ONLINE COMPUTER FARMS
CONFIGURATIONS AND CAPABILITIES

L. R. Fortney
Duke University
Durham, NC

Two computer farm architectures are described, each with the ability to accumulate all of the data for one event into the memory of a single processor. An argument is given to allocate a major fraction of the computing resources to online facilities at the intersection regions, mainly in the form of multilevel farms of graduated single processor capability.

INTRODUCTION

It is now widely recognized that a parallel arrangement of identical microprocessors, each executing identical code and driven by the incoming data, can process high energy physics events at higher rates than the most sophisticated single processor computer. With proper design, this farm of identical processors can have two fundamental advantages: high bandwidth and low cost. Its high bandwidth is derived by using a number of parallel busses, and its cost can be kept relatively low if the design is highly repetitive and simply connected.

The computer farm can be adapted to both online and offline applications, but because the SSC has the potential of essentially continuous event generation, the demarcation line between these two computing modes is likely to become blurred. Indeed, the farm concept offers the exciting possibility of full online reconstruction of events. We will assume here that each processor in the farm must be presented with all of the data for an event, so that it can effectively correlate different parts of the detector and possibly effect a complete event reconstruction.

Because of fundamental bus bandwidth limitations and the high throughput of large processor arrays, careful attention must be given to bus interconnection architecture. In the following discussion we assume 32-bit wide data busses operating at 10 KHz to give 40 Mbyte bandwidths. We take a worst case experimental requirement of 106 bytes/event and a 1 ms readout time. A 40 Mbyte maximum bandwidth implies that this data must be obtained from at least 25 different data channels, and must somewhere be assembled into a complete event. We examine two basic techniques, one compatible with the Fermilab Advanced Computer Program (ACP) design, and the other derived from the DO data acquisition system.

ACP Model

In an ACP compatible design the basic processor element is a single board computer which features a CPU, floating point coprocessor, and several megabytes of dual ported memory as shown in Fig. 1.

The VME bus is capable of the high bandwidth assumed here and can be used as a model. Speed, control, and handshaking specifications limit this bus to a single 19 inch crate, but multiple crates can be interconnected with an external data transfer bus of the same speed but more restrictive specifications. As shown in Fig. 2, a single crate would typically contain a bus control board, a DMA board, a spy or monitoring computer board, and perhaps an empty test slot. The remaining 16 slots can be filled with parallel processors, each with the capability of about one VAX-780. For discussion purposes we label this basic farm element a "row". The spy processor can be a typical ACP processor, but it would run a special program which monitors and summarizes the operation of the other processors in the row.

Event Builder

Since each ACP processor must be presented with the data for a complete event, it is necessary to build the event in an external device. A typical device is shown in Fig. 3. This figure assumes an idealized readout scheme where each channel conveys about the same amount of data, more or less synchronized in time.

The event and data channels are coupled at their crossover points by first-in first-out (FIFO) buffers. The FIFOs must have sufficient depth and control complexity to handle the rate and record length fluctuations of the input channels. If for simplicity we assume that full band width utilization of the event channels can be maintained, we would need only 25 event output channels. The control logic (not shown) of the event builder would switch incoming data onto the next available column of FIFOs. Each column of FIFOs empties onto an output event channel, and with the 25-in 25-out configuration, every event channel is active at all times.

Row Organization

Each event channel can be connected to one or more rows of processors, which to maintain the metaphor can be called a "garden". The output/control DMA from each row in a garden would be bussed onto a bidirectional garden channel as shown in Fig. 4. This channel would read processed events and summary information out of the garden and would also be used to download programs and control information.

If we assume each row is fully populated with 16 processors, the size of the farm is determined by the number of rows in each garden. If we are to run the event and garden channels at about the same data rate as the row (VME bus) bandwidth, then a minimum of two rows is required: one row saturated with read-in while the other interacts with the garden channel. A configuration where each garden has three rows yields a 1200 processor farm, and can be incremented in units of 400 processors (adding one row to each garden) subject only to physical space or event bus length limitations. At any instant, one row in each garden is receiving an event from its event channel, leaving the other rows free to interact with the garden bus.

For a variety of reasons, it may be desirable to add more layers to this farm model. If the volume of output events from each garden is large, it may not be possible to handle the combined data from several gardens on a single garden bus. In this case the gardens may need to be organized into clusters (tracts?) each with its own output bus as shown on Fig. 5. The need for this organization is clearly dependent on the event size and the rejection rate in each processor.

Speed Considerations

In the model just described, each processor would have about one second to process an event of about 10^6 bytes. Clearly a 1 MPS processor can only do justice to a small fraction of these bytes in one second. One second may be adequate to perform some type of full event filtering in the manner of a "quick and dirty" hardware trigger, but does not provide nearly enough time for online reconstruction of the event. Indeed, a reasonable estimate for reconstruction on a 1 MPS machine is 10^3 seconds.

The only rational model leading to reconstruction is a multi-layered process, where each level of processing eliminates at least 90% of the events. We note that if only 90% of the events are eliminated by the first layer of processor gardens, the output will still require a 100 Mbyte/second bandwidth and need at least a three tract organization.

Efficient online event reconstruction will likely require a very large processor memory, much larger than is justified on the first level of event filtering processors. A complete farm might therefore consist of processor gardens at several levels, with later levels populated by more powerful and more expensive processors. This design is sketched in Fig. 5.

DO Model

An alternate event building scheme is derived from the data acquisition system of the DO experiment. The basic processor element in this model is shown in Fig. 6. When extended to the SSC problem, the dual ported memory available to each CPU must be divided into 25 modules, with each module fed by a different data channel. The CPU modules would be organized into a row as shown in Fig. 7. Note that there is no separate event builder or event channels in this scheme. Instead, each processor module is directly connected to the data channels. The rows could still be organized into gardens and tracts as shown in Fig. 8.

While the DO model does not require a separate event builder, its bus architectural presents a serious problem on the SSC scale. Since it is rather difficult to imagine twenty-five 32-bit data busses terminating on a single board, the processor element would need to be spread over several boards. Each processor element would then occupy several slots of a crate, raising the overall cost per processor. However, if relatively few processors are involved the event builder savings might justify this approach, especially for the first level of a multilayer farm.

Online Verses Offline Analysis

If the processor farm architecture is successful and able to provide considerably more event processing power per dollar than conventional computers, then the division of resources between farms and conventional computers becomes an important question. Because of the inherently high bandwidth of the farm architecture, the point where farmed events are funneled into a conventional processor is likely to be a bottleneck in any analysis scheme.

One solution to this bottleneck is to fully reconstruct all accepted events online, saving for most events only the reconstructed four-vectors. Since this approach produces much smaller event records (a data summary record), a given bandwidth to the conventional computer and its storage medium will handle more events/second and permit a looser trigger. Since there would be no second or third pass over the raw data to recover incorrectly reconstructed events, this method would require continuous equipment calibration and very tight control of the online software.

However, these very factors also insure the best possible detector operation. The extensive online software, available with the same priority as the detector and trigger hardware, would immediately flag subtle detector failures which would otherwise only be seen offline. The fact that the SSC will be a nearly DC machine with an increasing luminosity in its early years also argues for treating the reconstruction software as simply an extension of the detector.

Most succinctly, the four-vector method saves less physics per event but more events for physics.

Summary

There appear to be no inherent limitations to applying the processor farm concept to the data acquisition system needed by the SSC. Presently available 10 MHz busses connected in parallel are able to provide the necessary bandwidth for 10^9 byte/second data acquisition, and the sophistication of the software event filter would appear to be limited only by the resources allocated to the online system.

If the possibility of the four-vector approach is to be preserved, the online facility should receive a large share of available computing resources. Because of the inherent bandwidth limitations of any long distance communication link, this argument favors the allocation of major computing resources to each intersection region (IR). Since online and offline event processing strategies both favor the high bandwidth farm architecture, an IR computing complex consisting of several levels of locally connected processor farms of varying capability would provide the most flexible arrangement of computing power per dollar.

For such a scheme to be successful, considerable effort must be expended on the operating system driving the farms. If this is sufficiently evolved, the farms appear capable of handling all experimental computing jobs with the exception of the rapid thru put of a single event.

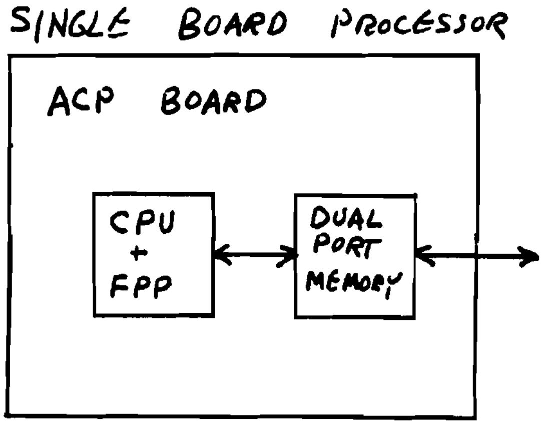
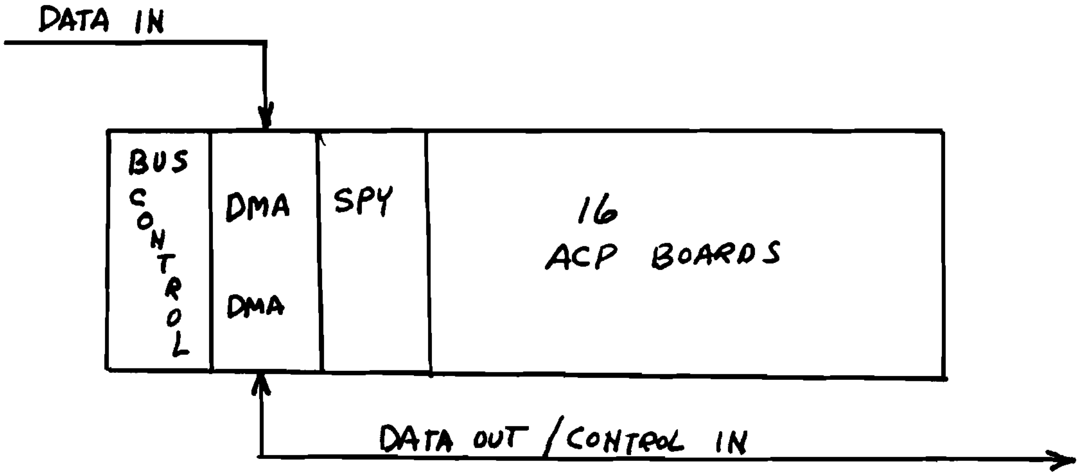


FIGURE 1



ONE ROW OF A COMPUTER FARM

FIGURE 2

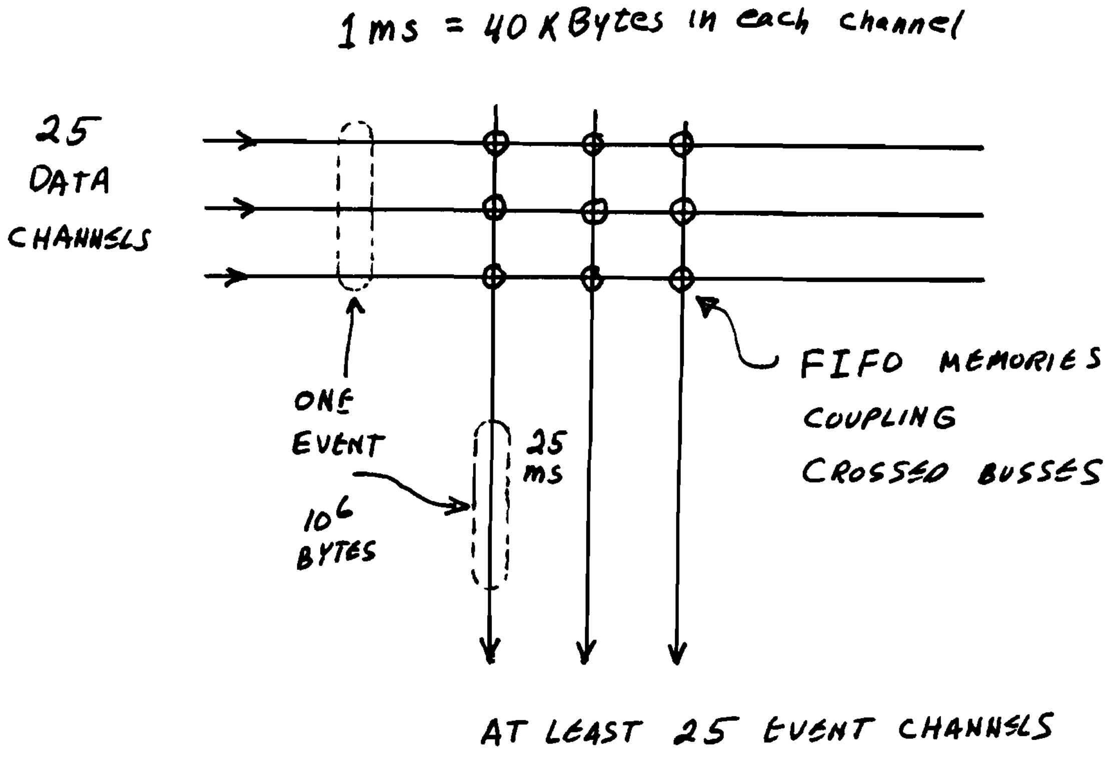

1 ms = 40 K Bytes in each channel
25 DATA CHANNELS
ONE EVENT
10^6 BYTES
25 ms
FIFO MEMORIES COUPLING CROSSED BUSSES
AT LEAST 25 EVENT CHANNELS

FIGURE 3

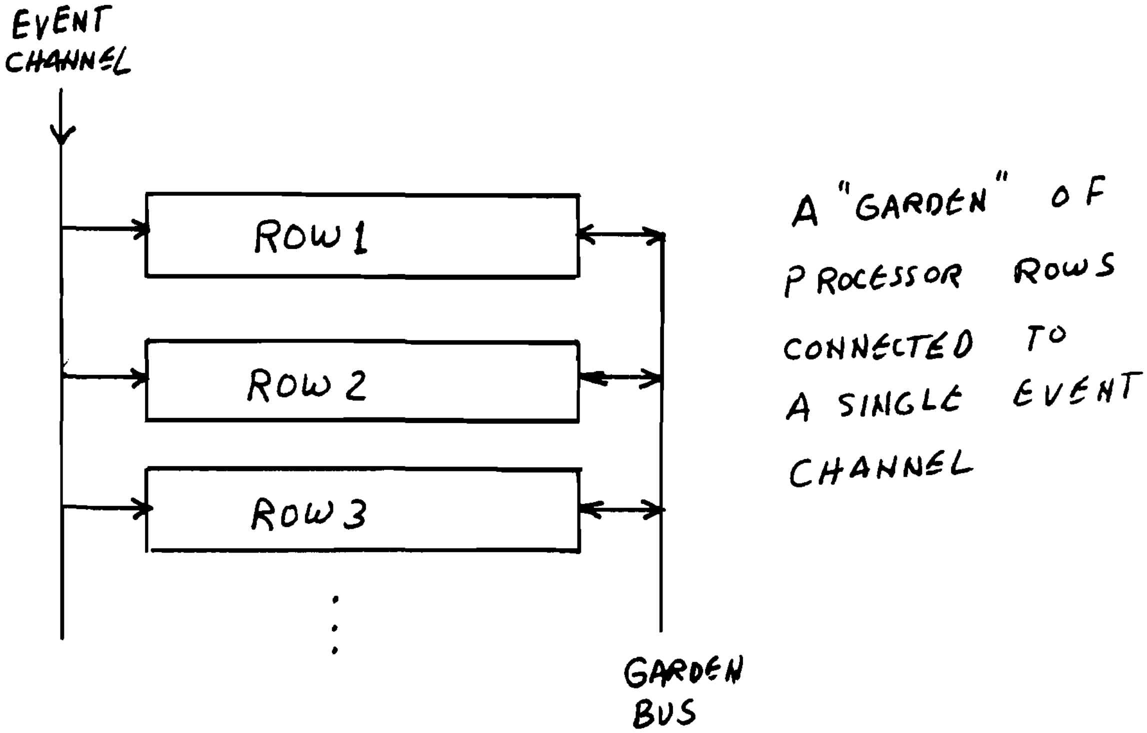

EVENT CHANNEL
ROW 1
ROW 2
ROW 3
GARDEN BUS
A "GARDEN" OF PROCESSOR ROWS CONNECTED TO A SINGLE EVENT CHANNEL

FIGURE 4

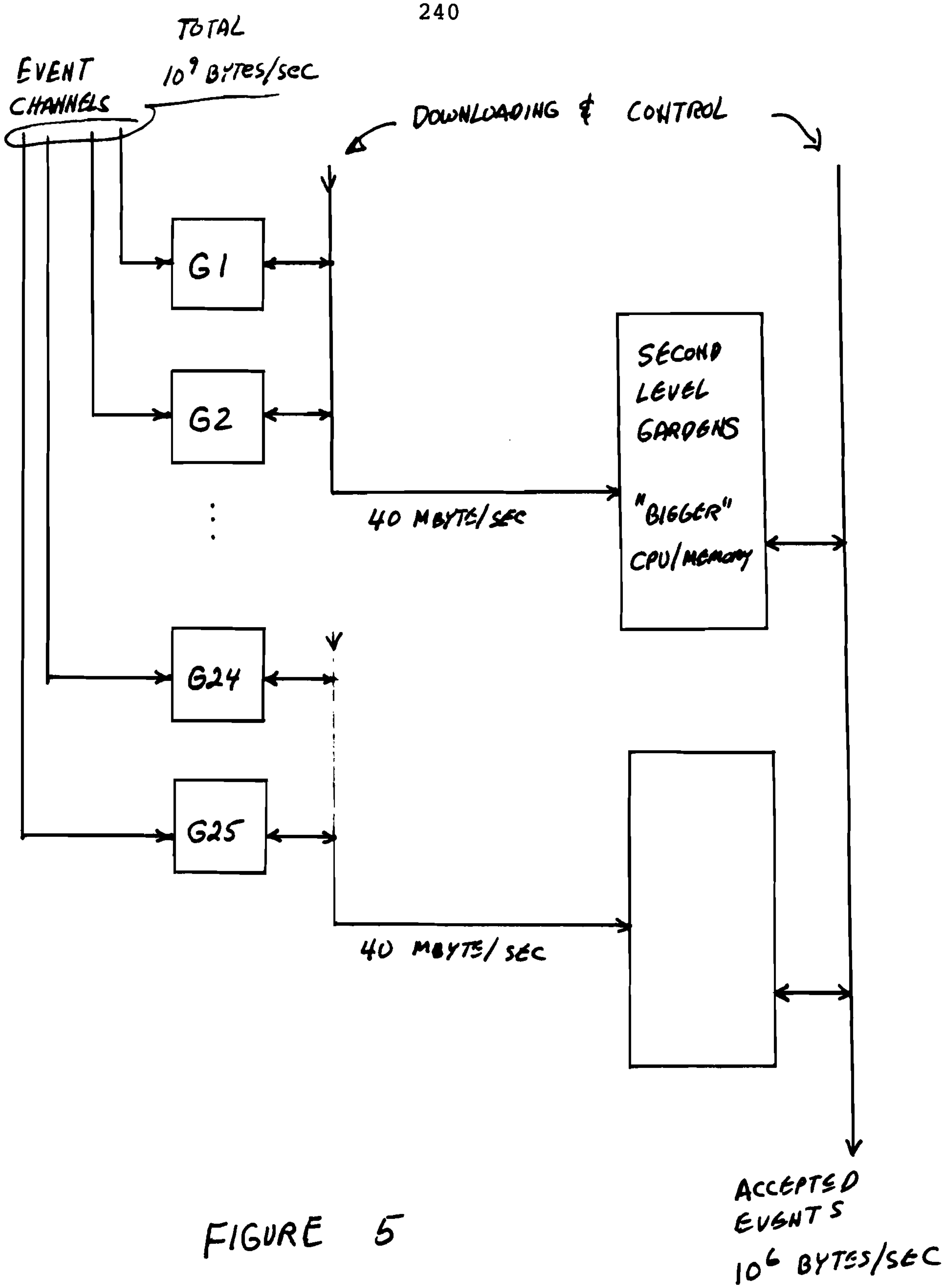

TOTAL
10⁹ BYTES/SEC
EVENT CHANNELS
DOWNLOADING & CONTROL
G1
G2
G24
G25
SECOND LEVEL GARDENS
"BIGGER" CPU/MEMORY
40 MBYTE/SEC
40 MBYTE/SEC
ACCEPTED EVENTS
10⁶ BYTES/SEC

FIGURE 5

DØ MODEL

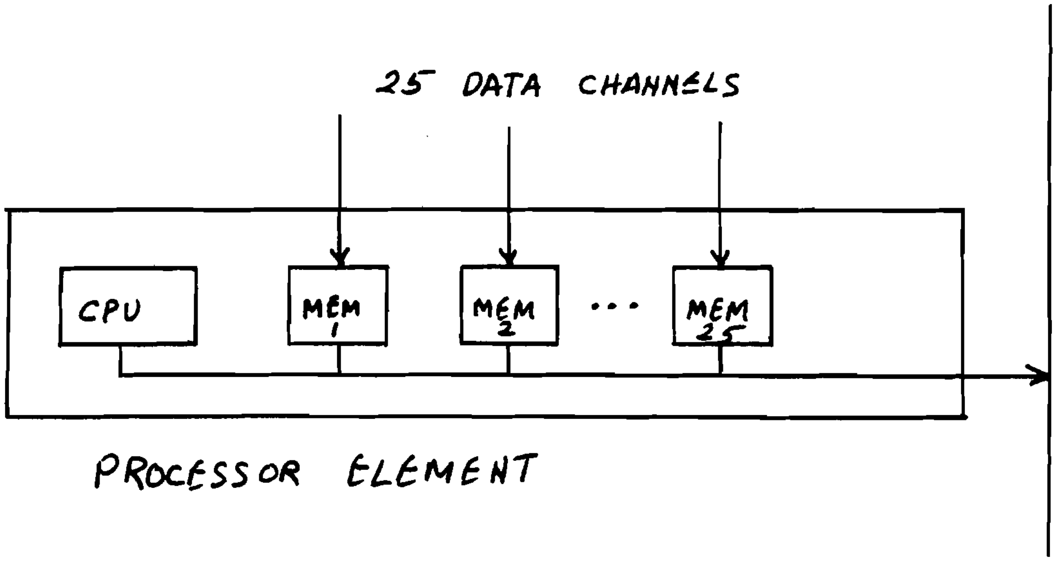


FIGURE 6

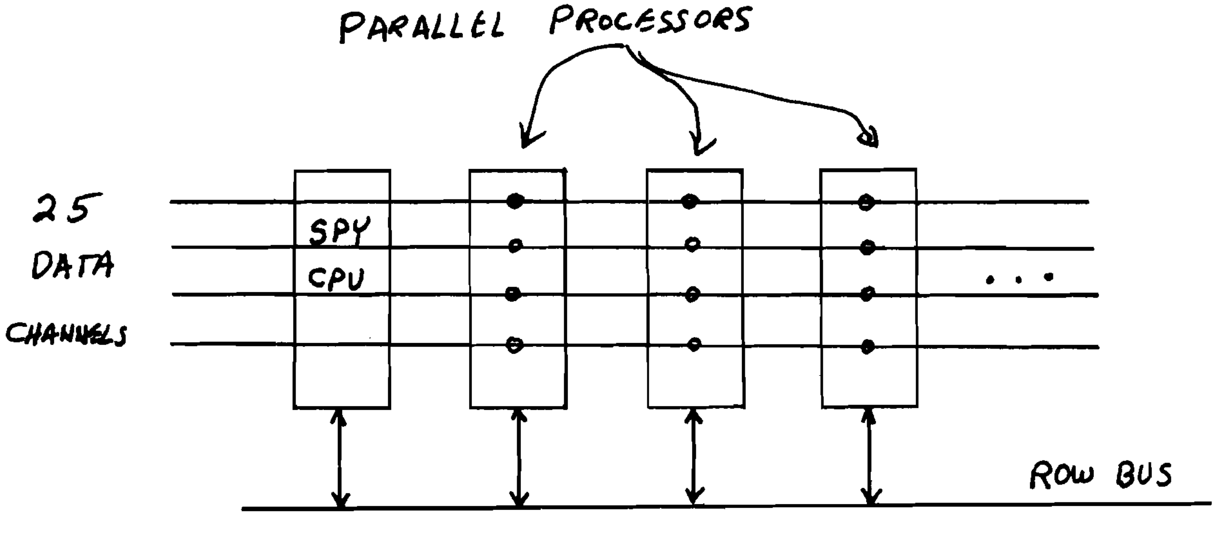


FIGURE 7

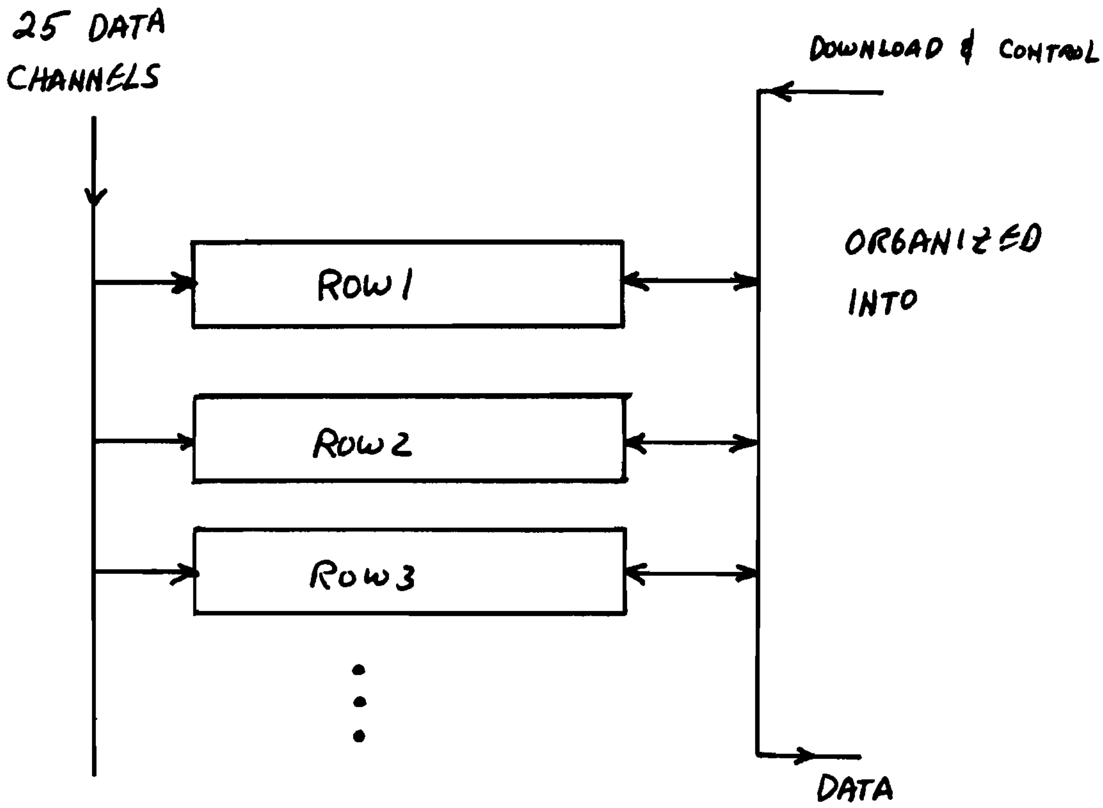
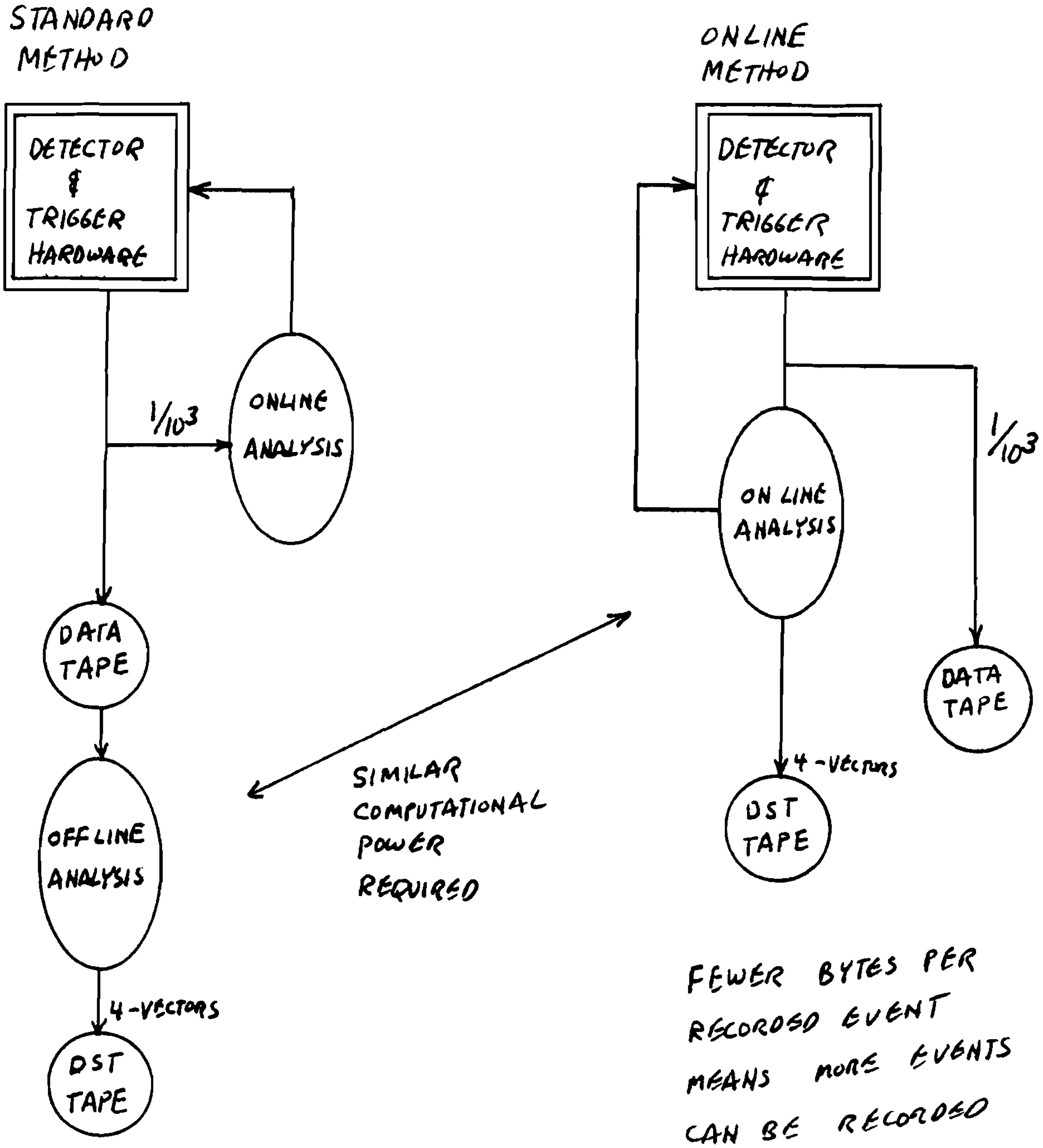

25 DATA CHANNELS
DOWNLOAD & CONTROL
ROW 1
ROW 2
ROW 3
ORGANIZED INTO
DATA

FIGURE 8



FEWER BYTES PER RECORDED EVENT MEANS MORE EVENTS CAN BE RECORDED

FIGURE 9