Performance and scaling of WRF and CCSM on Ranger, Hydra and Bluefire

Davide DelVento, Mike Page, Juli Rew, National Center for Atmospheric Research (NCAR)
Raffaele Montuoro, Texas A&M University (TAMU) - Supercomputing Facility

What are WRF and CCSM?
- Weather Research and Forecasting (WRF) Model v3.6.1.1
  - A next-generation mesoscale numerical weather prediction system
  - Serves both operational forecasting and atmospheric research needs
- Community Climate System Model (CCSM) v4.0a37
  - A fully-coupled, global climate model
  - Includes atmosphere, land surface, ocean, and sea ice
  - Simulates the Earth's past, present, and future climate states

Benchmarked Supercomputers

**Texas Advanced Computing Center (TACC):**
- Ranger
  - 2.3GHz AMD Barcelonas processors – quad core – 64bit
  - 32GB RAM per node
  - Infiniband interconnect
  - 2 microseconds latency
  - 50Gbps peak bandwidth per node (full duplex)
  - 3.936 16-way SMP nodes
  - Linux CentOS kernel 2.6.9
  - IBM XL Fortran Enterprise Edition for AIX, V10.1
  - Compiler and Runtime Options:
    - `-O3` `-qhot` `-qarch=auto` `-qtune=auto` `-qcache=auto`
    - `4.7GHz IBM Power6 processors – dual core – 64bit`
    - `4GBs peak bandwidth per node` (full duplex)
    - `127 32-way SMP nodes`
    - `AIX 5.3`
    - `IBM XL Fortran Enterprise Edition for AIX, V11.1`

**NCAR System: Bluefire**
- 4.7GHz IBM Power6 processors – dual core – 64bit
- 64GBs peak bandwidth per node
- IX Infiniband interconnect
- 1 microsecond latency
- 50Gbps peak bandwidth per node (full duplex)
- 1.6MSMs (multi core) latency
- 48 16-way SMP nodes
- AIX 5.3
- IBM XL Fortran Enterprise Edition for AIX, V10.1

**TAMU Supercomputing Facility’s System: Hydra**
- 1.9GHz IBM Power5+ processors - dual core – 64bit
- 32GB RAM per node
- 2X HPS links
- 16 microsecond latency
- 4 GBs peak bandwidth per node (full duplex)
- 48 16-way SMP nodes
- AIX 5.3
- IBM XL Fortran Enterprise Edition for AIX, V11.1

WRF Model

Multi-level parallelism:
- WRF can run in shared memory (OpenMP), distributed memory (MPI), or hybrid OpenMP+MPI - we used MPI only

Representation of the atmosphere:
- Variables of state discretized over regular Cartesian grids
- 2-D horizontal domains (X, Y) decomposed over processors (no Z)

Integration and interprocessor communication:
- Implicit high-order Runge-Kutta time-step integration scheme (in X, Y)
- Implicit solvers in the vertical coordinate (Z)
- Communication is logically nearest-neighbor
- 144 two-way messages between neighboring processes per time step

Compiler and Runtime Options

Hydra and Bluefire:
- `-O3` `-qhot` `-qarch=auto` `-qarch=auto` `-qarch=auto`
- Processor binding
- 64KB page size for Bluefire
- 4KB page size for Hydra
- no SMT

**Ranger**
- `-fast` `-stl` `-barcelona-64` `-Mvect=nosse` `-Mvect=fastsse` `-Mvect=nosse`
- `-O3` `-qhot` `-qarch=auto` `-qtune=auto` `-qcache=auto`

WRF Run Case 1 – low resolution

**CONUS 12km**
- CONUS (continental US) October 24th, 2001
- Uses the Eulerian Mass (EM) dynamics
- Full case is 48-hour, 12km resolution case
- Benchmark is 3 hours, starting from the 24th one
- 425 by 300 grid with 35 vertical levels (Z)
- 150 time steps, 72 sec each
- 30 Giga Floating Point Operations per time step
- Total memory: ~ 5 GB (increases with proc count)

WRF Run Case 2 – high resolution

**CONUS 2.5km**
- CONUS (continental US) June 4th, 2005
- Uses the Eulerian Mass (EM) dynamics
- Full case is 6-hour, 2.5km resolution case
- Benchmark is 3 hours, starting from the 3rd one
- 1500 by 1200 grid with 35 vertical levels (Z)
- 720 time steps, 15 sec each
- 412 Giga Floating Point Operations per time step
- Total memory: ~ 60 GB (increases with proc count)

WRF Runs Analysis

- WRF speed is an adimensional number that represents simulated seconds per elapsed wallclock second
- WRF scales fairly well
- Scaling for larger datasets is better
- For each integration time step, with the same processor count, **Hydra uses almost half the wallclock of Ranger**
- For each integration time step, with the same processor count, **Bluefire uses almost one third the wallclock of Ranger**
- A given speed may be achieved with a given processor count on range
- The same speed may be achieved with about half processor count on Hydra and about a quarter processor count on Bluefire

CCSM = The Model

CCSM contains four model components and a coupler
- Community Atmospheric Model (CAM), normally the "dominant" component in the configuration
- Community Sea Ice Model (CISM)
- Community Land Model (CLM)
- A version of the Parallel Ocean Program (POP)
- The coupler (CPL), which controls the execution of CCSM, controlling the flow of data between the various components
- CCSM is a simple-executable version of the model (compared with CCSM4, which is MPI-based)

Multi-level parallelism:
- Shared memory (OpenMP)
- Distributed memory (MPI)
- Hybrid OpenMP+MPI

Representation of the atmosphere:
- We used a resolution of 0.47 x 0.63 degrees
- The atmospheric grid is 766 x 286 in the horizontal and 26 levels in the vertical

CCSM4 on Bluefire and Ranger

WRF speed is an adimensional number that represents simulated seconds per elapsed wallclock second
- WRF scales fairly well
- Scaling for larger datasets is better
- For each integration time step, with the same processor count, **Hydra uses almost half the wallclock of Ranger**
- For each integration time step, with the same processor count, **Bluefire uses almost one third the wallclock of Ranger**
- A given speed may be achieved with a given processor count on range
- The same speed may be achieved with about half processor count on Hydra and about a quarter processor count on Bluefire

Conclusions

- We are able to run WRF and CCSM on Ranger
- WRF scales pretty well on Ranger
- CCSM v4 scales fairly well, especially when load balance is tuned
- Better than CCSM v3
- Large datasets scale better for both WRF and CCSM

Acknowledgment, References, and Contacts

- **Acknowledgments:**
  - We thank John Micikevics for providing the CONUS benchmarks
  - We thank Rich Loft for providing encouragement in this study and including us in the team of early Ranger users
  - We thank Texas Advanced Computing Center, who provided the study under TerraGrid project TG-CCS050010
  - We also thank John R. Boisvert for providing early user access and Karl W. Schultz for providing performance tips and initial hand holding needed to compute on Ranger

- **References:**
  - http://www.tacc.utexas.edu
  - http://www.cscs.ucar.edu
  - http://www.wrf-model.org

- **Contacts:**
  - Davide DelVento: ddelvento@tamu.edu
  - Mike Page: mpage@ucar.edu
  - Juli Rew: julian@ucar.edu
  - Raffaele Montuoro: montuoro@tamu.edu