сообщения
объединенного
института
ядерных
исследований
дубна

E10-87-378

I.F.Kolpakov, A.E.Senner, V.A.Smirnov

A SUPERCOMPUTER FOR PARALLEL DATA ANALYSIS

1987

1. INTRODUCTION

A dramatic gap between the real time event registration rate of HEP spectrometers and moderate data processing possibilities has stimulated extensive studies in this field.

Event patterns from elementary particle physics experiments are sequentially processed in a traditional computer. Technological and physical limitations of hardware modules do not permit one to overcome essentially a speed of 10 Mops/s for computers in the nearest future.

A nontraditional parallel supercomputer with a 50 Mops/s equivalent speed is proposed. This supercomputer would allow one to satisfy JINR demands for data processing from elementary particle physics spectrometers of a new generation and specially from DELPHI.

The supercomputer makes it possible to solve also some actual tasks of accelerator and elementary particle physics, which allow one to separate a main algorithm into many parallel subtasks, in particular Monte-Carlo simulation and calculation of accelerator orbits.

2. SUPERCOMPUTER CONCEPT

An expert analysis of user's demands of the JINR Central Computing Facilities over a period of 1986-1990 has revealed that 52% of computing power demands at JINR are used by the groups of physicists involved in research on elementary particle physics spectrometers. Their requests over this period can be satisfied by computers with a total speed of 50 Mops/s.

The proposed modular supercomputer is based on a set of commercial 32-bit processors with a speed of approximately 1 Mops/s each. All the processors are liaisoned on the basis of VME-bus structures. The system as a whole is controlled by a microVAX II computer. Data input and output are performed using a standard set of computer peripherals. The system software is based on FORTRAN-77. The supercomputer is integrated with the JINR Central Computing Facilities through the port of the existing net to provide access of all the users to it.

A similar computer system was put into operation at Fermilab in 1986 [1].

Other alternative approaches of high speed parallel event processing were also taken into consideration. The first approach assumes a configuration of the microVAX with a set of IBM PCs. An evident disadvantage of the system is a poor speed-to-cost ratio as compared to the set of processor boards. Another disadvantage of the PC system is a relatively small throughput of I/0 channels for information exchange. The time of data transmission exceeds the one of their processing. This situation complicates the use of IBM PC processor power for most experiments during 1986-1990.

The second approach represents the development of an emulator of a traditional powerful computer. This approach requires a lot of men's power for the development of hardware and software. This is because at the present time there are no available emulators, even they are not expected in the near future. Moreover in this case again the speed of the system in principle cannot exceed an order of magnitude.

Both these circumstances also lead to the poor speed-to-cost ratio of the system.

The kernel of the proposal assumes the application of parallel processing of data from elementary particle physics spectrometers insted of sequential one. Since events from the spectrometers are statistically independent, a number of independent processors can be used in parallel to process these data. In traditional computers (Figure 1a) events are analysed step by step, and the processing rate is limited by the CPU performance. In the proposed supercomputer, event data are distributed over multiple processors. Each processor performs an analysis in a fully independent way using the same software (Figure 1b). A relatively small volume of transmitted data in comparison with their analysis time permits one to involve tens and hundreds of parallel processors in the supercomputer system proportionally increasing the system performance.

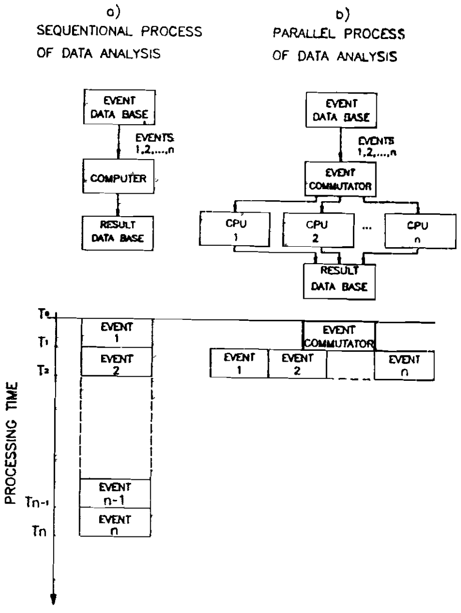


Fig. 1. Data analysis concept in a) a traditional computer, b) the proposed supercomputer.

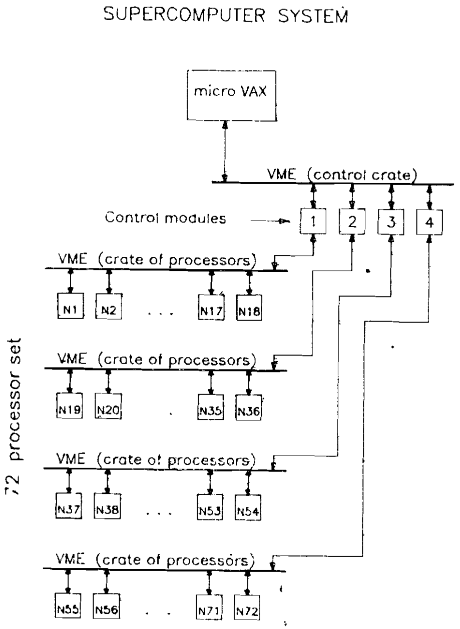


Fig. 2. Supercomputer architecture (N1 ÷ N72 – processor modules).

3. SUPERCOMPUTER ARCHITECTURE

The modular supercomputer system includes three main parts: a set of processor modules, a microVAX II computer with peripherals and interfaces to the JINR Central Computing Facilities. The introduction of the microVAX II as a host computer into the supercomputer system will also make compatible the data analysis with the DELPHI collaboration requirements. The microVAX computer provides the preparation and loading of computing tasks for the processor modules, readout of experimental information from data summary magnetic tapes or through the JINR LAN port and logging it into the processors, acquisition and evaluation of data analysis results from the processor set. Figure 2 demonstrates the supercomputer architecture which is based on 72 processor modules distributed over 4 VME crates as a stage of project implementation. Each VME crate contains 18 processors. The processor crates are radially connected to a VME control crate interfaced to the microVAX II. In principle, the system architecture provides the integration of 255 processor modules.

The main processor module (Figure 3) of the supercomputer is based on a MC 68020 32-bit microprocessor, a MC 68881 floating point coprocessor, a MC 68851 paged memory management unit, a 8-MByte RAM, an EPROM, and a VME-bus interface. The floating point coprocessor performs a complete set of floating point operations with the words having a 64-bit mantissa, a sing bit and a 15-bit signed exponent. The paged memory ma-

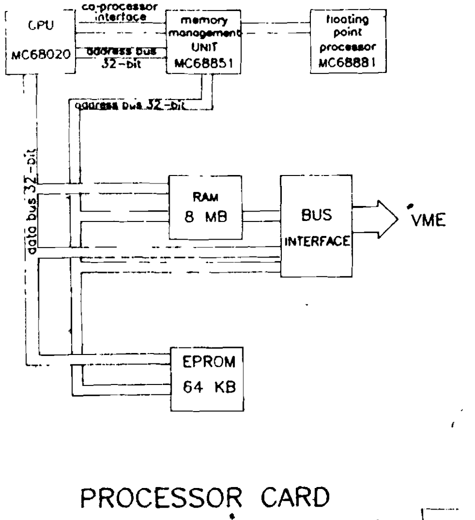


Fig.3. Processor module architecture.

nagement unit provides a microprocessor interface to the floating point coprocessor through an auxiliary bus to obtain an optimal performance of processor module.

The rate of data exchange via two direct memory channels of the microVAX computer with the control crate is 0.6 MBytes/s. The first direct memory channel of the microVAX computer will be used for program loading into the processor modules and for experimental data transfer to the processors. The second channel is planned for the transfer of data analysis results from the processor set to the microVAX II.

4. SUPERCOMPUTER PERFORMANCE

The supercomputer performance is influenced by the following factors:

- the number of main processor modules;
- performance of one processor unit;
- thoughput of the link channel between the host computer and the processor set.

The performance[2] was estimated during the experimental data analysis with a set of 53 processor modules. The performance of one processor was found to be 0.7 in VAX 11/780 units. Thus, with a reasonable degree of confidence the performance of the basis supercomputer processor module based on a MC 68020 microprocessor and a MC 68881 processor is 0.6 ÷ ÷ 0.7 of the VAX-11/780 performance (2 Mops/s) in the analysis of data obtained from elementary particle physics spectrometers.

The throughput (C) of the host computer liaisons to the set of processor modules has no direct influence on the supercomputer performance. However, the underloading of some processor modules and some deterioration of the supercomputer performance can occur for a certain class of tasks. It depends on the information volume of one event and the time of event analysis by one processor module (T_{an}).

A maximal number of parallel processors (N_{max}) which should be used to obtain the highest efficiency of the supercomputer is evaluated as

$$N_{max} = \frac{T_{an} \cdot C}{m + K},$$

where m is the number of bytes from information sources and k is the number of resulting information bytes obtained for one event after its analysis and transferred from the processor modules to the host computer via the link channels.

The main contribution to C is due to the interface between the microVAX and the VME control crate the speed of which is 0.5 Mbytes/s. The VME-bus throughput has no influence on C because its theoretical speed is 40 MBytes/s.

The evaluation of requirements of experimental data analysis at JINR during 1986-1990 shows that $N_{max} \approx 70$ satisfies most tasks. For example, during the data analysis of DELPHI the supercomputer performance should be close to the optimal one at $N_{max} = 82$ according to expression (1). The total supercomputer performance for the proposed set of 72 processor will be equal to the power of 40 VAX-11/780 computers and will be no less than 50 Mops/s for the worst case estimations.

5. SOFTWARE

Software efficiency of the supercomputer is based on the following necessary data processing algorithm attributes:

- data fragments are independent of one another;
- the problem can be easily distributed over several parts;
- there exists an algorithm which needs a large computing time:
- the time of data transfer is much less than the processing time of these data.

As mentioned above, the following classes of tasks satisfy these attributes in the field of physical research at JINR:

- high energy physics data processing;
- simulation of particles and nuclear interactions and their registration;
- simulation of charged particles and nuclei transport through magneto-optic channels;
- lattice calculations.

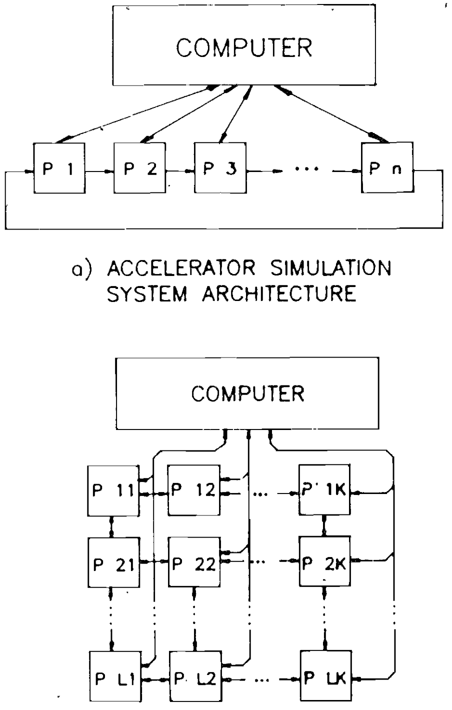
The sequence of tasks is arranged according to intensity of their use at JINR.

In the case of experimental data processing, an event is an information fragment for analysis by means of identical microprocessor software.

Each processor module independently executes an identical simulation algorithm during the simulation of particle and nuclear interactions.

An individual processor module imitates a single element (magnetic lens, bending magnet, space gap and so on) in the case of accelerator simulation. The calculated results of one processor module are transferred to the next one which imitates the following element of the simulated system (Figure 4).

Each processor module excutes a certain required algorithm for a lattice task. Input data for it are boundary conditions and the results obtained in neighbouring processors during the last iteration (Figure 4).



a) ACCELERATOR SIMULATION SYSTEM ARCHITECTURE

b) LATTICE GAUGE CALCULATION SYSTEM ARCHITECTURE

Fig. 4. System architecture for a) an accelerator simulation, b) a lattice gauge calculation.

Basic software components are the following: an interaction subsystem, software development tools (in particular, a FORTRAN-77 compiler), general purpose libraries (CERNLIB, HBOOK, ZBOOK, HPLOT, GEANT and so on) and a supercomputer imitator. Most of these components are available and need no resources to be involved.

Below some system and applied software problems and the organization of user program execution by the supercomputer are described in more detail.

6. USER PROGRAM MANAGEMENT

A user program is divided into two parts for executing by the proposed system (Figure 5).

The first part is processed by the host microVAX computer. The functions of the host computer are the following: service of all I/0 operations (magnetic tapes, printers, graphic units and so on), data preparation for the processor modules, data transmission to the processor modules and acquisition of results from them.

The second part of the user program is situated in the processor modules. Each processor accepts input data, stores them and accumulates statistics in its operative memory. There are no input-output operations in this software part. After completion of event processing, the processor module software sends, if necessary, (for example, in a DST generation mode) results to the microVAX and becomes ready for operations with a new event.

The parameters of the user program for current execution are provided by loading certain constants during the initiation program phase.

When a stream of input data finishes, the host computer collects all statistical results from the operative memory of the processor modules, accumulates them and sends the results to line printers and graphic units.

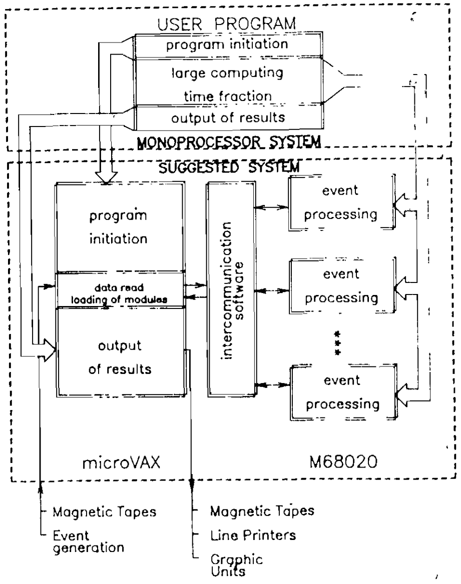


Fig. 5. Software managent in a) a traditional computer, d) the proposed sypercomputer.

- The above-mentioned classes of tasks are easily realized in the frame of the user program organization.

7. PROCESSOR INTERCOMMUNICATION

The intercommunication of the processor modules is accomplished by an intercommunication software subsystem. The subroutines of this subsystem provide just a limited number of functions. These functions allow one to broadcast initial values of constants, to send data from the host computer to the processor modules, to retrieve results from them, to identify their status, to collect and to accumulate statistics from all processor modules.

Such an approach supposes a high level of flexibility of the proposed system with respect to a user program. Its main advantage is the opportunity to formalize an unavoidable procedure of user program modification for executing it by the multiprocessor system. Note that an additional number of FORTRAN statements does not exceed a few tens and is practically independent of the user program size.

The limited number of functions of the intercommunication software subsystem also simplifies learning and usage problems.

8. SOFTWARE DEVELOPMENT TOOLS

The microVAX computer is supposed to be a basic one for software development. It provides high level language compilers (in particular, FORTRAN-77), a source editor, a file system, a linker, a loader and other traditional tools for software development.

There are two ways to manage the microprocessor software development. The first one is to use cross-tools (a cross-compiler, a cross-linker, a cross-loader and so on) for the processor modules on the host computer.

The other way implies the application of a FORTRAN-77 compiler of the processor module software. In this case the object code is transmitted to the host computer, and then it can be loaded into all processor modules.

The development of special software which imitates logical interlinks in the proposed system seems to be desirable. Such an imitator also serves as a preliminary test tool of the modified user program.

9. USER ACCESS TO THE SUPERCOMPUTER

There are three types of user access to the supercomputer.

The first of them is microVAX terminal access. In this case the user is serviced by an operation system of the host computer, and no software installation or development is needed.

Intercommunication between the microVAX and the EC-1055M provides access to the supercomputer through the High Energy Laboratory LAN. The realization of this option requires the use of certain computer-computer data transmission software.

Access to the supercomputer for other JINR users is provided by the integration of the microVAX with the JINR LAN and by the installation of the intercommunication deck KERMIT on the host computer.

10. MULTIUSER SERVICE

The proposed system is oriented to solve a wide range of the problems listed in section 5, and any user can have access to the supercomputer.

The execution of user programs is organized in a batch mode. A new user program begins to be processed by the system only when a previous user program is completed. Thus, at each instant of time the system services only one user. It is not expedient to support a multiprogram mode. Such an approach decreases substantially the development of the required software and its complication.It allows the development time to be diminished and reliability of the system to be increased.

A user through the above-mentioned access means directs his job to the microVAX. The job stays in an input queue according to its priority. This job starts to be executed when the current job is completed and there are no jobs of higher priority in the input queue. The user program load time to the supercomputer depends on the size of the program. The estimates show that for a 1 Mbyte program size the load time is no more than 3 minutes.

After the job is completed, the results are transmitted to the printer and graphic units of the microVAX or are sent to a remote user through the JINR LAN. Optionally the results can be written on a magnetic tape and then transferred to other research laboratories.

11. USER PROGRAM DEBUGGING

User programs for supercomputer processing have undergone some formal modifications. Despite a moderate scale of these modifications (a few tens of FORTRAN statements), it is desirable to have debug and test tools for the execution of a user program on the supercomputer.

The first test stage of the user program is its execution by the supercomputer imitator. This allows one to check the validity of division of the program into two parts: for microVAX and processor module executions.

The basic tool for debugging the processor module software is a source debugger which belongs to the TMS operation system.

12. DELPHY DATA PROCESSING

For the first year after LEP commissioning, the total data processing time including event filtration, simulation, calibration and data analysis is evaluated as 32000 CPU hours in IBM-168 units/4/.

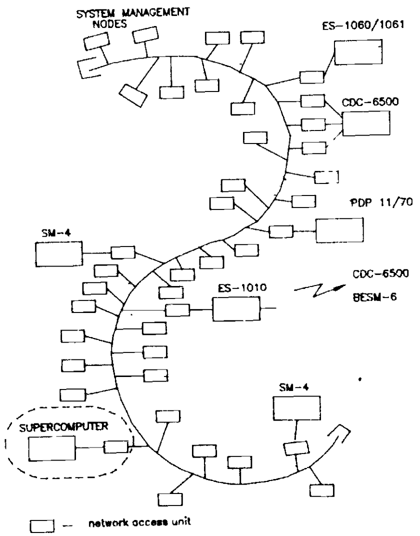
This experiment is characterized by a significant size of one event buffer (up to 1 Mbyte). In this case the operative memory of the processor module should be 8 Mbytes. At present the use of a larger memory volume is not required.

The speed of data analysis of the supercomputer allows one to organize a cluster of DELPHI data processing/3/ at Dubna. The use of the proposed supercomputer makes it possible to solve problems of total raw information processing, DST generation, DST analysis and Monte-Carlo simulation. The solution of these problems demands 93% of the net CPU time/4/. Thus, the JINR physicists have a unique chance of a true partnership in the DELPHI data analysis and participation in the development of its software.

The modular architecture of the system allows the processor module memory to be enlarged, if necessary, in the future.

13. SUPERCOMPUTER INTEGRATION WITH THE EXISTING JINR LAN

Supercomputer coupling to the JINR Central Computing Facilities can be accomplished through the existing JINR LAN (Figure 6). It is provided by an asynchronous RS232C standard interface of the microVAX II host



Supercomputer in the structure of the JINR Network

Fig. 6. Integration of the supercomputer to the existing JINR LAN.

computer. The maximum exchange rate of the supercomputer with the LAN is 9600 baud. A higher exchange rate (up to 0.2 Mbytes/s) provides coupling of an ES computer channel to the microVAX II bus because the supercomputer will be situated in the vicinity of it (20 m). This link also allows one to integrate the ES computer resources with the supercomputer and to accelerate the task execution though the Laboratory LAN. Access to the supercomputer for all the users of JINR is available by coupling the microVAX computer to the JINR LAN and by the KERMIT communication software adoption. The users access to the proposed system througth the microVAX II terminals is completely based on its operational media and so the design or the adoption of some additional software are eliminated.

MicroVAX II to ES-1055M coupling provides access to the supercomputer through the Laboratory LAN ports. Its realization requires some adoption of the existing software to provide data transfer between the computers.

It should be noted that the JINR LAN provides an efficient use of the supercomputer because the real transfer time between the computers does not exceed that from a magnetic tape.

14. CONCLUSION

The realization of the supercomputer solves one of the principal JINR problems, namely it satisfies demands for the computing power for data analysis and opens up long-term opportunities for future computer requirements. The modular architecture permits one to expand the computer in the future by adding new modern processor modules thus improving its performance.

The desing of the supercomputer allows one to watch the modern level in the analysis of physical experimental data[1], to use the ready software in the frame of interational cooperation and to provide prospects for increasing the performance of the Central Computing Facilities up to 200 Mcycles/s during 5 years.

The principle of the supercomputer concept and availability of its constituents (processor modules, microVAX II, VME bus and software created during the project realization) make it possible in the future to develop powerful centers for data analysis of electronic experiments at physics institutes of the JINR member-countries just as the film information analysis centers were organized two decades ago.

REFERENCES

1. Gaines I. et al. The ACP Multiprocessor System at Fermilab. Fermilab-Conf–87/21, FNAL, Batavia, 1987.

2. Nash T. et al. The ACP Multiprocessor System at Fermilab. Fermilab-Conf-86/32, FNAL, Batavia, 1986.

3. Papel H. Off-Line Interactive Centre. CERN/DELPHI 85-43, PROG-27, 21 May 1985.

4. DELPHI. Technical. CERN/LEPS83-3, LEPS/P 2, 17 May 1983.

Received by Publishing Department
on Juune 2, 1987.