UTILIZING INTEL XEON PHI COPROCESSORS CONCURRENTLY WITH INTEL XEON PROCESSORS TO ACCELERATE WRF SIMULATION THROUGHPUT

Daniel Dietz
Kimberly Hoogewind

January 8, 2015
CONTE COMMUNITY CLUSTER

• Built June 2013
  – 580 compute nodes
  – Intel Xeon-E5 Processors
  – Intel Xeon Phi Coprocessors
  – 64 GB Memory
  – 40Gbps FDR10 Infiniband
  – Lustre Scratch Filesystem

• Priority access to number of cores purchased

• Standby access to the rest of the cluster
• Coprocessor board from Intel
• Many Integrated Core (MIC)
  • 60 Intel x86 cores, 4 threads per core
  • 8GB memory
  • Runs Linux OS instance on each board

• WRF-ARW code (since version 3.5) supports running natively on Phi coprocessors
  • Only one available microphysics scheme (WSM5) optimized for Phis

• How can Phis and host processors be fully utilized?
  • Trivial solution: run two cases at once
When a large number of simulations are required

Example:
- Regional climate modeling
- Multi-decadal sequence of short, daily re-initialized forecasts
- 20 years = 7300 WRF runs
• WRF-ARW version 3.6
  • CONUS domain
  • 5 km horizontal grid spacing
  • 50 vertical levels, 5 hPa model top
  • 604 x 999 x 50 = 30,169,800 grid points
• Thompson MP (mp_physics=8)
• IC/BCs provided by GFDL-CM3 global climate model
• No intermediate nesting despite large resolution jump
SCALING STUDY

TESTING CONFIGURATION

• Tested two microphysics (MP) schemes
  – Phi Optimized WSM 5-Class scheme (mp_physics=4)
  – Un-optimized Thompson scheme (mp_physics=8)

• Intel 13.1.1.163 compilers
• Intel MPI 4.1.1.036

• Hybrid MPI+OpenMP strategy
  – 2 MPI tasks per node/phi
  – 8 OpenMP threads per node MPI task
  – 90 OpenMP threads per Phi MPI task
    – 3 threads per Phi core
    – 3x30 tiling strategy
CONSIDERATIONS

• Needs to fit within 4 hour standby queue wallclock limit
  – Including pre- and post-processing

• File I/O a significant problem with Phis
  – 1 hourly history output – 30 history files per run
    – 60GB+ output per run
  – Parallel-netcdf required at minimum
  – Host runs: ~10% of run time
  – Phi runs: ~45% of run time

• Solution: Use I/O quilting
  – 2 quilting nodes (4 phis)
  – Brings file I/O time in line with host nodes
SCALING STUDY

CONCURRENT EXECUTION

Launch job → Pre-process → WRF run on Phis → Post-process

WRF run on host

Wait for each run to finish
RESULTS

SCALING

Time step comparison (Thompson)

- Red line represents CPU
- Blue line represents Phi

The graph shows the time step comparison for different numbers of compute nodes, with CPU and Phi performance indicated by the respective lines.
RESULTS

SCALING

Run time comparison (Thompson)

- CPU
- CPU (quilting)
- Phi
- Phi (quilting)
RESULTS

THE RIGHT WIDTH

• 6 total nodes (4 compute, +2 for quilting)
  – Host processors: ~135 minutes
  – Phi coprocessors: ~135 minutes
  – Minimizes idle time waiting around for the slower run

• Wallclock considerations
  – Fully utilize 4 hour limit while minimizing nodes
  – 30 minutes for pre-processing on host node
  – 30 minutes for post-processing after runs complete
  – ~3.5 hours of walltime – 30 minute safety buffer
RESULTS

LIMITATIONS AND PRACTICAL CONCERNS

• NetCDF history output doesn’t work with quilting and Thompson MP
  – Used binary output – may not work for everyone

• Phis increase Conte’s node price by 66%

• Must be able to wait during busy periods

• Doesn’t help with real-time forecasts
CONCLUSIONS

• Run two WRF cases concurrently to fully utilize host processors and Phi coprocessors

• Optimized versus un-optimized microphysics
  - No surprise WSM5 completes faster than Thompson
  - Less complex MP scheme, and optimized

• Quilting I/O is a must to overcome poor Phi file I/O performance

• At right scale, this solution gives “BOGO” throughput
  - Can cut time to complete high-throughput project in half
• Figure out NetCDF with Thompson when quilting

• Implement code to optimize Thompson MP? (Mielikainen et al. (2014))
QUESTIONS?

Daniel Dietz
ddietz@purdue.edu

Kimberly Hoogewind
khoogewi@purdue.edu
RESULTS

SCALING

Time step comparison (WSM5)

- **CPU**
- **Phi**

![Graph showing time step comparison with CPU and Phi lines for different compute nodes.](image-url)
RESULTS

SCALING

Run time comparison (WSM5)

Time (minutes)

Compute nodes

CPU
CPU (quilting)
Phi
Phi (quilting)