A Software Tool for Accurate Estimation of Parameters of Heterogeneous Communication Models

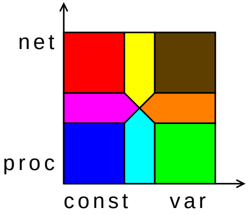
Alexey Lastovetsky Vladimir Rychkov Maureen O'Flynn
{Alexey.Lastovetsky, Vladimir.Rychkov, Maureen.OFlynn}@ucd.ie

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

The 15th European PVM/MPI Users Group conference
September 8, 2008, Dublin, Ireland

- MPI-based applications require optimization for heterogeneous platforms
- Model-based optimization of communication cost
  - Analytical predictive communication performance models
  - Heterogeneous switched clusters

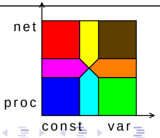
- MPI-based applications require optimization for heterogeneous platforms
- Model-based optimization of communication cost
  - Analytical predictive communication performance models
  - Heterogeneous switched clusters
- Ideal intuitive communication performance model
  - Point-to-point parameters:
    constant and **variable** (*message size*)
    contributions of **processors** and **network**
  - T_{coll} = combination of
    max (*parallel part*) and $\sum$ (*sequential part*)
    of the point-to-point parameters
  - There is a set of communication experiments
    that allow to estimate the parameters



Model	p2p	Experiments (series)
Hockney	$\alpha + \beta M$	$r \times (i \xleftrightarrow[0]{0} j + i \xleftrightarrow[M]{M} j)$ or $\sum_{k=0}^{r} i \xleftrightarrow[M_k]{M_k} j$
LogP	$L + 2o$ $(L + 2o + gM)$	$r \times (i \xleftrightarrow[M]{M} j + i \xleftrightarrow[0]{M} j + i \xleftrightarrow[M]{0} j) + \sum_{x=0}^{s} i \xleftrightarrow[0]{\overbrace{M \ldots M}^{2^x}} j$
LogGP	$L + 2o +$ $G(M - 1)$	$r \times (i \xleftrightarrow[M]{M} j + i \xleftrightarrow[0]{M} j + i \xleftrightarrow[M]{0} j) + \sum_{x=0}^{s} i \xleftrightarrow[0]{\overbrace{\bar{M} \ldots \bar{M}}^{2^x}} j$
PLogP	$L + g(M)$ $[o_s(M), o_r(M)]$	$r \times i \xleftrightarrow[0]{0} j + \sum_{k=0}^{m} [r \times (i \xleftrightarrow[0]{M_k} j + i \xleftrightarrow[M_k]{0} j) + \sum_{x=0}^{s} i \xleftrightarrow[0]{\overbrace{M_k \ldots M_k}^{2^x}} j]$

Traditionally designed for homogeneous platforms

- the same values of parameters for each pair of processors
- the parameters are found *statistically* from the communication experiments *between any two processors*



Communication performance models of heterogeneous clusters

Homogeneous models

the parameters are found by averaging values for all pairs of processors

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

Communication performance models of heterogeneous clusters

Homogeneous models

the parameters are found by averaging values for all pairs of processors

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

Heterogeneous models

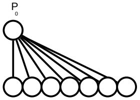
different link- (and processor-) specific parameters

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

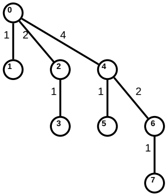
Hockney	Linear scatter/gather	Binomial scatter/gather
Homogeneous	$(n-1)(\alpha+\beta M)$ - sequential $\alpha+\beta M$ - parallel	$(log_2 n)\alpha+(n-1)\beta M$ - parallel/sequential
Heterogeneous	$\sum_{i=0,i\neq r}^{n-1}(\alpha_{ri}+\beta_{ri}M)$ - sequential $\max_{i=0,i\neq r}^{n-1}(\alpha_{ri}+\beta_{ri}M)$ - parallel	$\overbrace{\alpha_{ri}+\beta_{ri}2^{log_2(n-1)}M+\max(S(log_2(n-1)-1))}^{S(log_2(n-1))}$

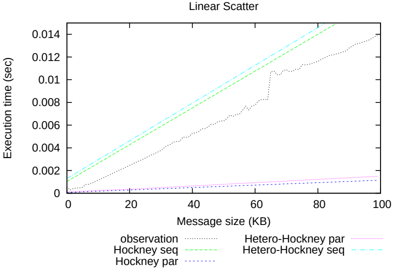
$$\alpha_{04}+4\beta_{04}M+max\begin{cases}\alpha_{02}+2\beta_{02}M+...\\ \alpha_{46}+2\beta_{46}M+...\end{cases}$$

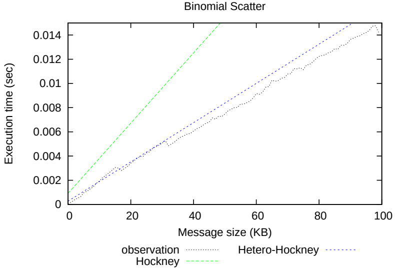
$$\begin{cases}...+max(\alpha_{01}+\beta_{01}M,\alpha_{23}+\beta_{23}M)\\ ...+max(\alpha_{45}+\beta_{45}M,\alpha_{67}+\beta_{67}M)\end{cases}$$

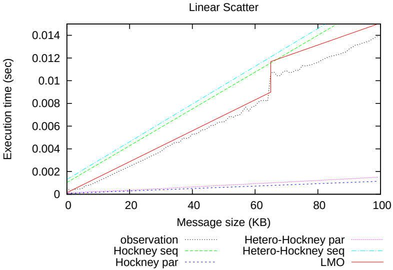


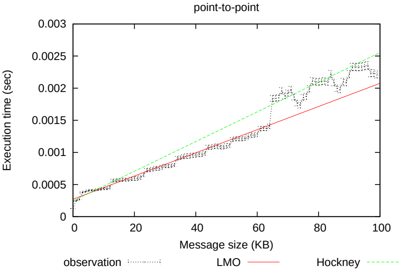
M - a recv/send buffer size




Linear Scatter
Execution time (sec)
Message size (KB)
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
20
40
60
80
100
observation
Hockney seq
Hockney par
Hetero-Hockney par
Hetero-Hockney seq

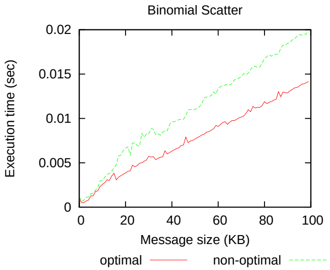

Binomial Scatter
Execution time (sec)
Message size (KB)
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
20
40
60
80
100
observation
Hockney
Hetero-Hockney

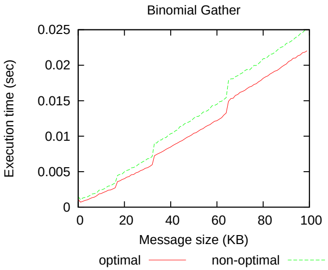

Linear Scatter
Execution time (sec)
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
20
40
60
80
100
Message size (KB)
observation
Hockney seq
Hockney par
Hetero-Hockney par
Hetero-Hockney seq
LMO


point-to-point
Execution time (sec)
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
0
20
40
60
80
100
Message size (KB)
observation
LMO
Hockney

Model-based optimization for binomial scatter/gather:

- r is a root, n is a number of processors
  $2^k M$ is the biggest message to send $(k = log_2(n-1))$
- send/recv to/from the processor i: $T(r \xrightarrow{2^k M} i) = min$
- repeat for the subtrees of order $k-1$, the root processors of which are already known: i (and r, if $n-1 = 2^k + 2^{k-1} + ...$)





How to achieve an accurate prediction?

- More link- and processor-specific parameters
- More natural expressions for the execution time

Problems

- How to estimate parameters of the heterogeneous models?
- Design of communication experiments?
- Efficiency?

Lastovetsky, A., Mkwawa, I., O'Flynn, M.: **An Accurate Communication Model of a Heterogeneous Cluster Based on a Switch-Enabled Ethernet Network**. *In: Proceedings of ICPADS 2006*, Minneapolis, MN, pp. 15-20 (2006)

Heterogeneous cluster with a single switch

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

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

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

link parameters: transmission rate (β_{ij})

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

$2n + C_n^2$ parameters

Lastovetsky, A., Mkwawa, I., O'Flynn, M.: **An Accurate Communication Model of a Heterogeneous Cluster Based on a Switch-Enabled Ethernet Network**. *In: Proceedings of ICPADS 2006*, Minneapolis, MN, pp. 15-20 (2006)

Heterogeneous cluster with a single switch

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

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

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

link parameters: transmission rate (β_{ij})

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

$2n + C_n^2$ parameters

- More than two linear parameters
- Not only link-specific, but also processor-specific parameters

Approach

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

Approach

- Select the communication experiments and express their execution time via the point-to-point parameters
- Measure the execution time and solve the system of equations, using the times as a right-hand side values
- **The number of linearly independent equations should be** $\geq 2n + C_n^2$
- **The execution time should be statistically reliable**

Approach

- Select the communication experiments and express their execution time via the point-to-point parameters
- Measure the execution time and solve the system of equations, using the times as a right-hand side values
- **The number of linearly independent equations should be** $\geq 2n + C_n^2$
- **The execution time should be statistically reliable**

Building

- A single roundtrip with empty message $i \overset{0}{\underset{0}{\longleftrightarrow}} j$: $T_{ij}(0) = 2(C_i + C_j)$

  How to find processor-specific parameters?

Approach

- Select the communication experiments and express their execution time via the point-to-point parameters
- Measure the execution time and solve the system of equations, using the times as a right-hand side values
- **The number of linearly independent equations should be $\geq 2n + C_n^2$**
- **The execution time should be statistically reliable**

Building

- A single roundtrip with empty message $i \underset{0}{\overset{0}{\longleftrightarrow}} j$: $T_{ij}(0) = 2(C_i + C_j)$

  How to find processor-specific parameters?
- Empty roundtrips in triplets $i < j < k$

  $$\begin{cases} T_{ij}(0) = 2(C_i + C_j) & i \underset{0}{\overset{0}{\longleftrightarrow}} j \\ T_{jk}(0) = 2(C_j + C_k) & j \underset{0}{\overset{0}{\longleftrightarrow}} k \\ T_{ik}(0) = 2(C_i + C_k) & i \underset{0}{\overset{0}{\longleftrightarrow}} k \end{cases}$$
- As C_i can be found from the equations for different triplets including i, find the average fixed processor delays: $\bar{C}_i = \frac{\sum\limits_{j,k \neq i} C_i}{C_{n-1}^2}$

- A single roundtrip with non-empty message $i \underset{0}{\overset{M}{\longleftrightarrow}} j$: $T_{ij}(M) = 2(C_i + C_j) + M(t_i + \beta_{ij} + t_j)$
  How to separate link- and processor-specific parameters - $n + C_n^2$ unknowns?
  Roundtrips in triplets are not enough: n independent equations

- A single roundtrip with non-empty message $i \overset{M}{\underset{0}{\longleftrightarrow}} j$: $T_{ij}(M) = 2(C_i + C_j) + M(t_i + \beta_{ij} + t_j)$
  How to separate link- and processor-specific parameters - $n + C_n^2$ unknowns?
  Roundtrips in triplets are not enough: n independent equations
- **Collective communications**

- A single roundtrip with non-empty message $i \xleftrightarrow[0]{M} j$: $T_{ij}(M) = 2(C_i + C_j) + M(t_i + \beta_{ij} + t_j)$
  How to separate link- and processor-specific parameters - $n + C_n^2$ unknowns?
  Roundtrips in triplets are not enough: n independent equations
- **Collective communications**
  - Consecuitive point-to-point communications in ring:
    $T_{ijk}(M) = 2(C_i + C_j + C_k) + M(2(t_i + t_j + t_k) + \beta_{ij} + \beta_{jk} + \beta_{ki})$
    fail: C_n^3 not linearly independent equations
    $T_{ijk}(M) = T_{ij}(M) + T_{jk}(M) + T_{ki}(M) - T_{ij}(0) - T_{jk}(0) - T_{ki}(0)$

- A single roundtrip with non-empty message $i \overset{M}{\underset{0}{\longleftrightarrow}} j$: $T_{ij}(M) = 2(C_i + C_j) + M(t_i + \beta_{ij} + t_j)$
  How to separate link- and processor-specific parameters - $n + C_n^2$ unknowns?
  Roundtrips in triplets are not enough: n independent equations
- **Collective communications**
  - Consecuitive point-to-point communications in ring:
    $T_{ijk}(M) = 2(C_i + C_j + C_k) + M(2(t_i + t_j + t_k) + \beta_{ij} + \beta_{jk} + \beta_{ki})$
    fail: C_n^3 not linearly independent equations
    $T_{ijk}(M) = T_{ij}(M) + T_{jk}(M) + T_{ki}(M) - T_{ij}(0) - T_{jk}(0) - T_{ki}(0)$
  - **Roundtrips and parallel one-to-two with not very large message**

$$
\begin{cases}
T_{ij}(M) = 2(C_i + C_j) + M(t_i + \beta_{ij} + t_j)) & i \overset{M}{\underset{0}{\longleftrightarrow}} j \\
T_{jk}(M) = 2(C_j + C_k) + M(t_j + \beta_{jk} + t_k)) & j \overset{M}{\underset{0}{\longleftrightarrow}} k \\
T_{ik}(M) = 2(C_i + C_k) + M(t_i + \beta_{ik} + t_k)) & i \overset{M}{\underset{0}{\longleftrightarrow}} k \\
T_{ijk}(M) = 2(2C_i + Mt_i) + \max_{x=j,k}(2C_x + M(\beta_{ix} + t_x)) & i \overset{M}{\underset{0}{\longleftrightarrow}} jk \\
T_{jik}(M) = 2(2C_j + Mt_j) + \max_{x=i,k}(2C_x + M(\beta_{jx} + t_x)) & j \overset{M}{\underset{0}{\longleftrightarrow}} ik \\
T_{kij}(M) = 2(2C_k + Mt_k) + \max_{x=i,j}(2C_x + M(\beta_{kx} + t_x)) & k \overset{M}{\underset{0}{\longleftrightarrow}} ij
\end{cases}
$$

- $T_{ijk}(M) = 2(2C_i + Mt_i) + \max_{x=j,k}(2C_x + M(\beta_{ix} + t_x)) = 2C_i + Mt_i + \max_{x=j,k} T_{ix}(M)$

$$\begin{cases} t_i = \frac{T_{ijk}(M) - \max_{x=j,k} T_{ix}(M) - 2C_i}{M} & \frac{1}{\beta_{ij}} = \frac{T_{ij}(M) - 2(C_i + C_j)}{M} - (t_i + t_j) \\ t_j = \frac{T_{jik}(M) - \max_{x=i,k} T_{jx}(M) - 2C_j}{M} & \frac{1}{\beta_{jk}} = \frac{T_{jk}(M) - 2(C_j + C_k)}{M} - (t_j + t_k) \\ t_j = \frac{T_{kij}(M) - \max_{x=i,j} T_{kx}(M) - 2C_k}{M} & \frac{1}{\beta_{ik}} = \frac{T_{ik}(M) - 2(C_i + C_k)}{M} - (t_i + t_k) \end{cases}$$

- $T_{ijk}(M) = 2(2C_i + Mt_i) + \max_{x=j,k}(2C_x + M(\beta_{ix} + t_x)) = 2C_i + Mt_i + \max_{x=j,k} T_{ix}(M)$

$$\begin{cases} t_i = \frac{T_{ijk}(M) - \max_{x=j,k} T_{ix}(M) - 2C_i}{M} & \frac{1}{\beta_{ij}} = \frac{T_{ij}(M) - 2(C_i + C_j)}{M} - (t_i + t_j) \\ t_j = \frac{T_{jik}(M) - \max_{x=i,k} T_{jx}(M) - 2C_j}{M} & \frac{1}{\beta_{jk}} = \frac{T_{jk}(M) - 2(C_j + C_k)}{M} - (t_j + t_k) \\ t_j = \frac{T_{kij}(M) - \max_{x=i,j} T_{kx}(M) - 2C_k}{M} & \frac{1}{\beta_{ik}} = \frac{T_{ik}(M) - 2(C_i + C_k)}{M} - (t_i + t_k) \end{cases}$$

- As t_i can be found from the equations for different triplets including i, find the average variable processor delays: $\bar{t}_i = \frac{\sum_{j,k \neq i} t_i}{C_{n-1}^2}$
- As β_{ij} can be found from the equations for different triplets including $i \leftrightarrow j$, find the average variable processor delays: $\bar{\beta}_{ij} = \frac{\sum_{k \neq i,j} \beta_{ij}}{n-2}$

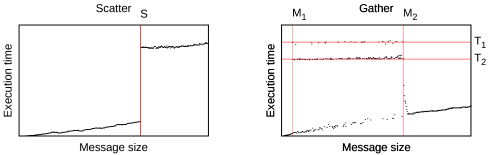
- $T_{ijk}(M) = 2(2C_i + Mt_i) + \max_{x=j,k}(2C_x + M(\beta_{ix} + t_x)) = 2C_i + Mt_i + \max_{x=j,k} T_{ix}(M)$

$$\begin{cases} t_i = \frac{T_{ijk}(M) - \max_{x=j,k} T_{ix}(M) - 2C_i}{M} & \frac{1}{\beta_{ij}} = \frac{T_{ij}(M) - 2(C_i + C_j)}{M} - (t_i + t_j) \\ t_j = \frac{T_{jik}(M) - \max_{x=i,k} T_{jx}(M) - 2C_j}{M} & \frac{1}{\beta_{jk}} = \frac{T_{jk}(M) - 2(C_j + C_k)}{M} - (t_j + t_k) \\ t_j = \frac{T_{kij}(M) - \max_{x=i,j} T_{kx}(M) - 2C_k}{M} & \frac{1}{\beta_{ik}} = \frac{T_{ik}(M) - 2(C_i + C_k)}{M} - (t_i + t_k) \end{cases}$$

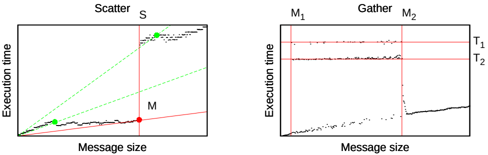
- As t_i can be found from the equations for different triplets including i, find the average variable processor delays: $\bar{t_i} = \frac{\sum_{j,k \neq i} t_i}{C^2_{n-1}}$
- As β_{ij} can be found from the equations for different triplets including $i \leftrightarrow j$, find the average variable processor delays: $\bar{\beta_{ij}} = \frac{\sum_{k \neq i,j} \beta_{ij}}{n-2}$
- **Theoretically OK**

- Measure the execution time to obtain statistically reliable results
  The solution depends on the selection of message size

- Measure the execution time to obtain statistically reliable results
  The solution depends on the selection of message size
- $i \overset{M}{\underset{0}{\longleftrightarrow}} jk = i \overset{M}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ (linear scatter + linear gather)

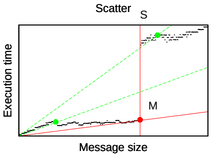


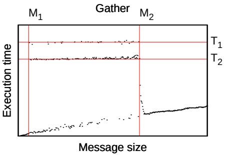
- Measure the execution time to obtain statistically reliable results
  The solution depends on the selection of message size
- $i \xleftrightarrow[0]{M} jk = i \xrightarrow{M} jk + i \xleftarrow[0]{} jk$ (linear scatter + linear gather)



- Selection of message sizes for one-to-two experiments: $i \xleftrightarrow[0]{M} jk$

- Measure the execution time to obtain statistically reliable results
  The solution depends on the selection of message size
- $i \overset{M}{\underset{0}{\longleftrightarrow}} jk = i \overset{M}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ (linear scatter + linear gather)





- Selection of message sizes for one-to-two experiments: $i \overset{M}{\underset{0}{\longleftrightarrow}} jk$
- Estimation of the execution time of linear scatter/gather

$$T_{scatter} = (n-1)(C_0 + Mt_i) + \begin{cases} \max_{i=1}^{n-1}(C_i + M(\beta_{0i} + t_i)) & M < S \\ \sum_{i=1}^{n-1}(C_i + M(\beta_{0i} + t_i)) & M \geq S \end{cases}$$

$$T_{gather} = (n-1)(C_0 + Mt_i) + \begin{cases} \max_{i=1}^{n-1}(C_i + M(\beta_{0i} + t_i)) & M < M_1 \\ \sum_{i=1}^{n-1}(C_i + M(\beta_{0i} + t_i)) & M > M_2 \end{cases}$$

- Statistical linear models $\{M^i, T(M^i)\}, M^{i+1} = M^i + s$, s - stride
  How to find the threshold parameters?

- Statistical linear models $\{M^i, T(M^i)\}, M^{i+1} = M^i + s$, s - stride
  How to find the threshold parameters?
- **The R environment for statistical computing**
  A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber, **Strucchange: An R package for testing for structural change in linear regression models**, *Journal of Statistical Software*, vol. 7, pp. 1-38, 2002

- Statistical linear models $\{M^i, T(M^i)\}, M^{i+1} = M^i + s$, s - stride
  How to find the threshold parameters?
- **The R environment for statistical computing**
  A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber, **Strucchange: An R package for testing for structural change in linear regression models**, *Journal of Statistical Software*, vol. 7, pp. 1-38, 2002
- Locate the break in the execution time of scatter, S, and the range of large messages for gather, M_2

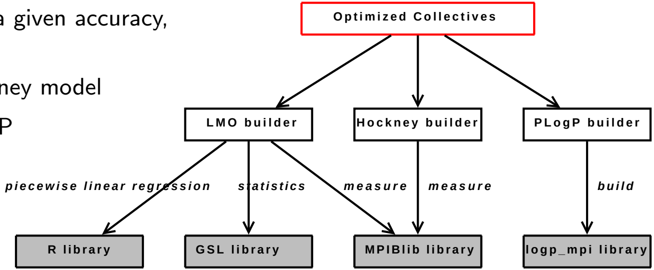
- Statistical linear models $\{M^i, T(M^i)\}, M^{i+1} = M^i + s$, s - stride
  How to find the threshold parameters?
- **The R environment for statistical computing**
  A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber, **Strucchange: An R package for testing for structural change in linear regression models**, *Journal of Statistical Software*, vol. 7, pp. 1-38, 2002

- Locate the break in the execution time of scatter, S, and the range of large messages for gather, M_2
- Separate small and large messages in the gather data row
  - Find the minimum message size $M^k = \bar{M}_1$: $T(M^{k+1})/T(M^k) > 10\%$
  - Decrease stride and perform the gather benchmark:
    $\{M^i, T(M^i)\}, M^{i+1} = M^i + s/2 < \bar{M}_1$
  - Repeat until s reaches 1 byte
  - $M_1 = \bar{M}_1$

The library

Builds heterogeneous communication performance models with a given accuracy, in parallel

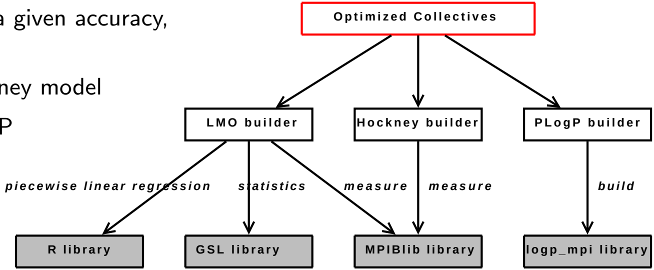
- Heterogeneous Hockney model
- Heterogeneous PLogP (and LogGP) model
- LMO model



The library

Builds heterogeneous communication performance models with a given accuracy, in parallel

- Heterogeneous Hockney model
- Heterogeneous PLogP (and LogGP) model
- LMO model

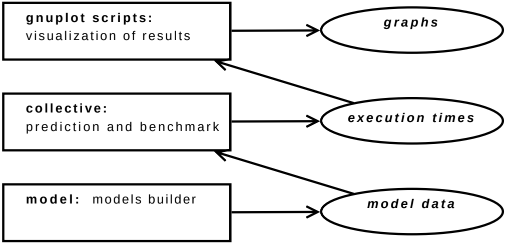


The third-party software

- MPI Benchmarking library (MPIBlib): measurement of the execution time of communication operations
- GNU Scientific Library (GSL): statistics
- The R statistical environment: piecewise linear regression
- The logp_mpi library: building heterogeneous PLogP (and LogGP) model

The tools

- Predicts the execution time of point-to-point and collective communication operations
- Provides model-based optimized implementations of collectives
- Implemented as a library - can be reused in applications
- Provides a basis for a run-time optimization



Model	Number of measurements	Estimation time, sec
hetero-Hockney	rC_n^2	0.17
hetero-PLogP	$(1 + m(2r + s))C_n^2$	63.11
LMO	$2rC_n^2 + cC_n^3$	0.33

Model	Linux	Processor	Bus(MHz)	L2 cache(MB)	#
Dell Poweredge SC1425	2.6	3.6 Xeon	800	2	2
Dell Poweredge 750	2.6	3.4 Xeon	800	1	6
IBM E-server 326	2.4	1.8 AMD Opteron	1000	1	2
IBM X-Series 306	2.4	3.2 P4	800	1	1
HP Proliant DL 320 G3	2.6	3.4 P4	800	1	1
HP Proliant DL 320 G3	2.6	2.9 Celeron	533	0.256	1
HP Proliant DL 140 G2	2.4	3.4 Xeon	800	1	3

- The approach to the estimation of parameters of heterogeneous communication performance models was suggested.
- This approach was applied to heterogeneous switched clusters.
- The solftware tool that automates the estimation was developed.

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

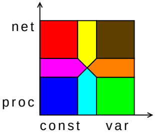
point-to-point execution time: $C_i + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_i)$

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

link parameters: transmission rate (β_{ij})

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

$2n + C_n^2$ parameters



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

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

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

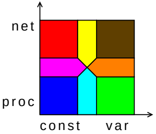
link parameters: latency (L_{ij}) and transmission rate (β_{ij})

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

$2(n + C_n^2)$ parameters

Hockney: $\alpha_{ij} = C_i + L_{ij} + C_j$

$\beta_{ij} = t_i + \frac{1}{\beta_{ij}} + t_j$



University College Dublin

Science Foundation Ireland

IBM Dublin CAS