并行高分辨率有限区预报系统在 IBM SP 上的建立*

朱政慧　闫之辉
(国家气象中心数控室 ,北京 100081)
Zaphiris Christidis
(IBM TJ Watson Research Center , New York 10598 , USA)

从理论上讲 ,在多节点多处理器并行计算机系统上用 MPI/ Open MP 进行混合并行编程 ,即在节点间用 MPI 通信 ,而节点内用 Open MP ,应该取得比用纯 MPI 并行化的应用程序更好的并行性能。目前 ,各国的应用程序开发者和气象工作者们都在致力于这方面的研究。可以说 ,这种方法的可行性是不容置疑的 ,它应当是今后并行编程框架设计的共同趋势。本文介绍了在 IBM SP 并行计算机上优化和并行化高分辨率有限区同化预报系统 (HLFAS) 中模式所采用的这种混合编程方法和技术。

1　HLAFS 系统

高分辨率有限区同化预报系统包括资料分析同化部分、预报模式部分和后处理部分。本文主要介绍模式部分。

作为我国自行开发的一个区域天气预报系统 ,近 10 年来 ,其分辨率、物理过程和参数化方案几经改进 ,至 1998 年运行水平分辨率为 0 .5° × 0 .5°的业务系统 ,对范围在 15° ~ 64° N ,70° ~ 145° E 之间的我国大部分地区作 60 h 的短期天气预报 ,并下发 48 h 以内的产品。对该系统的详细介绍请查阅参考文献[1]和[2]。

当前模式分辨率提高到 0 .25° × 0 .25° ,并引入新的包括冰相过程的显式降水方案 ,模式计算格点量扩大到原来的 4 倍 ,为 $I \times J \times K = 361 \times 237 \times 20$,双精度串行版本在 SP 机上用单 CPU 运转完成 48 h 预报墙钟时间大于 180000 s。可见 ,为满足业务运转需求(低于 9000 s) ,并行处理至关重要。

2　IBM SP 并行机

中国气象局配置有 10 个计算节点的 IBM RS6000 SP 系统。每个节点内有 8 个 POWER3 222 MHz 处理器 ,由一个共享内存交叉开关相连 ,采用了对称多处理器结构 (SMP ,Symmetric Multi-Processor) 。不同的节点间通过高性能开关(HPS , High Performance Switch) 通信。在每个 SMP 节点内处理器之间的通信可通过共享内存完成 ,而一个 SMP 节点不能直接访问另外一个节点的内存 ,必须通过消息传递实现。

3　HLAFS 模式的优化和并行化

3 .1　高速缓存的优化使用

最初的 0 .5° × 0 .5° HLAFS 模式是在 CRAY 向量机 C92 上作为业务系统运转的 ,鉴于共享内存型向量机具有高带宽、大内存的特性 ,原模式代码的编写也适用于向量机大吞

* 2001-11-28 收到 ,2002-01-08 收到修改稿。

吐量的特性。而 SP 机作为一种基于高速缓存(CACHE) 的 RISC 芯片商用机,需要充分考虑 CACHE 的访问问题才能获得较好的应用程序性能。因此,源程序代码需要进行极大的调整。首先,我们将所有的三维数组索引顺序由(I , J , K) 转换为(K , J , I) ,改变了循环的访问顺序,依据对数据块访问的空间局部性,CACHE 的命中率提高,使用高效,加快了计算速度。

我们还将该 64 位模式改为 32 位模式来完成计算。尽管 POWEER3 芯片的浮点运算是 64 位的,后者比前者完成 48 h 预报所需的墙钟时间短。这是因为单精度(4 个字节) 的三维数组最内层 $K = 20$ 共占用 80 字节,能一次性载入高速缓存行(cache lines) 内(128 字节) 。而双精度则需 160 字节,不能一次装入,因而 CACHE 的利用性能不如单精度版本。为了保证降水场的预报精度,我们使整个模式单精度运转,但在计算与降水过程有关的预报量时仍使用双精度。由此得到的 24 及 48 h 降水场预报图表明,其结果基本上同双精度模式所得的预报场相同。

另外,在优化源代码的过程中,考虑到不同的算术运算所需的时钟周期不同,我们注意将除法运算转换成乘法。同时,引入了饱和水汽压的查表法计算,将整个双精度模式的内存需求由 1 .7 G 减少到约 690 M 字节,而单精度模式只需 370 M,极大地改善了模式性能。

3 .2 HLAFS 模式的并行处理

3 .2 .1 并行计算方案

考虑到 IBM RS6000 SP 机多节点多处理器的结构特点,我们沿经圈方向(J) 进行了分区映射,分配给不同的 MPI 任务(task) 共同完成。而沿纬圈方向的计算量通过 Open-MP 并行分配给同一节点内的多个 CPU 共同处理,称为多线程(thread) 。I/ O 由一个 CPU 来完成,为保持较好的负载平衡,分配给该 CPU 的计算量要相对少一些。以 4 个 MPI 任务和 4 个 Open MP 线程为例,分区映射方案如下图所示:

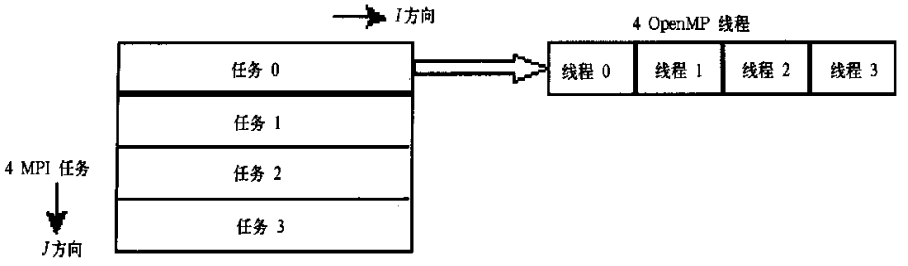

图 1 分区映射方案

3 .2 .2 分布 共享相结合的并行处理方法

由于模式格点为 C 型跳点,我们定义了数组 js0(mypid ,1) , js0(mypid ,2) …,js0(mypid ,7) 和 je0(mypid ,1) , je0(mypid ,2) , …,je0(mypid ,7) 来记录不同物理量的计算格点始止值,为不同的 MPI 任务(mypid) 指定各自所要操作的数据集。

用 MPI 编写并行程序由于需要分解数据结构,编写大量的消息传递程序,比较复杂且容易出错,但在确定了其正确并行运算后,注意减少通讯开销是一关键问题。通常一次

性传递大量的消息比多次传送小消息要高效得多,所以合并小消息、减少不必要的消息传递是减少通讯开销的首要任务。

IBM SP 机上的 OpenMP 库还只是其工业标准建议的子集,在本模式中,我们仅采用了在循环语句上加并行指导语句的方法。为减少分叉/结合的并行方式造成的过多的并行开销(生成并行区、结束并行区等),尽量将一些小循环合并成为大循环。

我们设定编译选项"-qsmp"赋值"noauto",即关闭编译器自动识别可并行区并添加指导语句的功能,全部工作由人工完成,这样可以获得较好的并行性能。即使是在有关消息传递的过程中,我们也消去了原有的数据依赖,加入了指导语句。使得消息传递前的数据收集过程也能多线程并行执行。

总的来说,整个并行模式采用多任务、多线程并行方式运行。除了对输入输出的操作由零号任务完成外,所有的计算由多个任务并行完成,它们执行同一指令代码,但要根据分区映射的结果对不同的数据集进行操作。每个任务在运行过程中一旦遇到由 OpenMP 指导语句标识的循环体,即生成多线程并行区来完成计算,而随着循环体的结束,该并行区也自动结束。

4　结果

现在,整个系统已在 SP 机上连接建立,除后处理部分仍是串行执行外,同化和模式部分都并行运转(使用 4 个节点共 32 个 CPU 计算)。表 1 给出了单精度并行模式多任务、多线程运转作 6 h 预报所需的墙钟时间,可以看出,我们已取得了较好的 MPI 和 OpenMP 并行性能,混合编程的并行模式性能明显优于纯 MPI 方式。

根据单、双精度模式综合试验结果,我们发现,在只可用 4 个节点时,最快的并行运行方式是采用每个节点内分配 2 个 MPI 任务,每个任务由 4 个 OpenMP 线程来完成。表 2 是该配置下并行模式的运行特征指标。

表 1　单精度模式并行运转所需墙钟时间对比　s

总任务数	线程数			
	1	2	4	8
1 × 1	15624	7959	4238	2313
1 × 2	8181	4236	2309	-
1 × 4	4227	2298	-	-
2 × 1	8235	4136	2208	1300
2 × 2	4198	2199	1205	-
2 × 4	2319	1231	-	-
4 × 1	4225	2178	1160	725
4 × 2	2311	1216	706	-
4 × 4	1258	690	-	-

注:总任务数 = 节点数 × 每个节点内的 MPI 任务数

表 2　并行运行性能指标

	64 位	32 位
每秒浮点运算量(MFLOPS/节点)	560	730
48 h 预报墙钟时间(s)	7368	5210

参 考 文 献

1　Yan Zhihui, Guo Xiaorong, Zheng Guoan, et al. The limited area analysis and forecast system and its operational application. *Acta Meteorological Sinica*, 1996, **10**(3): 295 ~ 308.

2　闫之辉.一个载水预报模式的业务预报应用试验.应用气象学报, 1999, **10**(4): 453 ~ 461.