Topology-aware placement of MPI processes for clusters with a communication network Angara

Mikhail Khalilov, Alexei Timofeev

HSE

During the seminar, one of the methods for optimizing the launch of MPI parallel programs on computational clusters using the Angarsk interconnection will be considered, and the results of using the proposed optimization will be shown.

First, in order to optimize the mapping, an analysis of the communication pattern of the message exchange between the processes of this MPI program is performed. The communication pattern can be represented in the form of a weighted graph, where the vertices correspond to MPI processes, the weights on the edges represent the intensity of the exchanges between the processes. The main idea of the method for optimizing the process mapping is to split the communication pattern graph into disjoint subsets of the intensively exchanging processes and to bind these subsets to nodes/processor cores connected by the fastest communication channels. The partition is performed to minimize the sum of edges joining different subsets of the partition. The split is recursively performed first for the inter-node level describing the exchanges between the compute nodes, and then for the inter-node level describing the exchanges within each of the nodes in the event that several processors are installed inside the node. The goal of such a partition is to minimize the execution time of the program.

A scheme for implementing the procedure for clusters using the Angara communication network will be presented. The decrease in the execution time of MPI programs is shown when using the optimal display algorithm. The results of the analysis of the dependence of the execution time of MPI programs on cluster parameters are considered.