4.8 总结

表 4.1 总结了本章讨论的各种集体通信操作的通信时间。One-to-All Broadcast、All-to-One Reduction和All Reduce操作的时间是两个表达式的最小值。这是因为，根据信息大小 m 的不同，第 4.1 和 4.3 节中描述的算法或第 4.7 节中描述的算法速度更快。表 4.1 假设选择了最适合给定报文大小的算法。表 4.1 中的通信时间表达式是本章前几节在超立方体互连网络的背景下，以直通路由方式推导出来的。不过，这些表达式和相应的算法适用于任何具有 $\Theta(p)$ 截面带宽的架构（第 2.4.4 节）。事实上，表 4.1 中列出的所有操作的表达式中，除了All-to-All Personalized Communication和循环移位外，与 t_w 相关的术语即使在环网和网状网络（或任何 k-d 网状网络）上也保持不变，前提是逻辑进程被适当映射到网络的物理节点上。表 4.1 的最后一列给出了在表中第二列给出的时间内执行一项操作所需的渐近截面带宽，假定进程与节点的映射最优。对于大型报文，只有全对全个性化通信和循环移位需要全部 $\Theta(p)$ 截面带宽。因此，如第 2.5.1 节所述，在横截面带宽较小的网络上应用这些操作的时间表达式时，t_w 项必须反映有效带宽。例如，p 节点方形网格的横截面宽度为 $\Theta(\sqrt{p})$，p 节点环形网格的横截面宽度为 $\Theta(1)$。因此，在方形网格上执行All-to-All Personalized Communication时，每个字的有效传输时间是 $\Theta(\sqrt{p})$ 乘以单个链路的 t_w，而在环形网格上，则是 $\Theta(p)$ 乘以单个链路的 t_w。

表4.1 第4.1-4.7节讨论的各种操作在超立方互连网络上的通信时间汇总。每个操作的信息量为m，节点数为p

Operation	**Hypercube Time**	**B/W Requirement**
One-to-All Broadcast / All-to-One Reduction	$\min\left((t_s + mt_w)\log p, 2(t_s \log p + mt_w)\right)$	$\Theta(1)$
All-to-All Broadcast / All-to-All Reduction	$t_s \log p + t_w m(p-1)$	$\Theta(1)$
All Reduce	$\min\left((t_s + mt_w)\log p, 2(t_s \log p + t_w m)\right)$	$\Theta(1)$
Scatter / Gather	$t_s \log p + t_w m(p-1)$	$\Theta(1)$
All-to-All Personalized	$(t_s + t_w m)(p-1)$	$\Theta(p)$
Circular Shift	$t_s + t_w m$	$\Theta(p)$

本章讨论的集体通信操作在许多并行算法中经常出现。为了便于快速、可移植地设计高效的并行程序，大多数并行计算机供应商都提供了用于执行这些集体通信操作的预打包软件。这些操作最常用的标准 API 被称为消息传递接口（Message Passing Interface）或 MPI。表 4.2 列出了与本章所述通信操作相对应的 MPI 函数名称。

表4.2 本章讨论的各种操作的MPI名称

Operation	**MPI Name**
One-to-All Broadcast	MPI_Bcast
All-to-One Reduction	MPI_Reduce
All-to-All Broadcast	MPI_Allgather
All-to-All Reduction	MPI_Reduce_scatter
All Reduce	MPI_Allreduce
Gather	MPI_Gather
Scatter	MPI_Scatter
All-to-All Personalized	MPI_Alltoall