Model-Based Optimization of MPI Collective Operations for Computational Clusters

Alexey Lastovetsky

School of Computer Science and Informatics
University College Dublin, Ireland

Computational clusters based on switched networks are widely used by the academic research community for high performance computing, with MPI being the primary programming tool for development of parallel applications. Contribution of communication operations into the execution time can be quite significant for many parallel applications on this platform. Therefore, minimization of the execution time of communication operations is an important optimization technique for high-performance computing on computational clusters.

In theory, analytical communication performance models can play an important role in optimization of communication operations. Their predictions could be used at runtime for selection of the optimal algorithm of one or the other collective operation depending on the message size and the computing nodes involved in the operation. In practice, this approach does not work well when traditional models are employed. The analytical predictions given by these models appear rather inaccurate, often leading to wrong optimization decisions.

In this talk, we analyze the restrictions of the traditional models affecting the accuracy of analytical prediction of the execution time of different algorithms of collective communication operations. The most important restriction is that the constant and variable contributions of processors and network are not fully separated in these models. We show that the full separation of the contributions that have different nature and arise from different sources would lead to more intuitive and accurate models of switched computational clusters. At the same time, we demonstrate that the traditional estimation methods based on point-to-point communication experiments cannot estimate parameters of such models because the total number of independent point-to-point experiments will always be less than the number of parameters. Thus, this is the traditional estimation methods that limit the design of the traditional communication models and make them less intuitive and less accurate than they might be.

We describe a recent solution to this problem, which is a new estimation method based on a set of independent communication experiments including not only point-to-point but also point-to-two operations. We also outline one non-traditional intuitive communication model, the LMO model, detail communication experiments for estimation of its parameters and demonstrate how this model can be used for accurate prediction of the execution time of collective algorithms.

In the second part of the talk, we extend our analysis to computational clusters consisting of heterogeneous processors, probably the most common parallel platform available to the academic research community. We consider and compare three different types of communication performance models that can be used for the heterogeneous

M. Ropo et al. (Eds.): EuroPVM/MPI 2009, LNCS 5759, pp. 4–5, 2009.

clusters: the traditional (homogeneous) models, heterogeneous extensions of the traditional models, and a heterogeneous extension of the non-traditional LMO model. We demonstrate that while the heterogeneous extensions of the traditional models are more accurate than their original homogeneous versions, they are not accurate enough to compete with the heterogeneous LMO model, which much more accurately predicts the execution time of collective algorithms. We also show that the optimised collective operations based on the heterogeneous LMO model outperform those based on the traditional models, whether homogeneous or heterogeneous.