神威机上集合数值天气预报系统中文件资料的管理*

张　贻
(北方计算中心 ,北京 100091)

提　要

从文件资料的命名 、存放与输入输出管理三个方面介绍了在神威集合数值天气预报系统中对大规模文件资料的管理实现 。

关键词 : 集合数值天气预报　文件资料管理　环绕访问

引　言

文件资料在神威集合数值天气预报系统中出现的频率高 、数量大 ,是贯通集合预报各子系统的主要途径之一 ,所以组织好文件资料的管理 ,并根据各子系统文件资料的特点选择相适应的输入输出方式 ,对系统的时效和运行管理的灵活配置有着直接的影响 。

1　文件资料概况

神威集合数值天气预报系统的主体部分由前处理 、资料同化 、集合预报初始场生成 、32 个样本的 10 天模式预报 、640 个时次的模式后处理 、产品生成等六个子系统构成 。集合预报系统运行时 ,所有文件资料所需的总存储量约为 120 GB 左右 ,文件个数近万个 ,分有格式文件和无格式文件两种 ,文件长度最短的为 24 字节 ,最长的达 75 兆字节 。

每个子系统中 ,都有三种不同功能的文件资料 ,即输入文件 、工作文件 、输出文件 。输入文件又包含有两类 ,一类除前处理子系统的输入文件是二进制报文 BFR 码外 ,其它子系统的输入文件是前一子系统的输出文件 ,即各子系统对前一子系统输出的资料进行处理 ,经过科学运算输出资料供下一子系统处理 ;另一类是每个子系统特有的 、固定不变的常数输入文件 。此外 ,各子系统运行时 ,还要产生一些工作文件 ,用于存储在计算过程中不适合保留在内存中的临时变量 。各子系统间主要输入输出资料流向如图 1 所示 。

图 1 中只标出了各子系统间流动的主要输入输出资料 ,各子系统运行时所需要的固定输入文件和运行过程中产生的另外一些输出文件和工作文件数量很大 ,无法在图中一一给出 。此外 ,除图 1 中所示的子系统间输入输出资料外 ,前处理子系统中还包含有报文格式转换 、解码 、数据生成 、探空报合并和报文重排序五个部分 ,集合预报初始场生成子系

* 本文由国家“863”高科技研究发展计划(863-306-ZD11-03-02)项目资助 。
2001-04-28 收到 ,2001-07-16 收到修改稿 。

统中则包含有气候场降阶、谱系数降阶、T21 L19 模式预报、奇异向量生成和谱系数升阶等五部分,子系统中的各部分之间也有大量的资料输入输出,同时也要生成许多工作文件。

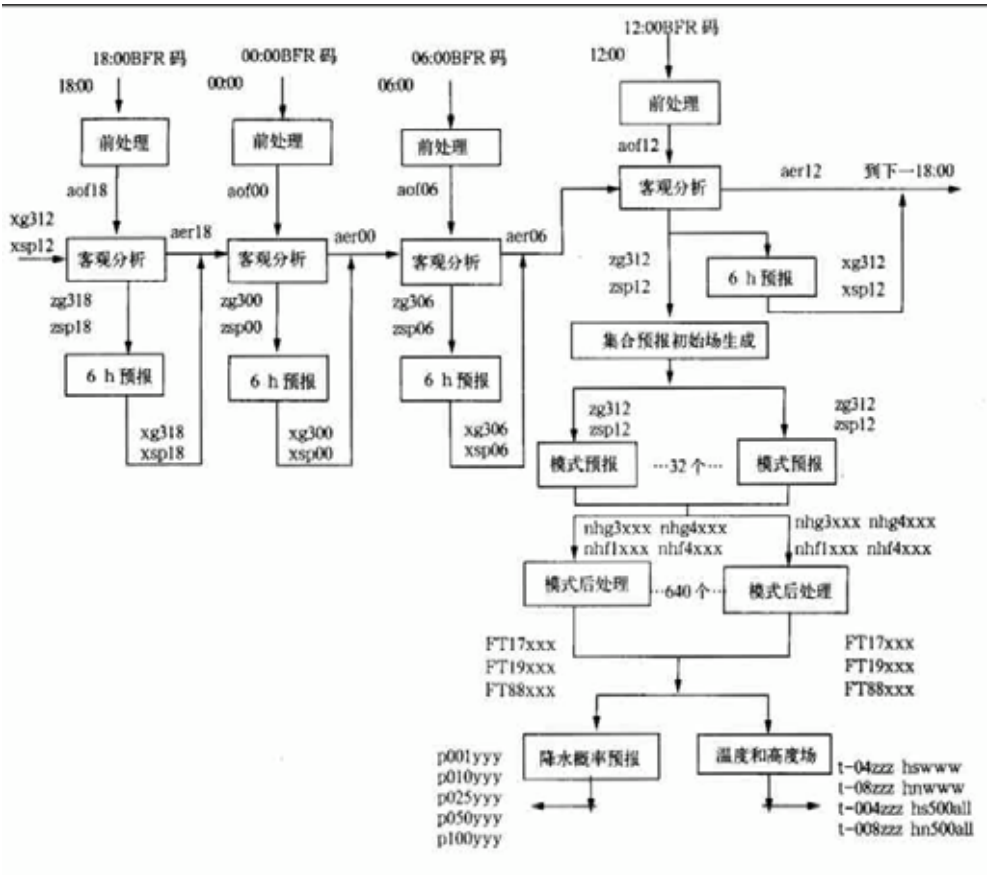

图 1　神威集合数值天气预报系统中主要输入输出资料流向图

(xxx 为 012 024 …240, yyy 为 012 036 …228, zzz 为 024 048 …240, www 为 01 02 …10)

对于这样大存储量、多文件数的资料,如果不对其输入输出方式、存放位置等进行合理的组织管理,势必会影响系统的运行效率,并给系统管理人员造成不必要的麻烦。

2　文件资料的管理

在神威集合数值天气预报系统中对文件资料的管理主要有三个方面,一是文件资料的命名,二是文件资料的存放;三是文件资料的输入输出管理。

2.1　文件资料的命名

集合预报系统中文件资料数目多,使用频繁,合理地命名文件资料不仅可以简化各子系统间和各时次资料滚动处理间的衔接,也给运行管理带来极大的方便。命名时,可将集合预报系统中文件资料分为两类;一类是固定名称的文件资料,即文件资料的名称一旦确

定后,在任何时次的任何子系统中都可以固定不变也不至于引起混乱,命名时用表示其功能含义的英文字母组合即可,运行管理时一目了然;另一类是可变名称的文件资料,即为了系统的正常运转和资料的保存,某些子系统中的文件资料在不同时次运行时,它们的名字不同。对于这类文件,在系统中设计了这样的命名规则:文件资料名只有在系统运行时才产生,程序中读入现在所运行的时次后,用表示其功能含义的英文字母组合拼接上读入时次组成所需要的文件名。这种命名方式为系统的运行管理带来了极大的方便。

2.2 文件资料的存放

资料的存放既要综合考虑软件系统和机器的特点,还要保证业务运行管理的简洁方便。显然,120 GB 的资料不可能存放在前端机的硬盘中,一是空间不够,二是读写速度慢。在神威机上,配置了 16 个大容量的外部存储器(简称 IOP),每个磁盘的存储量为 80 GB。因此,系统把除产品以外的所有资料都存放在大容量的 IOP1 至 IOP16 中。同时,资料存放时,充分考虑到集合预报多样本运算的特点,将前处理、客观分析、6 h 预报、集合预报初始场生成等与多样本处理无关的各部分运算所需的资料存放在一个 IOP 上,各子系统固定不变的输入资料集中存放在一个目录下,并在前端机硬盘中留有备份,以利于 IOP 故障时资料的转移。当集合预报初始场生成子系统生成 32 个样本的初始资料后,这些资料被均匀的散存于 16 个 IOP 上,即模式预报和模式后处理子系统运行时,每个 IOP 上都将存放与相应两个样本有关的输入输出及工作文件资料。这样多样本运算时各 IOP 上的读写量负担均衡,既充分地利用了机器的资源,又不至于造成输入输出的拥堵,有效的增强了并行效率。为方便与气象局在用数据库的连接,前处理的输入资料和产品生成的输出资料直接存放于前端机硬盘中,每天通过网络定时自动存取。

2.3 文件资料的输入输出管理

集合预报系统中资料的数量多、规模大,文件格式也不尽相同,这些特点使得在文件资料的输入输出管理方面应解决三个问题:一是输入输出方式的选取;二是资料的输入输出路径管理;三是对有特殊输入输出需求的子系统的特殊处理。

神威机提供了四种文件输入输出方式供用户使用:普通输入输出方式、快速输入输出方式、并行输入输出方式和大块文件输入输出方式。通过分析各子系统 I/O 的特点,测试机器所提供的几种 I/O 方式,综合考虑各子系统的特点,在集合预报系统中采用了快速 I/O 方式。

为增强系统的灵活性,将原先各子系统中输入输出文件路径在程序中明确指定的方式修改为从文件中灵活读取的方式,并编写了一套自动根据机器所提供的资源状况生成路径文件的程序。集合预报系统启动时,先用这套程序按照一定的顺序检查机器资源状况,然后生成一套存放各子系统中用到的输入输出文件路径的专用文件,各子系统运行时在相应的文件中取出所需的路径。这样系统中所有 I/O 路径可以很方便的根据机器环境、资源可用状况进行调整修改,某些磁盘发生故障时,系统自动转移运算环境而不影响系统的正常运行,而且 I/O 路径的改变与各子系统程序无关,从而增强了系统的容错可靠性,给子系统间的连接和多样本运行带来极大的灵活性,也给运行管理带来了极大的便利。

并行化产品生成子系统中温度与高度场产品生成部分使用了 10 个处理器,每个处理

器处理生成一天的产品,所以每个处理器都要访问所有 16 个 IOP 上 32 个样本相应时次的后处理资料,如果每个处理器都按常规顺序依次访问各个 IOP,势必会造成输入输出拥堵。为减少拥堵的产生,我们采用了不同处理器以不同 IOP 为起点环绕访问的优化算法,即 1 号处理器从 IOP1 起开始访问,依序至 IOP16 结束,2 号处理器则从 IOP2 开始,依序到 IOP16 后再环绕到 IOP1 结束,依次类推。从而使得每个处理器对每个 IOP 的访问时间完全错开,从真正意义上最大限度的发挥了并行的效用。表 1 给出了采用这种优化算法前后的测试结果,从中可以看出,优化后这部分程序的运行时间缩短了 1276 s,效果十分明显。

表 1　优化前后温度与高度场产品生成部分效率比较

系统名称	使用 PE 数	优化前用时 (s)	优化后用时 (s)
温度与高度场	10	1734	458

综上所述,对于像集合数值天气预报这样含有大量文件资料处理的系统,有效的资料组织不仅必要,而且十分有意义。它可以简化日常的业务运行管理,减轻业务管理人员和软件系统维护人员的负担,提高系统运行的灵活性,进一步增强整个系统的容错性,还可以使并行系统的运行效用得以最好的发挥。

参 考 文 献

1　国家气象中心编译. 资料同化和中期数值预报. 北京:气象出版社,1991. 96 ~ 130.

2　国家并行计算机工程技术研究中心. 操作系统用户指南. 神威计算机系统技术资料(7).1999. 24 ~ 35.

MASS DATA MANAGEMENT IN SW ENSEMBLE NUMERICAL WEATHER PREDICTION SYSTEM

Zhang Yi

(*North Computation Center, Beijing* 100091)

Abstract

The implementation of mass data management in SW ensemble numerical weather prediction system in respect of data file naming, data file access and input/output management is described.

Key words: Ensemble numerical weather forecast　Files and data management　Cycle access