Contents

• Simple Message Transfer
• One-to-All Broadcast and All-to-One Reduction
• All-to-All Broadcast, Reduction
• Prefix Sums
• One-to-All Personalized Messages
• All-to-All Personalized Messages*
All-Reduce

- Use all-to-all broadcast communication pattern
- Add data at each step instead of doubling message size
- Result: all processes have the global sum
- The total communication time (hypercube):

\[ T = (T_s + mt_w) \log p \]

- Note that this is the same cost as one-to-all broadcast

Prefix Sums

- A related pattern is prefix-sums
  - Each node computes a partial sum rather than all numbers
- Example
  - Original sequence
    - \(<3, 1, 4, 0, 2>\)
  - The sequence of Prefix sums
    - \(<3, 4, 8, 8, 10>\)
- Compute sums

\[ s_k = \sum_{i=0}^{k} n_i \quad 0 \leq k \leq p - 1 \]
Prefix Sums

(a) Initial distribution of values
(b) Distribution of sums before second step
(c) Distribution of sums before third step
(d) Final distribution of prefix sums

Figure 4.13 Computing prefix sums on an eight-node hypercube. At each node, square brackets show the local prefix sum accumulated in the result buffer and parentheses enclose the contents of the outgoing message buffer for the next step.

Prefix Sum Algorithm

1. procedure PREFIX_SUMS_HCUBE(my_id, my_number, d, result)
2. begin
3. result := my_number;
4. msg := result;
5. for i := 0 to d − 1 do
6. partner := my_id XOR 2^i;
7. send msg to partner;
8. receive number from partner;
9. msg := msg + number;
10. if (partner < my_id) then result := result + number;
11. endfor;
12. end PREFIX_SUMS_HCUBE

Algorithm 4.9 Prefix sums on a d-dimensional hypercube.
Scatter and Gather

- Scatter (one-to-all personalized communication)
  - The source node sends p unique messages, each destined for each node.
  - Different from one-to-all broadcast. Why?
- Gather
  - A single node collects a unique message from each node.
  - No combination or reduction of data

\[
\begin{align*}
M_0 & \\
M_1 & \\
M_p & \\
\vdots & \\
\end{align*}
\]

Figure 4.14 Scatter and gather operations.

Scatter on a Hypercube

(a) Initial distribution of messages
(b) Distribution before the second step
(c) Distribution before the third step
(d) Final distribution of messages

Figure 4.15 The scatter operation on an eight-node hypercube.
**Scatter/Gather Cost Analysis**

- Communication pattern is like broadcast
- The message sizes shrink at each step
- Communication time:

\[
T = \sum_{i=1}^{\log p} (t_x + 2^{\log p-i} mt_x)
\]

\[
= \sum_{i=1}^{\log p} (t_x + 2^{i-1} mt_x)
\]

\[
= t_x \log p + mt_x (p - 1)
\]

**All-to-All Personalized Communication**

- Each node sends a distinct message of size \(m\) to every other node.
- Each node sends different messages to different nodes

\[
\begin{align*}
M_{0,0} & \quad M_{1,0} & \quad \ldots & \quad M_{p-1,0} \\
M_{0,1} & \quad M_{1,1} & \quad \ldots & \quad M_{p-1,1} \\
m & \quad m & \quad \ldots & \quad m
\end{align*}
\]

\[
\begin{align*}
M_{0,p-1} & \quad M_{1,p-1} & \quad \ldots & \quad M_{p-1,p-1} \\
M_{0,0} & \quad M_{1,0} & \quad \ldots & \quad M_{p-1,0} \\
m & \quad m & \quad \ldots & \quad m
\end{align*}
\]

\[
\text{All-to-all personalized communication}
\]

\[
\begin{align*}
0 & \quad 1 & \quad \ldots & \quad p-1
\end{align*}
\]

\[
\begin{align*}
0 & \quad 1 & \quad \ldots & \quad p-1
\end{align*}
\]

**Figure 4.16** All-to-all personalized communication.
Matrix Transposition

- Map matrix onto n processors
- Process i sends all values other than i to others

Figure 4.17  All-to-all personalized communication in transposing a $4 \times 4$ matrix using four processes.

Summary of Communication Time

Table 4.1  Summary of communication times of various operations discussed in Sections 4.1–4.7 on a hypercube interconnection network. The message size for each operation is $m$ and the number of nodes is $p$.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Hypercube Time</th>
<th>B/W Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-to-all broadcast, All-to-one reduction</td>
<td>$\min((t_s + t_w)m \log p, 2(t_s \log p + t_wm))$</td>
<td>$\Theta(1)$</td>
</tr>
<tr>
<td>All-to-all broadcast, All-to-all reduction</td>
<td>$t_s \log p + t_wm(p - 1)$</td>
<td>$\Theta(1)$</td>
</tr>
<tr>
<td>All-reduce, Scatter, Gather</td>
<td>$\min((t_s + t_wm) \log p, 2(t_s \log p + t_wm))$</td>
<td>$\Theta(1)$</td>
</tr>
<tr>
<td></td>
<td>$t_s \log p + t_wm(p - 1)$</td>
<td>$\Theta(1)$</td>
</tr>
</tbody>
</table>
CS630/CSC730: Parallel & Distributed Computing
Questions?