EFFICIENCY OF MULTIPLE BUS STRUCTURE

I. Rozman, M. Colnarič, B. Stiglic*
TEHNIŠKA FAKULTETA MARIBOR
* ISKRA AVTOMATIKA LJUBLJANA

UDK: 681.3.02

ABSTRACT - Analysis of efficiency two or more buses linked withabus linker is shown in this article. A queueing theory is used. Analysis exactly valid only for exponential distributions for both λ and μ. It is shown how the linking of two buses influences the mean bus response time in comparision with the architecture with consists of one bus and the same number of computers that are connected to two buses.

INTRODUCTION

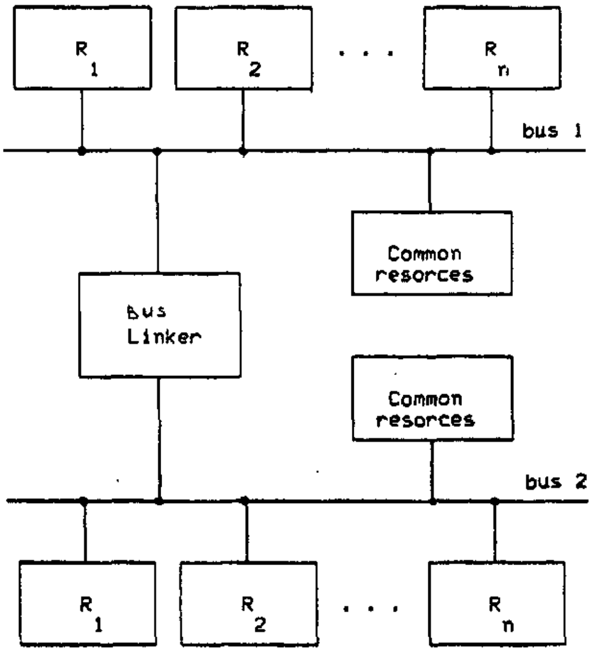
Analysis of efficiency of multicomputer architectures with a common bus is well know in literature /1/, /2/, /3/, /4/. A model of these architectures is derived. An anlytical treatment of this model is based on a queueing theory or better on already derived equations for mean time exsistence in the system which is described by the queue M/G/1/N. With the aid of introduced approximations is shown that results obtained without major tolerance are valid also in cases where distributions are unexponential i. e. in such cases which can be treated by a queue. For this queue an exact mathematical solution is not known. This statement is also valid for the calculation of throughput /but not for the mean bus response time W_N/ in the cases where the arbiter is not FCFS. But in all these cases, individual processors which are connected on a common bus perform a statistically equal work. In literature /3/ is shown the approximation which transforms a model with a statistically unequal work into a model with a statistically equal work.

The problem of conneting two or more buses where computers are connected to each bus still remains an open question. The most important problem is how the linking of two buses influences the mean bus responce time in comparision with the arhitecture which consists of one bus and of the same number of computers that are connected to two buses.

DESCRIPTION OF ARCHITECTURE

The linking of two buses to each other is made for the following reasons:

- the increase of the throughput of the whole architecture is greater if we add one more bus with connected computers;
- the realization price for a bus linker is low in coparison with the price of the whole system;
- the fault tolerance is existing.



Picture 1: Two-Bus-Linking

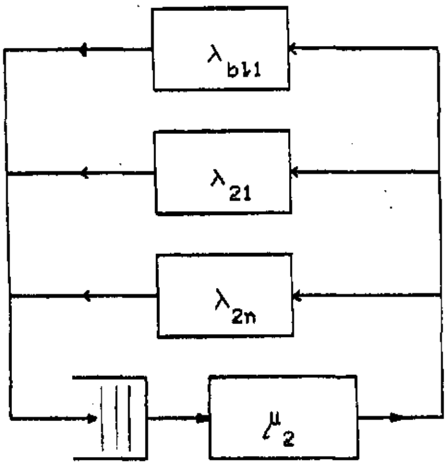
Bus linker is an active interface which enables two-direction communication between two buses. In its inherent structure, it must contain a memory with enough capacity. Into this memory computers write messages for the computers which are located on the other bus. A too small memory causes an increase of the response time beacuse the messages have to wait to obtain bus and also have to wait that the memory is empty. Therefore it is convenient that the memory is great enough for a two-side transmission in order to permit the transmission of all computers from one bus to the computers in the other bus in the same moment. The bus linker also packs messages into a block. When the bus linker obtains another bus, it permits the transmission of all existing messages in the block to the computers to which the messages are addressed.

THE MODEL

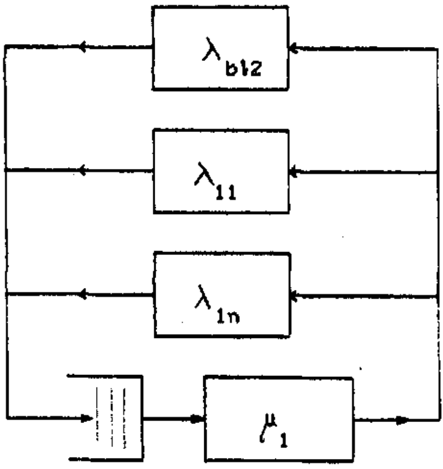
Searching for the suitable model the following suppositions are made:

- bus linker acts undependently for each direrection of transmission;
- the time which is necessary for the transmission of message through the bus linker is short in comparisson with the bus occupation time. Therefore it can be neglected;
- the arbiter at each bus is FCFS;
- the bus linker acts in the sense of bus occupation always then when the message enters its empty memory. In cases there are still messages in the memory (the bus linker is waiting for bus occupations), the new message is loaded into the memory.

On the basis of the above mentioned suppositions and the queueing theory the model is formed.



Picture 2(a): The model for transmission of messages in the following directions: bus 1 - bus 2



Picture 2(b): The model for transmission of messages in the opposite direction

In the model shown on picture 2 the bus linker is devided into two parts - separately for each direction. The influence of one bus upon the other is expressed by the source which generates the bus occupations λ_{BL1} or λ_{BL2} according to the direction of transmission. $\lambda_{11}\ldots\ldots\lambda_{1n}$ are processors which generate bus occupations for bus 1 with the mean time between bus occupations $\frac{1}{\lambda 1i}$. μ presents a bus with the mean bus occupation time $\frac{1}{\mu_1}$. The same impretertation is valid also for bus 2.

THE ANALYSIS

In the analysis of the model only exponential distributions are taken into account for both λ and μ. Other distributions cannot be taken into account exactly. In such cases some approximation methods should be applied. The exponential distribution leads to the solution of the quene Mi/Mi/1/N.

The mean time of retardance in the system is solved by Ferdinand /5/,/6/.

$$u_i = \frac{\lambda_i}{u_i} \qquad (1)$$

$$Z_N = \sum_{d_1 \ldots d_i \ldots d_n} \left(\sum_{k=1}^{N} (1-d_k)\right)! \cdot \prod_{k=1}^{N} \mu_k^{1-d_k} \qquad (2)$$

$$d_k = \begin{cases} 1, & \text{request from source k waiting for or being serviced} \\ 0, & \text{request source k is in operational state} \end{cases} \qquad (3)$$

$$W_i = \frac{1}{\mu_i} + (1 - \frac{U_N}{L_q}) . \sum_{k=0}^{N} \frac{1}{\mu_k} \cdot u_k \cdot \frac{d}{du_k} \ln Z_N$$

$$U_N = 1 - \frac{1}{Z_N} \qquad (4)$$

$$L_q = \sum_{i=1}^{N} u_i \cdot \frac{d}{du_i} \ln Z_N$$

The calculation Wi according to the equation (4) is difficult. When the value N is high, the calculation of Z_N is not simple. In the calculation Wi the derivation $l_n Z_N$ is necessary to be calculated which additionaly complicates the whole procedure.

Ferdinand /5/, /6/ presents the efficiency of each element $U_N^{(i)}$ as a probability that i^th element is not waiting for or being serviced:

$$U_N^{(i)} = Z_{N-1}^{(i)} / Z_N \qquad Z_{N-1}^{(i)} = (Z_N)_{u_i = 0} \qquad (5)$$

$U_N^{(i)}$ can be expressed with the following expression:

$$U_N^{(i)} = \frac{\frac{1}{\lambda i}}{\frac{1}{\lambda_i} + W_i} \qquad (6)$$

from which the mean bus responce time Wi for the element i can be derived (7).

$$W_i = \frac{\frac{1}{\lambda_i}(1 - U_N^{(i)})}{U_N^{(i)}} \qquad (7)$$

If we wont to calculate the throughput of the whole architecture according to the equation (8),

$$T_p = \sum_{i=1}^{N} \frac{1}{W_i + \frac{1}{\lambda_i}} \qquad (8)$$

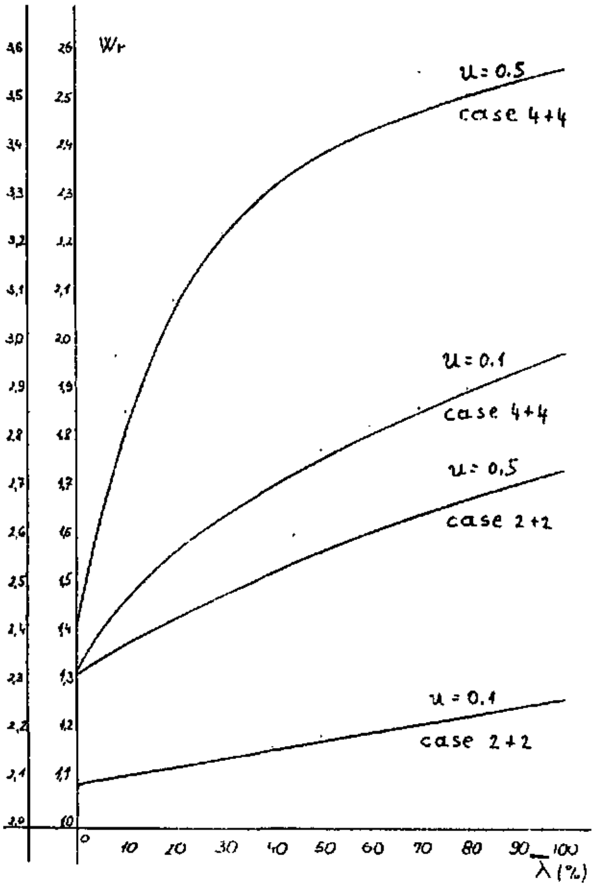
the correct result will not be obtained. The equation (8) is valid only in such cases in which each task obtains the bus. In our case a part of tasks is concluded on the level of one bus. The equation (8) involves only those tasks which occupy another bus but does not include the local ones.

THE NUMERICAL RESULTS

In the presentation of a numerical calculation two examples are shown (picture 3). They clearly present all features which are typical for a link with two buses. In the first example two computers are connected to each bus. In the second example four computers are connected to each bus.

The parameters u_1, i = 1 ... 4 are equal in the first example.

The parameters u_1, i = 1 ... 8 in the second example are equal, too.



Picture 3: Wr vs. % $\bar{\lambda}$

Wr is normative value of the mean bus responce time of the bus:

$$W_r = \mu . W_i \qquad (9)$$

$\bar{\lambda}$ presents the mean arrival of demands in one bus:

$$\bar{\lambda} = \lambda\ (N - L) \qquad (10)$$

In the equation (10), L stands for the mean time of retardance in the system for the queue M/M/1/N. This parameter can be simply calculated.

From the picture 3 we can see that the increase of traffic through the bus linker approximately parabolically prolongs Wi. The parabolicity becomes more sharp with the saturation.

If we compare Wi for one processor in the two-bus architecture with that in one bus architecture with that in one bus architecture, under the same conditions, we notice that Wi is always smaller in the two-bus architecture. In the first case, in the two bus architecture, when u_i, $i = 1 \ldots 4 = 0{,}1$, $\lambda = 0{,}1$ is $w_r = 1{,}090$ while $w_r = 1{,}320$ in one bus arhitecture.

In the second case we can notice a similar difference in favour of two-bus architecture. This difference becomes more stressed in higher density of traffic (higher u_i). This difference is caused by the packing of messages into a block.

CONCLUSION

From the results obtained we can conclude that the link of two or more buses is especially effective in such cases where one bus becomes saturated. Also the price for the bus linker is not so high that it cannot justify the realization of the linker. The price is approximately 5 % of the price of the whole architecture. When we deal with a very highly coupled system, one bus can be devided into several buses which are connected by bus linkers. This leads to the so called cluster architecture.

REFERENCES

/1/ Rozman, I., Colnarič, M.: Model multiprocesorske/multiračunalniške arhitekture s skupnim vodilom, Jugoslovensko savetovanje o mikroprocesorskim sistemima MIPRO 83, Rijeka 1983, str. 2.109-2.111.

/2/ Colnarič, M., Rozman, I.: Simulacija multiprocesorskega računalnika s skupnim vodilom, Jugoslovensko savetovanje o mikroprocesorskim sistemima MIPRO 83, Rijeka 1983, str. 2.32-2.38.

/3/ Rozman, I., Colnarič, M., Bonačič, D.: Analiza učinkovitosti arhitekture s skupnim vodilom pri statično neenakem zasedanju vodila, Jugoslovensko savetovanje o mikroprocesorskim sistemima MIPRO 84, Rijeka 1984, str. 3.80-3.84.

/4/ Rozman, I., Colnarič, M.: Modeliranje MP/MC arhitektur s skupnim vodilom, Jugoslovanski mednarodni simpozij za računalniško tehnologijo in probleme informatike, Informatica 83, Ljubljana 1983, str. 14-18.

/5/ Ferdinand, A.E.: A Statistical Mechanical Approach to Systems Analysis, IBM J. Res. Develop. Sept. 1970, pp 539-547.

/6/ Ferdinand, A.E.: An Analysis of the Machine Interference Model, IBM Sistem J., No 2, 1971, pp 192-202.