LAPORAN PENELITIAN

“WRF Performance Analysis and Scalability on Multicore High Performance Computing Systems”

Didin Agustian P. Ph.D

INSTITUT TEKNOLOGI NASIONAL
BANDUNG - 2020

18

WRF Performance Analysis and Scalability on Multicore High Performance Computing Systems

Weather Research and Forecast (WRF) is one of the most commonly used numerical weather prediction models that has superior scalability and computational efficiency. Its performance and scalability was tested on the Computing on Kepler Architecture (COKA) cluster, a recent multicore high-performance computing (HPC) system installed at the University of Ferrara, Italy. Two major experiments were designed: single domain with varied grid size (E1) and nesting domain (E2) WRF configuration. As expected, simulation speed decreased when domain grid size increased, while an increase in the number of computing nodes used in the simulation would increase the simulation speed under E1. We run WRF with several combinations of Message Passing Interface (MPI) tasks and threads per MPI task. Optimum performances for E1 and E2 were achieved either using 4 computing nodes with 8 MPI per node (MPN) and 2 threads per MPI (TPM) or 4 nodes with 2 MPN and 8 TPM. Most often, time was spent on computational processing (89–99%) rather than other processes such as input processing and output writing. The WRF model domain configuration was an important factor for simulation performance: for example, the nesting domain would slow down the simulation by 100 times in comparison to the single domain configuration. Further works can be done for testing the performance and scalability for other WRF applications, forecasting and air quality simulation on the newly installed TORUS cloud at the Asian Institute of Technology.

Chapter written by Didin Agustian PERMADI, Sebastiano Fabio SCHIFANO, Thi Kim Oanh NGUYEN, Nhat Ha Chi NGUYEN, Eleonora LUPPI and Luca TOMASSETTI.

TORUS 1 – Toward an Open Resource Using Services: Cloud Computing for Environmental Data, First Edition. Edited by Dominique Laffly.

18.1. Introduction

The Numerical Weather Prediction (NWP) model is an important tool for both research and operational forecast of meteorology which can be used for various other purposes such as weather aviation, agriculture, air pollution dispersion modeling, etc. A fundamental challenge is to understand how increasingly available computational power can improve modeling processes, in addition to the reliability of the NWP output itself [MIC 08]. The Weather Research and Forecast (WRF) model is one of the most commonly used NWP that is designed to run on a variety of platforms, either serially or in parallel, with or without multi-threading [SKA 07]. In light of the rapid development of the WRF model, a successor of the previously well-known mesoscale meteorological model (MM5), other NWP models exist such as the Regional Atmospheric Modeling (RAMS) System, the Regional Climate Model (RegCM), etc.

WRF model performance benchmarking has been done within different environments to demonstrate the scalability of the computational environment and considerations for higher productivity [HPC 15]. It is well understood that the WRF model is widely adopted for the assessment for at least the following two reasons: 1) its superior scalability and computation efficiency [CHU 17] and 2) the last generation of the NWP which is equipped with current developments in physics, numerics and data assimilation [POW 17]. We conducted WRF performance analysis and scalability on multicore a High Performance Computing (HPC) system using our own benchmarking configuration. We used WRF version 3.7 for testing its application for a tropical domain in Southeast Asia (SEA), dominated by convective meteorology conditions. First, we tested performance and scalability using a WRF single domain configuration for different grid sizes, followed by a two-way nesting configuration. In this study, we have run the code enabling both Message Passing Interface (MPI) to exploit parallelism among different node-processors, and Open-MPI to exploit parallelism within each node-processor.

18.2. The weather research and forecast model and experimental set-up

18.2.1. *Model architecture*

WRF, a mesoscale NWP, is designed for both atmospheric and forecasting research. It was initially developed in the 1990s under a collaborative partnership of the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the

Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma and the Federal Aviation Administration (FAA). The source code has been made available[1]. The model consists of two dynamical cores: the Advanced Research WRF (ARW) and Nonhydrostatic Mesoscale Model (NMM); the first is commonly used for research and application [SKA 08]. WRF equations are formulated using a terrain-following hydrostatic-pressure vertical coordinate (sigma pressure level). WRF solver currently supports four projections to the sphere: the Lambert conformal, polar stereographic, Mercator and latitude–longitude rojections.

The WRF modeling system includes three components, with pre-processing (known as WPS), main solver (ARW) and post-processing of outputs which can be used for another model (e.g. air quality model) or visualization. The initial conditions for the real-data cases are pre-processed through a separate package called the WRF Preprocessing System (WPS) which takes terrestrial and meteorological data (lateral boundary conditions). Static geographical data such as terrain and land-cover can be taken from the WRF website repository[2]. Lateral meteorology boundary conditions can be taken from the National Center for Environmental Prediction FNL (Final) Operational Global Analysis data which are available on a resolution of 1° (approximately 100 km) for every six hours. The main solver ARW includes two classes of simulation, with ideal initialization and using real data. In this research, we used the real-data which require pre-processing through WPS to generate initial and lateral boundary conditions. The final run is done in a main solver of WRF to generate 3-dimensional (3D) simulated fields of meteorological parameters. This final run is recognized as the most time-consuming process. The system is shown in detail in Figure 18.1.

Multiple physical parameterizations are offered in WRF, included in physics categories, i.e. microphysics, cumulus parameterization, planetary boundary layer (PBL), land surface model (LSM) and radiation. Each parameterization is used to resolve certain specific physical atmospheric processes. For example, water vapor, cloud particle formation is resolved by microphysics, while the effect of sub grid scale clouds is represented by cumulus parameterization schemes. PBL schemes provide flux profile due to eddy transports in the whole atmospheric column. The radiation schemes provide the atmospheric heating profiles and estimation of net radiation for the ground heat budget. The surface layer (SL) schemes are used to calculate the friction velocity and exchange coefficients that enable the estimation of heat, momentum and moisture fluxes by the LSMs.

1 http://www2.mmm.ucar.edu/wrf/users/download/get_source.html.
2 http://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html.

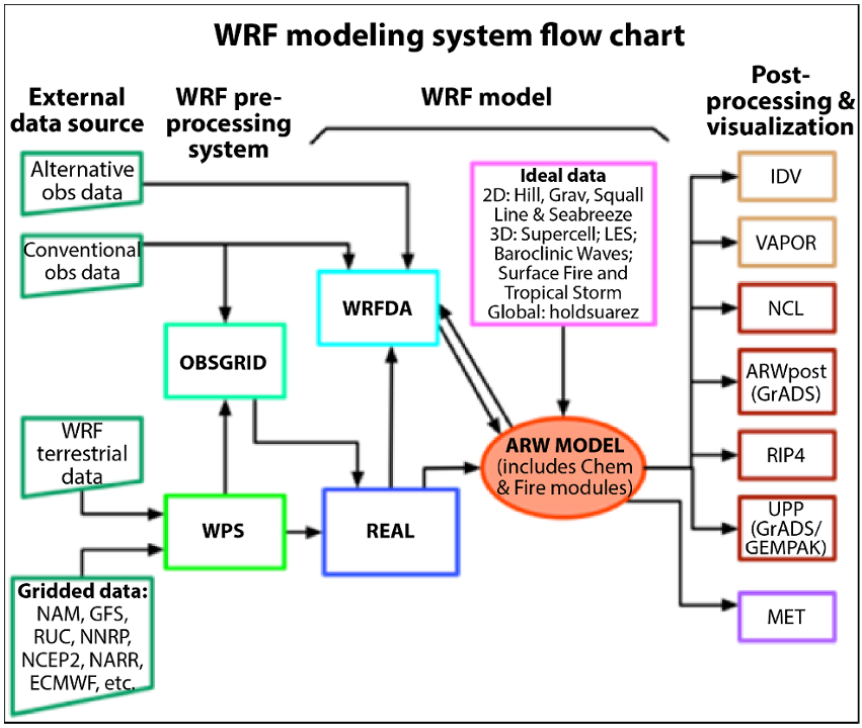

Figure 18.1. *WRF–ARW modeling system flow chart (source: Wang* et al. *2016). For a color version of this figure, see www.iste.co.uk/laffly/torus1.zip*

18.2.1.1. *Experimental set-up for performance and scalability of the system*

To assess performance and scalability, we prepared two benchmarking configurations: i) experiment 1 (E1), single domain with a different grid number (constant grid resolution of 18 × 18 km^2 is presented in Figure 18.2), and ii) experiment 2 (E2), nesting domain, i.e. 3 domains with a resolution of 18, 6 and 2 km respectively (as presented in Figure 18.3). The detailed geographical configuration of E2 WRF experiment is presented in Table 18.1.

For E1, WRF was used to simulate a 6-hour period (approximately 21,600 seconds), from January 1, 2007, 00:00 AM to January 1, 2007, 06:00 AM. For E2, WRF was used to simulate a 5-day period (January 1, 2016, 00:00 AM to January 5, 2016, 00:00 AM) for all cases of grid sizes. For both experiments, we performed simulations using different combinations of number of nodes, N (1, 2 and 4), number of MPI tasks per node (MPN) and number of threads per MPI task, TPM (1, 2, 4, 8 and 16).

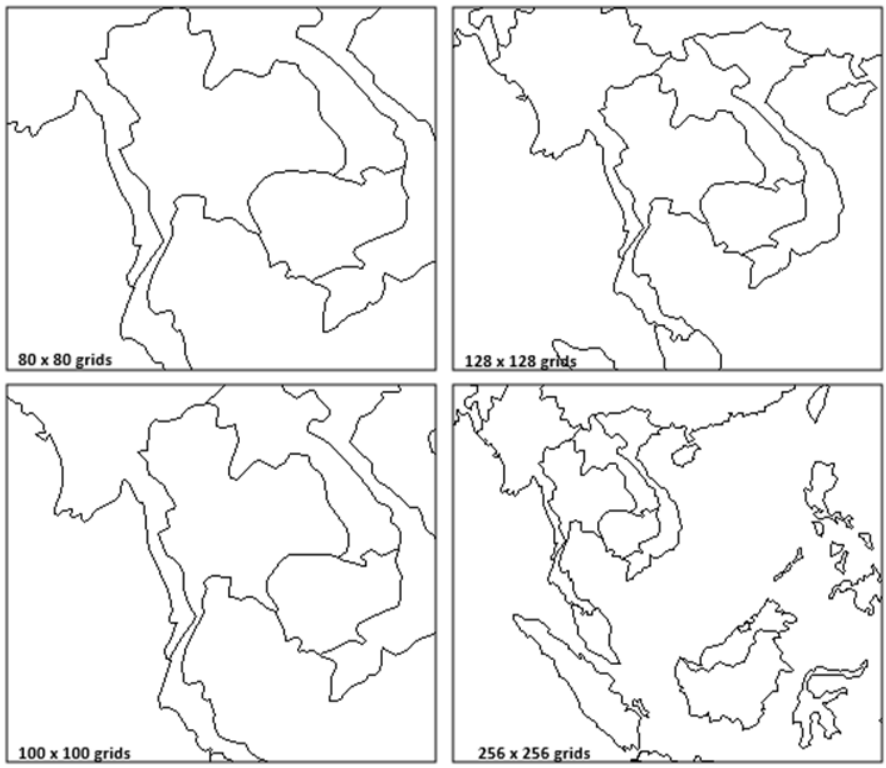

Figure 18.2. *Domain configuration for E1*

We then monitored the following parameters:

– time (T) required to finish each simulation (sum of computation, writing output processing input, and processing boundary conditions);

– simulation speed (S) that is defined as the ratio of the actual period simulated by WRF to the time required to finish simulation;

– total core (TC) that is estimated by the following equation:

$$\text{Total cores} = N \times MPN \times TPM \quad [18.1]$$

with N, MPN and TPM defined above.

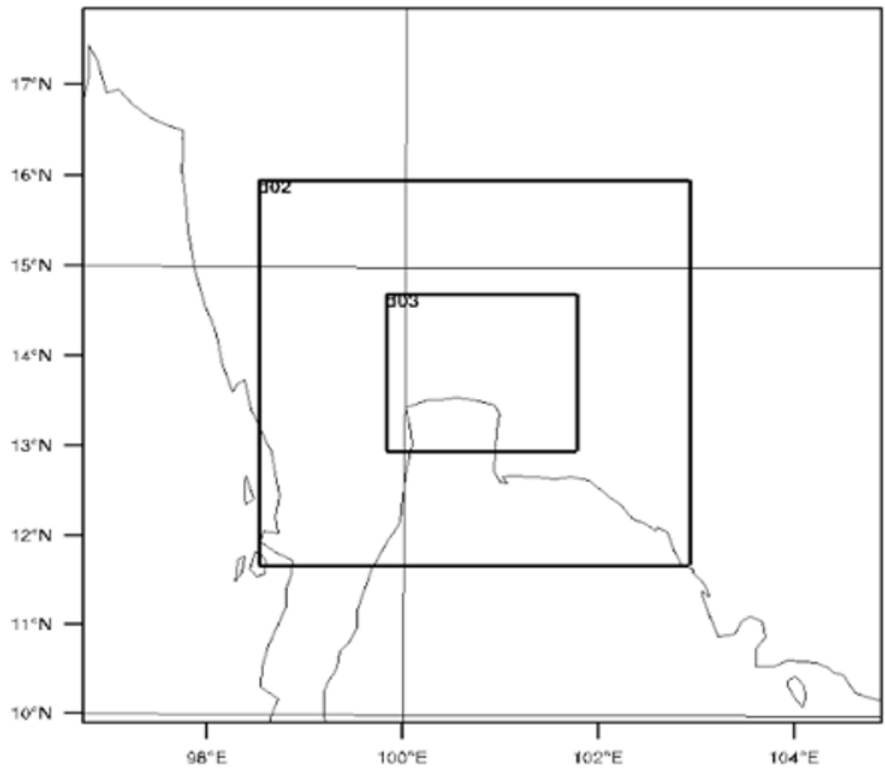

Figure 18.3. *Domain configuration for E2*

The coarsest domain for the E2 experiment comprised 50 × 50 grid cells with a grid resolution of 18 km. The second and the finest domains consisted of 81 × 81 and 96 × 99 grid cells with a grid solution of 6 km and 2 km, respectively (Table 18.1).

Selection of physics options is important as the incorporated physics options would affect the time required for simulation. We selected meteorological physics options, namely microphysics, cumulus, radiation, boundary layer and land surface interaction from previous publications that are suitable for this region [CHO 11, PRA 13]. The WRF experiment E1 (single domain) used the options presented for domain 1 in Table 18.2, while experiment E2 (nesting domain) used all of the options presented in the same table.

Domains	X	Y	Lon	Lat
Domain 1: dx = 18 km				
Number of Dot Points	51	51	–	–
Starting Lower Left i, j of Parent Grid	1	1	–	–
Domain 2: dx = 6 km				
Number of Dot Points	82	82	–	–
Starting Lower Left i, j of Parent Grid	12	12	–	–
LCP of the SW Dot Point (km)	–243	–243	98.5	11.6
LCP of the NE Dot Point (km)	243	243	102.99	16.0
Domain 3: dx = 2 km				
Number of Dot Points	97	100	–	–
Starting Lower left i, j of Parent Grid	24	24	–	–
LCP of the SW Dot Point (km)	–105	–105	99.7	12.8
LCP of the NE Dot Point (km)	87	93	101.5	14.7

Table 18.1. *Geographical configuration of "two-way" nesting domains*

Physics options	Domain 1 (18 km)	Domain 2 (6 km)	Domain 3 (2 km)
Microphysics	WDM6	WSM3	WSM3
Cumulus parameterization	BMJ	BMJ	BMJ
Short-wave radiation	Dudhia	Dudhia	Dudhia
Long-wave radiation	RRTM	RRTM	RRTM
Planetary Boundary Layer	YSU	YSU	YSU
Surface Layer	Monin-Obukhov	Monin-Obukhov	Monin-Obukhov
Land Surface Layer	Noah LSM	5-layer thermal diffusion	5-layer thermal diffusion

Table 18.2. *Physics options selected for the WRF experiment E1 (domain 1) and E2 (all 3 domains)*

The combinations of N, MPN and TPM used in all experiments are presented in Table 18.3. Simulations for each domain were done for all 15 combinations.

No.	N	MPN	TPM
1	1	1	16
2	1	2	8
3	1	4	4
4	1	8	2
5	1	16	1
6	2	1	16
7	2	2	8
8	2	4	4
9	2	8	2
10	2	16	1
11	4	1	16
12	4	2	8
13	4	4	4
14	4	8	2
15	4	16	1

Table 18.3. *Combinations of N, MPN and TPM used in experiments. N = number of node; MPN = number of MPI tasks per node; TMP = number of threads per MPI task*

18.3. Architecture of multicore HPC system

The results we show in this chapter were taken from an initial run on the COKA installed cluster at the University of Ferrara in Italy. The COKA cluster has 5 computing nodes, each node hosting 2 Intel Xeon E5-2630v3 CPUs, 256 GB of memory and 8 dual-gpu NVIDIA K80 boards. The Intel Xeon E5-2630v3 processor embeds an 8-core, each supporting the execution of 2 threads, and 20 MB of L3-cache. It runs at a 2.40 GHz frequency that can be boosted up to 3.20 GHz under specific conditions of workloads. Nodes are interconnected with 56 Gb/s FDR InfiniBand links, and each node hosts 2 Mellanox MT27500 Family [ConnectX-3] HCA, allowing *multirail networking* for a doubled inter-node bandwidth. The two InfiniBand HCAs are connected respectively to the two PCIe root complexes of the two CPU sockets. This allows for a symmetric hardware configuration, where each processor has one local InfiniBand HCA, connected to the same PCIe root complex; so, data messages do not need to traverse the Intel Quick Path inter-socket communication link.

Since processors of this cluster are multi-core, the WRF code has been configured to use both the OpenMP and MPI libraries. In practice, our application launches one or more MPI processes, and each process spawns several threads associated to a physical core of the processor. This allows us to fully exploit all the levels of the parallelism offered by the machine: node-level parallelism is handled through the MPI library, while core-level parallelism is handled through the OpenMP library.

Under the TORUS project, cloud cluster infrastructure (computing node and storage) was installed at the Asian Institute of Technology (AIT), Pathumthani, Thailand. To exploit the running of WRF on cloud, 10 nodes (each consisting of 8 cores) were installed over virtual machine (VM) configuration as part of total 20 × proc Intel Xeon E5 2650 10 cores or 200 cores. This hardware infrastructure has total 640 GB of RAM, 10 nodes, on X6800 multi-node rack of Huawei brand. The system has a hardware infrastructure storage of total 114 TB storage, 6 TB high-speed storage.

18.4. Results

This section will report the results we have measured on the cluster about the performance and scalability of the system. Even though time (T) was recorded in our experiments, the results presented here are mainly focused on the simulation speed (S) of each combination for both E1 and E2.

18.4.1. *Results of experiment E1*

For each grid-size case, the speed (S) is presented against MPN and TPM as plotted in Figure 18.4. We observe that using a single node, the highest value of S was achieved with 8 MPN and 2 TPM for all grid size configurations. However, this pattern was not seen if we used 2 nodes that seem to be dependent on the grid-size configuration.

In this experiment, using 4 nodes, the optimum performance was achieved when we used 2 MPN and 8 TPM for the grid-size configuration corresponding to 80 × 80 and 100 × 100 cells; however, for the larger number of grid cells, the highest simulation speed was achieved using 8 MPN and 2 TPM (Figure 18.4). Overall, the increase in the number of grids was associated with lower simulation speed. The increase in the number of nodes used in the simulation was associated with increased performance (higher simulation speed). In Table 18.4, we report the simulation speed achieved for a different grid size as a function of the total number

of cores (TC) used. TC is defined as the product of N, MPN and TPM as expressed in [18.1]

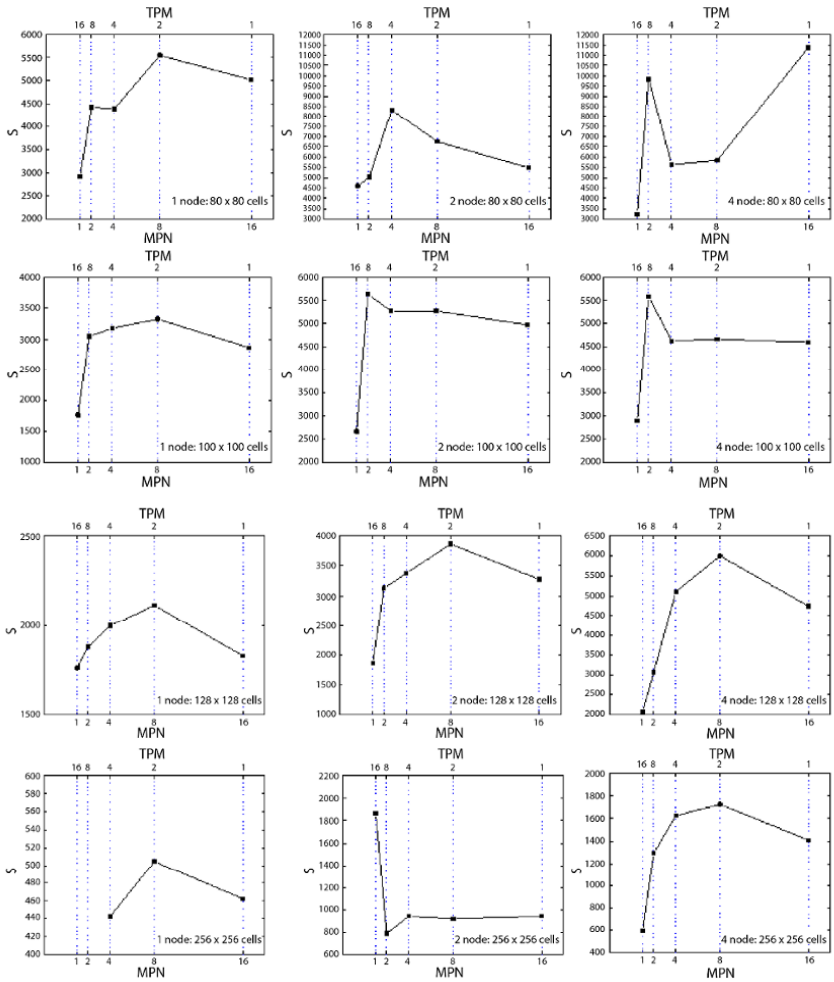

Figure 18.4. *Simulation speed measured for all combinations under E1. The lower horizontal axis reports MPN (the number of MPI task per node), and the upper is the corresponding TPM (the number of threads per MPI task). The vertical axis reports the simulation speed (S)*

Horizontal grid size (number of cells)	Total number of core (TC)	Simulation speed (S)
80 × 80	16	5,553
	32	8,308
	64	11,368
100 × 100	16	3,328
	32	5,640
	64	5,596
128 × 128	16	2,114
	32	3,864
	64	6,000
256 × 256	16	463
	32	1,864
	64	1,728

Table 18.4. *Relationship between grid size, total core and simulation speed for experiment E1*

It was found that the increase in the total number of cores would increase the simulation speed; however, the grid-size configuration would also affect this relationship. For the grid size corresponding to 80 × 80 and 128 × 128 cells, a clear relationship was seen. However, for the grid sizes corresponding to 100 × 100 and 256 × 256 cells, the optimum performance was achieved for the total number of cores of 32 rather than 64, although the difference was not large. This indicated that the WRF domain configuration during the parallel computation, plays an important role in its computation performance.

In Table 18.5, we analyzed the time required for different processes, including the processing input, writing output, processing boundary conditions and simulation and present the results. Computation dominated the share of the total time required, i.e. 89–98% with the average value of 94.3% followed by the output writing of 1.2–8.5% (averaged at 4.3%). Other processes collectively shared only 0.13–4.2% (averaged at 1.38%). It is obvious that the simulation speed is largely affected by the time required for computation which, in turn, is affected by the model configuration.

Process	% of total time		
	Average	Max	Min
Computation	94.30	98.40	89.05
Writing output	4.32	8.47	1.16
Processing input	0.60	1.61	0.10
Processing lateral boundary	0.78	2.68	0.03

Table 18.5. *Share of different processes to the total time required for experiment E1*

18.4.2. *Results of experiment E2*

The relationship between MPN, TPM and S is presented in Figure 18.5 for WRF experiment E2 over 3 domains. It was found that when using a single node, optimum performance (highest S value) was achieved using 8 MPM and 2 TPM. Similar results were seen for all the experiments with 4 nodes. However, a different pattern was observed when we used 2 nodes which had the optimum performance achieved using 4 MPM and 4 TPM. Overall, using 2 nodes would increase performance (1.4 times faster) than only a single node. However, increasing the number of nodes to 4 would only slightly improve the performance (Table 18.6). There is a need to further investigate a bottle neck, especially when we used 2 and 4 nodes.

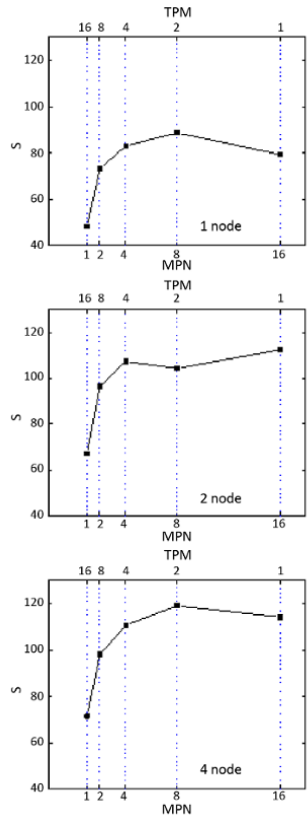

Figure 18.5. *Simulation speed measured for all combinations under experiment E2. The lower horizontal axis reports MPN (the number of MPI task per node) and the upper is TPM (the corresponding number of threads per MPI task). The vertical axis reports the simulation speed (S)*

Configuration*	Number of total core	Simulation speed
1-1-16	16	48
1-2-8	16	73
1-4-4	16	83
1-8-2	16	89
1-16-1	16	79
2-1-16	32	67
2-2-8	32	97
2-4-4	32	108
2-8-2	32	104
2-16-1	32	113
4-1-16	64	72
4-2-8	64	98
4-4-4	64	111
4-8-2	64	119
4-16-1	64	114

Table 18.6. *Relationship between the total number of cores (TC) and the simulation speed (S) based on the results of experiment E2. *Number of node = MPI task per node thread per MPI task*

Similar to experiment E1, we also analyzed the time required for different processes, and the results are presented in Table 18.7. The computation consumed the most, i.e. 98.7–99.6%, of the total computation time with the average of 99.13%, followed by the output writing of 0.32–1.14% with an average of 0.81%. Other processes collectively shared only 0.03–0.1% with an average of 0.06%. As compared to the results of experiment E1, the time required for computation in experiment E2 is larger. This is because in experiment E2, the 2-way nesting simulation was simultaneously applied for all three domains and hence more time was required.

Process	% of total time		
	Average	Max	Min
Processing input	0.05	0.08	0.03
Writing output	0.81	1.14	0.32
Processing lateral boundary	0.01	0.02	0.00
Simulation	99.13	99.64	98.77

Table 18.7. *Share of time consumed for different processes to the total time required for experiment E2*

18.5. Conclusion

WRF model version 3.7 was applied for a tropical convective dominated domain of SEA, using multicore HPC at the University of Ferrara, Italy, for performance and scalability analysis. Single domain with different grid sizes (experiment E1) and nesting domain (experiment E2) configurations were simulated and evaluated in terms of time and simulation speed under different node set-ups, MPI task and thread. E1 results showed that the simulation speed decreased when the number of grid cells in the domain increased, which was expected, and the increase in the number of nodes used in the simulation would increase the simulation speed. In E1, when using a total of 64 cores, the highest speed was obtained for the domains 80×80 and 128×128 cells, and this was better than the results of 16 and 32 cores. E2 results showed the optimum performance when using 4 nodes, 8 MPN and 2 TPM which was only slightly better than using 2 nodes. Overall, the time required for computation contributed the most (89–99%) for both experiments compared to that consumed for input processing and output writing. Simulation speed of nesting domain configuration experiment (E2), when two-way nesting was applied for simultaneous simulation on 3 domains, was 100 times slower than the one-way nesting simulation for single domain (E1); thus, WRF model domain configuration (one-way or two-way nesting) was an important factor for simulation speed, in addition to the computing core configuration.

Future work should be firstly extended to cover a more extensive performance analysis and scalability for ideal cases of WRF simulations, such as hurricane/typhoon and meteorology forecasting. Secondly, similar tests should be conducted for WRF applications to drive air quality dispersion models. Thirdly, simulations with different computing configurations using GP–GPUs need to be conducted to significantly reduce the execution time of most computing intensive kernel routines. Fourthly, GP–GPUs can be used by new and recent programming frameworks such as OpenACC, which would allow reduced execution time of several scientific applications. Lastly, a scalability study needs to be performed with similar WRF benchmarks using the newly installed TORUS cloud cluster at AIT.

18.6. References

[BON 18] BONATI C., CALORE E., D'ELIA M. *et al*., Portable multi-node LQCD Monte Carlo simulations using OpenACC. *International Journal of Modern Physics C*, 29(1), doi: 10.1142/S0129183118500109, 2008.

[CAL 16] CALORE E., GABBANA A., KRAUS J. *et al*., Performance and portability of accelerated lattice Boltzmann applications with OpenACC, *Concurrency Computation*, 28(12), 3485–3502, doi: 10.1002/cpe.3862, 2016.

[CHU 17] CHU Q., XU Z., CHEN Y. *et al.*, Evaluation of the WRF model with different domain configurations and spin-up time in reproducing a sub-daily extreme rainfall event in Beijing, China, *Hydrology and Earth System Sciences*. Available at: https://doi.org/10.5194/hess-2017-363, 2017.

[HPC 15] HPC AC, WRF 3.7.1: Performance Benchmarking and Profiling. Report, High Performance Computing Advisory Council, 2015.

[MIC 08] MICHALAKES J., HACKER J., LOFT R. *et al.*, WRF nature run. *Journal of Physics: Conference Series*, 125(1), 2008.

[POW 17] POWERS J.G., KLEMP J.B., SKAMAROCK W.C., *et al.*, The Weather Research and Forecasting (WRF) model: Overview, system efforts, and future directions. *Bulletin for the American Meteorological Society*. Available at: https://doi.org/10.1175/BAMS-D-15-00308.1, 2017.

[SKA 08] SKAMAROCK W.C., KLEMP J.B., DUDHIA J. *et al.*, A description of the advanced research WRF version 3. Technical Note, NCAR, Boulder, Colorado, USA, 2008.