Two algorithms of irregular scatter/gather operations for heterogeneous platforms

Kiril Dichev Vladimir Rychkov Alexey Lastovetsky
Kiril.Dichev@ucdconnect.ie, Vladimir.Rychkov@ucd.ie, Alexey.Lastovetsky@ucd.ie

Heterogeneous Computing Laboratory
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland
http://hcl.ucd.ie

September 10, 2010

Model-based optimization of collectives

Motivation

- A communication model can capture the network heterogeneity (latency, bandwidth)

Model-based optimization of collectives

Motivation

- A communication model can capture the network heterogeneity (latency, bandwidth)
- A collective operation can be dynamically constructed to take this heterogeneity into account

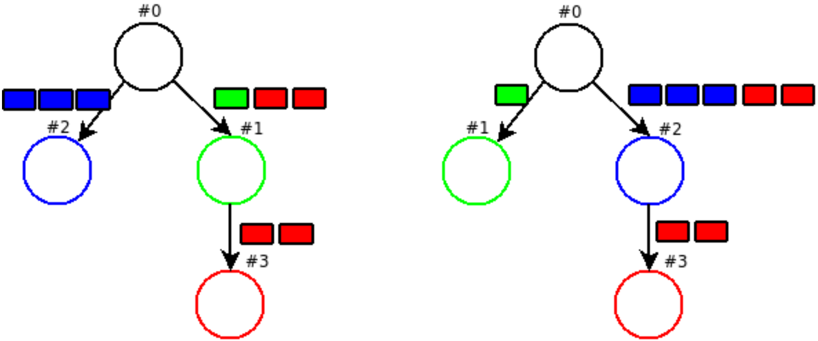
Model-based optimization of collectives

Motivation

- A communication model can capture the network heterogeneity (latency, bandwidth)
- A collective operation can be dynamically constructed to take this heterogeneity into account
- We use communication models to build faster communication trees on heterogeneous platforms

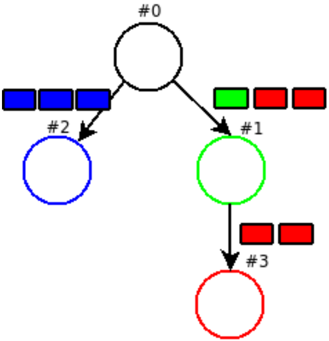
Example

- Same scatter operation, different communication trees
- Which tree would perform better ?
- Answer is not trivial



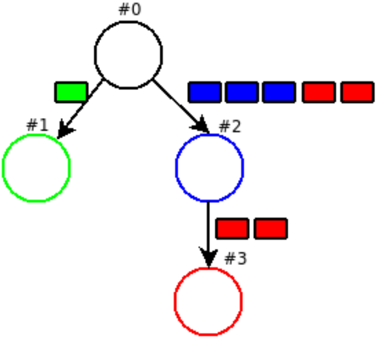
Example

- Same scatter operation, different communication trees
- Which tree would perform better ?
- Answer is not trivial
- This tree may be faster if 0-2 link is slower than 0-1 (less messages propagated through 0-2 link



Example

- Same scatter operation, different communication trees
- Which tree would perform better ?
- Answer is not trivial
- ... but this tree may be faster if 0-2 link is faster than 0-1 (more messages propagated through 0-2 link



Model-based irregular Scatter/Gather

We introduce communication models to two algorithms for the operations MPI_Scatterv and MPI_Gatherv

- The first algorithm is a basic binomial tree algorithm
- The second algorithm is a sophisticated scatterv/gatherv algorithm
- In both cases, our heuristics check communication properties between processes
    - provided by a prediction function

Model-based binomial MPI_Scatterv/MPI_Gatherv

- We don't change the structure of a binomial tree
- We choose a process mapping with heuristics based on communication properties

A more sophisticated model-based MPI_Scatterv/MPI_Gatherv algorithm

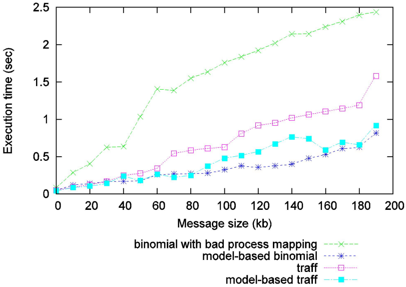
An algorithm was developed particularly for irregular scatter/gather operations:

- During the tree construction, we partition a set of nodes similarly to Träff[1]
- Our heuristics consider both *message size* and *network properties*

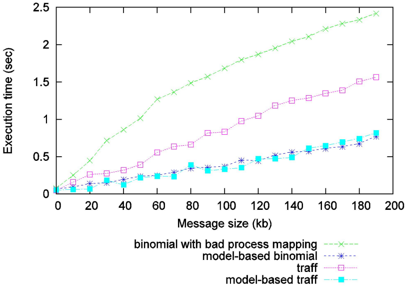
[1] Träff, J.L.: Hierarchical Gather/Scatter Algorithms with Graceful Degradation. In: IPDPS04, vol. 1, pp. 80–89. IEEE (2004)

Experimental results

We performed tests on the cluster Grid5000 with heterogeneous network



Scatterv



Gatherv

Thank you!