文章编号：1001－2486(2002)04－0033－04

# 并行计算的数据重分配*

孙安香，张理论，宋君强

（国防科技大学计算机学院，湖南 长沙　410073）

摘　要：为提高算法的并行计算性能，许多并行程序必须进行数据重分配。数据重分配是在并行计算过程中实现的，其开销影响算法的并行性能，高效的数据重分配对提高并行计算的性能有重要意义。本文阐述了数据重分配的环形算法；提出了数据重分配的蝶网算法，并证明了其正确性；设计了结构性数据交换方法；通过理论和数值实验分析了两种算法的性能。

关键词：并行计算；数据重分配；环形算法；蝶网算法；结构性数据交换

中图分类号：TP301.6　　文献标识码：A

# Data Redistribution of Parallel Computing

SUN An-xiang，ZHANG Li-lun，SONG Jun-qiang

（College of Computer，National Univ. of Defense Technology，Changsha 410073，China）

**Abstract**：Data redistribution is necessary to enhance the algorithm performance in many parallel programs. Since data redistribution is performed at run-time，the cost of redistributing data among processors affects the performance of algorithm. High performance data redistribution is important for the parallel program. Butterfly algorithm for data redistribution of parallel computing is put forward in the paper. We have proved the correctness of the algorithm. Structured data transposition is designed. The performance is analyzed both theoretically and numerically.

**Key words**：parallel computing；data redistribution；circle algorithm；butterfly algorithm；structured data transposition.

数据并行的程序设计模式在多处理机系统环境下的科学计算应用问题中得到广泛应用，数据分配方法对这种并行程序模式的性能影响很大。数据分配包括数据的分布和数据的排列，其中数据的分布为数组怎样分配到各个处理机，数据排列为数组在各个处理机内部的排列方式。研究数据分配方法是为了实现并行计算的最优负载平衡和减少通信开销，从而提高并行计算的效率。从并行计算性能的角度来看，Prylli、Barros 分别论证了解线性代数方程和解二维扩散方程的交替方向隐式法[1]、多维谱变换[2]等算法，固定的数据分配方式因为不能达到并行计算的最优负载平衡，并不是好的数据分配方法，必须进行数据重分配才能实现算法的高效并行计算。数据重分配是在并行计算过程中实现的，其开销将影响算法的并行性能，高效的数据重分配对提高并行计算的性能有重要意义。

科学计算问题的数据分配可归纳为相关变量的矩阵划分，相应的数据重分配可简化为矩阵的重新划分[3～5]。本文主要讨论二维数组的一维数据划分的数据重分配算法—矩阵转置算法，多维数组的多维数据划分可做类似处理。

## 1　数据重分配

并行计算的多处理机系统具有 $n$ 台处理机：$p_1$，$p_2$，$p_3$，…，$p_n$，并行计算的数组为二维数组 $A=(a_{ij})_{l\times m}$，为了简化讨论，假设 $l=nt$，$m=ns$（$m$，$n$ 均为整数），即 $l$ 和 $m$ 能被 $n$ 整除。$l$ 或 $m$ 不能被 $n$ 整除时，对整除的余数行或列略作处理即可。数据重分配后的二维数组为 $\boldsymbol{B}=(b_{ij})_{m\times l}$，$B$ 的第 $i$ 行第 $j$ 列元素与 $A$ 的第 $j$ 行第 $i$ 列元素相同，即：

$$b_{ij}=a_{ji}\quad(i=1,2,\cdots,m,\ j=1,2,\cdots,l)$$

* 收稿日期：2002－03－12
基金项目：国家 863 高技术项目资助（306－ZD01－03－4）
作者简介：孙安香（1965－），女，副研究员，硕士。

数据重分配时处理机 $p_i$（$i=1$，2，…，$n$）的元素不需要全部进行数据交换，有以下引理：

引理1 $A=(a_{ij})_{l\times m}$按列平均分配到$n$台处理机，当$(x-1)\ s+1\leqslant i\leqslant xs$，$a_{ij}$为$p_x$的数据。

对于 $B$ 的每一个元素 $b_{ji}$同样有引理2：

引理2 $B=(b_{ji})_{m\times l}$按列平均分配到$n$台处理机，当$(y-1)\ t+1\leqslant j\leqslant yt$，$b_{ji}$为$p_y$的数据。

由引理1和引理2可得到以下结论：

定理1 $A=(a_{ij})_{l\times m}$按列平均分配到$n$台处理机，当$(x-1)\ s+1\leqslant i\leqslant xs$，且$(y-1)\ t+1\leqslant j\leqslant yt$（$x\neq y$），按列分布的数据重新分配（即实现转置）时，$a_{ji}$的目的处理机为$p_x$、$a_{ij}$的目的处理机为$p_y$。

定理2 $A=(a_{ij})_{l\times m}$按列平均分配到$n$台处理机，当$(x-1)\ s+1\leqslant i\leqslant xs$，且$(x-1)\ t+1\leqslant j\leqslant xt$时，按列分布的数据重新分配前后，$a_{ij}$均为处理机$p_x$的元素。

## 2 数据重分配算法

### 2.1 环形算法

文献[4]等对环形算法作了深入研究。环形算法将数据交换分为$n-1$个阶段：在第$k$（$0\leqslant k<n$）阶段，处理机$p_i$（$0\leqslant i<n-1$）给处理机$p_{(i+k)\bmod(n)}$发送一个消息。当$l=m=n=4$时，$A=(a_{ij})_{4\times4}$按列分配到4台处理机：处理机$p_i$（$i=0$，1，2，3）拥有$A$的第$i$列数据。将数据交换分成3个阶段完成，其中第$j$（$j=1$，2，3）阶段第k台处理机将数据$a_{k,(k+i)\bmod(4)}$送给$(k+i)\bmod(4)$，这样即可完成$A$在4台处理机上按列行分配的数据重分配。

从环形算法的描述可以看出，消息的发送和接收交替进行，每个处理机都以$\max(g,2o)$的时间间隔发送或接收消息，整个网络传输被组织成一个流水线作业，这样某个消息在节点的发送时间和另一个消息在网上的延迟时间可以重叠。Logp模型下环形算法实现$A=(a_{ij})_{n\times n}$的数据重分配的时间为：

$$T=(n-1)\ n\ [\max(2o,g)+L-g]$$

### 2.2 蝶网算法

$A=(a_{ij})_{n\times n}$，处理机数目为n，且$n=2^r$，处理机$p$（$p=0$，1，2，…，$n-1$）拥有$A$的第$p$列数据$\{a_{ip}:i=0,1,2,\cdots,n-1\}$。为方便讨论，以下给出几个记号：

(1) 处理机号$p$的二进制表示为：$p=(x_0x_1\cdots x_{r-1})_2$，其中

$$x_k=0\text{ 或 }1\ (k=0,1,2,\cdots,r-1),\ x_k\text{ 的反为 }\bar{x}_k;$$

(2) 处理机$p$的数据$a_{ip}$简为$a(j)$；

(3) send$(a(j),h,q,p)$表示将数据$a(j+iq)$（$i=0$，1，2，…，$h-1$）送给处理机$p$；

(4) recv$(a(j),h,q,p)$表示从处理机$p$处接收数据$b(j+iq)$（$i=0$，1，2，…，$h-1$）。

每个处理机的发送数据和接收数据分$r$个阶段进行，每个阶段给特指的一台处理机发送数据，然后从另一台特指的处理机接收数据。对于处理机$p=(x_0x_1\cdots x_{r-1})_2$算法描述如下：

```
do   k = 1, r-1
     p_k = (x_0 x_1 ... x_{k-1} x̄_k x_{k+1} ... x_{r-1})_2
     send (a (n/2), n/2, 1, p_k)
     if (p_k = 0) then
        do   i = (n-1)/2, 0, -1
             a (2i) = a (i)
        enddo
     recv (a (1), n/2, 2, p_k)
```

else

do $i=\frac{n}{2}$，$n-1$

$b(2i+1-n)=b(i)$

enddo

fecv（$a(0)$，$\frac{n}{2}$，2，$p_k$）

endif

enddo

**引理** 3 $A=(a_{ij})_{n\times n}$ **按列分配到** n **台处理机上，按蝶网算法实现的数据重分配，在第** $k$ **步，处理机** $p=(x_0x_1\cdots x_{r-1})_2$ **的数据** $a_{ij}$ **满足：**

**(1) 下标标号** $j$ **的不同个数为** $\mu=2^{k+1}$（$j_0<j_1<\cdots<j_{\mu-1}$，**且** $j$ **有形如：**

$$(x_0'x_1'\cdots x_k'x_{k+1}\cdots x_{r-1})_2$$，**其中** $x_k'$ **和** $x_h$ **为** 0 **或** 1

**(2) 标号** $i$ **的不同个数为** $\lambda=2^{r-k-1}$，**且** $i_0=\sum_{v=0}^{k}p_v2^{r-v-1}$，$i_v=i_0+v$

**(3) 处理机** $p$ **的数据的下标排列为：**

$$\begin{array}{ll}(i_0,\ j_0) & (i_0,\ j_1)\cdots(i_0,\ j_{\mu-1})\\ (i_1,\ j_0) & (i_1,\ j_1)\cdots(i_1,\ j_{\mu-1})\\ & \vdots\\ (i_{\lambda-1},\ j_0) & (i_{\lambda-1},\ j_1)\cdots(i_\lambda,\ j_{\mu-1})\end{array}$$

**可用归纳法证明引理** 3 **的结论成立。当** $k=r$ **时，由引理** 3 **可得到以下定理：**

**定理** 3 $A=(a_{ij})_{n\times n}$ **在** $n$ **台处理机上按列分布，处理机** $p$ **的数据为** $\{a_{ip}:i=0,1,2,\cdots,n-1\}$，**按蝶网算法进行数据重分配后，处理机** $p$ **的数据为** $\{a_{pi}:i=0,1,2,\cdots,n-1\}$。

**蝶网算法实现** $A=(a_{ij})_{n\times n}$ **在** $n$ **台处理机的列行数据重分配，实际上是将每个处理机的发送和接收数据分成** $r=\log_2 n$ **个阶段来完成，每个阶段发送和接收一次数据。根据** Log$p$ **模型，蝶网算法在超立方体网络结构**[6] **多处理机系统实现数据重分配的时间为：**

$$T=[\max(2o,\ g)+L-g]\times\frac{2n}{\sigma}\log_2 n$$

$\sigma$ **为超立方体的每一级处理机数；在蝶网网络结构**[6] **机制的多处理机环境下实现数据重分配的时间为：**

$$T=[\max(2o,\ g)+L-g]\times 2\times\log_2 n$$

## 3 结构性数据交换

**讨论** $A=(a_{ij})_{l\times m}$ **在** $n$ **台处理机按列分布的数据重分配实现，其中** $l$ **和** $m$ **均能被** $n$ **整除。**

**将** $A=(a_{ij})_{l\times m}$ **剖分为** $n^2$ **个矩阵块** $A=(A_{IJ})_{n\times n}$，$A_{IJ}=(a_{\alpha\beta}^{IJ})_{\frac{l}{n}\times\frac{m}{n}}$，**其中**

$$\alpha=i-[(I-1)\times\frac{l}{n}+(J-1)\times\frac{m}{n}],\ \beta=j-[(I-1)\times\frac{l}{n}+(J-1)\times\frac{m}{n}]$$

$A$ **在** $n$ **台处理机按列分布，处理机** $p$ **的数据为矩阵集** $\{A_{IP}:I=0,1,\cdots,n-1\}$ **中的数据。由定理** 1 **和定理** 2 **可知：实现** $A=(a_{ij})_{l\times m}$ **在** $n$ **台处理机按列分布的数据重分配，矩阵块** $A_{pp}$（$p=0,1,\cdots,n-1$）**的元素只需在本处理机** $p$ **内部进行数据重排；源处理机** $I$（source）**的矩阵块** $A_{IJ}$（$I\neq J$；$I$，$J=0,1,\cdots,n-1$）**的元素具有同一目的处理机** $J$（dest）。

$A_{IJ}$（$I\neq J$；$I$，$J=0,1,\cdots,n-1$）**的元素** $a_{\alpha\beta}^{IJ}$ **具有同一目的处理机** $J$，**但是在** $A$ **中不完全连续。为完成** $A=(a_{ij})_{l\times m}$ **按列分布的数据重分配，一种直接的方法是将** $A_{IJ}$ **分成**$\frac{m}{n}$**个消息，传送给处**

理机 $J$，其中第 $\beta$ 个消息为数据$\left\{ a_{\alpha\beta}^{IJ}: \alpha = 1, 2, \cdots, \frac{1}{n} \right\}$。完成 $A_{IJ}$的发送或接收的时间为：

$$t_{ns} = \frac{m}{n} \left[ \max (2o, g) + L - g \right]$$

我们称这种方法为“非结构性数据交换”方法。

根据定理1和定理2，我们设计了“结构性数据交换”方法。“结构性数据交换”方法利用消息传递环境的打包功能，将 $A_{IJ}$（$I \neq J$；$I$，$J = 0, 1, \cdots, n-1$）的元素打成一个包；共分成$\frac{m}{n}$次打包，第 $\beta$ 次的元素为$\left\{ a_{\alpha\beta}^{IJ}: \alpha = 1, 2, \cdots, \frac{1}{n} \right\}$，然后源处理机（source）$I$ 的矩阵块 $A_{IJ}$（$I \neq J$；$I$，$J = 0, 1, \cdots, n-1$）的元素作为一个消息发送给目的处理机（dest）$J$。完成 $A_{IJ}$的发送或接收的时间为：

$$t_s = \frac{m}{n} \times 2o + \max (2o, g) + L - g$$

由于$\frac{m}{n} > 1$、$o \ll L$，所以 $t_{ns} > t_s$；而且结构性数据交换比非结构性数据交换减少了消息数目，减轻网络通信的拥挤。由此可见结构性数据交换可减少数据通信的总时间。

## 4 数值实验和结论

采用结构性的数据交换方法，设计了环形算法和蝶网算法的数据重分配实验程序。实验环境为分布式存储的多处理机系统 MPP，网络结构为超立方体，程序语言为 FORTRAN，消息传递库为标准的 MPI。数值实验结果如附表所示，与第3部分的理论分析一致。

附表 两种数据重分配算法的实现时间（单位：s）

| 处理机 | 2 | | 4 | | 8 | | 16 | |
|---|---|---|---|---|---|---|---|---|
| 规模 \ 算法 | 环形 | 蝶网 | 环形 | 蝶网 | 环形 | 蝶网 | 环形 | 蝶网 |
| $2^{10} \times 2^{10}$ | 9.23 | 8.92 | 11.32 | 9.32 | 12.56 | 10.89 | 14.44 | 11.12 |
| $2^{11} \times 2^{11}$ | 11.56 | 10.94 | 12.57 | 10.65 | 13.69 | 12.15 | 15.78 | 12.89 |
| $2^{12} \times 2^{12}$ | 13.05 | 11.33 | 14.48 | 12.07 | 15.71 | 13.97 | 16.53 | 14.15 |

理论分析和数值实验表明：超立方体网络结构的分布式存储环境下，所设计的蝶网数据重分配算法所占开销较环形算法少，可提高需要数据重分配的算法的并行计算效率。

## 参考文献：


[1] Prylli L, Tourancheau B. Fast runtime block cyclic data redistribution on multiprocessors [J]. Journal of Parallel and Distributed Computing, 1997, 45 (8): 63 – 72.

[2] Barros Sau R M. On the paralleliation of spectral eulerian shallow models [J]. Parallel surpercomputing In Atmosspheric Science, 1992, 11: 37 – 43.

[3] SUN Anxiang, SONG Junqiang, KONG Jingzhu. Parallel Spectral Method on Distributed Memory Computer [C]. An International Workshop on Computational Physics. Aug, 1999.

[4] CHING – HSIEN Hsu, YEH – CHING Chung. Efficient methods for $kr \rightarrow r$ and $r \rightarrow kr$ array redistribution [J]. The Journal of Supercomputing, 1998, 12: 253 – 276.

[5] 孙安香，宋君强，李晓梅. 数值气象预报中的并行计算研究 [J]. 高技术通讯，2001 (12).

[6] 李晓梅等. 面向结构的并行法—设计与分析 [M]. 长沙：国防科技大学出版社，1996.