第6章、 微机集群的任务管理

一个大规模的微机集群，它的任务调度管理也是一项非常重要的工作。要使这样一个由很多微机连接在一起的大规模微机集群能够有效地工作，充分发挥它的威力，或者说，能够象一个传统超级计算机一样有效地工作，还必须以某种方式对这个微机集群进行有效管理。这样的管理方式至少要使一个并非是计算机专家的一般用户使用起来也比较方便，而且最好使他们并不感觉到他们是在使用一大群机器。使用的情况最好与他们使用单个独立的计算机接近。否则，使用不方便，微机集群的使用效率就会大打折扣。

1. OpenPBS 概述

OpenPBS（Portable Batch System）最初是由美国航空航天局（NASA）开发的任务排队调度管理系统，已经成为微机集群优先采用的任务管理软件。

1.1. 任务管理的必要性

在一个大型并行计算机中，管理计算任务的递交，控制计算任务的运行，区分用户运行的权限等等是一个十分困难的工作。没有一个好的管理，并行计算机利用效率就很低：或者互相挤塞，或者是十分空闲。

如果仅仅靠人工地进行任务递交，比如递交定时的后台作业（at 命令），对一个分布式并行计算机来说是远远不够的。因为就算是自己的作业也很难预计什么时候会结束。如果是他人的作业，就根本不可能准确估计结束的时间。如果后台作业提前执行或延迟执行都不好：提前不但可能使一个处理器上运行多个任务，问题最大的是如果前一个作业的输出数据需要作为下一个作业输入数据；而延迟会造成计算机资源浪费。同时，如果用户需要学习很烦琐的任务递交才能有效地使用并行计算机，不但很花费时间，也不一定能用好。对于不同系统，不同命令，用户无所适从，不能把主要精力集中在所研究的核心问题上。

由于分布式并行工作方式，微机集群的使用效率与管理水平有直接的联系。

为什么这样说？在第 2 章，我们已经介绍了微机集群的并行计算方式：将一个串行计算任务分成几个阶段，每个阶段分割成几份，让所有计算节点机共同承担。完成一个阶段的计算任务后，进行数据交换，以得到下一阶段所需的数据，然后才进入下一阶段的计算，如此往复进行。

因此，这样的分布内存式并行计算机最有效的利用方式应该是，让所有计算节点的计算任务基本平衡。就是所有节点机最好能够在相同的时间里完成同一阶段的计算任务，然后进行数据交换，再进入下一阶段的计算。在每个节点机完成本阶段的计算任务前，即使已经先完成任务的节点机也必须等待其他节点机完成这一阶段的计算，而不能提前进入下一阶段的计算。因为，下一阶段所需的数据要通过前一计算阶段并行计算后，交换数据后才能得到。而且，由于分布式并行计算机的特点，已经完成的本节点计算任务的计算节点也不能分担同一计算阶段其他计算节点的计算任务，这与共享内存的并行计算机不同。否