tiers. A safety factor is needed.

Certainly, other analysis-time/cut-factor profiles can be imagined and a corresponding calculation performed. One unrealistic example may serve to show the range of possibilities: if the first three tiers are eliminated ($f_a = f_b = f_c = 1$, $f_d = 1/640$), and the fourth still takes an average of 100 Vsec, then that tier needs 64,000 Vax equivalents!

One result is invariant to the change in profile. If, indeed, online reconstruction is desired for a final event rate of 1 Hz, then 1000 Vsec of analysis time per event requires 1000 Vax equivalents in the farm working on that task over and above the processors devoted to earlier tiers. This does little to settle any debate about whether full reconstruction should be done in the online data stream as envisioned here, locally offline at the interaction region, or in some central computing facility.

TABLE 1

Example: Rates, Times and Computing Power

Needed in a 5-Tiered Processor Farm

Tier	Average Event Analysis Time (Vax sec)	Event Rate R	Computing Capacity Required (Vax Equiv.)
a	0.1	640	64
b	1.0	128	128
c	10	25.6	256
d	100	5.12	512
e	1000	1.024	1024

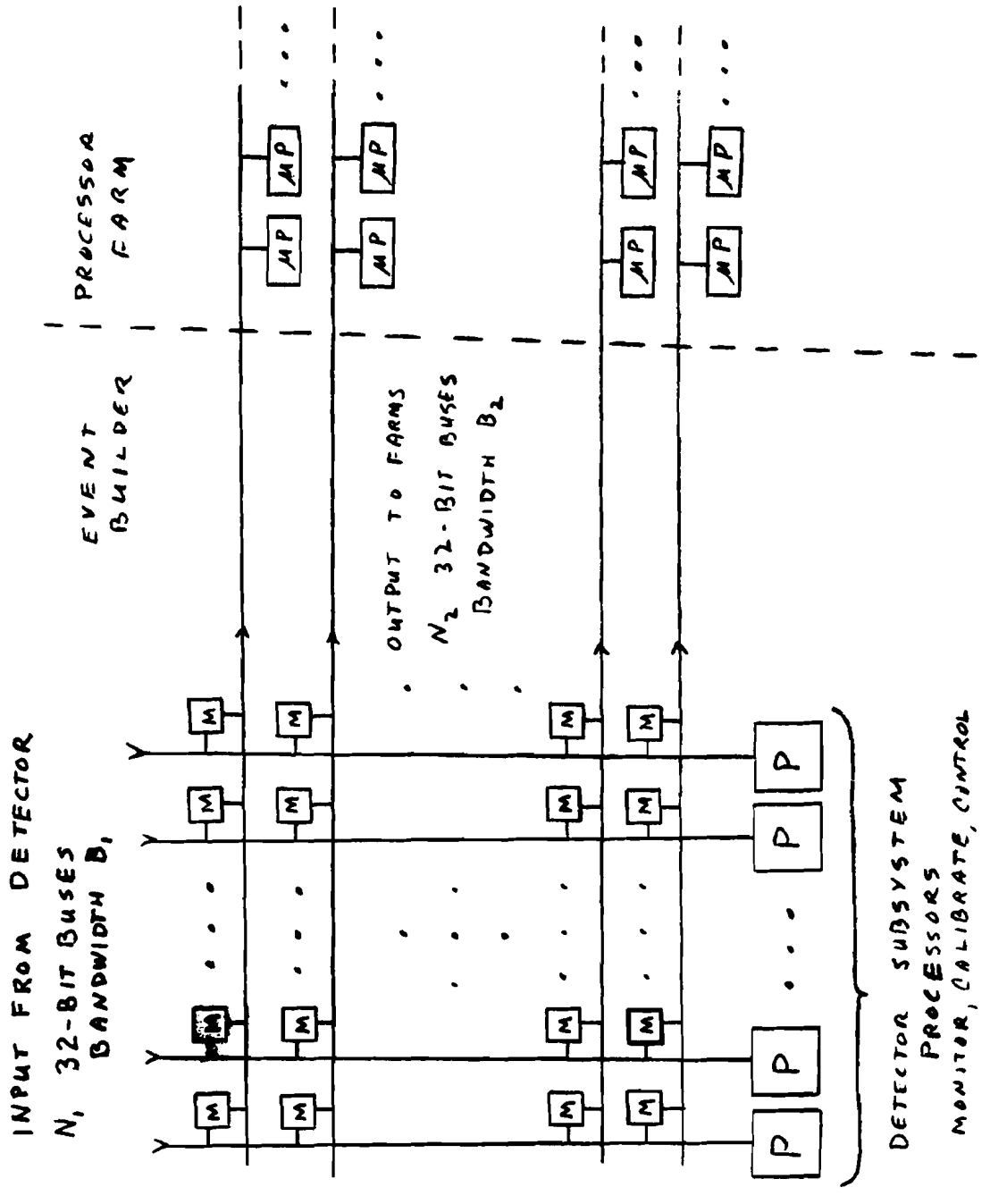
INPUT FROM DETECTOR
N_1 32-BIT BUSES
BANDWIDTH B_1
EVENT BUILDER
PROCESSOR FARM
OUTPUT TO FARMS
N_2 32-BIT BUSES
BANDWIDTH B_2
M
MP
P
DETECTOR SUBSYSTEM PROCESSORS
MONITOR, CALIBRATE, CONTROL

FIG. 1

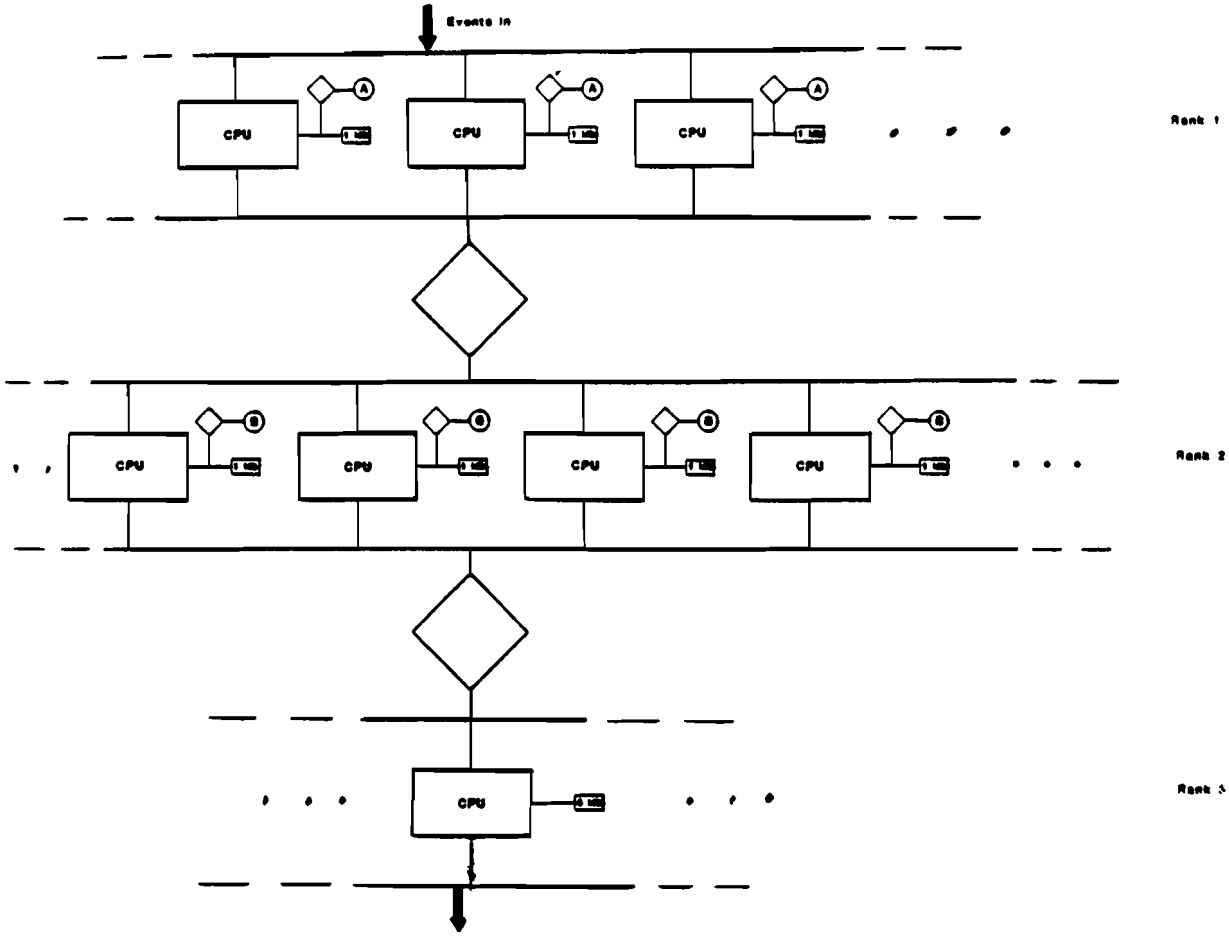

FIG. 2