异构平台上数值天气预报系统业务化运行环境*

田 浩 张建春
(国家气象中心,北京 100081)

提 要

该文阐述了在CRAY/VAX分布式系统和异构网络平台上建立数值天气预报系统业务化运行环境过程中遇到的困难和解决的方法,并指出目前业务运行环境中仍存在的一些问题.

关键词:数值天气预报 运行环境 数据传输 作业控制

引 言

国家气象中心于1994年秋安装了CRAY巨型机.为了尽早投入数值天气预报业务运行,首先要解决CRAY C92巨型机通过CRAY EL98与VAX6320之间的数据传输问题,以便获得数据源及分发产品.这需要解决连接EL98和VAX的以太网速度(10Mb/s)能否满足数值预报业务系统需求的问题.

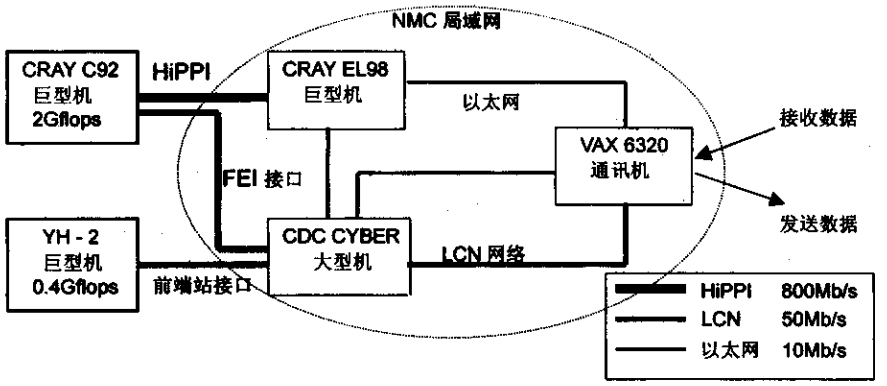

图1 NMC数值天气预报业务相关网络示意图

图1为国家气象中心(NMC)数值天气预报业务相关网络示意图.从图1可以看出,T63、台风、暴雨等数值预报系统要在C92-EL98-VAX平台上实施业务化运行,还必需解决分布于不同计算机系统预报程序有关模块间的数据传输方式、作业传输、异机间作业的

* 国家气象中心ZXYWSZ-96-07课题资助.
1997-05-14收到,1997-12-02收到再改稿

相互启动和监控手段等技术问题，即在连接 C92、EL98 间的 HiPPI 网(800Mb/s)和连接 EL98、VAX 间的以太网上要解决作业和数据如何传输、数值预报系统在 3 台机器上的各部分如何衔接、控制以及监控程序如何设计，所采用的技术方法和实现技巧等，以实现在异构平台上建立数值天气预报系统业务化运行环境.

1 以太网用于数值天气预报业务的可行性

在尽可能考虑到诸多影响传输速度的情况下，在控制传输(TCP/IP)协议下用网络应用程序 ftp 测试以太网用于业务化数据传输的实际平均传输速度. 重点测试大数据量文件的传输时效和超时(time out)错误的出现频率及时段. 测试考虑到了不同的时间段(计算机系统和网络的不同负荷条件下)，不同文件大小，不同格式的数据文件(二进制文件和文本文件)，文件所在的不同磁盘(磁盘 I/O 的繁忙程度)，不同方式(分前、后台或批方式)和不同的传输方向(CRAY EL98→VAX6320，VAX6320→CRAY EL98). 经不同测试重点的大量传输试验，测试结果虽然达不到 10Mb/s 的理想速度，但实际平均速度仍能够接受. 传输的稳定可靠性方面，除白天极少数情况下出现大数据量文件传输时的 time out 错误外，其他情况均正常. 考虑到 T63 程序大量数据的传输是在北京时间的后半夜进行，网络的资源竞争和碰撞相对白天要少得多. 根据 CRAY 和 VAX 间传输的数据量、C92 上模式运行时间和产品发送时间综合考虑，认为以太网的传输速度能基本满足预报的时效要求.

2 批处理环境下文件传输手段的实现

ftp 应用程序的命令使用和参考手册仅提供了该命令的交互式使用方法. 在这种方式下的数据文件传输只能是交互式的，即在用户投入 ftp 命令后，对方主机的 ftp server 程序会提示你输入帐号和口令，经过合法性验证后，进一步提示你可键入 ftp 程序的子命令，如 put 或 get 命令. 虽然你可通过在主目录下建立. netrc 文件，描述远程系统的合法认证而进行自动注册或登录(login)，但你仍必须键入 put 或 get 命令进行文件的传输. 因此预报系统的业务流程靠这种人机交互式的数据传输方法显然是不现实的，除非运行过程中有故障出现而必须采用这种方法.

为了能在批处理(batch)环境中使用 ftp 应用程序，反复研究了 ftp 命令的各种连接方法(connect)和所有参数的各种具体用法[1]，最终在 Korn 和 POSIX shell 中找到了方法[2]. 就是利用 shell 中的 here document 功能，将 ftp 命令的输入方式从标准输入改为紧随其后的即时文件输入方式，从而解决了这一问题. 使交互式的 ftp 应用满足了数值预报模式业务运行所需的实时性及非人工干预要求. 批处理方式的文件传输方法为：

```
ftp -nv vax632<<@>output.msg 2>&1
    user vax 用户名  口令
    put(get) cray_ file . /vax632/vax_ file
quit
@
```

这里 ftp 命令的-n 参数为限制 ftp 命令自动 login,-v 参数为报告远程 ftp server 的所有响应信息及数据传输的统计信息.@为即时文件的结束符. output. msg 为接收-v 参数产生的信息和 ftp 命令的错误信息之文件. ftp 的 user 命令要求用户提供 VAX632 上的合法用户名和口令,中间用空格分开.

如把上面这些命令存放在一个文件(shell script) ftpjob 中,可将该作业卡以批处理方式提交:

```
qsub ftpjob
```

或以后台方式提交:

```
chmod u+x ftpjob
nohup . /ftpjob &
```

且传输的文件不会因为传输过程中的会话退出(exit)或关机下电而丢失或不完整.

还有另一种非交互式的 ftp 实用方法:

```
ftp -vn vax632 <ftp_ cmd. script> output. msg 2>&1
```

ftp_ cmd. srcipt 内容为:

```
vax632 用户名　口令
put(get) cray_ file . /vax632/vax_ file
quit
```

该问题的解决为随后的工作奠定了基础,是后续一系列研究和试验工作的先决条件. 本方法也可为 UNIX 工作站用户非交互使用 ftp 命令时所借鉴.

3　使用并发技术提高产品时效

这项研究内容的出发点是假设以太网的传输速度不能满足数值天气预报业务的时效,即使能基本满足当时的时效要求,但如果对模式进行了改进(如增加模式的分辨率和精度)或升级新模式(如 T106),势必将大大增加网络的数据传输量,如果等实时业务运行后再来解决该问题未免被动而且影响很大.

数值天气预报模式每做完一天的预报即产生一次输出,与其等到产生 7 天或 10 天的预报结果一起传输到 VAX 机上,不如每产生一天的预报结果随即就发送到 VAX 上. 这样就需要模式具有内部控制功能,即每做完一天的预报,该预报作业本身要能启动另一批作业来完成一天预报结果的传输. 预报作业在传输作业发送一天的预报数据的同时继续执行,也就是预报作业和传输作业是并行执行的,否则对 VAX 来讲总体时效是一样的,起不到节省时间的作用. 该方案是完全可行的,因为预报程序本身的并行程度不是特别高,而 NMC 的 CRAY 机至少有两个 CPU,有运行数据传输作业的 CPU 空闲;数值预报模式(T63)本身又有数据分层(天)输出的特点.

经过反复试验,发现 CRAY 的批处理软件 NQS(Network Queuing System)允许用户的作业卡(shell script)中出现多个 qsub 命令,每个 qsub 命令提交一个批处理作业. 可以把作业卡中的每个 qsub 命令看作 IBM MVS 系统作业卡中的一个作业步,但与其不同的是,NQS 中每个 qsub 提交的作业是并行执行的,只要 NQS 系统允许你同时运行多个作业(当时就遇到了这个困难,因为作为一般用户,CRAY 的每个批队列(batch queue)只允许用户提交一个作业,而在所有的批队列中用户同时只能运行两个作业,所以必须保证

并行的两个作业运行在不同的批队列中，否则提交的并行作业只会串行运行）. 这对用户而言当然是方便的，但在数值预报业务的作业卡中却无法确定模式何时产生了 1 天的预报结果，然后并行执行一个 ftp 批作业，把一天的预报结果传输到 VAX 系统上. 问题的关键是如何在模式的 FORTRAN 程序中提交一个 ftp 批作业与模式并行执行以提高数据的传输时效.

在对 NQS 作业的运行机制[1]、shell 命令的执行过程（fork 和 exec 系统调用）以及一些 CRAY 的 FORTRAN 库过程进行深入研究后，编写了一个库子程序 libpgm. a，放在 CRAY 的/home/u3/tianh/sa 下. 用户也可以用 bld 命令把相同目录下的 libpgm. o 目标模块放入自己的目标库中. 具体用法为：

在执行 FORTRAN 程序的作业卡中，指定：

cf77 -L /home/u3/tianh/sa -l pgm Myprogram. f

如果用户把 libpgm. a 放入到了自己的子程序库中且重新命了名，则-L 参数为存放该子程序的库的全路径名，-l 为该库子程序的新文件名. 其中 Myprogram. f 为要编译运行的 FORTRAN 程序名.

在 FORTRAN 程序内部的任何位置，如 write 语句后，指定：

CALL CONCURRENT('qsub My_ jobrequest')

引号中的 qsub 命令的句法和用法同以前完全一样，My_ jobrequest 是存放并行作业卡的文件名. CONCURRENT 为库子程序的入口名.

用此功能模块可供 T106 等数值预报模式程序在运行的不同时段，内部分批以 ftp 批作业的形式把模式产生的逐日数据产品和结果传输到 VAX 机上，解决网络传输速度给预报系统的业务化造成的潜在时效问题.

原来获得第 1 天与第 7 天的预报结果时间相差不多，即第 1 天的预报结果也要等模式 7 天预报全作完后才能获得. 使用并发技术后，只要模式作完 1 天的预报，即可输出进行后处理获得该天的产品. 由于 CRAY 是多 CPU 主机，所以在后处理该天预报结果的同时，模式照常进行下一天的运算.

设模式从启动运行到 7 天预报生成产品的时间为 n，每天的预报提前时效为 T，i 为预报天数，则 $T_i=n(7-i)/7 \quad (i=1,2,\cdots,7)$

该功能模块的研制还使“模块间的相互控制”（本文第 6 部分）成为可能.

4 传输数据的一致性

网络传输的时效问题和数据传输的批处理方式解决之后，就面临着传输的数据是否能正确使用的问题. 从 NMC 数值预报业务的主线来看，问题集中在 EL98 和 VAX 之间. 由于 CRAY 和 VAX 具有不同的硬件结构及不同的操作系统，数据在机器中的表达方式和存取方法根本不同. 原则上要屏蔽 CRAY/UNICOS 和 VAX/VMS 给数据传输带来的变异和对数据的不同解释，可用国际标准高级语言的有格式数据类型；否则就要象 UNICOS 一样，在其高级 I/O 功能里提供与 DEC、IBM、CDC 等各种主机间的数据进行格式转换的专用软件包，保证数据的格式在应用水平上的一致[3,4].

NMC 数值预报模式的开发语言主要为 FORTRAN,因此在数据传输前利用发送方和接收方事先约定 F 格式的数据传输方法,保证传输和接收的数据在 FORTRAN 应用程序水平上的一致.经反复试验,对于 CRAY EL98 和 VAX 6320 间传输的有格式数据,得出重复因子为 200 的 F13.6 格式数据传输效率高,实践证明也是行之有效的.但在 FORTRAN 的 OPEN 语句中要加入 RECL 参数:

OPEN (UNIT=..,FILE=..,STATUS="..",RECL=2600)

5　异机间作业互启和分布环境下系统控制的方法

NMC 初期建立的数值预报系统在 CYBER 和 VAX 平台上运行,CYBER 和 VAX 由 CDC 的高速局域网 LCN 连接,该网的高层应用程序支持文件(MFLINK)和作业(MFQUEUE)两种方式的传输.MFQUEUE 命令提供的在本地主机上直接提交批处理作业到远程主机上执行的功能,是模式在当时的分布式系统环境下的业务化运行基础.新的业务系统方案结构上虽然与以前相似,但缺少异机间直接提交批处理作业的业务化运行基础,即 TCP/IP 不支持网络的作业传输.

为此研究了 TCP/IP 的另一个高层应用程序 telnet,套用 ftp 命令的解决方法,使 telnet 命令能够在 NQS 批处理环境中使用,以便直接把作业提交到远程主机上运行.又考虑在 VAX/VMS 上安装 RQS (Remote queuing system),该远程队列环境可在本地系统上把作业直接提交到远程的 CRAY 主机上运行,但 RQS 只能单方面解决问题,问题的另一面 CRAY 向 VAX 直接提交作业还是不能解决.甚至考虑了利用 LCN 网的方法,都不能彻底解决这一问题.最终采取的是消息树方法,即用 ftp 批作业方式先把消息文件发送到对方主机,通知对方主机要马上执行一个作业;随后把需要对方主机执行的作业(卡)作为数据文件(照顾用户过去使用 MFQUEUE 的习惯),用 ftp 批作业发送到对方主机;当然需要远程执行的此作业卡本身也可放在对方主机的磁盘上,这要视具体情况而定.最后分别在 CRAY EL98 和 VAX6320 上建立监控程序,定时监视是否有消息文件到达.如有,对接收的作业文件进行可用性检查后,提交后续有关作业.

CRAY EL98 和 C92 间作业的提交不需用此方法,UNICOS 的 NQS 可以直接把批作业通过 HiPPI 网从一 CRAY 系统传输到另一 CRAY 系统上.具体方法是在构成 NQS 时,把一 CRAY 系统的 batch queues 定义成另一 CRAY 系统上 pipe queue 的目的队列:

- 在 EL98 和 C92 的 NQS 构成文件中分别定义网络管道队列(pipe queue)sn4207 和 sn5425.这两个管道队列自动会把一 CRAY 系统的作业传输到另一 CRAY 系统上运行.当作业运行结束,作业的输出结果、作业 log 及错误信息会返回到原始主机.
- 如想在提交作业时自动进行口令认证,省掉每次提交作业时键入另一系统上用户名和口令的麻烦,需在你的主目录下建立.nqshosts 或.rhosts 文件,当.nqshosts 文件不存在时,NQS 检查.rhosts.用户在提交作业时仅需指明所用的网络管道队列即可.

结合这两种方法,解决了该业务系统的作业调度和异机间作业的相互控制问题.

6 模式间的相互控制

台风和暴雨是破坏力巨大的灾害性天气，因此 T63 同台风和暴雨模式之间的关系应该是，如果有台风和暴雨等灾害性天气情况出现，那么 T63 在运行到能满足台风或暴雨模式的边界条件时，需要立刻启动这两个预报业务. 换句话说如果有台风或暴雨报文到来，台风或暴雨预报的运行优先级别应高于 T63 模式.

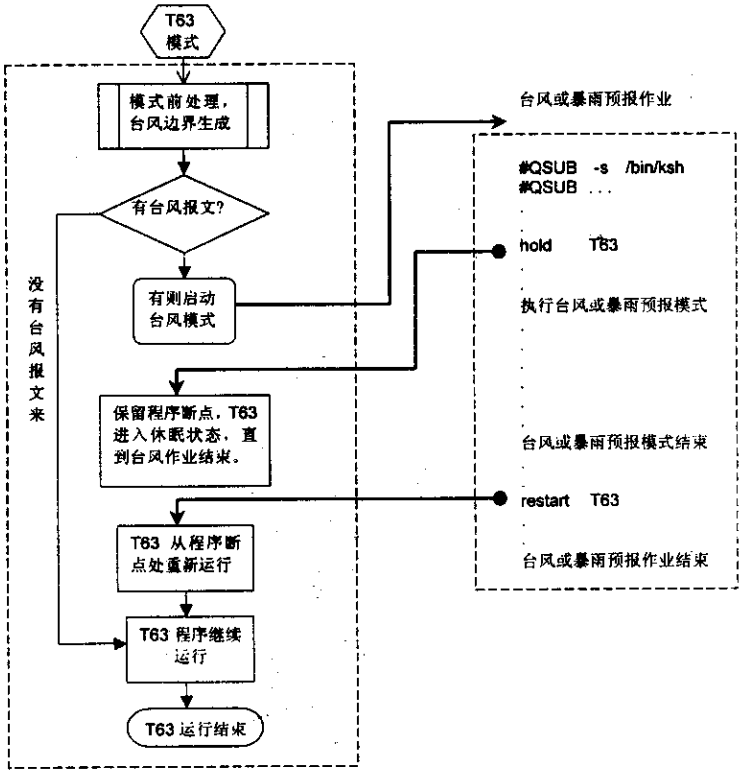

图 2 模式间运行关系示意图

这种实际上的需求给这些数值天气预报模式提出了相互控制和事件驱动的作业调度要求. CRAY 机系统上普通用户是没有特权进行有关作业挂起(hold)和释放(restart)等调度工作的. 结合 CRAY NQS 的构成和管理，作业的并发，系统功能调用和 awk 实用程序等方面的内容进行研究，现已为一般用户(特别是 NWP 业务用户)开发了在(模式的)批作业运行中启动、挂起和释放另一(预报模式)批处理作业的功能机制，使模式间可以传递作业的运行情况和动态地相互控制，实现模式间事件驱动的自动作业调度，提高模式的运行效率. 以图 2 模式间运行关系示意图为例，当需要作台风预报时，T63 应在台风模式启动所需的边界条件满足时启动台风批作业，同时挂起 T63 作业本身以全力保证台风模式的优先运行；当台风作业结束时要能释放 T63 作业，让其继续执行.

7 存在的问题和展望

用消息树手段进行作业互启可能是一隐患,当时是不得已而为之,因为中间环节多可能导致系统稳定性下降.即使是今后VAX6000系列机升级为Alpha而且又安装了RQS,也只能解决VAX到CRAY的单向作业传输问题,CRAY到VAX的作业传递问题仍需进一进研究.

在数值预报业务系统中起重要作用的VAX6320在Alpha机到货之前不会与CRAY C92一起连在FDDI环上,它通过以太网与FDDI环上的一个交换机连接.如果T63升级为T106的话,VAX到交换机这段以太网的速度不知是否能满足时效的要求,但可用本文第3部分的方法作些改进.有关提高T106运行效率的问题,有关人员可用第6部分的方法进行试验.

致谢:对课题组的施培量、季京英、陈建军、石曙卫、杨学胜等同志的工作以及解决了无格式数据传输一致性问题的NMC数控室的张德新同志和有关领导及单位的热心帮助支持,表示诚挚的感谢.

参考文献

1 Cray Research,Inc. UNICOS System Administration. Volume 1~4 Cray publication SG-2113 8.0/9.0,1994.

2 Bill Rosenblatt. Learning the Korn Shell. O'Reilly & Associates,Inc. 1990.

3 Cray Research,Inc. I/O User's Guide. Cray publication SG-3075 8.0,1993.

4 Cray Research,Inc. Advanecd I/O User's Guide. Cray publication SG-3076 8.0,1993.

OPERTIONAL ENVIRONMENT FOR NWP PRODUCT ON HETEROGENEOUS PLATFORM

Tian Hao Zhang Jianchun

(*National Meteorological Centre,Beijing* 100081)

Abstract

The problems and technique methods of solving them during setting up a production operational environment for NWP product on heterogeneous CRAY/VAX computing platform are described and some problems still existing in the present production operational environment are pointed out.

Key words: Numerical Weather Prediction (NWP) Operational environment Data transmission Job control