Bandwidth Management For Distributed Systems

Andrzej Tucholski, Agnieszka Francesson, Arkadiusz Majka, Joanna Sasin
Research and Development Department of Polish Telecom,
Obrzezna 7 02-691 Warsaw, Poland,
telephone: (+48 22) 699 51 99
e-mail:andrzej.tucholski@telekomunikacja.pl

Marcin Pilarski
Faculty of Mathematics and
Information Science,
Warsaw University of Technology,
pl.Politechniki 1,00-669 Warsaw, Poland

***Abstract*—The method of finding point of bending on efficiency curve in a distributed systems has been presented. The performance of distributed environment was studied as a cost function of mean network throughput and memory buffers availability. The method can be used then to build up an autonomic bandwidth management. The bounding point on efficiency curve was found as well analytically as for testing facility of a distributed laboratory environment. The influence of data replication and scheduling strategy on final results has been included. The computer simulation has been supported by analytical calculation. Two regions of different behavior on universal hyperbolic like dependence of the performance versus mean network throughput was shown up. A method of finding bounding point has been discuss in order to optimize efficiency of the considered system by such parameters like memory buffers and network bandwidth.**

I. Introduction

Recently several studies have been made in order to optimize distributed systems and reduce network cost as well as increase data availability[4], [5], [6]. Data replication[1], [2], [3] was constituted one of the methods of increasing efficiency of grid environment. In our paper we established dependence between performance of grid versus network bandwidth for several values of data replication parameter (the number of replicas of a certain data set in a system). In particular we have shown a hyperbolic like dependence of that function. We have shown point of bending where farther increase of bandwidth is inefficient. Heterogeneous environment and a flat topology have been used to make our studies more universal. The aim of the studies was to get a cost function of the performance of grid in relation to memory buffers and network bandwidth. Once the above is achieved, the optimal algorithm can be realized and global optimization of a grid efficiency becomes possible. What is more, network parameter can also be optimized and an autonomic bandwidth allocation can be constructed. This will be discussed latter on. In II there is a model description and an analytical approach presentation. The computer simulation of the model is shown in III. The results and discussion in context of analytical approach and other parameters are set forth in IV. In particular a data replication parameter dependence will be analyzed.

II. Model description and analytical approach

We model Grid System as a network of computing elements and storage elements (nodes) and links (edges). Computing elements comprise a finite number of processors, while storage elements may contain unlimited amount of data. Links are set to have certain, constant bandwidth. A system constructed this way can be represented mathematically as a direct, complete graph. To obtain heterogeneous character of the system computing elements have different computing power taken as real number from the range between (1,3).

We considered two different scheduling algorithms named: data-to-job and job-to-data. In the first one, a job was submitted to the most powerful node e.g. with maximum available power. This is, of course, a dynamic variable which changes in time. Jobs were characterized by two parameters: computing power and execution time. When job is entering the queue it cannot be overreached by any of the following ones unless they do not interfere in its expected execution time (modified FIFO algorithm). If required data is present on the node, there is no data transfer, if not, it is transferred.

In order to be more mathematically precise, let us define m_n as the data set requested by job n, $m_n = 1, 2, , M$. The job enters the local queue on the node where requested data can either be already present or not. In the first case no transfer occurs. It happens with probability equal to $P(t^{tr}_{m_n} = 0) = \frac{1}{L}$, where L denotes the number of nodes. In the second case, when a job and requested data m_n are not located in the same node, a transfer takes place. It happens with probability equal to $P(t^{tr}_{m_n} = \frac{S_{m_n}}{B}) = \frac{1}{M-1}(1 - \frac{1}{L})$, where S_{m_n} is the size of data set m_n and B is bandwidth.

As a performance measure we took the total mean time spent by jobs in a system called often in the literature *mean response time*. This time includes execution time, data transfer and time spent in a queue. Since the last two times usually overlap the n-thy job's response time reads

$$T_{n,m_n} = \max(t^n_q, t^{m_n}_{tr}) + t^n_{ex} \tag{1}$$

where t^n_q is queue waiting time of job n, $t^{m_n}_{tr}$ is transfer time of data set m_n equal to 0 (no transfer) and $\frac{S_{m_n}}{B}$- if data set must be transferred, t^n_{ex} is execution time of job n.

The mean over all jobs is the following:

978-1-4244-2390-3/08/$25.00 ©2008 IEEE

$$\langle T_{m_1,m_2,,m_N}\rangle = \frac{1}{N}\sum_{n=1}^{N}\max(t_q^n, t_{tr}^{m_n}) + \langle t_{ex}^n\rangle \qquad (2)$$

where the last term denotes mean execution time.

The above formula was used in a computer simulation. We had 16 nodes and 6 data sets with sizes chosen randomly from the range [125MB, 1,25GB]. Average over jobs reflects average over time as we assume constant streaming of new jobs. $\langle T_{mean}\rangle$ is a good measure of a total grid performance (the smaller the better). One can easily see hyperbolic like dependence to mean bandwidth of a network, recalling that $t_{tr}^{m_n} = \frac{S_{m_n}}{B}$.

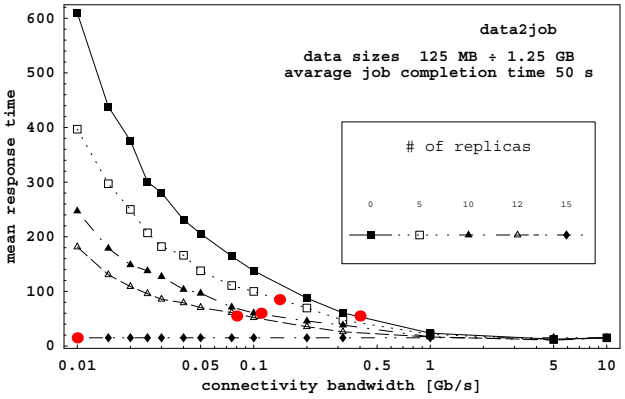

Fig. 1.

III. Numerical simulation

For simplicity we assumed full mesh topology. For real networks it can be taken exact. The bandwidths between nodes were constant for each cycle of simulation and ranged from 0 to 10Gb/s for all simulations.
On fig1 the cost function of grid performance versus network bandwidth has been shown. Different curves relate to different parameters of data replication which differ from 0 (no replication) to 15. To represent performance we have used the time a job is spending in a grid environment, which was extensively discussed in chapter II One can easily notice hyperbolic like dependence of grid performance versus network bandwidth.

IV. Results

To discuss this hyperbolic dependence one can use ϵ defined as:

$$\epsilon = \frac{\partial \ln T(B)}{\partial lnB} \qquad (3)$$

where ϵ ranges from 0 to 1.
We can notice two specific regions of B. One of the regions describes the situation where small changes of B lead to big changes of network bandwidth. The other region describes the situation where big changes of B leads to rather small changes in network bandwidth. If we assign B^* a boundary point between these two regions, one can draw up the dependence on replication number.

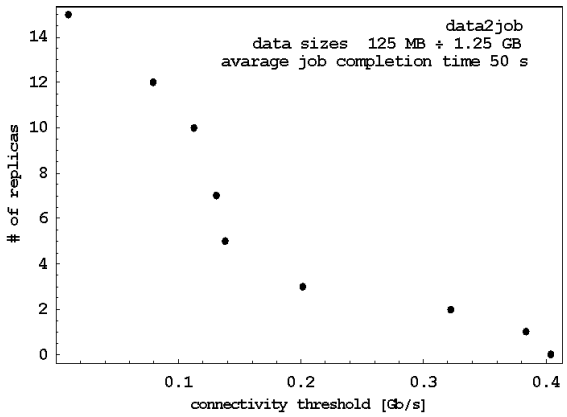

Fig. 2.

On the fig2 such a dependence is shown and exactly the number of replica versus previously defined connectivity threshold B^*. As one can see replication of data in our system is important below 150Mb/s of a the network throughput. The result was obtained for the range of data sizes between 125MB and 1.25GB. We have checked this for other data sizes, too, and there have been no qualitative changes in a result. There have been only obvious quantitative changes which are not important for our discussion.
If we are above 150Mb/s of the network throughput the replication of data is not effective. There are only small changes in performance even for big changes of B.

V. Conclusions

We have shown the hyperbolic dependence of the cost function of grid performance versus mean network bandwidth. Mathematical properties of the hyperbole lead to a separation of two regions: one with the strong dependence on the network bandwidth and the other with the weak dependence. Explicitly we have shown a point of breaking on efficiency curve when farther increase of a network bandwidth becomes inefficient. The result can be easily applied into to the convergence of grid and IP network to construct an autonomic bandwidth allocation.

Acknowledgment

The work presented in this paper was supported by France Telecom grant number 023/2006.

References

[1] W. B. David. *Evaluation of an economy-based file replication strategy for a data grid.* In International Workshop on Agent based Cluster and Grid Computing, pages 120-126, 2003.
[2] M.M. Deris, Abawajy J.H., and H.M. Suzuri. *An efficient replicated data access approach for largescale distributed systems.* In IEEE International Symposium on Cluster Computing and the Grid, April 2004.

[3] K. Ranganathan and I. Foster, *Decoupling Computation and Data Scheduling in Distributed Data-Intensive Applications, Proceedings of 11th IEEE International Symposium on High Performance Distributed Computing (HPDC-11)*, Edinburgh, Scotland, July 2002, IEEE CS Press, USA.

[4] W. Hoschek, F. J. Janez, A. Samar, H. Stockinger, and K. Stockinger. *Data management in an international data grid project.* In In Proceedings of GRID Workshop, pages 77-90, 2000.

[5] K. Ranganathan, A. Iamnitchi, and I.T. Foster. *Improving data availability through dynamic modeldriven replication in large peer-to-peer communities.* In In 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, pages 376-381, 2002.

[6] H. Lamehamedi, B. Szymanski, Z. Shentu, and E. Deelman. *Data replication strategies in grid environments.* In In Proceedings of 5th International Conference on Algorithms and Architecture for Parallel Processing, pages 378-383, 2002.