Profiling of the operational and experimental code of the National Air Quality Forecasting Capability (12km horizontal grid spacing) running on NCEP's p6 Supercomputer

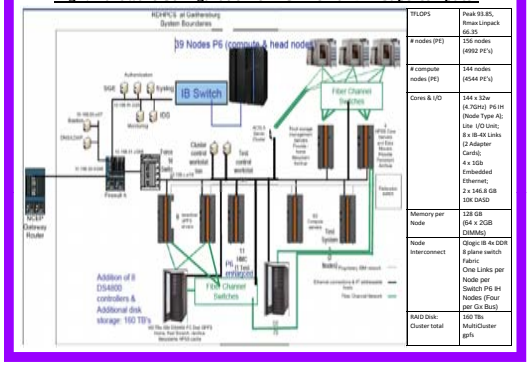
Pius Lee[1*], Daewon Byun[1], Hsin-Mu Lin[2], David Wong[3], Daniel Tong[2], Tianfeng Chai[2], Fantine Ngan[2], Rick Saylor[1], Ariel Stein[4], Hyun-Cheol Kim[2], Youhua Tang[5], Jeff McQueen[6], Marina Tsidulko[5], Jeff Young[3], Ho-Chun Huang[5], Sarah Lu[5], Dongchul Kim[5], Caterina Tassone[5], Ken Carey[7], Ivanka Stajner[7], and Paula Davidson[8]

INTRODUCTION

▪ **National Air Quality Forecasting Capability (NAQFC):** NOAA's Air Resource Laboratory (ARL) of the Office of Atmospheric Research (OAR) is leading the research effort to spearhead the science excellence of the NAQFC. The National Centers for Environmental Prediction (NCEP), National Weather Service Weather Research and Forecasting Non-hydrostatic Mesoscale Model (WRF/NMM) has been coupled with the EPA Community Multi-scale Air Quality (CMAQ) model to form the NAQFC. NAQFC running at NCEP's operational machine timely and reliably provides numerical guidance to the local forecasters with its state-of-the-art model science but layman-term guidance. In anticipation of extending the forecasting time and geographical coverage, finer horizontal and vertical grid spacing, optimizing NAQFC with respect to the operational machine is important. Last fall, NCEP had upgraded its supercomputer to IBM's p6 hardware (see Fig. 1). Profiling of CMAQ with figurations shown in Table 1, on this new machine aims to better optimize the code by addressing:

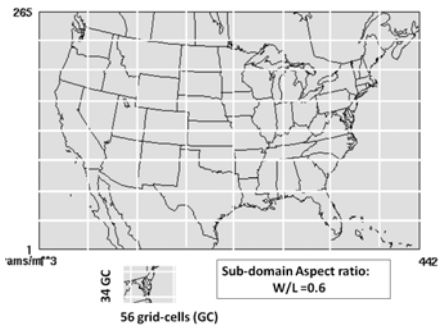
- diminishing speedup with respect to increasing resource allocation; and
- fine granularity analyses for I/O and communication over-burdens for the science processes.

Fig 1. Hardware Configuration of NCEP's P6 IBM-Supercomputer



TFLOPS	Peak 93.85, Rmax Linpack 66.35
# nodes (PE)	156 nodes (4992 PE's)
# compute nodes (PE)	144 nodes (4544 PE's)
Cores & I/O	144 x 32w (4.7GHz) P6 IH (Node Type A); Lite I/O Unit; 8 x IB-4X Links (2 Adapter Cards); 4 x 1Gb Embedded Ethernet; 2 x 146.8 GB 10K DASD
Memory per Node	128 GB (64 x 2GB DIMMs)
Node Interconnect	4Gxpic IB 4x DDR 8 plane switch Fabric One Links per Node per Switch P6 IH Nodes (Four per Gx Bus)
RAID Disk Cluster total	160 Tbs MultiCluster gpfs

Fig.2 Sample decomposition with 8x8 PE's for AQ-CONUS 12km (5x)

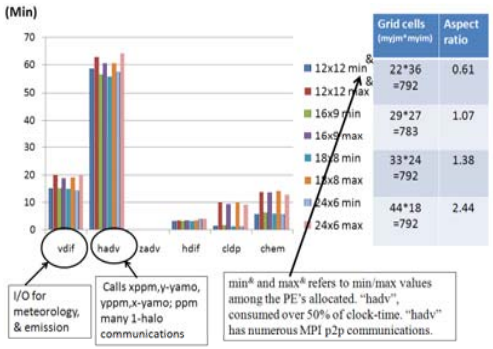


ACKNOWLEDGEMENTS

The authors are grateful to Drs. Rohit Mathur, Jon Pleim, Tanya Otte, Shaocai Yu, George Pouliot, and Ken Schere of the Atmospheric Modeling Division, USEPA , for their technical input and discussion. Advice from Dr. Daiwen Kang of CSC , Ms. Carolyn Pasti of IBM and Doris Pan of Raytheon is deeply appreciated.

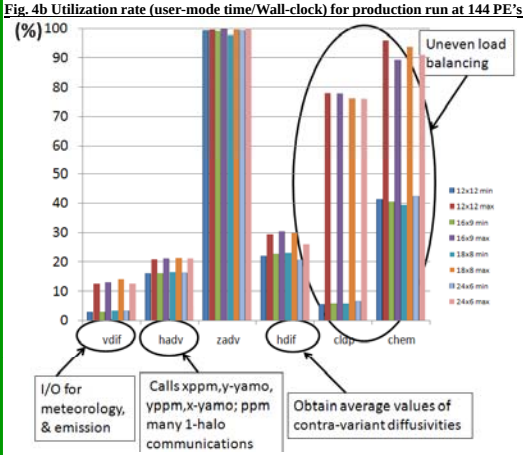
[1]NOAA/OAR/ARL, 1315 East West Hwy, Room 3316, Silver Spring, MD 20910; *pius.lee @noaa.gov
[2]Science and Technology Corporation, Hampton, Virginia.
[3] EPA, Research Triangle Park, NC.
[4] Earth Resources Technology, Annapolis Junction, MD.
[5]Scientific Applications International Corporation, Camp Springs, Maryland.
[6]National Centers for Environmental Prediction, Camp Springs, Maryland.
[7]Noblis, Inc., Falls Church, VA.
[8]Office of Science and Technology, National Weather Service, Silver Spring, Maryland.

Fig. 4a Wall-clock for a 60 h production run at 144 PE's level



Grid cells (myjm*myim)	Aspect ratio
22*36 =792	0.61
29*27 =783	1.07
33*24 =792	1.38
44*18 =792	2.44

Fig. 4b Utilization rate (user-mode time/Wall-clock) for production run at 144 PE's



The huge disparity in utilization rates between the minimum and maximum values stems from the disparity of cloudiness, chemical regimes, and PBL structure attributing to the intrinsically uneven balanced problem in parallelizing "cldp", "chem", and "vdif", respectively.

Fig. 4c Flop counts for production run at 144 PE's level

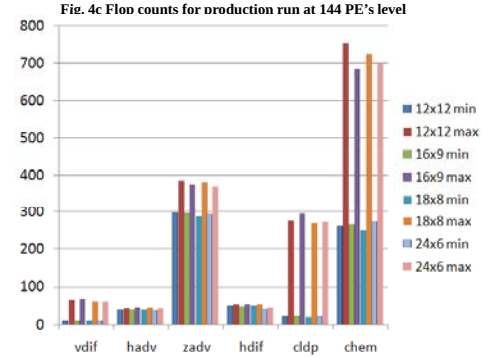
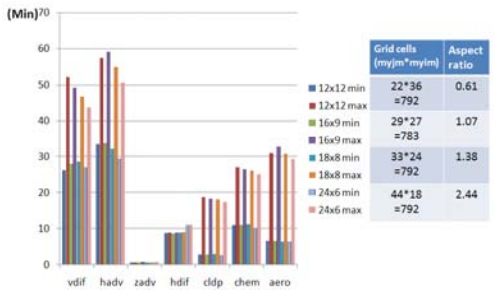


Fig. 5a Wall-clock for a 60 h experimental run at 144 PE's level



Grid cells (myjm*myim)	Aspect ratio
22*36 =792	0.61
29*27 =783	1.07
33*24 =792	1.38
44*18 =792	2.44

Fig. 5c Utilization rate for experimental run at 144 PE's level

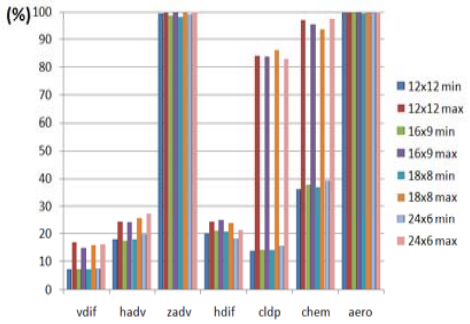


Fig. 5c Mflop count for experimental run at 144 PE's level

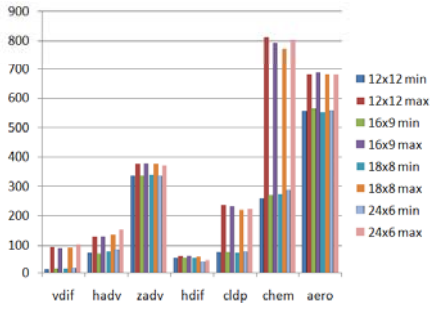


Table 2. CPU breakdown of aerosol process calculations (16x9 PE's)

Code segments (computational details)	Minimum CPU (Min)	Maximum CPU (min)
Subroutine GETPAR: aerosol density, Dg,.., etc	11	13
Subroutine ORGAER3 for SOA	3.8	5.6
Subroutine EQL3C for heterogeneous N_2O_5 reaction	6.7	23.2
Moments, number-concentration and partition calculation, mean free path, viscosity, modal mass	29.0	36.1
Size-dependent condensation growth factors for sulfate and organics	3.1	3.8
New particle due to binary nucleation of H_2SO_4 and H_2O	3.4	3.8
Coagulation (analytical) within and between modes	3.5	3.7l
Update number concentration of Aitken (Riccati Eq.)	3.2	3.4

Table 1. CMAQ configurations

Run slot	Gas-chemistry mechanism	Aerosol module
Production	CBM-IV	aero3
Experimental	CB05	aero4

Fig. 3 Wall-clock time for 60 h CONUS (5x) runs

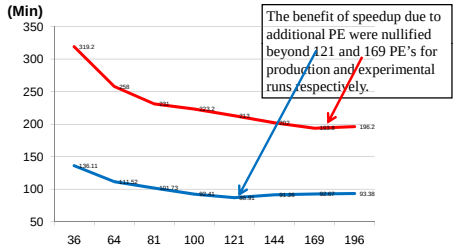


Table 3. MPI breakdown for PE=0 for a 10*10 PE's run (60h)

Data for MPI rank 0 of 101:
Times and statistics from MPI_Init() to MPI_Finalize().

MPI Routine	#calls	avg. bytes	time(min)
MPI_Comm_size	2	0.0	0
MPI_Comm_rank	2	0.0	0
MPI_Recv	36000	1724934.5	7.49
MPI_Iprobe	1035523260	0.0	79.7
MPI_Barrier	2	0.0	0.01

6 calls per hour but much more time in probing readiness of result from slave PE's --- At current resolution, use of additional PE's for I/O may not reduce wall-clock time

The master PE is the dedicated "output" PE. It was not a major factor in holding up wall-clock time as shown in Table 3. Most of its time is spent waiting for the slave PE's to deliver results.

Summary

•Increase of diminishing return for speedup by allocating more PE's is noticed. Additional resource does not gain speedup beyond 121 and 169 PE's for the production and experimental runs respectively.

•Horizontal advection with Piecewise Parabolic Method (PPM) entails large number of MPI calls; and each has rather large data-size (~650Mbytes). This is a major challenge to efficiently parallelize "hadv" and "hdiff".

•Large load uneven-balancing for cloud, chemical, and PBL processes. This is a design challenge for CMAQ.

•In terms of computational efficiency (Flops counts), vertical advection, chemistry, and aerosol processes are doing well. Further speed-up can come from OpenMP and threading. These options should help other processes too.

• When assigning ncol and nrow, one should aim for sub-domain aspect ratio as close to unity as possible to obtain optimal computational efficiency.

•On 5x CONUS domain with 12 km horizontal grid spacing (Table 3), the PE dedicated to I/O is idle most of the time.