Shared Memory Multistage Clustering Structure: An Efficient Structure for Massively Parallel Processing Systems

H.S.SHAHHOSEINI, M.NADERI, R.BUYYA*

Electrical Engineering Department
Iran University of Science & Technology
Narmak, Tehran, 16844, Iran.

***** School of Computer Science and Software Engineering
Monash University, Melbourne, Australia.

EMAIL : h_s_shahhoseini@hotmail.com

Abstract

In this paper a new structure is proposed for the massive parallel processing systems. This structure is expandable in vertical and horizontal manner and cover many of the previous computer designs. The queuing theory and Jackson queuing network is applied for constructing an analytical model for the proposed structure. This model gives a closed-form solution for the system performance metrics, such as processor's waiting time, system processing power, and so on. Two novel points in development of these analytical models are: application of open queuing network rules for analyzing a closed queuing network, and calculation of the input rate of each service center as a function of the input rate for previous center.

The model can be used for evaluating the MPP system or optimizing its specification on design space.

Keywords: MPPs, Multistage clustering structure, Shared memory, Analytical modeling, Queuing theory.

1. Introduction

The design of the massive parallel processing (MPP) systems has been the subject of the sustained research by many persons in recent years[1-7]. One of the primary goals in designing these systems is scalability, or linear increase in processing power by increasing the number of processing units [8,9]. The main problem for constructing a scaleable system, is the conflict over the common resources such as memory modules and interconnection networks, I/O units and so on. The limited service capacity of these common resources cause an increase in the waiting time of the processors when the number of the processors increases. The high waiting time causes lower the processor utilization and consequently the system would become non-scaleable. Using more powerful common resources is the conventional method for decreasing the waiting time, but the capacity of servicing of the resources such as the effective memory access time and the interconnection network bandwidth is saturated by the technology and their structures. This problem would be more important in MPPs that utilize more than one thousand processors.

In this paper a new structure for overcoming to the above problems in the design of MPPs is proposed. In the proposed structure, processors are divided into groups or clusters, and organized in several stages. So the new structure is called MSCS (multistage clustering structure). MSCS is expandable in vertical and horizontal manner, and is so flexible that it covers many of the previous parallel machine designs. This structure can be used by the shared variable based and the message passing based systems [10] . In this paper we investigate only the shared variable based structure.

In the next section the MSCS for the shared variable based MPPs is introduced. In section 3 an analytical model is constructed for the proposed structure and in section 4, on base of the analytical model, an example of the shared variable based system is studied and the performance graphs of the system are depicted. Conclusions are presented in the last section.

2. Multistage Clustering Structure

To introduce the MSCS, consider a basic cluster that include several processing units, memory units, two interconnection networks and some other units that will be described later. This basic cluster is depicted in Figure 1. Each processing unit has a local memory for its own computation, and there is a shared memory for facilitating the communication between processors. A horizontal communication network (HCN) is used for transmitting data between processors and shared memory. Moreover the basic cluster include a unit for I/O operations and a unit for supervisory and managing the processors. A vertical communication network (VCN), is used for transmitting control signals, and vertical expansion of the system.

The basic cluster can be expanded in two ways: increasing the number of the processing units, or using several basic clusters with one additional memory that is shared by those clusters. By applying the second way, a two stage system is constructed (Figure 2). It must be noted that in the second level of the system, there is a HCN that connect the VCN of each basic cluster to SM_2. The units that are located inside the basic clusters are indicated by index 1 (such as SM_1 , HCN_1 ,...), and the units that are located outside of the cluster are indicate by index 2 (such as SM_2, HCN_2 ,...).

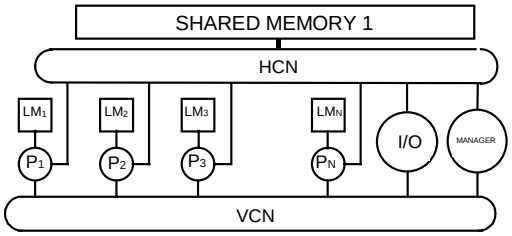

Figure 1. Basic cluster

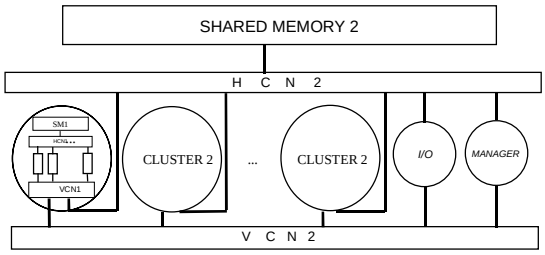

Figure 2. Two stage system

By the similar method one can expand the system vertically and constructing s-stages system. A cluster in i-th stage of the s-stages system is depicted in Figure 3. This cluster include some processing cluster (or PCs), one I/O cluster, and one managing cluster. Also there are two interconnection networks, HCN_i and VCN_i , that transmitting data inside and outside of the cluster respectively. This system can expand vertically by increasing the number of stages or horizontally by increasing the number of PCs in each level.

MSCS can adapted to previous structure, by choosing different interconnection network. For example, if MSCS includes only one level and one cluster, it map to traditional multiprocessors, if it includes two level, it can be mapped to clustered parallel machine (such as CEDAR, UltraMax), and so on.

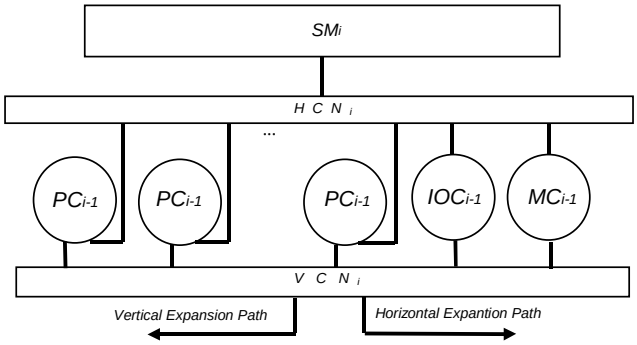

Figure 3. A cluster in i-th stage of s-stages system

In MSCS-based system, if the number of PCs that make a cluster will be equal for all clusters of i-th stage, the system is called homogenous at level i. If a system is homogenous in all level it will be called homogenous and if system will not be homogenous at least in one stage, it will be called non-homogenous or heterogeneous. In next section, the homogenous multistage cluster system will be discussed and analyzed.

3. Analytical Modeling

For evaluating the performance of the system, queuing theory and Jackson queuing networks is applied. Consider a system that is constructed based on homogenous MSCS. In this system any processor perform a piece of the main program, that is called processor's job. During job execution, it is probable that a job needs to communicate with the other jobs. Therefore several queues would be constructed for each interconnection networks and shared memories.

We assume the following assumptions for the analyzing a system with MCSC structure:

1. The number of PCs in i-th stage of system is C_i and the number of processor in each basic cluster is C_0.
2. The inter job communication requests are generated independently by processors.
3. The destination of each request will be uniformly distributed between other processors' jobs and the probability of outgoing request from i-th stage indicated by P_i.

4. The time between two consecutive requests have exponential statistical distribution with parameter of λ.
5. Access time to memory in i-th stage have exponential statistical distribution with parameter of μ_{m_i}.
6. The service time of the interconnection networks in i-th stage have exponential statistical distribution with parameter of μ_{h_i} and μ_{v_i} for HCN_i and VCN_i respectively.
7. Conflict over memory modules and interconnection networks will be resolved by queuing center with FCFS discipline.
8. Requester processors must be waited until they offer service as per the above scheme; and during waiting period, they cannot generate any other request.

The above assumption describe our system completely. So one can consider a Jackson closed queuing network to analyze this system [11]. For analyzing a Jackson queuing network, the input rate of each stage must be computed, and any service center should be assume as M/M/1 queuing center [11-13]. Since the results of the analyze should be used for design of MPPs with a large number of units, the volume of computation for closed queuing network will be very large. We apply an open queuing network rules for analyzing the closed queuing network, and also derive the input rate of each service center as a function of the input rate for previous center. These techniques reduce the volume of calculation and simulation time considerably. By investigating the situation of each request of an individual processor in this system, one can reach to the state diagram that depicted in Figure 4.

As shown in Figure 4, all of the requests that departs from HCN_i will pass through the SM_i with probability of one. So we only compute input request rates of VCNs and HCNs. The processor requests will be directed to service center HCN_1 and VCN_1 by probability of $1-P_1$ and P_1 respectively. If the request rate of a processor will be λ, the input rate of HCN_1 and VCN_1 that originate from that process or will be $\lambda(1-P_1)$ and λP_1. Since there are C_0-1 processors in each basic cluster, the requests that receive to HCN_1 and VCN_1 originating from other processor in the same cluster, that are indicated by γ_{h1} and γ_{v1}, will be $\lambda(1-P_1)(C_0-1)$ and $\lambda P_1(C_0-1)$ respectively.

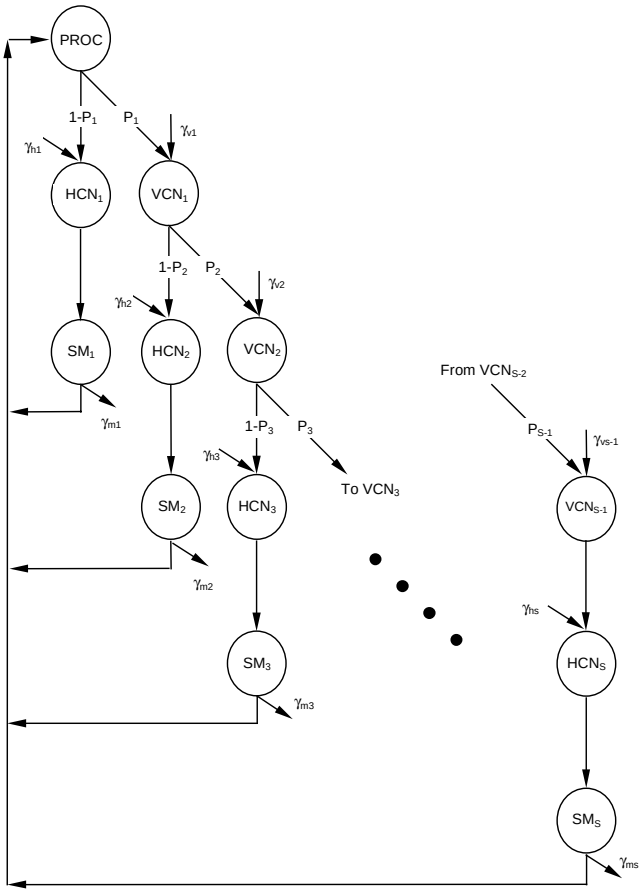

Figure 4. State diagram of a multi stage Clustering MPPs with s-stages

So the total requests of the processors that received to service centers in the first stage, can be computed by following equation:

$$\lambda_{v1} = P_1.\lambda + (C_0-1)P_1.\lambda = C_0.P_1.\lambda \quad (1)$$

$$\lambda_{m1} = \lambda_{h1} = (1-P_1)\lambda + (C_0-1)(1-P_1)\lambda = C_0.(1-P_1).\lambda \quad (2)$$

In the i-th stage of the system, the input request rate that originate from each PCs is λ_{vi-1}, so one can derived input rate of each service center by the similar method:

$$\begin{aligned}\lambda_{vi} &= P_i.\lambda_{v(i-1)} + (C_{i-1}-1)P_i.\lambda_{v(i-1)} \\ &= C_{i-1}.P_i.\lambda_{v(i-1)}\end{aligned} \quad (3)$$

$$\begin{aligned}\lambda_{mi} = \lambda_{hi} &= (1-P_i).\lambda_{v(i-1)} + \\ &\quad (C_{i-1}-1)(1-P_i).\lambda_{v(i-1)} \\ &= C_{i-1}.(1-P_i).\lambda_{v(i-1)}\end{aligned} \quad (4)$$

and in the last stage, there is no request for outer cluster so:

$$\lambda_{vs} = 0 \quad (5)$$

$$\lambda_{ms} = \lambda_{hs} = C_{s-1}.(1-P_s)\lambda_{v(s-1)} + C_{s-1}.P_s.\lambda_{v(s-1)} = C_{s-1}.\lambda_{v(s-1)} \quad (6)$$

Based on M/M/1 queue's equation [11,14,15], one can compute the queue lengths of each center for all stages ($k = 1,2,\ldots,s$). Then the average of total waited processors in the system can be computed base on the number of service center in the system and the number of the waited processor in each center:

$$L = \sum_{k=1}^{s} \{(L_{vk} + L_{hk} + L_{mk}) \prod_{i=k}^{s-1} C_i\} \quad (7)$$

It must be noted that according to assumption 8, the waited processors would not be able to generate request, and in such a situation the effective processor's request rate would be lower than the λ. The effective request rate will be decreased with the same ratio as the active to the total processor's number. L and λ are iteratively computed, till their changes in two consecutive steps will be negligible. After determining effective request rate and waited processor, the waiting time can be computed by the equations (8)-(11):

$$W_{mk} = \frac{1}{\mu_{mk} - \lambda_{mk}} \quad (8)$$

$$W_{hk} = \frac{1}{\mu_{hk} - \lambda_{hk}} \quad (9)$$

$$W_{vk} = \frac{1}{\mu_{vk} - \lambda_{vk}} \quad (10)$$

$$W = \sum_{i=0}^{s} [P_{vi}.W_{vi} + P_{hi}.W_{hi} + P_{mi}.W_{mi}] \quad (11)$$

In equation (11) P_{vi}, P_{hi}, P_{mi} are the probabilities of referring a processor request to VCN_i , HCN_i and SM_i respectively and can be computed by following equation:

$$P_{vi} = \prod_{j=0}^{i-1} P_{j+1}$$
$$P_{mi} = P_{hi} = \frac{(1-P_i)}{P_i} \prod_{j=0}^{i-1} P_{j+1} \quad (12)$$

By determining the average waiting time of a processor for each communication request, we can compute the processor utilization as follows:

$$PROCESSOR\_UTILIZATION = PU = \frac{1}{1+\lambda W} \quad (13)$$

Finally the most important metric for evaluating of the system's performance, i.e., total processing power of the system (TPP), can be computed on base of single processor power (SPP), by the following equation:

$$TPP = N \times PU \times SPP = \frac{SPP}{1+\lambda W} \prod_{i=0}^{s} C_i \quad (14)$$

4. Design Tradeoff

In this section the capibility of analytical model is investigated. The model can be used to determine the performance metrics, such as the processor utilization and the total processing power of the system. In fact the performance model is useful not only for evaluating the system performance for given configuration, but also for investigating the effect of different parameters' varation on the system performance. The last capibility of the performance model can be used during the system design. The follwing discussion illusterated these concept.

Consider a system that used from 2700 pieces of the 400MIPS RISC processors, the mulibus interconnection networks (with 100MB/Sec bandwidth for each single bus) and the memory modules with 20ns access time. It is assumed that the system is organized in a 3 stages by MCSC structure. The other assumptions regarding to the system specifications and parameters are indicated in Table 1.

At the first glance it may seem that the 2700 PCs of 400 MIPS processor, must give total processing power of 1'080'000 MIPS. This processing power is reachable if there is no overhead by parallelism. For reaching to the maximum processing power and finding the best structure, we study the processing power curves versus the number of clusters in each stages.

	Quantity	Unit
Processor's Power	400	MIPS
Total processor	2700	Pieces
1st stage Memory	90	Modules
2nd stage Memory	5	Modules
3rd stage Memory	30	Modules
Total Memory	125	Modules
1st stage Bus	85	Single bus
2nd stage Bus	10	Single bus
3rd stage Bus	20	Single bus
Total Bus	300	Single bus
Inter Job Communication Probability	0.2	%
Memory Reference Per Instruction	1.4	-

Table 1. System assumption

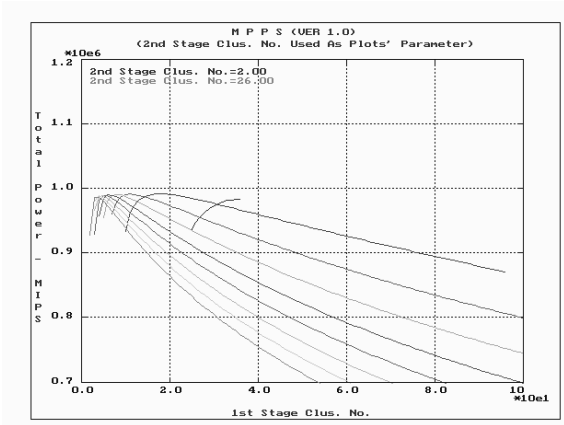

Figure 5. Total Processing Power vs. Cluster numbers for 3-stage MSCS system

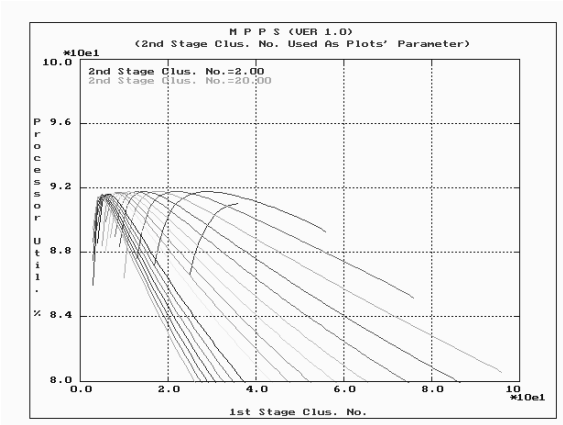

Figure 6. Processor's Utilization vs. Cluster numbers for 3-stage MSCS system

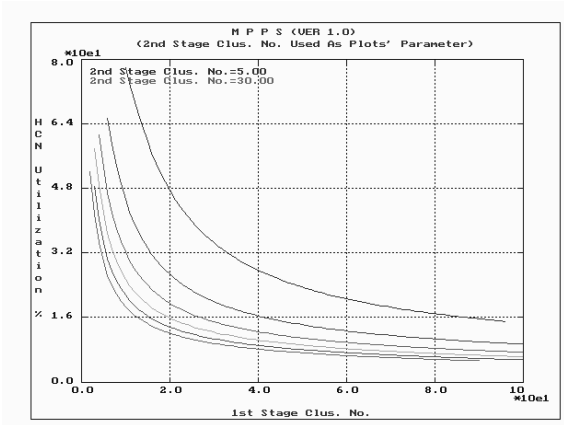

Figure 7. Bus Utilization vs. Cluster numbers for 3-stage MSCS system

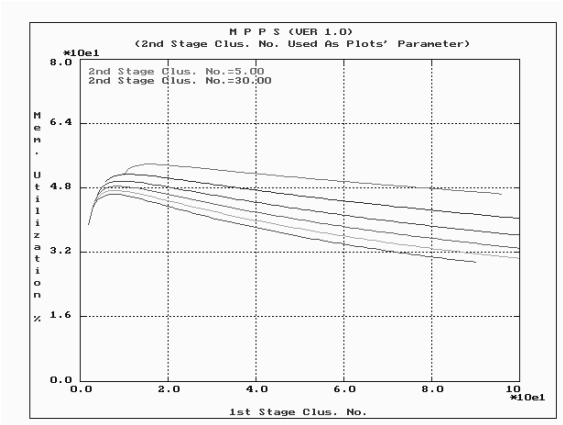

Figure 8. Memory Utilization vs. Cluster numbers for 3-stage MSCS system

Figure 5 depicted these curves when the first stage's cluster number varies from 1 to 100. The curves are plotted for some values for the second stage's cluster number from 2 to 26. Depicted curves show that the maximum processing power will be occurred on 18 clusters on the first and 5 clusters on the second stages. For this configuration the processing power will be equal to 991'200.62MIPS or 0.991TIPS.

Figure 6 shows the variation of the processor utilization. These graphs also confirm the previous results, i.e. the best configuration for a system with specification of the Table 1, in 3 stage construction, will be held on 5 clusters of 18 sub-clusters, that each sub-cluster consists 30 processors.

There are some other metrics that may be used for evaluating the system. For example Figures 7 and 8 show the bus and the memory utilization. Increasing the first stage cluster number, cause the request for common resources on next level, is increased, and the request for the common resources of the inside of the cluster will decreased. So there will be a trade off for utilization of the inner and the outer cluster resources in each stage. It must be noted that Figure 7 and 8 show the average values of resource utilization in the system. So these average metrics may be increased and decreased for different values of the cluster number.

As depicted in Figure 8 the memory utilization will increased by increasing the first stage cluster number and then it decreased. So there is an optimal point in each curves. The optimum point on curves regarding to second stage cluster of 5 is about 18 that confirms previous values.

The bus utilization curves are not agree with above discussion. The closer observation of Figure 7 shows that by increasing the inner or outer cluster number, the bus utilization is decreased. It means that the busses is a bottleneck of our example, and its total capacity in not sufficient or its allocation to different stage is inappropriate.

In this example it is showed how can evaluate the system that used MCSC structure. Performance of these system is dependent on several interrelated parameters and various constraints. It is extremely difficult to come up with an optimal design satisfying all the requirement. Designer must consider requirement upon their own priorities and optimized the design for their own propose.

5. Conclusions

The design and evaluation of the massive parallel processing system is a considerable interesting problem. In this paper we proposed a new structure and its analytical model for MPPs. Analytical model was constructed on queuing theory, and the system performance metric was expressed as mathematical equations. The model was used for plotting the system performance metrics.

The performance graphs, may be used by designer to find the optimum system configuration for reaching to maximum performance with fixed resources.

The future works focuses on improving the analytical model for heterogeneous system, or using neural net to determine the optimum point in design space. The other subject is improving the analytical model by applying software and scheduling features.

References

[1] A.Louri, B.Weech, C.Neocleous, "A spanning multichannel linked hypercube: A gradually scaleable optical interconnection network for massively parallel computing," IEEE Trans. on Parallel & Distributed System, pp. 497-511, May 1998.

[2] S.P.Dandamudi, D.L.Eager,"Hierarchical interconnection networks for multiprocessor systems," IEEE Trans. Comput., pp. 786-797, June 1990.

[3] P.Mohapatra, C.R.Das and T.Y.Feng, "Performance analysis of cluster based multiprocessors," IEEE Trans, on Comp. Vol. 43 pp. 109-114 Jan 1994.

[4] E.Deckker, "Architecture scalability of parallel vector computer with shared memory," IEEE Trans. On Comp. vol.47, No.5, pp.614-624, May 1998.

[5] W.T.Hsu, P.C.Yew, "The performance of hierarchical systems with wiring constraints," in Proc. Int. Conf. Parallel Processing, pp. i9-i16,Agu. 1991.

[6] W.S.Lacy, J.L.C.Rivera, D.S.Wills, "The offset cube: A three dimensional multicomputer network topology using through-wafer optics," IEEE Trans. On Parallel & Distributed System, pp. 893-907, Sep 1998.

[7] P.Mohaparta, C.R.Das, "Performance analysis of finite buffered asynchronous multistage interconnection networks," IEEE Trans. on Parallel & Distributed System, pp.18-35 Jan 1996.

[8] K.H.Hwang, Z.Xu, Scalable parallel computing. McGraw Hill, 1998.

[9] D.A.Patterson, John L.Hennessy, Computer architecture a quantitative approach, 2nd Ed., Morgan Kaufmann Publishers Inc. 1996.

[10] H.S.Shahhoseini, " Structural design & modeling of the multistage clustering parallel processing system," PhD Dissertation, Iran university of Science & Technology, March 1999.

[11] D.Gross & C.M.Harris, Fundamental of queuing theory, John Wiely& sons 1974.

[12] R.L.Disney, "Queuing networks," American Mathematical Society Proceeding of Symposia in Applied Mathematics, Vol.25, pp.53-83, 1981.

[13] W.J.Gordon, and G.F.Newell, " Closed queuing systems with exponential servers," Oper. Research, Vol.15, pp.254-265, 1967.

[14] L.Keleinrock, Queuing systems, vol I:Theory. NewYork Wiely, 1975.

[15] L.Keleinrock,Queuing systems, vol II, Computer Application. NewYork Wiely, 1975.

[16] H.S.Shahhoseini , M.Naderi , "An improved model for performance evaluation of multiple-bus multiprocessor system", Proc. of International Conference on Computer System and Application ,pp-151-154 Irbid, Jordan, April 1998.