Revisiting communication performance models for computational clusters

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

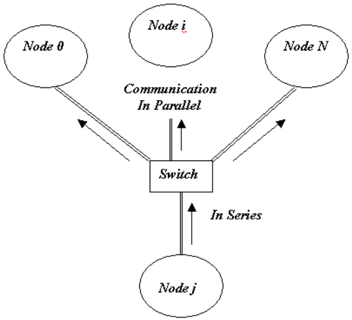
Heterogeneous Computing Laboratory
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland
http://hcl.ucd.ie

18th International Heterogeneity in Computing Workshop
May 25, 2009, Rome, Italy

- MPI-based applications require optimization for heterogeneous platforms
  - Minimization of communication cost
  - Analytical predictive communication performance models
  - Heterogeneous clusters with a single switch

- MPI-based applications require optimization for heterogeneous platforms
  - Minimization of communication cost
  - Analytical predictive communication performance models
  - Heterogeneous clusters with a single switch

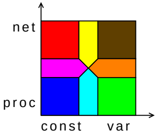
- Analytical predictive communication performance model
  - Point-to-point parameters
  - Prediction $T_{coll}(M, n) =$ combination of point-to-point parameters, message size, M, and number of processors, n



Ideal communication performance model

- Point-to-point parameters: **constant** and **variable** (*message size*) contributions of **processors** and **network**
- $T_{coll}(M, n)$ = combination of *max* (*parallel part*) and $\sum$ (*serial part*) of point-to-point parameters, message size and number of processors
- There is a set of communication experiments that allows for the accurate estimation of the parameters

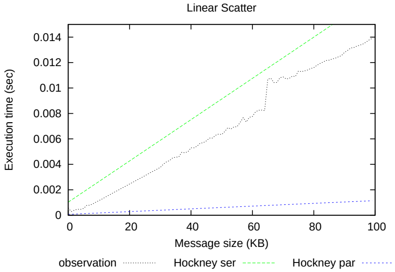
Model	p2p	Experiments
Hockney	$\alpha + \beta M$	$\left\{ i \xleftrightarrow[0]{0} j + i \xleftrightarrow[M]{M} j \right\}_{k=0}^{R}$ or $\left\{ i \xleftrightarrow[M_k]{M_k} j \right\}_{k=0}^{R}$
LogP	$L + 2o$	$\left\{ i \xleftrightarrow[M]{M} j + i \xleftrightarrow[0]{M} j + i \xleftrightarrow[M]{0} j \right\}_{k=0}^{R} + \{ i \xleftrightarrow[0]{\overbrace{M \ldots M}^{2^x}} j \}_{x=0}^{S}$
LogGP	$L + 2o + G(M-1)$	LogP experiments $+ \{ i \xleftrightarrow[0]{\overbrace{\bar{M} \ldots \bar{M}}^{2^x}} j \}_{x=0}^{S}$, large $\bar{M}$
PLogP	$L + g(M)$	$\left[\left\{ i \xleftrightarrow[0]{M_m} j + i \xleftrightarrow[M_m]{0} j \right\}_{k=0}^{R} + \{ i \xleftrightarrow[0]{\overbrace{M_m \ldots M_m}^{2^x}} j \}_{x=0}^{S} \right]_{m=0}^{N}$

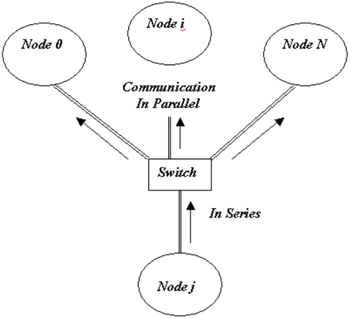


Traditionally designed for homogeneous platforms

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

Example: Hockney model of 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 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

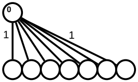
- Small number of parameters, compact formulas for collectives
- $O(n^2)$ communication experiments to estimate the parameters
- Significant heterogeneity = inaccurate prediction

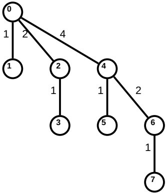
Heterogeneous models

different link- (and processor-) specific parameters

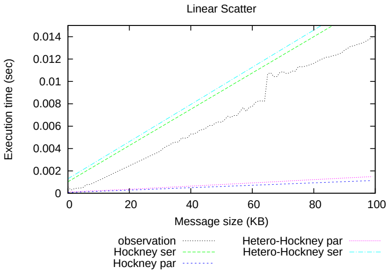
- $O(n^2)$ parameters, flexible formulas for collectives
- $\geq O(n^2)$ communication experiments to estimate the parameters
- **More natural expression of collectives = more accurate prediction**

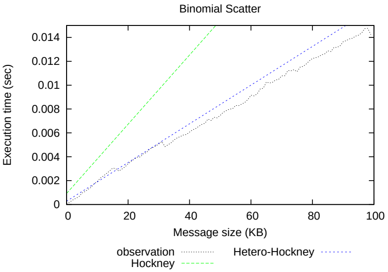
Hockney	**Linear scatter/gather** large-grained parallelism	**Binomial scatter/gather** fine-grained parallelism
Homogeneous	$(n-1)(\alpha + \beta M)$ - serial $\alpha + \beta M$ - parallel	$(log_2 n)\alpha + (n-1)\beta M$ - parallel/serial
Heterogeneous	$\sum_{i=0, i \neq r}^{n-1} (\alpha_{ri} + \beta_{ri} M)$ - serial $\max_{i=0, i \neq r}^{n-1} (\alpha_{ri} + \beta_{ri} M)$ - parallel	$T(k) = \alpha_{rs} + \beta_{rs} 2^{k-1} M + \max_{c \in C_{k-1}} T_c(k-1)$ $\alpha_{04} + 4\beta_{04} M + max \begin{cases} \alpha_{02} + 2\beta_{02} M + \ldots \\ \alpha_{46} + 2\beta_{46} M + \ldots \end{cases}$ $\begin{cases} \ldots + max(\alpha_{01} + \beta_{01} M, \alpha_{23} + \beta_{23} M) \\ \ldots + max(\alpha_{45} + \beta_{45} M, \alpha_{67} + \beta_{67} M) \end{cases}$





Example: Hockney model of heterogeneous cluster

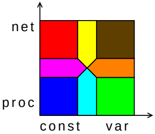




LMO heterogeneous communication performance model

$i \xrightarrow{M} j$:	$(C_i, t_i) \xrightarrow{(L_{ij}, \beta_{ij})} (C_j, t_j)$
point-to-point execution time:	$C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_i)$
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 $\beta_{ij} = \beta_{ji}$

$2(n + C_n^2)$ point-to-point parameters

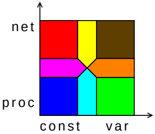


LMO heterogeneous communication performance model

$i \xrightarrow{M} j$:	$(C_i, t_i) \xrightarrow{(L_{ij}, \beta_{ij})} (C_j, t_j)$
point-to-point execution time:	$C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_i)$
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 $\beta_{ij} = \beta_{ji}$

$2(n + C_n^2)$ point-to-point parameters

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



Estimation of the point-to-point 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 the LMO point-to-point parameters

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

Estimation of the LMO point-to-point parameters

- Select the communication experiments and express their execution time via the point-to-point parameters
  - **The point-to-point communications are not enough**
- Measure the execution time of these communications
  - **The execution time should be statistically reliable**
- Build and solve the system of equations, using the times as a right-hand side values
  - **The number of linearly independent equations should be** $\geq 2(n + C_n^2)$

- Point-to-point communications, roundtrips: $i \overset{0}{\underset{0}{\longleftrightarrow}} j$, $i \overset{M}{\underset{M}{\longleftrightarrow}} j$

  $T_{ij}(0) = 2(C_i + L_{ij} + C_j)$ $\quad C_n^2$ equations

  $T_{ij}(M) = 2(C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_j))$ $\quad C_n^2$ equations

- Parallel point-to-two communications: linear scatter + linear gather

  $i \overset{0}{\underset{0}{\longleftrightarrow}} jk = i \overset{0}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ $\quad C_n^3$ equations

  $i \overset{M}{\underset{0}{\longleftrightarrow}} jk = i \overset{M}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ $\quad C_n^3$ equations

  How to express the execution time via the point-to-point parameters?

- Point-to-point communications, roundtrips: $i \overset{0}{\underset{0}{\longleftrightarrow}} j,\ i \overset{M}{\underset{M}{\longleftrightarrow}} j$

  $T_{ij}(0) = 2(C_i + L_{ij} + C_j)$ $\quad C_n^2$ equations

  $T_{ij}(M) = 2(C_i + L_{ij} + C_j + M(t_i + \frac{1}{\beta_{ij}} + t_j))$ $\quad C_n^2$ equations

- Parallel point-to-two communications: linear scatter + linear gather

  $i \overset{0}{\underset{0}{\longleftrightarrow}} jk = i \overset{0}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ $\quad C_n^3$ equations

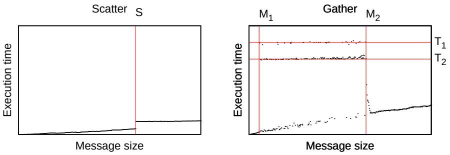
  $i \overset{M}{\underset{0}{\longleftrightarrow}} jk = i \overset{M}{\longrightarrow} jk + i \underset{0}{\longleftarrow} jk$ $\quad C_n^3$ equations

  How to express the execution time via the point-to-point parameters?

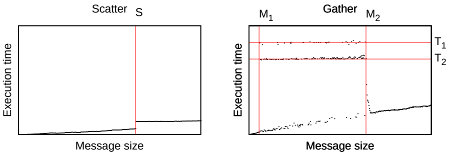
- In a triplet of processors: $i < j < k$

  12 unknowns 12 linearly independent equations

- Observation for the linear scatter/gather



- Observation for the linear scatter/gather

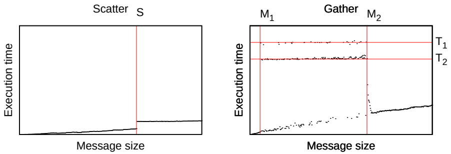


- Prediction for the linear scatter/gather

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

$$T_{gather} = (n-1)(C_r + Mt_r) + \begin{cases} \max_{i=0, i \neq r}^{n-1} \left(L_{ri} + \frac{M}{\beta_{ri}} + C_i + Mt_i\right) & M < M_1 \\ \sum_{i=0, i \neq r}^{n-1} \left(L_{ri} + \frac{M}{\beta_{ri}} + C_i + Mt_i\right) & M > M_2 \end{cases}$$

- Observation for the linear scatter/gather



- Prediction for the linear scatter/gather

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

$$T_{gather} = (n-1)(C_r + Mt_r) + \begin{cases} \max_{i=0,i\neq r}^{n-1}(L_{ri} + \frac{M}{\beta_{ri}} + C_i + Mt_i) & M < M_1 \\ \sum_{i=0,i\neq r}^{n-1}(L_{ri} + \frac{M}{\beta_{ri}} + C_i + Mt_i) & M > M_2 \end{cases}$$

- Selection of message sizes for the point-to-two experiments: $i \overset{M}{\underset{0}{\longleftrightarrow}} jk$:

$$T_{ijk}(M) = 2(2C_i + Mt_i) + \max_{x=j,k}(2(L_{ix} + C_x) + M(\frac{1}{\beta_{ix}} + t_x))$$

Efficiency of estimation

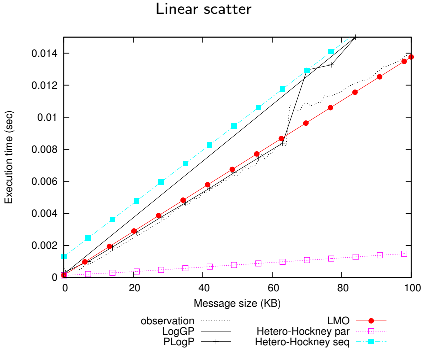
- Parallel estimation of the point-to-point parameters on nonoverlapped sets of processors (on clusters with a single switch)
- Average the values of parameters found independently from different independent experiments:
  - Average C_i and t_i from the equations for different triplets including i:

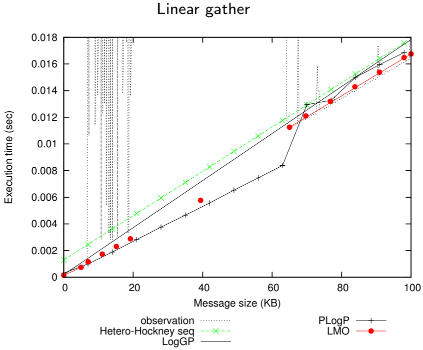
$$\bar{C_i} = \frac{\sum_{j,k \neq i} C_i}{C_{n-1}^2} \qquad \bar{t_i} = \frac{\sum_{j,k \neq i} t_i}{C_{n-1}^2}$$

  - Average L_{ij} and β_{ij} from the equations for different triplets including $i \leftrightarrow j$:

$$\bar{L_{ij}} = \frac{\sum_{k \neq i,j} L_{ij}}{n-2} \qquad \bar{\beta_{ij}} = \frac{\sum_{k \neq i,j} \beta_{ij}}{n-2}$$

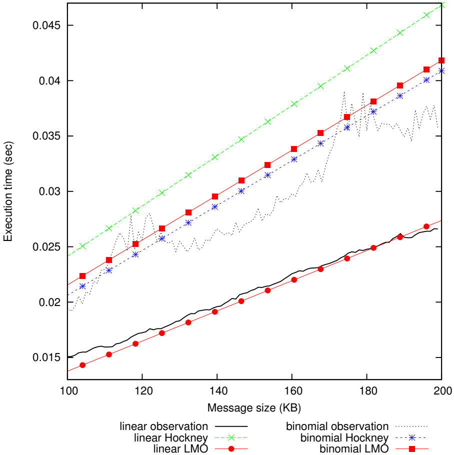
Models' predictions vs observations





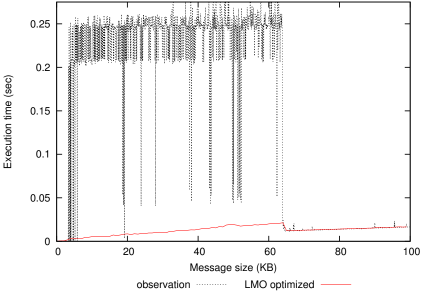
- LMO more accurately predicts the execution time of linear scatter/gather

Model-based switch for scatter



- Hockney: switch to binomial
- LMO: switch to linear

Optimized linear gather



- LMO: splitting the messages of meduim size

- The common problem of all traditional models is the combining of contributions of different nature and, therefore, non-intuiive expression of the execution time of collective communications.
- The LMO model separates the constant and variable contributions of the processors and the network. The execution time of any collective communication operation is expressed as a combination of maximums and sums of the point-to-point parameters and message size.
- The LMO parameters cannot be estimated from only the point-to-point experiments. The efficient technique for accurate estimation was proposed.
- The accuracy of the intuitive modelling of scatter and gather was validated experimentally.

University College Dublin

Science Foundation Ireland

IBM Dublin CAS