一种适用于有限差分模式的负载平衡区域分解方法*

金之雁 王鼎兴

(中国气象科学研究院, 北京, 100081) (清华大学计算机系, 北京, 100084)

摘 要

分布式内存并行处理在数值天气预报等超大规模科学计算中已经得到了广泛的应用。中尺度模式由于分辨率高, 计算量大, 需使用更多的处理机进行并行运算。另一方面, 由于复杂的物理过程的采用, 增加了不同天气的计算量的不平衡。但是, 目前所广泛使用的并行处理方法在处理机数量较多时不能很好地均衡计算负载, 引起并行计算效率的降低。本文提出了一种新的非规则区域分解负载分配方法。并与已有的负载分配方法进行了分析试验对比, 该方法能更有效地平衡负载, 取得更好的加速效果。

关键词: 并行计算, 负载平衡, 区域分解, 数值天气预报。

1 引 言

现代高性能计算机常采用分布式内存的并行多处理机体系结构, 由几百至几千个处理机构成, 达到每秒万亿次浮点操作的运算速度。但是, 这种计算机系统对应用程序的要求较高, 要求应用程序将计算负载均匀地分配到各个计算结点上进行计算。计算负载分配方法的好坏是影响这种计算机性能发挥的关键技术之一。参与计算的处理机数量越多, 负载平衡越困难。

在有限差分模式中, 并行处理方法是采用区域分解法, 即将离散化后的水平网格点分配到不同的处理机上进行计算。增加模式在每个处理机上的水平格点的数量可以提高模式并行计算的效率。但是在实际业务预报中, 这种方法是行不通的。数值预报分辨率提高 1 倍, 计算量增加 16 倍, 水平网格点数量增加 4 倍。如果运算时间不变, 需要增加处理机数量 16 倍, 每个处理机上的水平格点数量下降到原来的 1/4。因此, 高分辨率模式要求使用的处理结点越来越多, 但每个处理机点处理的水平格点数量却越来越少, 在这种情况下, 要求将计算负载十分均匀, 少量的不平衡也会产生严重影响。

在有限差分模式中被广泛使用的均匀区域分解方法(见图 1a) 在处理机数量较多时, 计算负载划分不均匀的现象愈来愈严重。中尺度模式常采用比较复杂的物理过程, 由于天气不同, 格点间计算量可以相差非常大, Edward James P[1] 发现在有微物理过程的模式中, 模式中的中尺度系统使模式计算速度下降 1/3。

负载平衡可以分为静态负载平衡和动态负载平衡, 静态负载平衡是实现动态负载平衡的基础。本文只讨论已知计算负载分布条件下的静态负载分配策略问题。

2 规则区域分解负载平衡方法

目前绝大多数有限差分数值预报模式采用的是一种基于规则网格点的计算负载分配方法[2~4], 该方法将处理机数量 P 分解为 $P = P_x \times P_y$, 将预报区域的 x, y 方向的格点分别等分近似为 P_x, P_y 组行和列。形成面积相等或近似相等的矩形(见图 1a, 简称方法 1)。它的特点是方便直观, 易于对以往大量程序进行并行化移植, 因而使用较普遍, 且在处理机数量较少时取得了比较好的效果。存在的问题是: 第一, 如果处理机数量与区域网格点数量不能很好地互相匹配时, 处理机间网格点数量会相差较大, 造成负载不均衡, 尤其是在处理机数量比较多时, 情

* 初稿时间: 2001 年 10 月 4 日; 修改稿时间: 2001 年 12 月 20 日。
资助课题: 国家自然科学基金项目(69933020)。

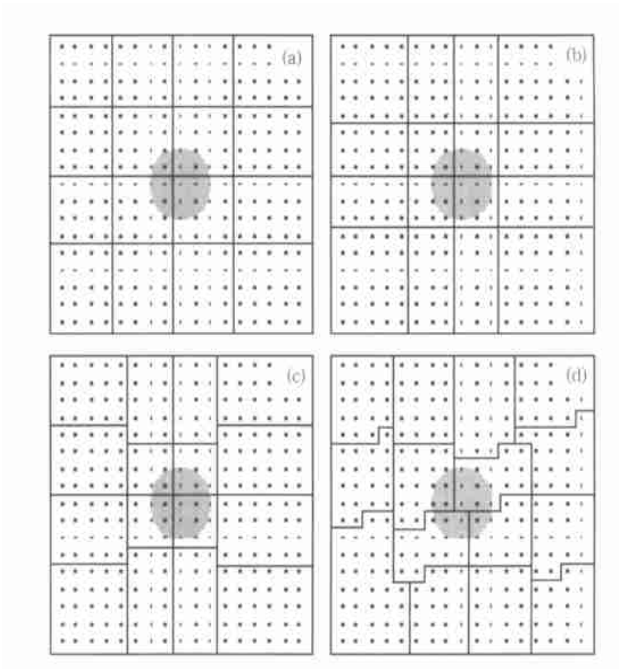

图 1　4 种区域分解方法示意图
(a, b, c, d 分别为方法 1, 2, 3, 4)

况更加严重; 第二, 这个方法只是根据网格点数量进行划分, 由于每个格点所处理的天气现象、地表状况不相同, 计算量相差很大, 也会造成结点间计算负载的不平衡, 影响并行计算效率。Hendrik Elbern[5] 提出了两种改进方法, 一是根据网格点负载的不同统一调整区域划分的行、列宽度(见图 1b, 简称方法 2); 二是统一调整区域划分的列宽度, 根据计算负载情况分别调整行宽度, 但仍然保持每一个分解后的子区域为矩形区域(见图 1c, 简称方法 3)。这两种方法都需要维持分解后的区域是矩形, 平衡负载能力受到限制, 尤其是在处理机数量较大时, 负载不平衡依然严重。本文提出一种非规则的区域分解方法(见图 1d, 简称方法 4), 能够将计算负载尽量均匀地分配到各个处理机上。这种方式不仅能够在处理机数量较多时依然能够达到负载平衡, 而且对每个格点计算量相差很大的非均匀负载问题也十分有效。

3　非规则区域分解负载平衡方法

定义预报区域 G 是矩形区域, 东西、南北方向的网格点数分别为 N_x, N_y。格点(i, j)的计算负载为 w_{ij}, 负载总和为 $W = \sum_{(i,j)\in G} w_{ij}$。参加运算的处理机数量为 P。如果每个处理机的计算负载为 $\overline{W} = \frac{W}{P}$, 则负载达到平衡。将处理机分成 N 个组, 第 k 组内有 N_k 个处理机:

$$N = \left[\sqrt{P \frac{N_x}{N_y}} \right] \tag{1}$$

$$N_k = \begin{cases} \left[\frac{P}{N}\right] + 1 & k \leqslant \mathrm{Mod}(P, N) \\ & k \in 1, N \\ \left[\frac{P}{N}\right] & k > \mathrm{Mod}(P, N) \end{cases} \tag{2}$$

其中, $\mathrm{Mod}(P, N)$ 是 P 除以 N 的余数, [] 表示舍去小数部分取整。

按以下方法将区域 G 划分成个 N 区域:

$k = 1$

do $j = N_y, 1, -1$

　　do 由 $i = 1, N_x$

将(i, j)点标为 k 计算 Σw_{ij}, 如果 $|\Sigma w_{ij} - \overline{W} \times \sum_{i=1}^{k} N_k|$ 达到最小值, $k = k + 1$

　　end do

end do

将 N 个区域的所有格点按照内循环 i 由小至大, 中循环 j 由小至大, 外循环是区域 k, 由 1 至 N 的顺序排列成一维数组, 设下标为 l, 按照以下方法将划分成 P 个子区域:

　　$s = 0$

　　$k = 1$

　　do　$l = 1, N_x \cdot N_y$

　　格点 l 标记为 k　　$s = s + W_l$

　　如果 $|s - \overline{W} \times k|$ 达到最小, $k = k + 1$

　　end do

　　设 G 内第 k 子区域的计算负载为 W_k, 显然有

$$|W_k - \overline{W}| < W_{\max}$$

其中, $W_{\max} = \max(w_{ij})$, $(i, j) \in G$, 所以对任意两个子区域有

$$|W_k - W'_k| < 2w_{\max} \tag{3}$$

由于该方法允许将行或列断开使得子区域的形状出现小台阶, 每一个台阶会使增加一个格点的通信量, 最多会增加 $P-1$ 个格点的通信量。所以, 通信的增加量是很少的。

已知 w_{ij}, 不计通信开销, 可给出某方法的并行加速比估算值 S:

$$S = \frac{\sum_{(i,j)\in G} w_{ij}}{\max_k \left(\sum_{(i,j)\in \text{子区域}\, k} w_{ij} \right)} \tag{4}$$

其中, max( ) 表示在 P 个处理机内取最大值。

式(4)可用来对分区方法的性能进行估计。

本方法根据式(3)得 $S = \dfrac{W}{\overline{W} + 2w_{\max}}$

如果 $\overline{W} \gg w_{\max}$, 则 $S \rightarrow P$。

数值预报计算主要集中在动力部分和物理过程部分。动力部分主要是计算偏微分方程中的非线性平流等绝热过程的影响, 物理过程部分主要是计算由于水汽的相变、地表与大气的相互作用、太阳辐射、长波辐射等对大气运动的影响。随着模式的细化, 物理过程的计算量越来越大, 可以达到动力部分的几倍到十几倍, 且分辨率越高, 所占比重越大。它也是负载不平衡的主要来源。为分析本方法平衡非均匀负载能力, 需要模拟物理过程。

我们分别在均匀负载和非均匀负载条件下计算。假定均匀负载: $w_{ij} = 1$ (5)

假定非均匀负载:

$$w_{ij} = \begin{cases} 1 & \sqrt{(i - i_c)^2 + (j - j_c)^2} > 10 \\ & (i, j) \in G \\ 10 & \sqrt{(i - i_c)^2 + (j - j_c)^2} \leqslant 10 \end{cases} \quad (6)$$

其中, $i_c = N_x/2$, $j_c = N_y/2$, 计算中设 $N_x = 101$, $N_y = 101$。它表示在 101 × 101 的区域中部, 有一个半径为 10 倍格距的区域计算量较大, 该区域计算量是周围地区的 10 倍。分别将式(5)和(6)所假定的负载情况按照各自的分区方法进行分区, 并计算各子区域的负载和, 根据式(4), 算出 S。

表 1 是在上述负载条件下不计通信开销时间各种方法算出的 S。在均匀负载情况下, 方法 1~ 3 是一样的。可以看出, 各种方法差别不大, 方法 4 比其它方法好一些; 在非均匀负载条件下, 各种方法有明显差别, 由低至高排列分别是方法 1, 2, 3, 4。可以看到, 方法 4 与均匀负载的情况相差不大, 表明这种方法有较强的负载平衡功能。方法 3 的情况稍差一些, 但是由于该方法是矩形的分区, 在对模式进行并行化改造时比较方便, 且具有一定的负载平衡功能。方法 2 效果较差。因此方法 1 适合于计算负载比较均匀的情况, 如分辨率较低、物理过程简单, 主要计算量为模式动力过程的模式。方法 3 比较适合于在对串行模式进行并行化改造, 同时又希望尽量减少模式改动量时使用。方法 4 在两种计算负载情况下, 较其它方法优越性明显, 是比较理想的方法。

表 1 各区域分解方法在均匀和非均匀负载下的 S

处理机数量	均匀计算负载				非均匀计算负载			
	方法 1	方法 2	方法 3	方法 4	方法 1	方法 2	方法 3	方法 4
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.98	1.98	1.98	2.00	1.95	1.95	1.95	1.99
4	3.92	3.92	3.92	4.00	3.82	3.82	3.82	3.99
8	7.69	7.69	7.69	7.99	6.14	7.40	7.70	7.98
16	15.09	15.09	15.09	15.99	8.80	8.07	13.78	15.90
32	30.18	30.18	30.18	31.98	10.30	19.52	28.12	31.61
64	60.36	60.36	60.36	63.76	18.58	35.15	50.80	62.3

4 试验研究及结果分析

数值预报程序量庞大, 直接试验困难比较大。因此, 我们利用一个简化的模型, 对该方法进行试验。二阶线性扩散方程虽只是数值预报模式的一小部分, 但是它与模式动力过程的计算特点是一样的。因此, 用它对模式动力过程进行模拟可以说明问题:

$$\frac{\partial F}{\partial t} = K_x \frac{\partial^2 F}{\partial x^2} + K_y \frac{\partial^2 F}{\partial y^2} + K_z \frac{\partial^2 F}{\partial z^2} \quad (7)$$

离散化形式为:

$$\frac{F_{i,j,k}^{t+\triangle t} - F_{i,j,k}^{t}}{\triangle t} =$$

$$K_x \frac{F_{i+1,j,k}^{t} - 2F_{i,j,k}^{t} + F_{i-1,j,k}^{t}}{\triangle x^2} +$$

$$K_y \frac{F_{i,j+1,k}^{t} - 2F_{i,j,k}^{t} + F_{i,j-1,k}^{t}}{\triangle y^2} +$$

$$K_z \frac{F_{i,j,k+1}^{t} - 2F_{i,j,k}^{t} + F_{i,j,k-1}^{t}}{\triangle z^2} \quad (8)$$

其中 F 是扩散变量, K_x, K_y, K_z 分别是 x, y, z 三方向上的扩散系数, $\triangle t$ 和 $\triangle x$, $\triangle y$, $\triangle z$ 是时间步长和 3 个方向上的格距。离散化后的网格点是 (N_x, N_y, N_z) 阶三维矩阵, 边界为固定边界条件。从式(8)可以看出, 在子区域上计算 $F_{i,j,k}^{t+\triangle t}$ 除需要本子区域上的数据外, 还需要与它相邻的网格点的 $F_{i,j,k}^{t}$, 这部分数据需要从其它处理机上用通信的方式获得。在实际实验中, 通信开销是不可忽略的。

该方程除边界外每个格点的计算量都是相同的, 边界点上的计算量为零。我们将每个计算结点上的数据用一维数组存放(图 2) 。

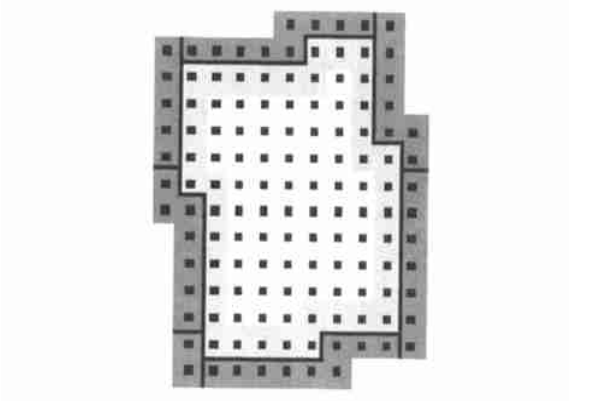

图 2　处理机内部数据排列示意图

(重阴影区是外边界区, 轻阴影是内边界区, 无阴影区是内区)

其中重阴影区是与周围处理机数据的重叠部分, 我们称为外边界区, 轻阴影区的计算需要外边界区的数据, 称为内边界区, 无阴影区称为内区。它的优点是增加了向量长度, 有利于向量运算, 内区的部分的计算所需的数据已经在该处理机上, 不需要通信。内边界区的计算必须在外边界区的数据从其它处理机上获取以后才能进行。

计算过程如为时间积分循环开始; 外边界区数据交换开始; 内区计算; 等待区数据交换结束; 内边界区计算; 时间积分循环结束。

在一些计算机上, 上述安排可以使通信与计算重叠进行, 隐藏通信开销。但是有些计算机做不到这一点。物理过程是负载不平衡的主要来源, 为考察各方法的平衡负载能力, 模拟物理过程部分是必须的。我们采用在每个格点(i, j)上增加计算三角函数的办法进行模拟, 使区域中心格点的计算时间为周围格点的 10 倍(见式 6) 。在上述试验中, 我们设定 $N_x = N_y = 101$, $N_z = 100$。

实验研究在 IBM 的 SP 计算机上进行, 它总共有 10 个计算结点, 每个结点由 8 个 CPU, 在结点内部各 CPU 之间共享内存, 结点间用高性能交换开关相连。在试验中, 我们没有使用结点内部共享内存的特性, 每一个 CPU 都作为单独的结点使用。

表 2 是用二阶扩散方程进行实际测试得到的结果, 对二阶扩散方程而言, 各种方法相差不大, 与前面的理论分析结果一致。在增加了模拟物理过程的计算之后, 各种方法效果相差明显。方法 1, 2 的加速比下降明显, 这显然是由于负载不平衡引起的。一般来讲, 增加计算量可以提高计算与通信的比率, 可以提高价速比, 因此方法 3, 4 在增加了模拟物理过程的计算之后, 加速比上升, 证明可以较好地平衡不均匀负载。由表 1、表 2 的对比可以看出, 除了由于通信开销, 试验加速比稍低外, 其余均与理论分析相吻合。方法 4 在所有情况下均优于其它方法。

表 2　4 种负载分配方法在 SP 计算机上的加速化

处理机数量	线性扩散方程				线性扩散方程+ 模拟物理过程			
	方法 1	方法 2	方法 3	方法 4	方法 1	方法 2	方法 3	方法 4
1	1. 00	1. 00	1. 00	1. 00	1. 00	1. 00	1. 00	1. 00
2	1. 98	1. 98	1. 97	1. 99	1. 96	1. 96	1. 96	1. 99
4	3. 88	3. 90	3. 88	3. 97	3. 82	3. 81	3. 83	3. 98
8	7. 51	7. 60	7. 68	7. 52	6. 10	7. 30	7. 79	7. 70
16	14. 39	14. 59	15. 67	14. 75	8. 60	10. 86	13. 40	15. 00
32	25. 50	26. 15	26. 73	26. 96	9. 97	18. 60	26. 90	29. 35
64	42. 18	44. 00	43. 73	45. 00	10. 40	33. 00	47. 10	55. 70

5　结　论

通过理论分析和试验结果可以得出: 方法 1 只适合于在均匀负载情况下使用。如低分辨率、简单物理过程模式, 这种模式的主要计算集中在模式的动力框架, 基本属于均匀计算负载情况。方法 3 实现起来比方法 4 简单, 比较适合于移植复杂模式时采用, 可以减少模式改动量, 也有一定的平衡负载的能力。但在结点数量较多时比方法 4 效果差。方法 4 是一种比较理想的方法, 具有很强的均衡负载的能力, 尤其是在处理机数量较多的情况下有较高的加速比, 但是它实现起来比较复杂。如果不考虑工作量, 则应当采用该方法。方法 2 实现的复杂程度与方法 3 相当, 效果却不如方法 3, 基本不宜采用。

参考文献

1 Edwards James P, Snook John S, Zaphiris Christidis. Use of a parallel sesoscale model in supportof the 1996 summer olympic games. In: Geerd-R Hoffmann, Norbert Kreitz, eds. Proceedings of the Seventh ECMWF Workshop on the Use of Parallel Processors in Meteorology. World Scientific Publishing Co Pte Ltd, Singapre. 912805. 1996. 342~ 353

2 Burton P, Dickinson A. Parallelising the unified model for the Cray T3E. In: Geerd-R Hoffmann, Norbert Kreitz, eds. Proceedings of the Seventh ECMWF Workshop on the Use of Parallel Processors in Meteorology. World Scientific Publishing Co Pte Ltd, Singapre. 912805. 1996. 68~ 82

3 Zhiyan J, Chritidis Z. Paralle implementation of YH limited area model on SP2. In: Geerd-R Hoffmann, Norbert Kreitz, eds. Proceedings of the Seventh ECMWF Workshop on the Use of Parallel Processors in Meteorology. World Scientific Publishing Co Pte Ltd, Singapre. 912805. 1996. 473~ 477

4 Rodrigues B. A library for the porttable parallelization of operational weather forecast models. In: Geerd-R Hoffmann, Norbert Kreitz, eds. Proceedings of the Sixth ECMWF Workshop on the Use of Parallel Processors in Meteorology. World Scientific Publishing Co Pte Ltd, Singapre. 912805. 1994. 148~ 161

5 Hendrik Elbern. Load balancing of a comprehensive air quality model. In: Geerd-R Hoffmann, Norbert Kreitz, eds. Proceedings of the Seventh ECMWF Workshop on the Use of Parallel Processors in Meteorology. World Scientific Publishing Co Pte Ltd, Singapre. 912805. 1996. 429~ 444

A LOAD BALANCING DOMAIN DECOMPOSITION METHOD FOR FINITE DIFFERENCE NUMERICAL WEATHER PREDICTION MODELS

Jin Zhiyan

(Chinese Academy of Meteorological Sciences, Beijing 100081*)*

Wang Dingxing

(Tsinghua University, Beijing 100084*)*

Abstract

Hundreds to thousands of nodes, which can reach T flops, compose modern parallel computers. But programming on such kind of system is difficult. A very important issue is load balancing. The more the nodes in the system, the more difficult to balance the load.

Domain decomposition is common technique in parallel processing of mesoscale weather prediction models. The different columns of the model are distributed on different nodes. One can expect to increase the speedup the model by increase the resolution of model. However, as the resolution of the model is increased, the grids of the model and the steps of iteration are increase. More nodes are needed if we want the model can be finished in the same periods of time. As results, less columns running on each node. A little of unbalance of the load can be a serious problem on highly paralleled models. At the same time, the physical process of higher resolution model can be more complex, which results more unbalance among processors. Many models use regular east-west north-south domain decomposition technique use n by m nodes, ignoring the load balancing problem completely, The advantage is its simple and the communication between processors is low. It is success if the grids and the number of the nodes is highly compatible and the physics is not very complex. However, when the grid points of the model is not highly compatible with the number of the nodes, which is often the case in very dynamic environments. For example, one processor has one role and column of grid points than others, the processor with more grids slow down very other processors as though the load on each grid point is the same, and it can be more serious if the load of each grid points is very different becaus the physics of the model is very different under different weather, or on different land surface. Some researches show that the speedup of the model goes down rapidly after the model runs several hours when the microphysics is turn on. The solution is using adjustable do-

main to catch up the variation of the load. Some researchers use adjustable rectangle domain, which is better than the fixed domain. But our results shows a nearly rectangle domain, adding a few steps on some sides of the rectangle domain, can balance the load more better than the rectangle domain with only a little increase of communication due to the steps.

The result shows the algorithm to partition the grid points of the forecast area and compare it with three other methods, one is fixed rectangles domain method and others are two adjustable rectangles domain method. In order to analysis the effects of the method three other method was also tested. In the experiment, a distribution of the computation load for each grid point is given, four methods were applied and the load unbalance can be calculated for each method, which shows that our method is the best one in balancing load.

This method is applied to a three-dimension diffusion equation model to test the feasibility in and real model. The experiment was on IBM SP machine. In order to test its ability to balance the load with sharp different of computation between grid points, a simulated physics process was added in the model. The computation time was measured for every grid point and it was used as the load of the grid points. The result is identical with previous one with only the difference that the speedup is lower due to the communication.

Key words: Distributed parallel computing, Load balancing, Domain decomposition, Numerical weather prediction.

第十二届全国热带气旋科学讨论会在宁波召开

在系列性的“全国热带气旋科学讨论会”创立 30 周年之际,第十二届全国热带气旋科学讨论会于 2002 年 4 月 9—12 日在浙江省宁波市召开。中国气象学会副理事长、天气与极地气象学委员会主任陈联寿院士主持了会议。原中国气象局局长温克刚到会指导并作了重要讲话,从五、六十年代就开始从事台风科研和预报工作、并对中国台风科研、业务、教学和管理工作作出了重要贡献的祝启桓等 6 名特邀代表和来自全国 18 个省、自治区(直辖市)有关气象科研、业务部门、大专院校、军队、民航气象部门共 43 个单位近 120 名代表出席了会议。

与会代表就热带气旋的登陆、热带气旋结构和强度变化、热带气旋的形成和短期气候预测、路径和风雨影响等问题进行了广泛的交流,反映了近几年来台风科研和业务预报工作获得的进展和创新:

(1) 探测手段更新。多普勒雷达,高分辨数字化卫星,加密的 AWS 等获取了新的探测资料,有助于中小尺度系统的识别和初始场的优化。

(2) 中尺度数值模式的使用单位扩展(已达地市级台站),用于热带气旋研究的数值模式的类型增多,时、空分辨率加大,几种新的 Bogus 方案提出,同化技术取得进展。研究对象更具有中尺度特征。

(3) 在更广的时空尺度谱的范围认识台风环流演变。从南极冰盖,ENSO 到南亚高压,直至中β尺度云团甚至小尺度龙卷,都可能与台风有联系。同时,把台风视为一个具有多重复杂精细结构的系统。对涡旋 Rossby 波的动力学本质予以澄清。

从事台风科研、业务、管理、教学的青年科技专家是中国台风界的未来和希望,因此,本次会议还组织了由青年科学家主持的两场青年论坛,主题分别为:(1) 如何将台风科研推向世界先进水平,(2) 如何提高台风预报准确率。青年代表们就主题问题展开了热烈的讨论,一些资深专家也参与了交流。

与会代表还对未来热带气旋的研究提出了一些应着重解决的科学问题:

(1) 研究台风登陆过程中结构和强度变化、路径偏折、风雨强度和分布特征的机理和预报技术,提高对登陆台风的预报能力。

(2) 应用卫星等非常规资料,改进热带气旋初始场的质量,提高对精细结构的描述能力;研究 Bogus 技术;针对精细结构特征完善模式物理过程和计算方案,提高模式分辨率。

(3) 加强理论研究。如强风条件下边界层特征的理论;台风精细结构理论;四维变分同化理论等。理论研究应与模式改进和预测实际紧密结合。

(4) 完善现行统计预报模式和动力统计释用方案,完善台风预报误差评价方法,加强台风各种预报方法的性能研究。尤其是加强对台风路径突变、近海台风突然加强和衰减以及登陆台风风雨强度和分布预报等的研究。

(刘宗秀)