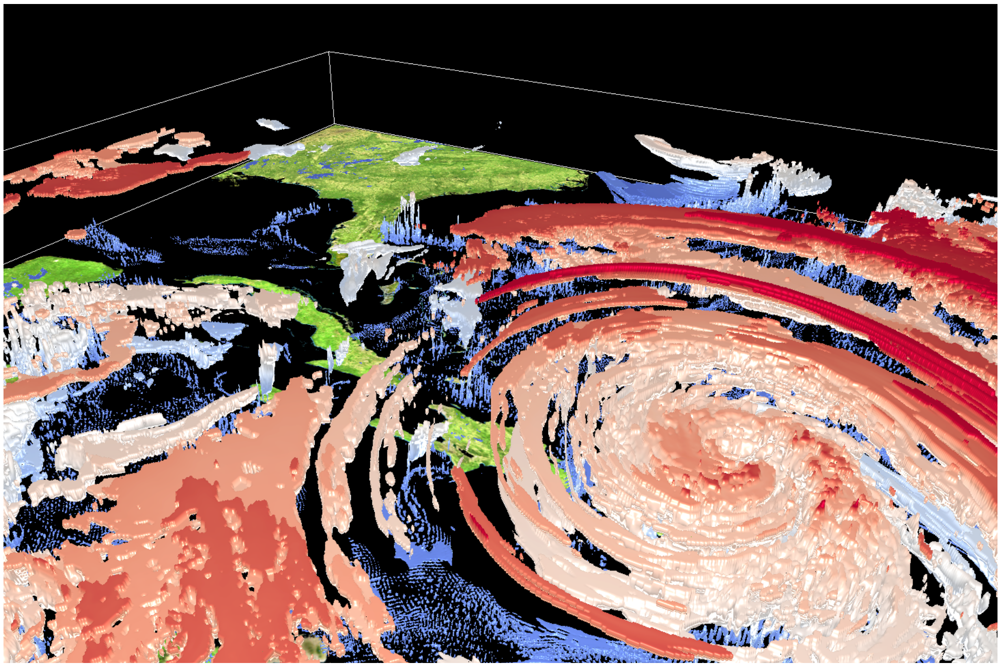
insideHPC Special Report

Accelerate WRF Performance

Expedite Predictions with In-Depth Workload Characterization Knowledge

by Michael Schulman



BROUGHT TO YOU BY

Contents

Introduction

Investment in High-Performance Computing (HPC) needs to show a return on investment from the start. Merely buying a bunch of servers and expecting your applications to run faster most likely won't work. Understanding the underlying hardware and how applications can take advantage of the critical computing power is one of the most critical aspects of using HPC. By partnering with a supplier that has the domain expertise and the technical know-how to work with various organizations will pay off in the long run.

Servers and More

A modern computer server contains several necessary components that need to work seamlessly together to deliver fast results. A necessary HPC infrastructure will include many servers, network interconnects, and storage that is suited to the capacities needed for the applications that will be executed.

Typical Server Resources:

- 1, 2 or 4 sockets
- 16+ cores per socket
- DRAM Memory – 512 GB or more
- Internal storage
- Networking capability

While many companies today design and manufacture servers that contain the above hardware, many of these companies do not have the expertise to design a full solution that meets demanding customers' performance requirements and cost constraints. Domain expertise is critically important when working with a server vendor. This expertise leads to a more optimized hardware and software architecture for the given workloads that a customer must execute. By understanding in great depth these workloads, a valued partner can experiment with sophisticated technologies to determine the ideal combination of software and hardware. This leads to impressive gains in the performance of the overall system, which can contain many thousands of servers or computing units. It is especially important to understand the interaction of the software and the underlying hardware architecture, at a level that many end-users will not understand. Development environments such as the compiler used, the libraries used, and the communication choices will all affect the performance of critical workloads.

The knowledge that a server and storage supplier bring to specific industry domains cannot be overlooked for several reasons, including:

- **Proper sizing of infrastructure –** Customer receives solution based on requirements
- **Expert advice on server components –** Customer gets the best matching of CPUs, Memory, Storage
- **Tuned software stack –** Customer can get productive faster
- **Knowledgeable Sales Engineers –** Customer can focus on science rather than optimizing the system

Understanding Climate Change

One of the most pressing, if not the most severe challenge of our time is to understand the consequences of climate change. Research is being performed around the world to determine the extent of damage, both physically and politically, that will are a result of climate change. Also, research into mitigation options and their effect on the climate continues to be an active topic.

> Since WRF is open-sourced, organizations can download the application code, make changes or add features and compile and deploy on their on-premises servers or work with others to use their new version.

The time to solution will typically decrease in time as faster processors become available. Still, there is also room for improvement in the programming of the physics involved and making better use of the underlying hardware infrastructure.

A popular application that simulates climate change is the Weather and Research Forecasting (WRF) model, which is a collaborative partnership of the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (represented by the National Centers for Environmental Prediction (NCEP) and the Earth System Research Laboratory), the U.S. Air Force, the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA). WRF is used by thousands of researchers worldwide in over 150 countries.

Many scientific application codes are written in FORTRAN, due to several reasons. These include:

- Easy to learn
- Optimized numerical libraries
- Extensive collection of supporting applications

Since WRF is open-sourced, organizations can download the application code, make changes or add features and compile and deploy on their on-premises servers or work with others to use their new version. The choice of compilers for the source code can have a significant effect on the ultimate performance of the application. Customers who download the source code typically want to see results and are not interested in testing many different compilers with many different flags set. There are many compilers available for FORTRAN today, and many end-users look for guidance as to which one of these compilers will result in the highest performance. The goal of the end-user is to obtain answers as quickly as possible and not experiment with different compilers or compiler options.

> There are many compilers available for FORTRAN today, and many end-users look for guidance as to which one of these compilers will result in the highest performance.

Quanta Cloud Technology

Quanta Cloud Technology (QCT) is a global datacenter solution provider. QCT combines the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation datacenter design and operation challenges. QCT serves cloud service providers, telecoms and enterprises running public, hybrid and private clouds.

QCT has the expertise to investigate which compilers work best with different applications. By compiling and then testing the different codes using the same hardware, the efficiency of the compiler can easily be determined.

Introduction to WRF

WRF is a regional weather model with users ranging from researchers to forecasters all over the globe. Noted for being a mature and sophisticated model for weather research, WRF produces initial weather conditions for environmental models, such as air quality models, small-scale Large Eddy Simulation(LES) models, and disaster assessment models. WRF is among one of the significant workloads in major High-Performance Computing (HPC) systems, thus understanding how WRF performs and behaves under different optimizations could increase the HPC efficiency and thus reduce operating costs.

> WRF is highly parallelized and takes advantage of the distributed-memory method using MPICH, the shared-memory method using OPENMP, or the combination of both techniques, a hybrid approach.

Similar to other weather and climate models, WRF discretizes the target simulated area into three-dimensional grids. The physics properties of each grid are then dispatched to computational threads to calculate their tendencies (the rate of change of the physical properties in the timestep). After each time step is finished, the computed results will propagate to the corresponding grids both horizontally and vertically, depending on the calculated direction of the wind. WRF is highly parallelized and takes advantage of the distributed-memory method using MPICH, the shared-memory method using OPENMP, or the combination of both techniques, a hybrid approach.

Characteristics of the workload

Because of the grid approach and the parallelization of WRF, there is a large amount of data that is transferred between grids after each timestep is completed. Thus, the overall performance is dependent on the high memory bandwidth and low latency of the interconnecting network. The output, which is a massive list of variables from all the grids, requires high-efficiency storage bandwidth. QCT investigated the WRF performance impact from the latencies and bandwidth from both inside the processors and the chosen interconnect.

Benchmark settings

QCT ran the WRF benchmarks on a total of three QuantaPlex T42D-2U servers. Each T42D-2U server consists of four dual-socket computing nodes in a 2U form factor.

In total, twelve nodes were used to evaluate the scalability of WRF performance. Each node consists of two Second-Generation Intel® Xeon® 8280 Scalable Processors (28 cores at 2.7Ghz base frequency) and 384GB DDR-4 2933 memory on each node, which results in a total of 56 cores and 384GB of memory in one node, or 224 cores and 1296GB of memory in each T42D-2U system. Each node connects with other computing nodes and storage nodes with 10 Gbits/s Ethernet and Infiniband HDR-100 100 Gbit/s networks. The BeeGFS parallel file system is used as the underlying file system to maximize storage throughput. The hardware specification is listed below.

Specification of T42D-2U Server for Benchmark	
Nodes per server (enclosure)	4
Processors each node	2 x Intel Xeon 8280
Total cores per node	56
Total memory per node	384GB (12 x 32GB DDR-4 2933)

WRF settings

QCT used WRF V4.1.5 for the benchmark investigation. QCT followed Kyle (2018)'s work and created a new CONUS 2.5km domain for version 4 of WRF. The Conus 2.5km domain as shown in Figure 1 below, consists of 1901 x 1301 grid points and 40 vertical layers. The results were measured by the averaged WRF-output computation time of each timestep. Also, the output benchmark was measured by the averaged WRF-output computation time of each output timestep.

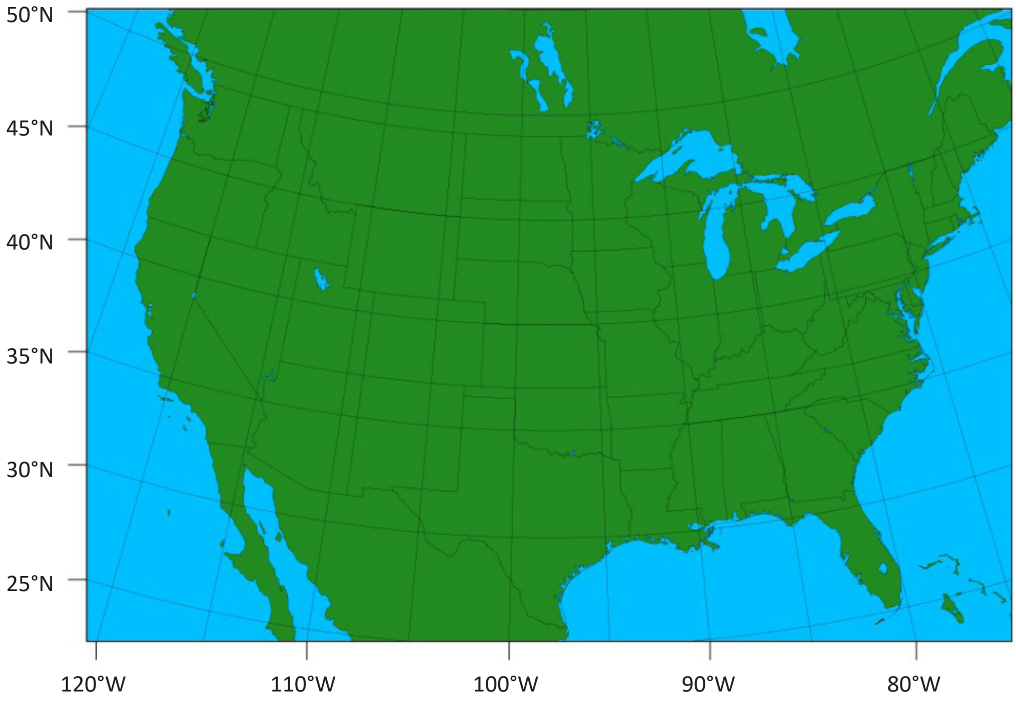


Figure 1. Domain setting of the CONUS 2.5km domain.

Source: QCT

Compiler Options

QCT used three different compilers with the latest version available to compile WRF and its dependent libraries (OpenMP/Mvapich, NetCDF, HDF5). The three compilers tested were GNU compiler v 9.2.0, AOCC compiler version 2.1.0 by AMD, and Intel® FORTRAN compiler (part of Intel® composer XE version 2020). The compiler flags other than the default WRF settings are listed below:

GNU compiler version 9.2.0 (gcc)

- -O3
  The default is -O2.

AOCC compiler version 2.1.0 (aocc)

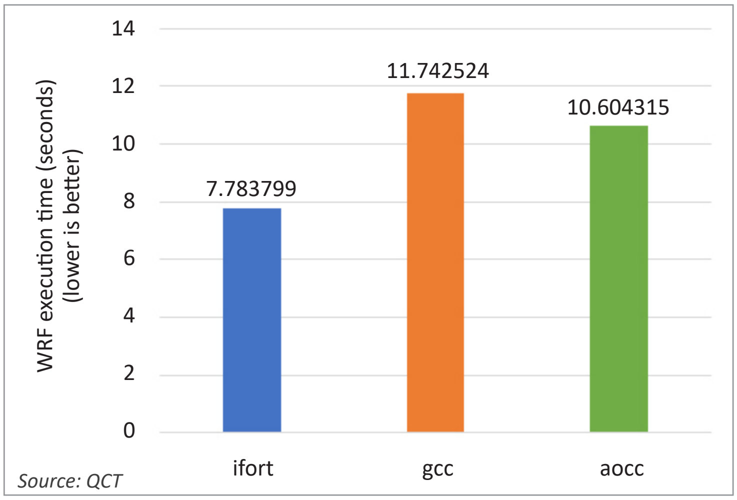
- -O3
  Adapted from the WRF default GNU compiler setting to CLANG/FLANG settings, and change -O2 to -O3.
- -Mbytwswapio
  Ensure the endianness of WRF input/output

Intel® compiler version 19.1.1 (v2020) (ifort)

- -xCORE-AVX512 (or -Xhost AVX512)
  Optimized for Cascade Lake Xeon® 8280, utilizing the full 512-bit SIMD instruction set.

Benchmark Results

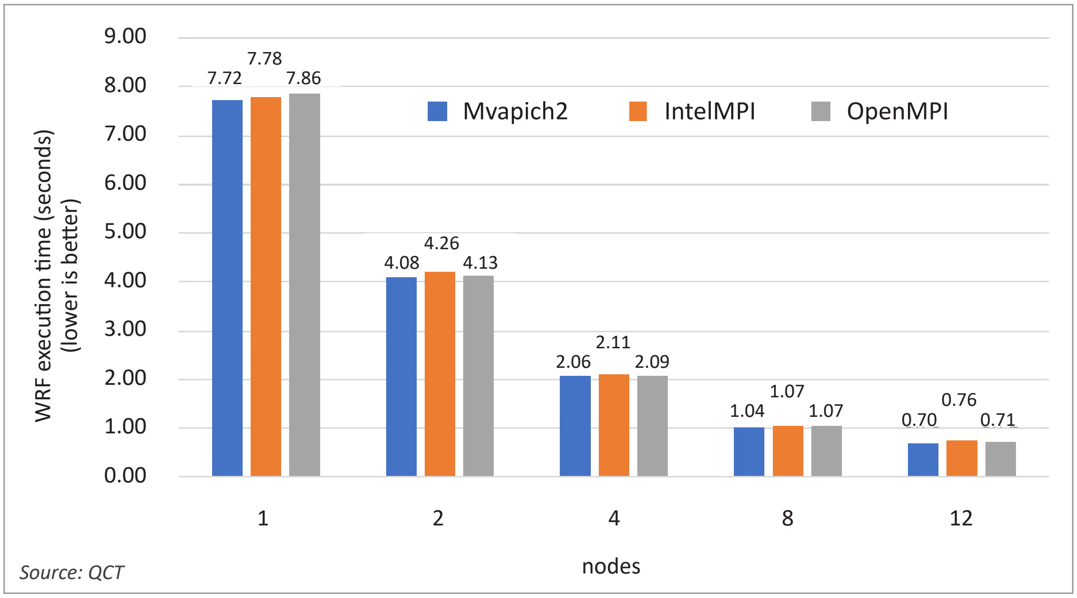
The first to be measured was WRF performance across popular compilers. Among the three compilers that were used to compile WRF and the corresponding libraries with, the Intel® compiler performs best, and leads other counterparts by more than 25%. Figure 2 shows the average execution time of each computation timesteps of WRF. Intel®-compiled WRF has ~ 25% less execution time compared to the other two. Figure 2 shows these results.



Source: QCT

Figure 2. WRF performance of intel, GCC, and AOCC compilers on one dual Xeon 8280 node.

Next to be investigated are the communication libraries. Figure 3 shows that with the integration of Infiniband Mvapich2 (v2.3.4) libraries decrease WRF execution time by ~ 5% as compared to the Intel® MPI (v2020 update 1) and the OpenMPI (v4.0.3) libraries.



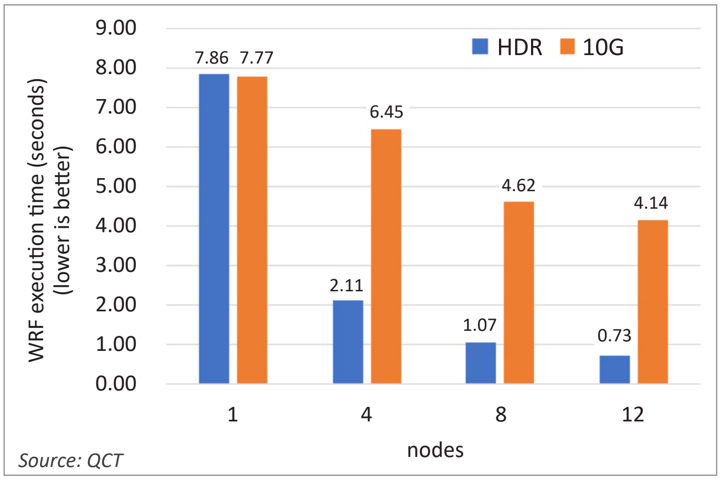
Source: QCT

Figure 3. WRF performance of OpenMPI and Mvapich libraries on 1-12 dual Xeon Xeon 8280 node. Mvapich2 and OpenMPI are compiled with the Intel compiler.

Impact of latency

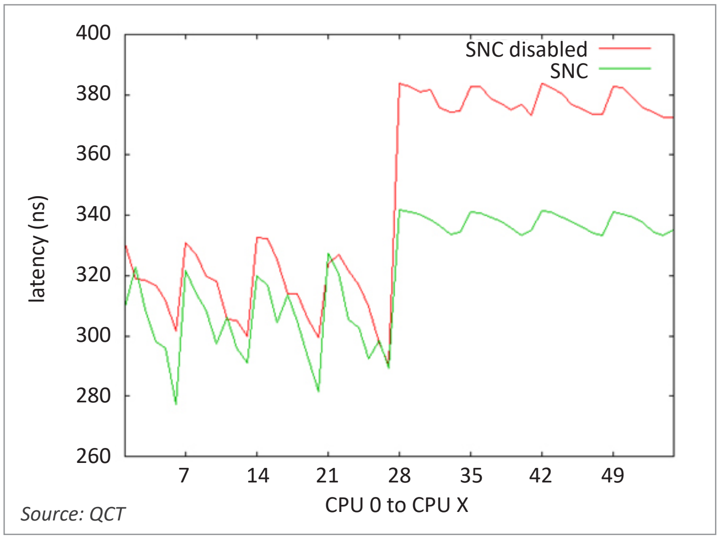
Next to be investigated was the comparison of Infiniband and Ethernet. WRF was executed on 1, 4, 8, and 12 nodes over Infiniband HDR and 10G Ethernet to examine the impact of interconnect latency on its performance. The node-to-node latency of Infiniband HDR starts from 1.01 microseconds of 1-Byte packet size, and 10G Ethernet starts from 8.7 microseconds. WRF performs three times better over Infiniband on four nodes than over Ethernet, and approximately six times better on 12 nodes. Figure 4 shows these results.

OpenMP allows different cores to share the same segment of memory. The performance of WRF is best with OMP_NUM_THREADS=4 and decreases more than 10 percent when OMP threads exceed four. The trend of increasing WRF performance is attributable to the four dips of latencies within the sockets (28 cores) as shown in Figure 5 shows a particular group of low-latency cores could improve the WRF performance. Also, WRF divides the sub-domains by the OMP Thread number. An OMP that cannot be wholly divided by 28 would result in a subdomain that needs to use cores on both sockets, which increases to core-to-core latency drastically. The decrease in performance when OMP_NUM_THREADS exceeds four shows the impact of the latency increase by crossing CPU sockets on WRF. One should take a careful arrangement of process affinity to CPU cores to avoid performance drop. Figure 6 below shows the performance as a function of the number of OMP threads.



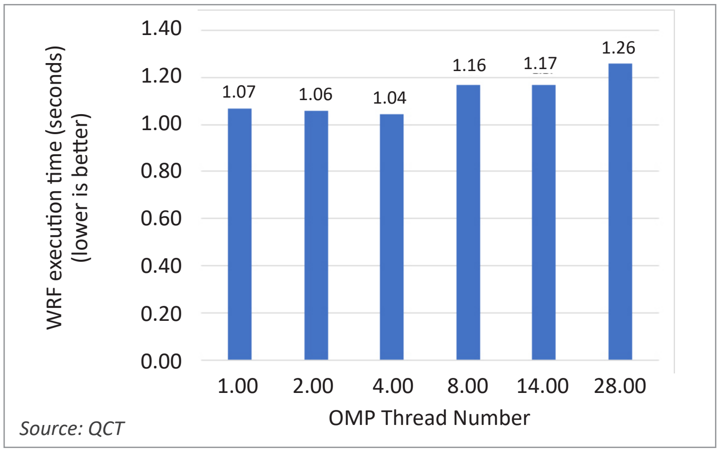
Source: QCT

Figure 4. WRF performance over Infiniband HDR and 10G Ethernet.



Source: QCT

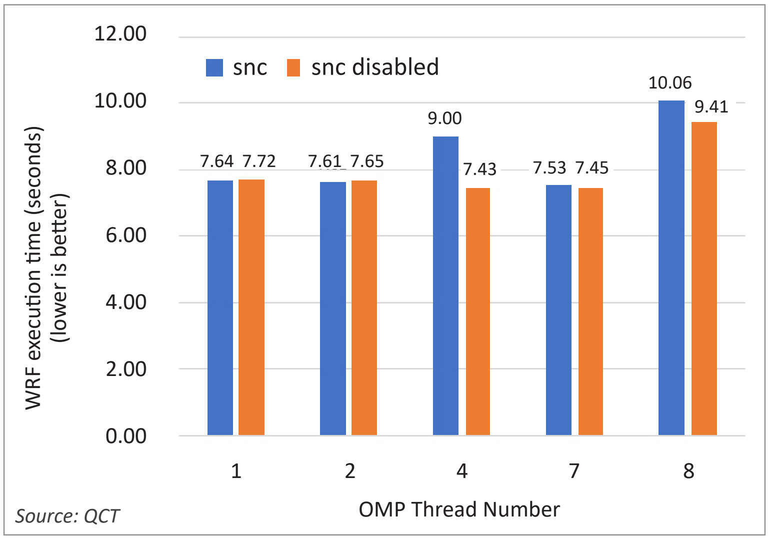
Figure 5. Core-to-core latencies of the dual Xeon 8280 system.



Source: QCT

Figure 6. WRF performance of OpenMP threads arrangements.

Next, QCT further investigated the WRF performance when sub-Numa clustering (SNC) was turned on. SNC allows a single Xeon® 8280 CPU to split into two groups of cores and thus decrease to core-to-core latencies within the sub-numa domain, as shown in Figure 7. QCT found turning on the SNC increases the WRF performance by 1-2 percent when OMP threads are less than 2. But the performance deteriorates drastically on threads number four because four cannot be divided evenly into 14 cores and has to run across two sub-Numa domains. The experiments show the importance of grouping the low-latency cores and avoiding the imbalanced OpenMP partitioning of WRF subdomains.



Source: QCT

Figure 7. WRF performance with sub-numa cluster(SNC) enabled.

Summary of WRF Benchmarks

The performance of WRF V4.1.5 highly relies on the compiler and the latencies between processors and interconnect. The Intel® compiler shows excellent execution performance for Fortran codes. The test on interconnect fabric and protocol, as well as the communication between CPU cores, shows the impact of increased latencies on WRF execution time. QCT highly recommends using Infiniband and group the adjacent OMP threads in low-latency memory (such as cache on each CPU) to reduce the impact on intercommunication.

QCT Expertise

QCT can work with leading research and commercial organizations to lower the Total Cost of Ownership by supplying highly tuned applications that are optimized to work on leading-edge infrastructure. By reducing the time to get to a solution, more applications can be executed, or higher resolutions can be used on the same hardware. QCT also has experts that understand in detail various HPC workloads and can deliver turnkey systems that are ready to use. For customers that wish to modify source code or that develop their own applications, QCT supplied highly tuned libraries and extensive guidance on how to get the most out of your infrastructure, that not only includes servers, but networking and storage as well.

For more information on how QCT can help you to maximize your HPC environments, please visit: https://go.qct.io/solutions/data-analytic-platform/qxsmart-hpc-dl-solution/

References

Kyle, A, 2018, “Weather Research and Forecast (WRF) Scaling and Performance Assessment NCAR SIParCS Program”, https://akirakyle.com/WRF_benchmarks/results.htm