GPU PERFORMANCE STUDY FOR THE WRF MODEL ON THE SUMMIT SUPERCOMPUTER

Jeff Adie, *NVIDIA*, Gökhan Sever, Rajeev Jain, *DOE Argonne NL*, and Stan Posey, *NVIDIA*

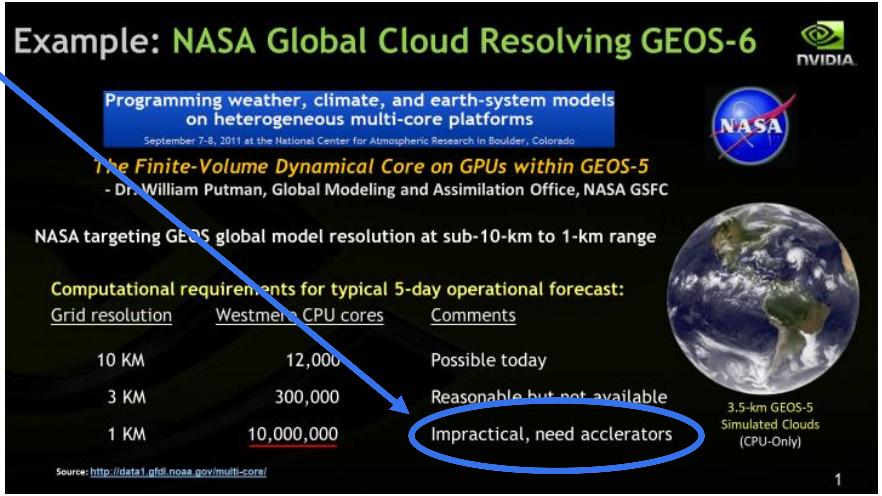
EARTH SYSTEM NEEDS KEEP GROWING

"Future Earth system models will need over 1000 times the computational power of today" – ESPC Position Paper on HPC Needs, 2017[1]

		Observations	Models
Today:	Volume	20 million = **2 x 10^7**	5 million grid points 100 levels 10 prognostic variables = **5 x 10^9**
	Type	98% from 60 different satellite instruments	physical parameters of atmosphere, waves, ocean
		Observations	**Models**
Tomorrow:	Volume	200 million = **2 x 10^8**	500 million grid points 200 levels 100 prognostic variables = **1 x 10^{13}**
	Type	98% from 80 different satellite instruments	physical and chemical parameters of atmosphere, waves, ocean, ice, vegetation
		Factor 10 per day	**Factor 2000 per time step**

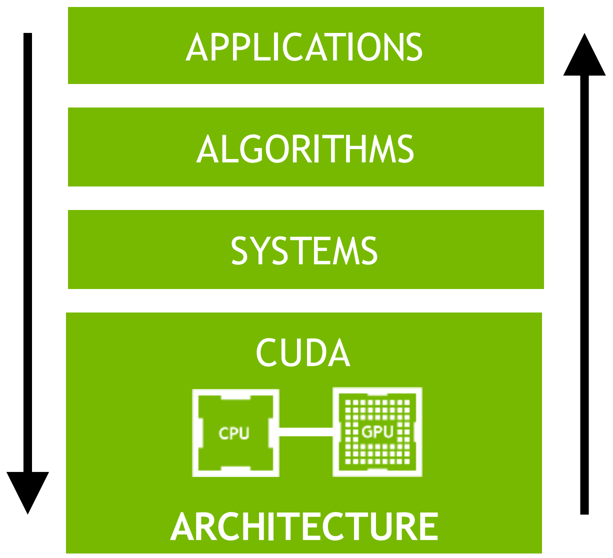
ECMWF – 2000x needed by 2032[2]

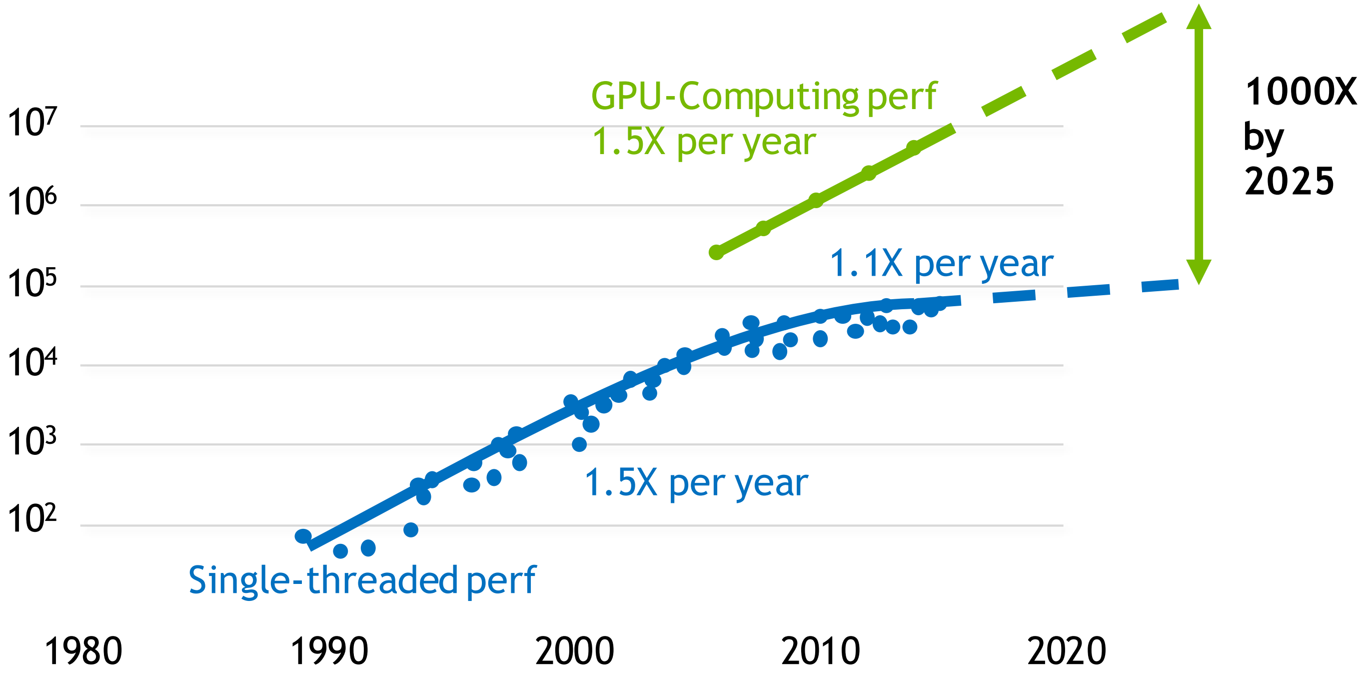
NASA – 10M CPU cores – "impractical"[3]



1:: https://doi.org/10.7289/V5862DH3 3,4 – refer to AVEC Benchmark report
2: Challenges of Getting ECMWF's Weather Forecast Model (IFS) to the Exascale – G. Mozdzynski, ECMWF 16th HPC Workshop
3: http://data1.gfdl.noaa.gov/multi-core/

RISE OF GPU COMPUTING

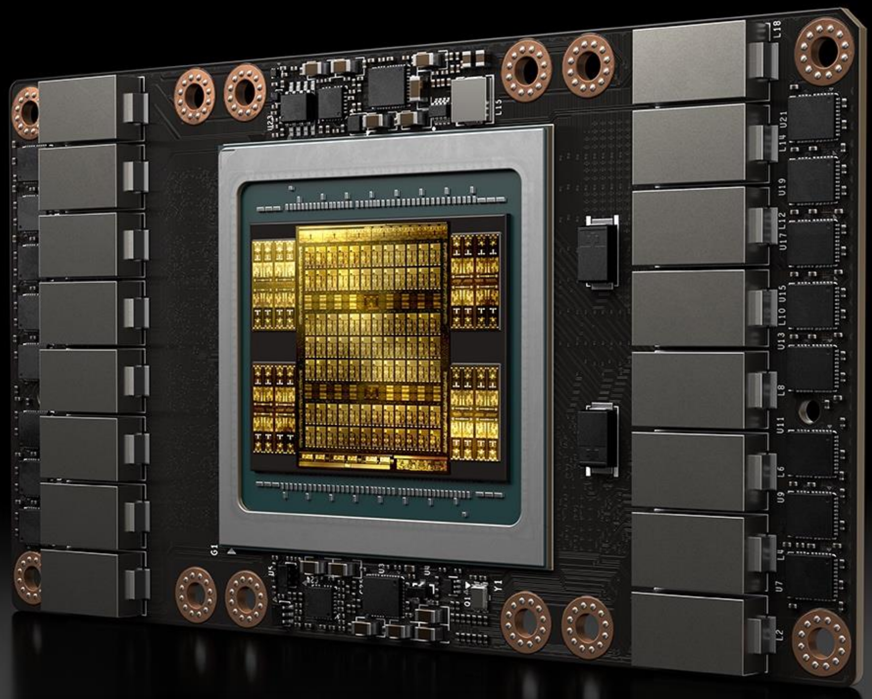




Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

SUMMIT - WORLD'S FASTEST SUPERCOMPUTER

Summit Becomes First System To Scale The 100 Petaflops Milestone



144 PF
HPC

3 EF
AI

4,608
Nodes

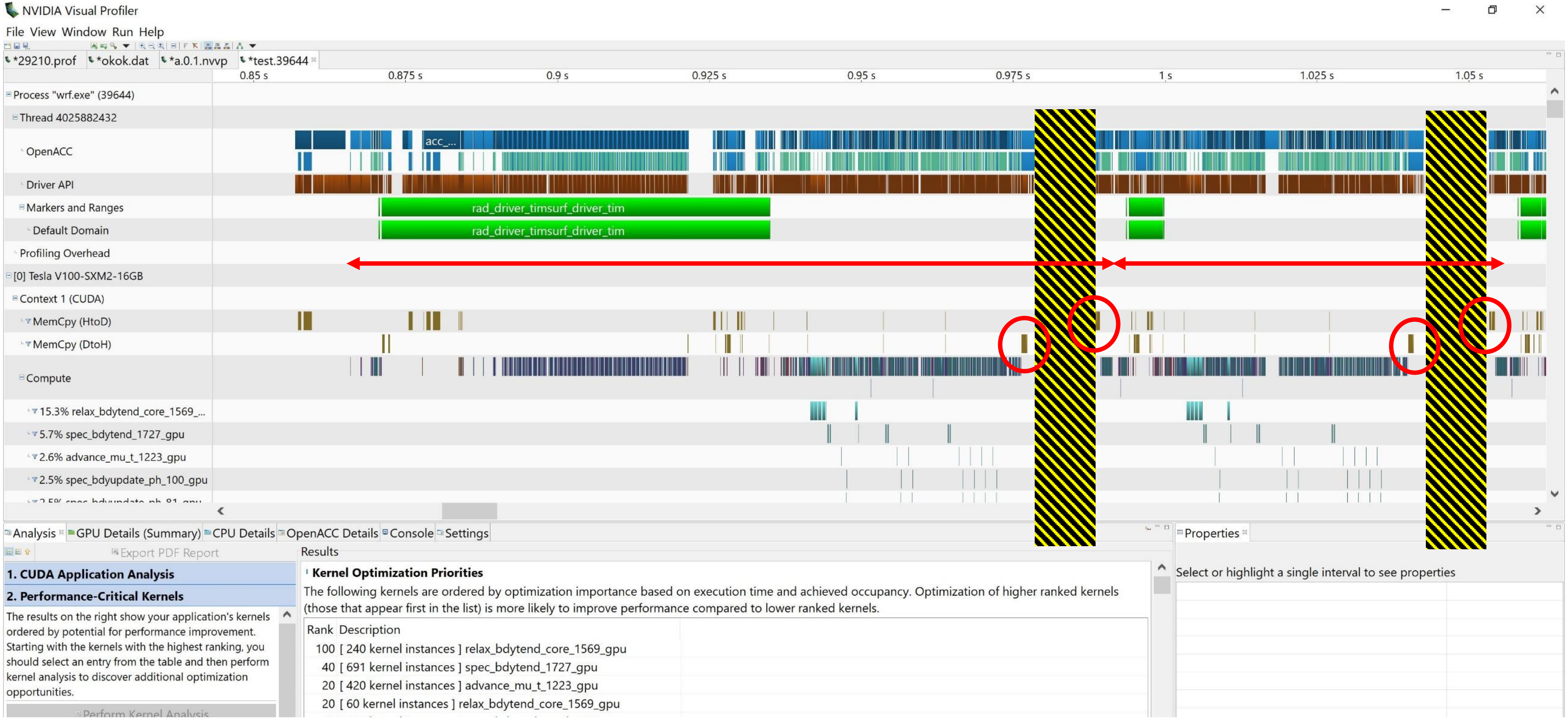
27,648
Volta Tensor Core GPUs

ACCELERATED COMPUTING CHALLENGES

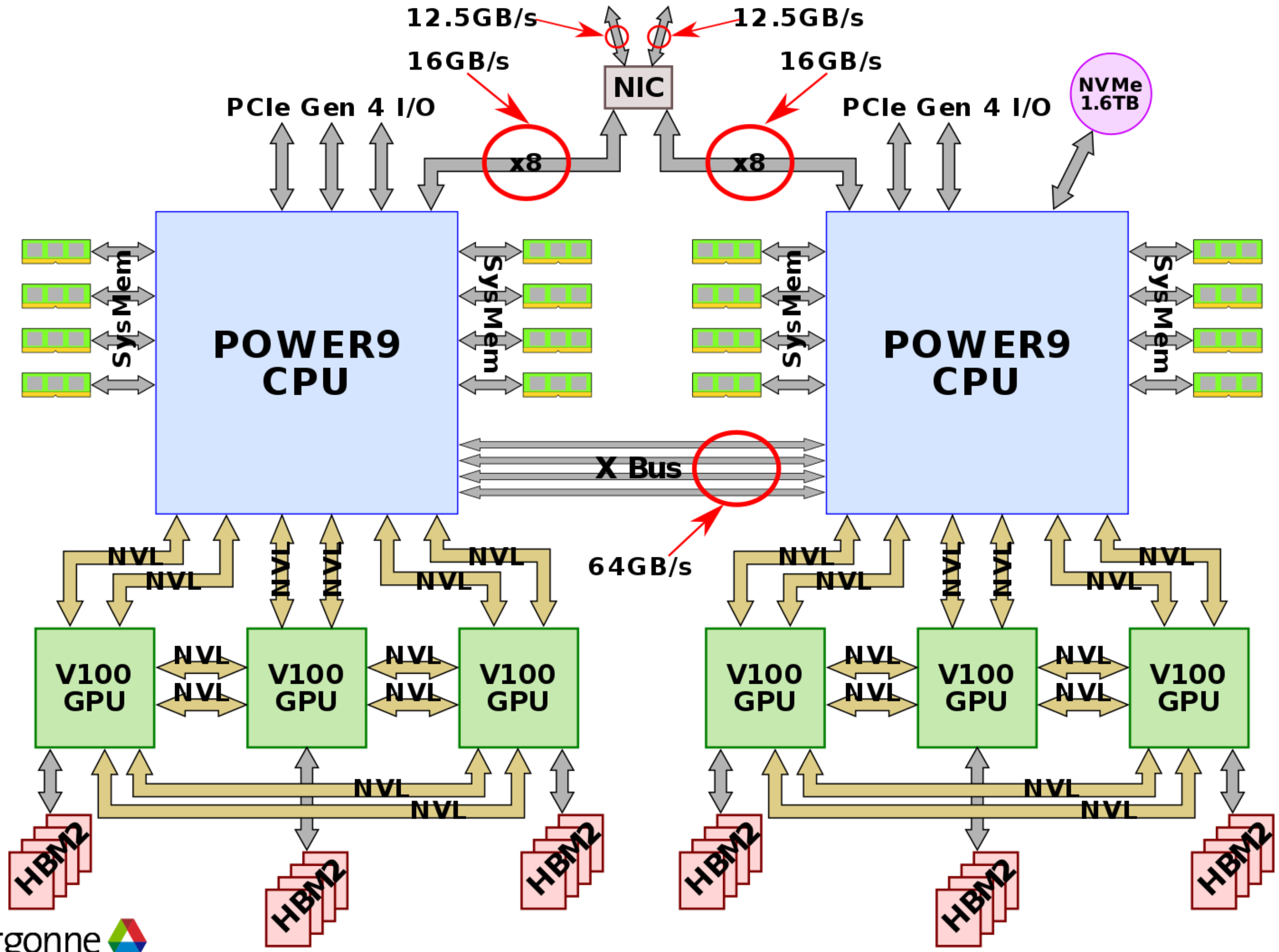
Not quite plug-n-play!

- Software needs to support accelerator (or be ported to accelerator)
  - OpenACC provides an ideal mechanism for portable, easy porting
- CPUs and accelerators have different physical memories. Data transfer has to occur, even if it is hidden from the programmer (CUDA unified memory, OpenACC managed data, CAPI, etc)
  - Minimise memory transfers (keep data on GPU for reuse wherever possible)
  - Faster memory transfers - NVLink

ACCELERATED COMPUTING CHALLENGES



SUMMIT NODE ARCHITECTURE



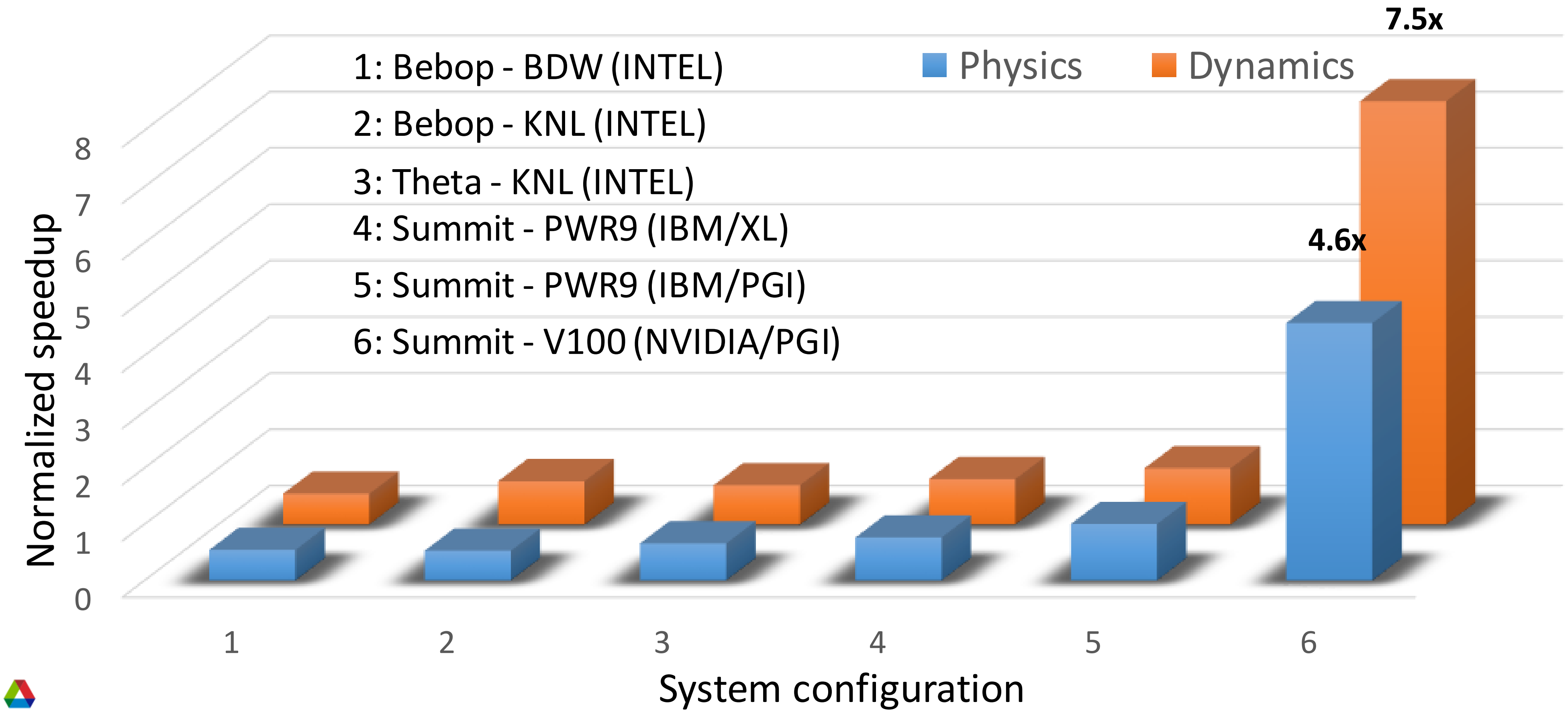
WRF PERFORMANCE STUDY

Methodology

- Ported NVIDIA OpenACC WRF code (based on WRF V3.7.1) to POWER9
  - PGI 19.1 Compiler
  - IBM Spectrum MPI (PAMI disabled)
- Model testing using em_les ideal case for dynamics only (384x384x42, ~6M cells)
  - Scaled vertically for GPU occupancy (384x384x252, ~37M cells)
  - Scaled horizontally for multi-node studies (3840x3840x252, ~3.7B cells)
  - Full model physics test with Thompson MP, RRTM/Dudhia, YSU PBL, Revised MM5+TDS

WRF SINGLE NODE PERFORMANCE

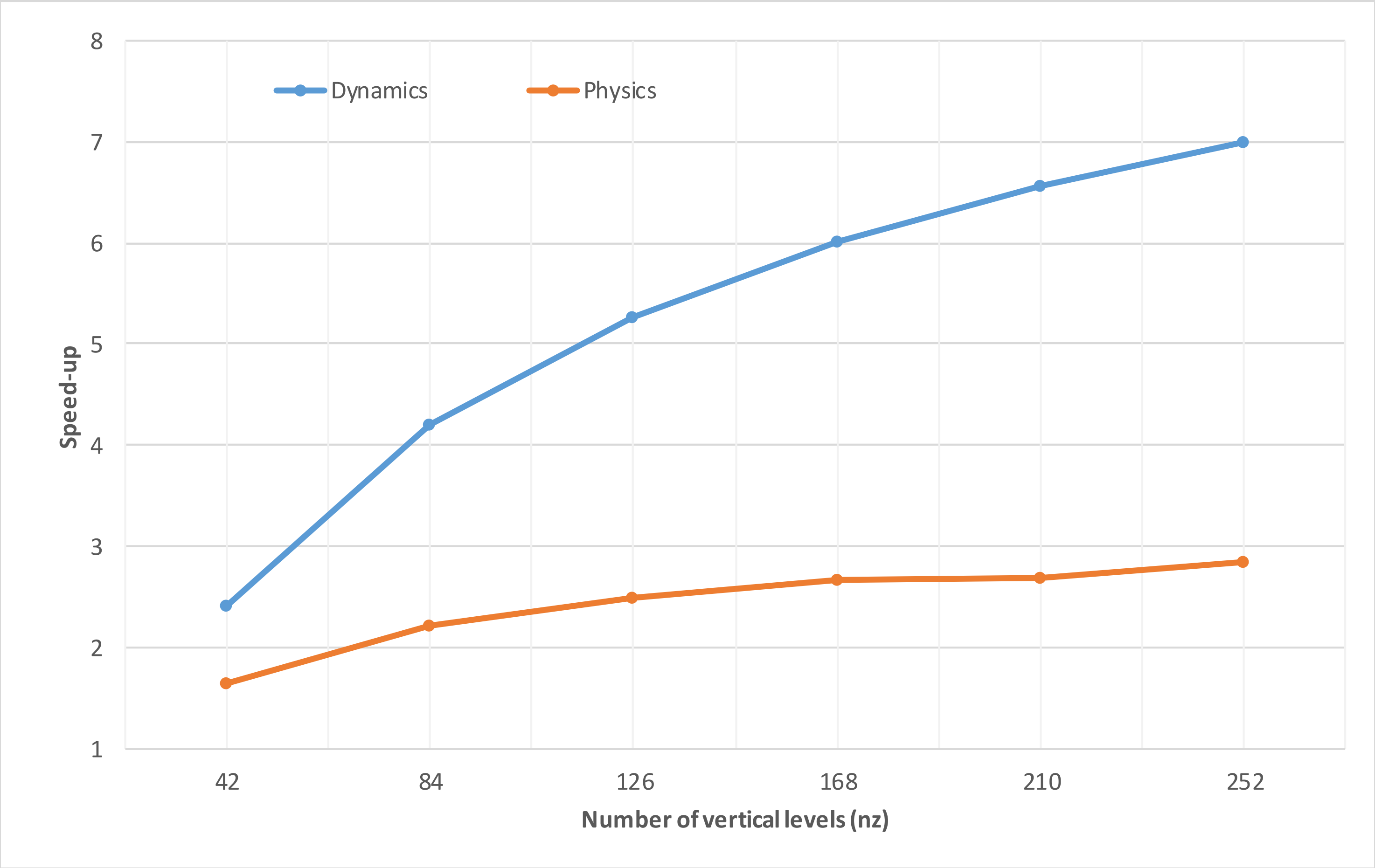
Benchmarks of single-node setups on DOE clusters



Argonne NATIONAL LABORATORY

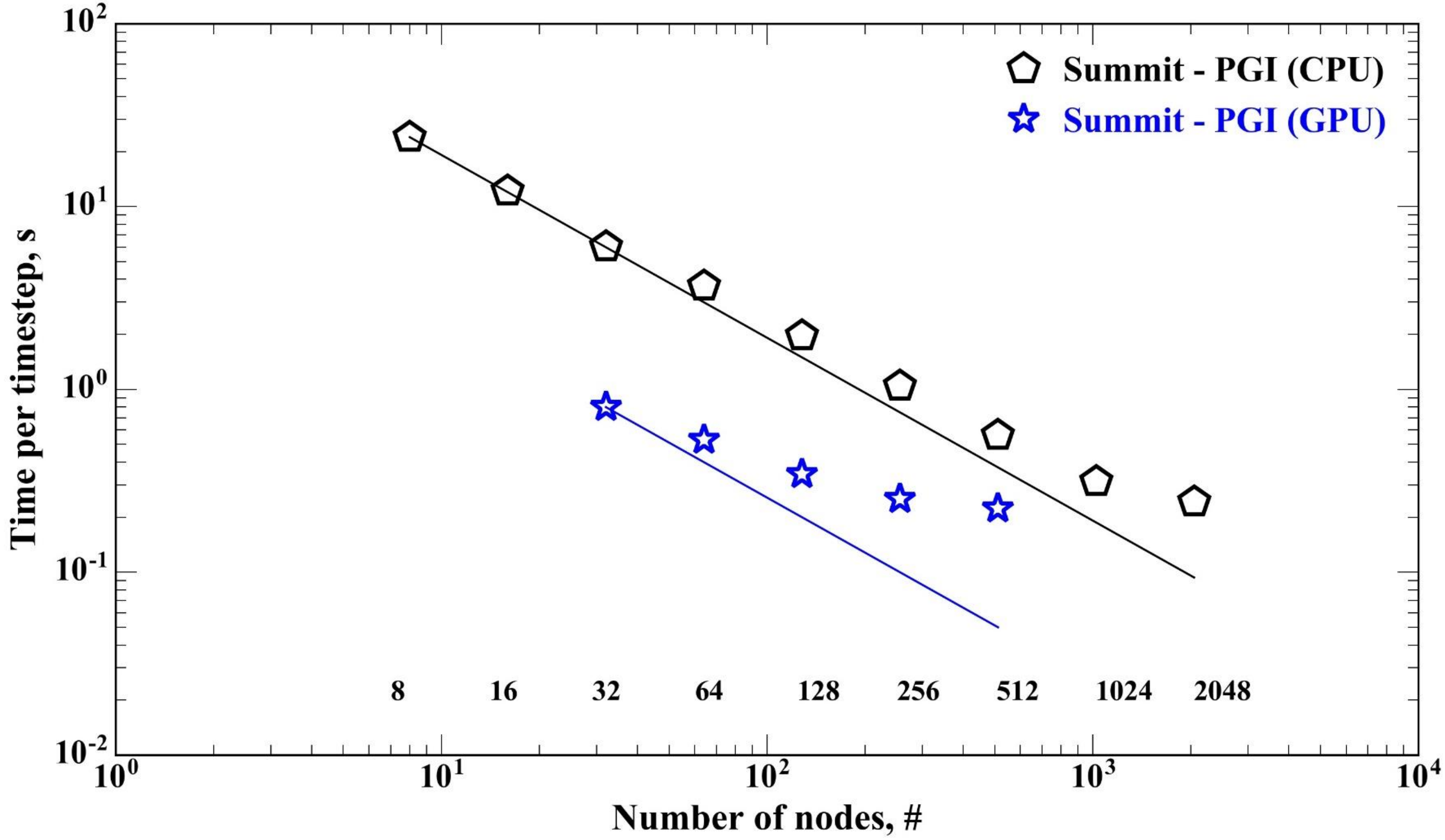
WRF SINGLE NODE PERFORMANCE

Speedup sensitivity to amount of work



WRF SCALING

Dynamics only, 4B cells

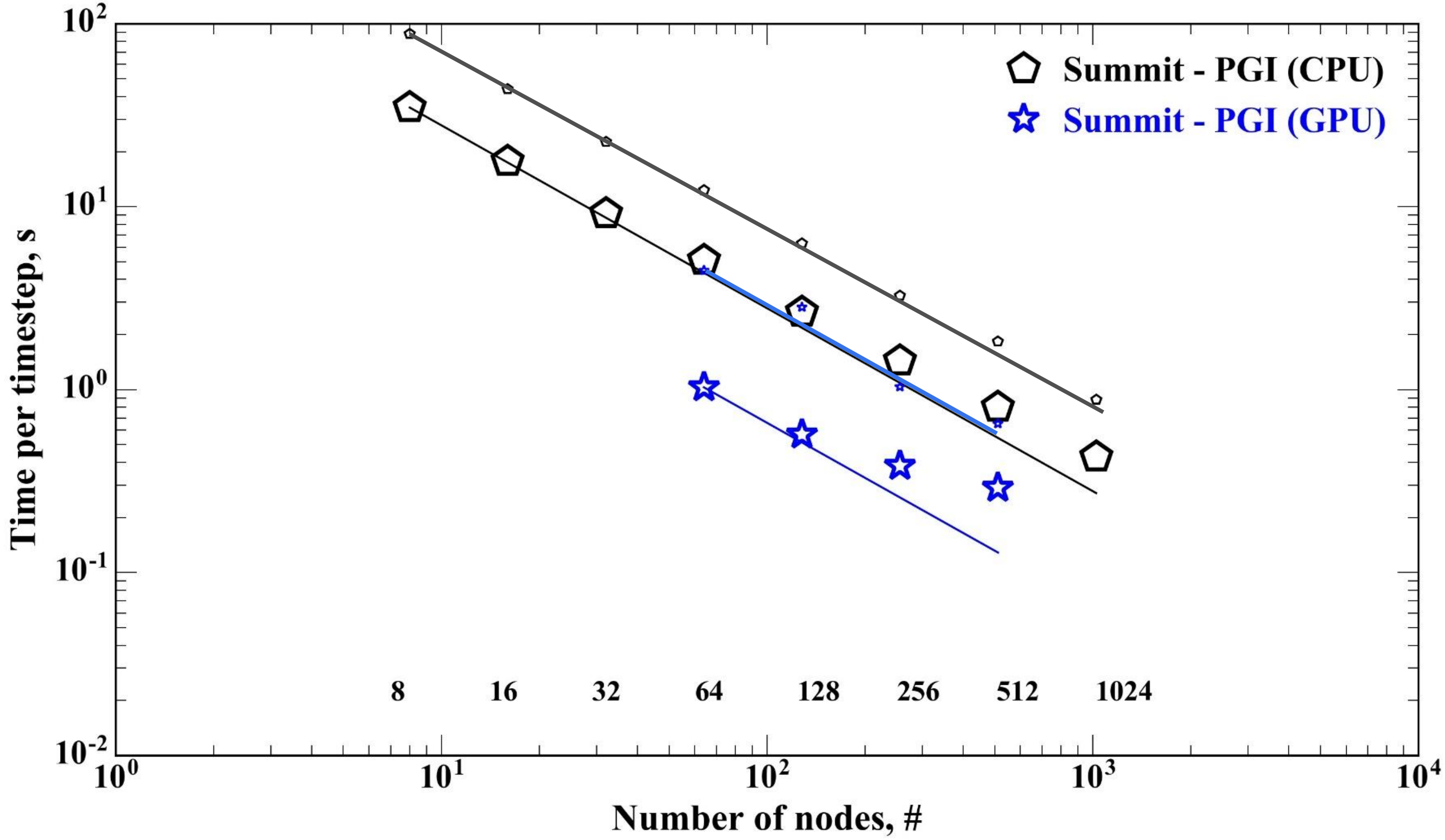


Argonne NATIONAL LABORATORY

 NVIDIA.

WRF SCALING

Full Model, 4B cells

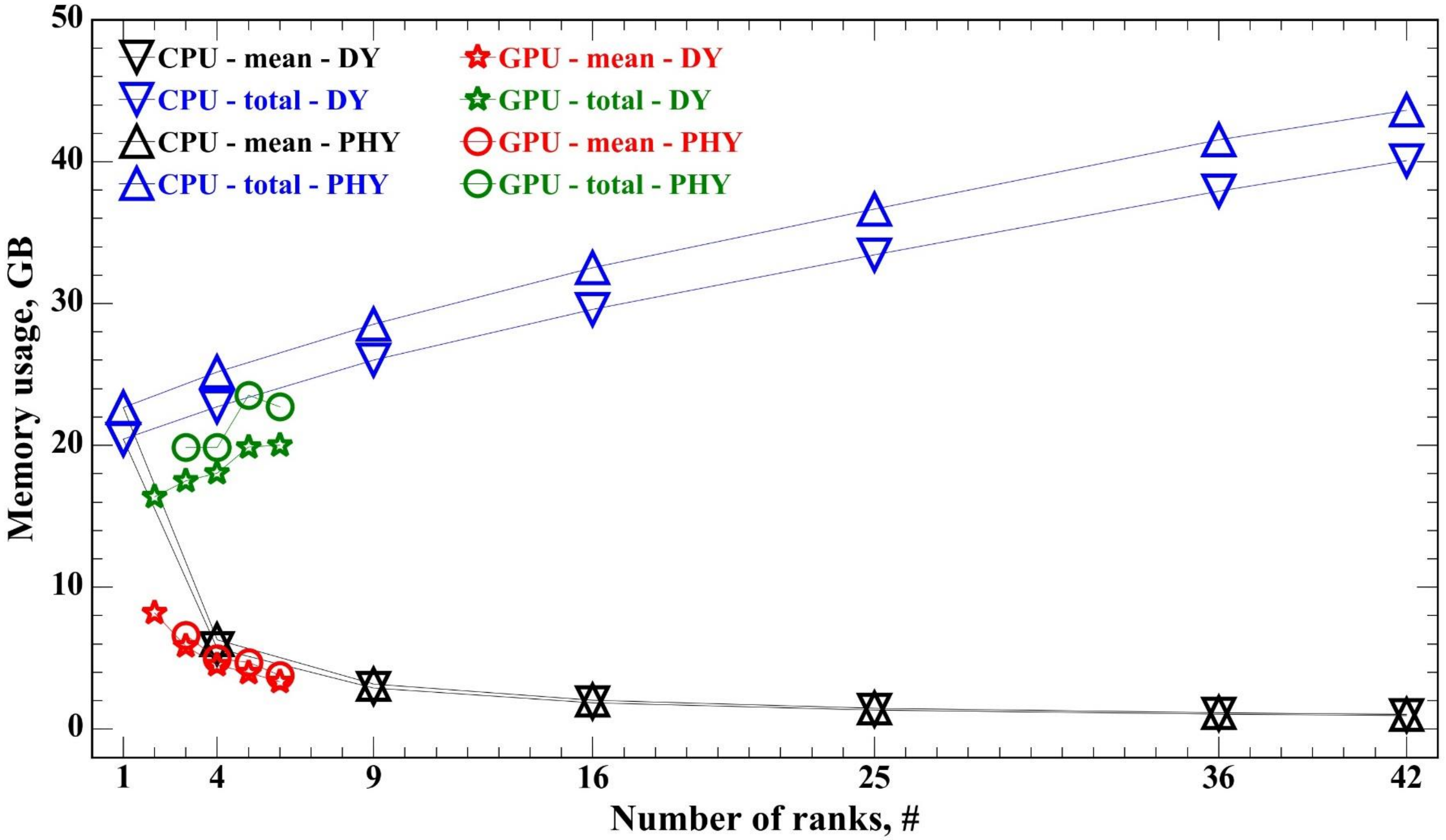


Argonne NATIONAL LABORATORY

NVIDIA

WRF MEMORY USAGE

Single Node



WRF PERFORMANCE STUDY

Preliminary Conclusions

- Single node shows 7.5x speedup on Dycore, 5x on full model with Physics
- Multi node scaling tested to 512 nodes (3,072 GPUs)
  - Scaling above 64 nodes (384 GPUs) limited by model size (GPU Occupancy)
- Currently not using CAPI between CPUs and GPUs
  - GPU memory limits model size
  - Cannot use PAMI for MPI acceleration
  - Requires code changes to the RSL_LITE comms

WRF PERFORMANCE STUDY

Future Work

- Perform strong and weak scaling tests
- Get CAPI working
  - Remove GPU Memory size limitation
  - Allow for PAMI support for faster communications
- Port UCM (Urban Canopy Model) to GPU
- Look at supporting later WRF versions (V4)

WRF PERFORMANCE STUDY

Acknowledgements

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

The computational resources on Summit for the account # CSC307 are provided via the UrbanECP project

Thanks to Bob Walkup (IBM) and Todd Hutchinson (TWC/IBM) for help with the WRF configuration for POWER.

We appreciate the help from the OLCF support team for various technical problems.

Akira Kyle's script (wrf_stats.py) is utilized to extract average benchmark timings.

NVIDIA

Argonne
NATIONAL LABORATORY