Research on Impact of Two Levels Switching in a LAN for Cluster Computing

Mamta* Research Scholar, CSE, RIMT Mandigobindgarh, PTU India Charanjit Singh Assistant Professor, RIMT Mandigobindgarh, PTU India

Abstract—In the past few years there is a change in the view of high performance applications and parallel computing. The demand of high performance computers is always a hunt for scientists and researchers. Cluster computing is widely used in the world of computing for computer intensive applications. However, the design and implementation are always depends on the desktop machines. In this approach the computer nodes are connected with Fast/Gigabit Ethernet in a local area network. The activities of the nodes are orchestrated by “clustering middleware”, a software layer that sits atop the nodes and allows the users to treat the cluster as by and large one cohesive computing unit. TCP/IP protocol is to be studied for performance issues of cluster computing and to optimize communication among the networked machines keeping in view of latency and bandwidth.

Keywords—Cluster Computing, Bandwidth measurement, Latency, TCP/IP, Network performance, Parallel computing.

I. INTRODUCTION

Parallel processing in a network of PC’s are attracted a boost of attention and becoming one of the most promising areas of large scale scientific computing. Desktop machines with two, four or eight processors connected together have emerged as common platform for designing large scale data intensive applications. A cluster is made up of identical computers connected by a fast network. Building an efficient cluster means studying the issues related to performance measures. A collection of computer nodes, interconnected by a LAN and/or a high-speed switching network, all nodes can be used individually or collectively as a cluster. All computers in the cluster are glued together with middleware support for collective usage as a single computing resource, in addition to the traditional usage as independent computers. The use of interconnected PC’s as single logical computational resources has become a widespread approach to a cost effective computer intensive problems.

Every modern operating system provides industry standard software for computer communication but these software’s are so inefficient that it actually appear to act as the real technology bottleneck of cluster computing. A number of research groups in universities and industry are building efficient communication hardware and software for parallel computing on the network of PCs. The other reason for such an interest is because price/performance advantage of the network of PCs in comparison to supercomputers. Cluster is a type of parallel or distributed processing system which consists of a collection of interconnected stand-alone/complete computers cooperatively working together as a single integrated computing resource and provides scalability by adding servers to it or by adding more clusters to the network as the need arises. Since most of the organisations are flooded with Desktop PC’s and most of the time these are idle. These idle CPU cycles can be used for high performance computing in cluster. Cluster offers high system availability and reliability due to the redundancy of hardware, operating systems and applications. It provides data support and high performance massive storage by running cluster enabled programs.

The power of Internet and Intranet can be used for integrating remote and heterogeneous computer into a single global computing facility for parallel and collaborative work. To gain control over the resources of Internet-based computers for parallel computing has introduced new difficulties and problems that have never addressed by parallel computing in LAN (Local Area Network) environment. Cluster computing is potentially able to deliver high performance at the unbeatable price/performance and thus providing a low-cost alternative to both shared memory multiprocessors and distributed memory Massively Parallel Processors. In a multi-programmed environment, processor may be running independent tasks and it is also possible to have multiple processors executing the single program and sharing the code which are referred as threads. The full benefit of MIMD multi-processors with n processors will be achieved by having at least n threads or processes. The decomposition process divides the problem into small units called tasks that may be executed in parallel and it is the smallest individual unit that is executed by only one processor. The parallel execution of multiple tasks, which may be of different sizes on different processors results in speeding up the execution. The tasks that are independent may be executed in any sequence but some tasks that may use data generated by others need to wait for completion of other tasks and such interdependencies of tasks decide the order of execution.

TCP/IP is a standard protocol for the communication performance among the computers and it is studied for performance issues of cluster computing. It provides end-to-end connectivity specifying how data should be formatted,
addressed, transmitted, routed and received at the destination. It runs at a very low efficiency rate on modern LAN’s. The optimal values of TCP/IP parameter affect the latency and bandwidth. In this paper, we studied the behaviour of parametric TCP/IP and optimize the communication among the network machines keeping in view latency and communication load on bandwidth. Switched based LANs improved a lot in late nineties. It is worthwhile to harness the idle cycles of computers available on the network. In LAN latency involved in sending the data from one node to other node depends on the software overhead of the message passing library. The parallel computation time can be reduced drastically. The tuning of TCP/IP for sending and receiving the message for local and non-local nodes is kept under BDP (Bandwidth Delay Product) so that there should be no congestion in the network. Ping Pong benchmark is used for standardising the message size to be used for communication over the network.

II. EMERGENCE OF NETWORK MACHINES

Modern high-end Personal Computers (PCs) connected in a high speed local area network are eventually replacing uni-processor in every application field, as the performance improvement exhibited by uni-processor architectures during the last decade will soon become unable to satisfy the ever growing demand for higher and higher performance. Modern high-end Personal Computers (PCs) provide computation speed as well as storage capacity at the best price/performance ever. Therefore an obvious way to obtain higher performance, parallel systems has to build a distributed memory platform out of a pool of PCs interconnected by a fast Local Area Network (LAN) hardware, commonly called a cluster of PCs. Local Area Network provides huge amount of unused computational power that can be tapped to solve large complex problems in parallel. The unused cycles of Personal Computers (PCs) in a LAN can be combined together and form a Cluster enabled PC Cluster. Cluster is potentially able to deliver high performance at the unbeatable price/performance and thus providing a low-cost alternative to both shared memory multiprocessors and distributed memory Massively Parallel Processors. The nodes on the network are connected to each other through routers or switches. These routers or switches are used to pass message to each other. The nodes on the network can be both homogeneous and heterogeneous and communication among the nodes is carried through message passing protocol to make the data exchange compatible.

However standard communication protocols like TCP/IP run at a very low efficiency rate on modern LANs. This has the immediate consequence of poor performance achieved by parallel programs on clusters. In this attempt is made to obtain good performance out of Cluster enabled PC cluster by optimized TCP/IP windows socket buffers based on the windows operating system. Gupta et .al Studied the effect of TCP socket size for the local nodes and non-local nodes in a grid enabled PC Cluster for parallel computing. It shows that the obvious answer is to use level 1 Ethernet network with socket size 64 KB along with other parameter TCP window size value and additional parameter of GlobalWindowSize. Desk one has performed better in peer to peer topology but when the client move across the switches them Desk one is not suitable as the bandwidth chokes.

It proposed a scheme in which multiple programs or tasks carried out simultaneously. Which makes programs or tasks run faster because there are more CPU’s or cores running it without interfering with each other.

This approach relieves the user from worry of bottlenecks. But as the demand of high performance computers is always a hunt so to achieve high speed, low latency interconnection networks. So we analyse that it could be better to have modifications. In this paper study of TCP/IP protocol is done and optimal value of parameters, which are going to affect the latency and bandwidth, will be purposed. The network design consists of three components

Master PC (job submitter)
Server (Job synchronizer and partitioner)
Cluster of nodes (Task execution)

Master PC is responsible for defining job and submitting the job to the server then Server breaks the job and allocates tasks to client nodes and sends back the final results to the Master PC to display the results. Now it can be anywhere in the Intranet and connected to server in the peer-to-peer network fashion. Cluster can be divided into two classes: A dedicated and non-dedicated system, in non-dedicated each PC executes its normal work such as word processing or internet browsing and only idle CPU cycles are used to execute parallel tasks. In network evaluation, latency and
bandwidth are two parameters which affect the performance of both data intensive and compute intensive applications, certain values of TCP/IP which improves the latency and bandwidth are suggested. Ping is used to check a one way communication. The TCP/IP window size, a parameter, is the amount of buffering allowed by the receiver before an acknowledgement is required. Data is sent by TCP in segments. In order to improve throughput the sender must transmit multiple segments. The formula that governs the optimum window size is

TCP Window Size >= Bandwidth (in Bytes) * Latency (RTT)

The time of sending and receiving the message should be analysed based on TCP/IP parameters defined above. In 2nd test the size of message kept constant 8kb for proposed design then we evaluate network performance with different TCP window Size.

III. METHODOLOGY

The execution time of a parallel algorithm depends on input size, number of processing elements and their computation speeds. The number of speed up laws have been proposed such as Amdahl’s law (1969) based on a fixed workload or problem size, Gustafson’s law (1985) based on skilled problem size where problem size increases with the increase in machine size and Sun and Ni(1993) law for scaled problems bounded by memory availability. In comparison to serial algorithm, parallel algorithm spends more time in inter processor interaction and idling due to load imbalance.

In network evaluation, latency and asymptotic bandwidth are the two parameters, which affect the performance of both data intensive and compute intensive applications. In the research, certain parametric values of TCP/IP are suggested which improves the latency and bandwidth. "Ping-Pong" is a standard tool to check one way communication and two-way communication in a network of PCs. A ping utility sends specifically marked packets from the local computer to a remote computer/device. Besides determining whether the remote computer is currently 'alive', ping also provides indicators of the general speed or reliability of the network connection. Ping utility will be used to see the network traffic, no. of hops, and zero-TTL conditions.

Effective bandwidth always varies and can be affected by high latency. Too much latency in too short a period of time can create a bottleneck and prevents data from "filling the pipe", thus decreasing effective bandwidth. In parallel system based on the networked PCs, the time of sending and receiving the message is analysed based on the TCP/IP parameters defined above and suitability of Ethernet/Fast Ethernet for coarse grain application is tested.

IV. RESULTS AND ANALYSIS

The parameters of TCP/IP affecting the throttling of the bandwidth were studied and design of the network cluster was tested and finalized. Then experiment was conducted with macro-benchmark i.e. Parallel Matrix Multiplication over the Cluster. Timing measurements were averaged for all the processors used in the cluster.

Testing Network Design

The network design is tested with Ping-Pong test. In the test, send and receive time is recorded for various message sizes ranging from 1024 to 8192 bytes. This test is conducted to see the variation between the send and receive time of message communication. As at the receiver end, the NIC issues a DMA operation to move the message data to temporary buffer memory. This test is crucial as most of the communication libraries support pack and unpack functions for sending non-contiguous data so that network design will be made generalized for all communication libraries. In pack-unpack, the application explicitly packs data into a contiguous buffer before sending it, and unpacks it from a contiguous buffer after receiving it. Results obtained using default TCP window size of 16kB, are summarized in the Tables 1 and 2

Table 1 represent the send time and receive time between Server and Client node through a switch.

<table>
<thead>
<tr>
<th>Message Size (B)</th>
<th>Send Time (ms)</th>
<th>Recv Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>0.862</td>
<td>0.956</td>
</tr>
<tr>
<td>2048</td>
<td>1.003</td>
<td>0.988</td>
</tr>
<tr>
<td>4096</td>
<td>1.201</td>
<td>1.354</td>
</tr>
<tr>
<td>8192</td>
<td>1.298</td>
<td>1.275</td>
</tr>
</tbody>
</table>

Table 2 Server-Master RTT

<table>
<thead>
<tr>
<th>Message Size (B)</th>
<th>Send Time (ms)</th>
<th>Recv Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>0.591</td>
<td>0.524</td>
</tr>
<tr>
<td>2048</td>
<td>0.681</td>
<td>0.532</td>
</tr>
<tr>
<td>4096</td>
<td>0.732</td>
<td>0.698</td>
</tr>
<tr>
<td>8192</td>
<td>0.745</td>
<td>0.759</td>
</tr>
</tbody>
</table>
Table 2 contains the timing between Server and Master PC, where Master PC is directly connected to Server through the multi-port Ethernet card installed in the Server.

Analysis of Test I.

Sending a message consists of two parts:
1. Overhead portion O(1) – Start up process time, Header transmits time over the network (Constant Factor).
2. Data portion, O(n) – copying and transmitting data bytes between buffers (depends on the size of message i.e. n).

Therefore, the time to send a message follows a linear regression model and is given by

\[ t = a + b * n \]

This implies that parallel program design should have

- Much more computation than communication i.e. O(n^3) computation vs. O(n^2) communication such as in case of Matrix Multiplication over network of PCs.
- There should be big messages to overcome the delay.

On analysing Table 1, receive time is on higher side as compared to Table 2. This is due to the slower CPU used as client node. The lower send time in Table 2 as compared to Table 1 is attributed due to the direct connection of the Master PC in the multi-port Ethernet card of Server.

Analysis of Test II.

In the 2nd test, the size of the message is kept constant i.e. 8kB for proposed design. Network performance is evaluated with different TCP window sizes. In the this test, the effect of TCP Window buffer size is studied so that losses due to synchronization overhead can be reduced for the local nodes (there is single hop between server and clients) and non-local nodes (more than single hop between server and clients).

This parameter is going to help in the cycle of shrinking and slowing expanding windows size and subsequently in the data intensive and compute intensive parallel applications. The amount of data that can be in transfer in the network, termed as “Bandwidth-Delay-Product” or BDP for short, is simply the product of the bottleneck link bandwidth and round trip time (RTT).

Table 3 summarizes the BDP of different TCP window buffer size for local and non-local level clusters of nodes. In the table, hop “0” refers to direct connection of server with clients through multi-port Ethernet card i.e. there is no fast switch in between. Congestion is determined after comparing the BDP of the network channel and actual TCP window size.

<table>
<thead>
<tr>
<th>No. of Hops</th>
<th>TCP Window Size</th>
<th>Congestion if BDP &gt;=TCPWinSize</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>32kB</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td>64kB</td>
<td>16.9</td>
</tr>
<tr>
<td></td>
<td>BDP (kB)</td>
<td>BDP (kB)</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>16.2</td>
</tr>
<tr>
<td>1</td>
<td>17.6</td>
<td>17.8</td>
</tr>
<tr>
<td>1</td>
<td>26.0*</td>
<td>28.4</td>
</tr>
</tbody>
</table>

From the Table 3, it is inferred that cluster can be built at the local level with default value of TCP windows size i.e. 16kB. TCP window size should be 32kB or 64kB for good throughput in the network of level of one or two. 4kB size is ok for level 1 and size 8kB can be used for level 0.

V. CONCLUSION

Some performance measures are evaluated so that local area network can be used for cluster computing. For this behaviour of TCP/ IP protocol is studied. It is concluded from the graphical presentation that parallel computing in a cluster of PC cluster is greatly influenced by the communication parameters of TCP/IP. To achieve the performance of non-local nodes as that of local nodes in a cluster, it is necessary to keep the socket buffer size under BDP measurement so that network congestion has no impact on the transfer of data. From the analysis, we can conclude that the performances of the two clusters are almost identical with the present setup of network design. From the analysis, it also reveals that latency varies a lot when message size is less. It becomes stable for large value of message size. It is very much clear that by increasing the buffer size for the non-local nodes in the cluster, there is much more percentage change in time as compared to the nodes of local cluster. It also concluded that, at higher buffer size, the maximum bandwidth in
both local and non-local nodes can be got. The performance of non-local node is upgraded because of the higher buffer size and contributes towards parallel computing in the cluster environment without degraded effect. It is also concluded for graphical representation that latency time greatly affects the performance but as the buffer size is increased, the overall time overtakes the latency time. There is decrease of latency in the network, if the buffer size is increased.

REFERENCES
[5] TCP/IP test: Ping - pong test. This is the classic latency and bandwidth testing URL www.jncasr.ac.in
[20] Impact of Two Levels Switching in a LAN for Cluster Computing -Mamta , Charanjit Singh IJETT volume 5 number 3- Nov 2013