Accurate Communication Performance Models of Heterogeneous Clusters: Estimation and Use

Vladimir Rychkov   Kiril Dichev

Heterogeneous Computing Laboratory
School of Computer Science and Informatics
University College Dublin

May 11, 2012

Introduction

- MPI-based applications require optimisation for heterogeneous platforms
    - Minimization of communication cost

Introduction

- MPI-based applications require optimisation for heterogeneous platforms
  - Minimization of communication cost
- Analytical predictive communication performance models
  - Traditionally designed for homogeneous platforms
  - Prediction $T_{coll}(M, n)$ = combination of point-to-point parameters, message size, M, and number of processors, n
- Model-based optimisation of collective communication operations
  - Switch between different algorithms implementing the operation
  - Construction of optimal communication tree for the operation

Introduction

- MPI-based applications require optimisation for heterogeneous platforms
  - Minimization of communication cost
- Analytical predictive communication performance models
  - Traditionally designed for homogeneous platforms
  - Prediction $T_{coll}(M, n) =$ combination of point-to-point parameters, message size, M, and number of processors, n
- Model-based optimisation of collective communication operations
  - Switch between different algorithms implementing the operation
  - Construction of optimal communication tree for the operation
- **Heterogeneous cluster: how to design, estimate and use communication performance model?**

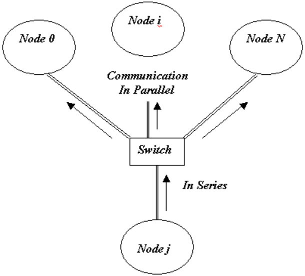
Traditional Communication Performance Models

- Point-to-point parameters, the same values for all links
- Point-to-point communication experiments to estimate parameters

Traditional Communication Performance Models

- Point-to-point parameters, the same values for all links
- Point-to-point communication experiments to estimate parameters
- **Unable to capture constant and variable contributions of processors and network**
- **Inaccurate to predict communication execution time**

Single-switch cluster

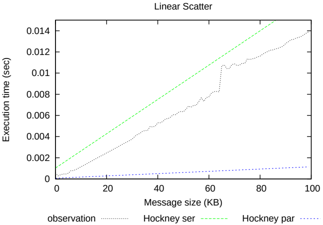


Hockney prediction for linear scatter

Serial: $T(M, n) = (n-1)(\alpha + \beta M)$

Parallel: $T(M, n) = \alpha + \beta M$

M - a message sent to each processor



Communication Models for Heterogeneous Clusters

Homogeneous models

parameters are found by averaging values for all pairs of processors

- Small number of parameters, compact formulas for collectives
- $O(n^2)$ p2p communication experiments to estimate params
- Significant heterogeneity = inaccurate prediction

Communication Models for Heterogeneous Clusters

Homogeneous models

parameters are found by averaging values for all pairs of processors

- Small number of parameters, compact formulas for collectives
- $O(n^2)$ p2p communication experiments to estimate params
- Significant heterogeneity $=$ inaccurate prediction

Heterogeneous models

different link- (and processor-) specific parameters

- $O(n^2)$ parameters, flexible formulas for collectives
- $\geq O(n^2)$ communication experiments to estimate params
- More natural expression of collectives $=$ more accurate prediction

Communication Models for Heterogeneous Clusters

Homogeneous models

parameters are found by averaging values for all pairs of processors

- Small number of parameters, compact formulas for collectives
- $O(n^2)$ p2p communication experiments to estimate params
- Significant heterogeneity = inaccurate prediction

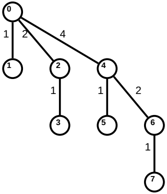
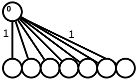
Heterogeneous models

different link- (and processor-) specific parameters

- $O(n^2)$ parameters, flexible formulas for collectives
- $\geq O(n^2)$ communication experiments to estimate params
- More natural expression of collectives = more accurate prediction

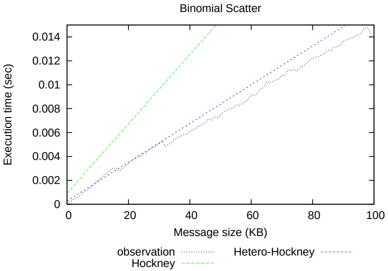
- **Straightforward heterogeneous extension of traditional models**
- **Design of new elaborated heterogeneous models**

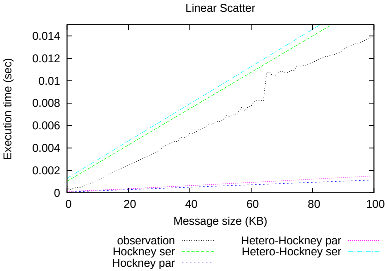
Heterogeneous Extension of Traditional Models

Hockney	**Binomial scatter/gather** fine-grained parallelism	**Linear scatter/gather** coarse-grained parallelism
Homogeneous	$(\log_2 n)\alpha + (n-1)\beta M$ - parallel/serial	$(n-1)(\alpha + \beta M)$ - serial $\alpha + \beta M$ - parallel
Heterogeneous	$T(k) = \alpha_{rs} + \beta_{rs} 2^{k-1} M + \max_{c \in C_{k-1}} T_c(k-1)$	$\sum_{i=0, i \neq r}^{n-1} (\alpha_{ri} + \beta_{ri} M)$ - serial $\max_{i=0, i \neq r}^{n-1} (\alpha_{ri} + \beta_{ri} M)$ - parallel
		

Implemented heterogeneous extensions: Hockney, LogGP, PLogP

Heterogeneous Extension of Traditional Models





LMO Heterogeneous Communication Model

- **Target platform:** heterogeneous cluster with a single switch

$i \xrightarrow{M} j$:	$(C_i, t_i) \xrightarrow{(L_{ij}, \beta_{ij})} (C_j, t_j)$
point-to-point execution time:	$C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_j)$
$2n$ processor parameters:	fixed (C_i, C_j) and variable (t_i, t_j) delays
$2C_n^2$ link parameters:	latency (L_{ij}) and transmission rate (β_{ij})
	we suppose $L_{ij} = L_{ji}$ and $\beta_{ij} = \beta_{ji}$

LMO Heterogeneous Communication Model

- **Target platform:** heterogeneous cluster with a single switch

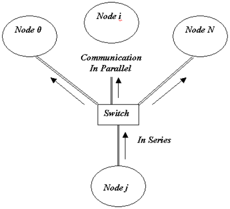
$i \xrightarrow{M} j$: $(C_i, t_i) \xrightarrow{(L_{ij},\beta_{ij})} (C_j, t_j)$

point-to-point execution time: $C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_j)$

$2n$ processor parameters: fixed (C_i, C_j) and variable (t_i, t_j) delays

$2C_n^2$ link parameters: latency (L_{ij}) and transmission rate (β_{ij})

we suppose $L_{ij} = L_{ji}$ and $\beta_{ij} = \beta_{ji}$



More intuitive and accurate predictive formulas:

$$T_{scatter} = (n-1)(C_r + Mt_r) + \max_{i=0,i\neq r}^{n-1}(L_{ri} + \frac{M}{\beta_{ri}} + C_i + Mt_i)$$

How to estimate these parameters?

Point-to-point experiments are not enough

Estimation of Parameters

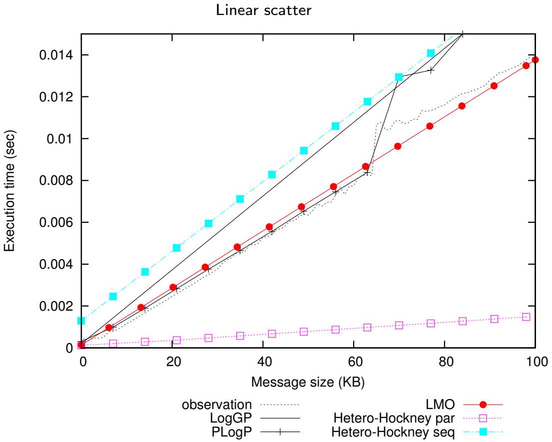
- Select the communication experiments and express their execution time via the point-to-point parameters
- Measure the execution time of these communications
- Build and solve the system of equations, using the times as a right-hand side values

Estimation of Parameters

- Select the communication experiments and express their execution time via the point-to-point parameters
- Measure the execution time of these communications
- Build and solve the system of equations, using the times as a right-hand side values

- In a triplet of processors $(i < j < k)$: 12 unknowns
- Point-to-point communications, roundtrips: 6 independent equations
  $i \underset{M}{\overset{M}{\longleftrightarrow}} j \quad T_{ij}(M) = 2(C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_j)) \quad M := 0, M$
- Linear scatter + linear gather: 6 independent equations
  $i \underset{0}{\overset{M}{\longleftrightarrow}} jk = i \overset{M}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$
  $T_{ijk}(M) = 2(2C_i + Mt_i) + \max\limits_{x=j,k}(2(L_{ix} + C_x) + M(\frac{1}{\beta_{ix}} + t_x))$
  $M := 0, M$

Model Prediction

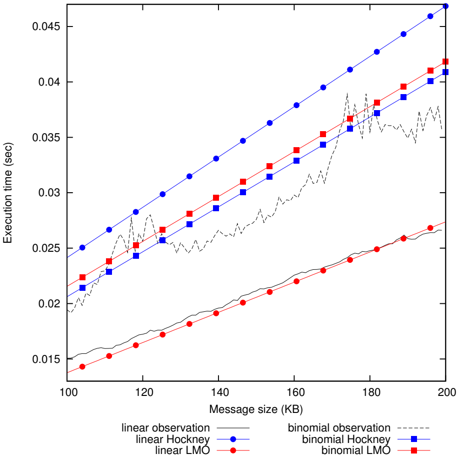


Model-Based Switch between Algorithms

- Which scatter algorithm is faster for a given message size on a heterogeneous cluster?

Model-Based Switch between Algorithms

- Which scatter algorithm is faster for a given message size on a heterogeneous cluster?
- Hockney: switch to binomial
- LMO: switch to linear



Model-Based Construction of Communication Trees

Main approaches

- Mapping of nodes to a tree of a given structure
- Constructing a tree of some structure

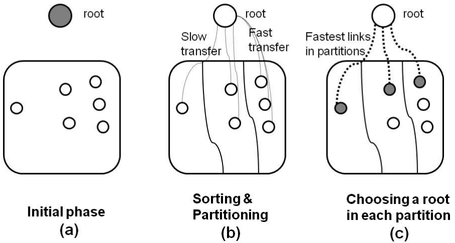
Model-Based Construction of Communication Trees

Main approaches

- Mapping of nodes to a tree of a given structure
- Constructing a tree of some structure

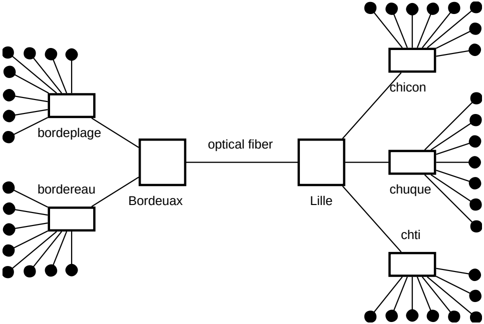
Example: MPI_Scatterv/MPI_Gatherv

- Binomial algorithm - message-unaware, binomial tree
  fastest-first mapping of nodes: depth-first traverse starting with the lowest-order subtrees
- Traff algorithm - irregular tree based on message sizes
  sorting and choosing the roots based on model prediction



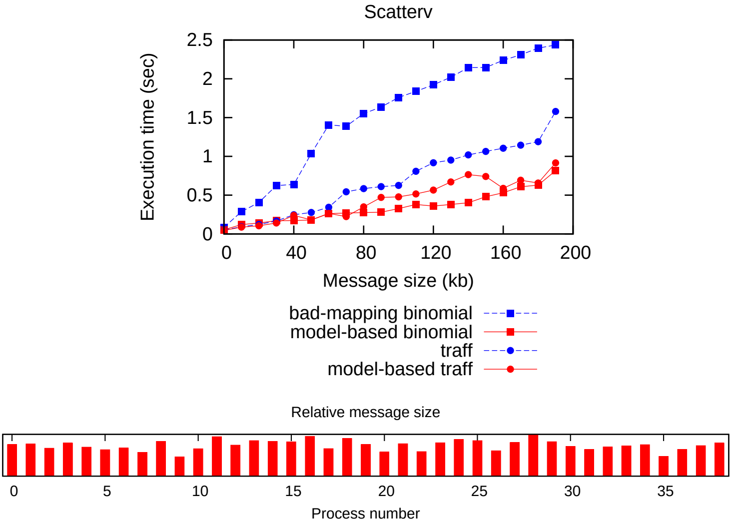
Experimental Platform: Grid'5000

2 sites, 5 clusters, 40 nodes



MPICH2 over TCP/IP

Experimental Results: Heterogeneous Hockney



Ongoing and Future Study

- Communication performance models of hierarchical heterogeneous platforms: interconnected heterogeneous clusters, multicore and multi-GPU clusters
- Optimisation of MPI communication operations on hierarchical heterogeneous platforms

- Application of the proposed approaches to the IBM Exascale platform
- Integration of the proposed approaches into development tools for parallel programming: HeteroMPI, Open MPI

Publications

2011 - 1 conference paper

- Dichev, K., Lastovetsky, A., Rychkov, V. "Improvement of the Bandwidth of Cross-Site MPI Communication Using Optical Fiber", EuroMPI 2011, vol. 6960, Santorini, Greece, Springer, September 18-21, 2011.

2010 - 1 journal article, 1 conference paper

- Lastovetsky, A., Rychkov, V., O'Flynn, M. "Accurate Heterogeneous Communication Models and a Software Tool for their Efficient Estimation", International Journal of High Performance Computing Applications, vol. 24, issue 1, pp. 34-48, 2010.
- Dichev, K., Rychkov, V., Lastovetsky, A. "Two Algorithms of Irregular Scatter/Gather Operations for Heterogeneous Platforms", EuroMPI 2010, vol. 6305, Stuttgart, Germany, pp. 289-293, Sep 12-15, 2010.

Output

Project web page: `http://hcl.ucd.ie/project/cpm`

Software

- 2 packages developed at HCL: MPIBlib, CPM
- Based on system and mathematical software: C/C++, MPI, Autotools, GNU Scientific Library, Boost C++ libraries

Applications

- Hyperspectral Image Processing (University of Extremadura, Spain)

Team

- 2 postdoctoral researchers: Vladimir Rychkov, Jun Zhu
- 2 PhD students: Kiril Dichev, Khalid Hasanov

Collaboration

Hardware

- Myrinet cluster (Innovative Computing Laboratory, University of Tennessee, USA)
- Grid'5000 (INRIA, CNRS, RENATER, France)

Collaboration

- David Valencia (University of Extremadura, Spain)
- Shaukat Ali, Rolf Riesen (Exascale Systems, IBM, Ireland)

Financial Support

Science Foundation Ireland

UCD CSI

IRCSET/IBM Exascale Collaboratory