Acceleration of WRF on the GPU

Daniel Abdi, Sam Elliott, Iman Gohari
Don Berchoff, Gene Pache, John Manobianco

TempoQuest
1434 Spruce Street
Boulder, CO 80302
720 726 9032
TempoQuest.com
Our Product

- TQI is a Weather Prediction Software and Analytics Company
- We Produce Micro-Weather Predictions for Custom Applications
- We Deliver and Support on Premise or via Software-as-a-Service
- Flagship Product: AceCAST-WRF
- The Breakthrough: 5X to 7X Acceleration Running the Weather Research Forecast (WRF) Model on Graphic Processing Units (GPU)
Our approach to Re-factoring

- WRF ported to run entirely on the GPU
- Profile and optimize most time consuming parts
- Avoid/minimize data transfer to/from GPU
- Leverage WRF registry to produce GPU code
- Pack halo data on GPU and send via infiniband
- Process multiple tiles and columns in a kernel
- Tiling to reduce memory consumption for radiation
Our approach to Re-Factoring

- Two branches: hybrid CPU + GPU vs pure GPU
- 7x difference in speedup between those two
- “Premature optimization is the root of all evil”
- Parallelize->Profile->Optimize->Rewrite & Repeat
- Try to avoid rewriting code->Harder to upgrade
Our approach

Physics code refactoring

• Existing code not suitable for GPU
  ○ Turn 1D processing to 3D processing manually -> fast but cumbersome + unmaintainable
  ○ Keep the 1D processing format -> convenient but slow most of the time.
  ○ Keep the 1D format but minimize data allocations in routines -> Efficient + maintainable
Horror code -> rewrite

Example horror code in nesting

CALL rsl_lite_to_child_info( ic, jc, flag)

DO WHILE (flag)

Pack hundreds of fields

....

CALL rsl_lite_to_child_info( ic, jc, flag)

ENDDO
Profile on P100 GPU - Before Optimization

Wrf dynamics profile: Before Optimization

- `memcpy_h_d`: 1.3%
- `rk_update_scalar`: 1.5%
- `calc_coef_w`: 1.5%
- `memcpy_d_h`: 1.8%
- `horizontal_press`: 2.4%
- `Other`: 37.0%
- `advance_w`: 26.4%
- `advance_mut`: 16.9%
- `advance_uv`: 7.8%
Profile on P100 GPU - After Optimization

Wrf dynamics profile: After Optimization

- advance_w: 14.7%
- advance_mut: 7.1%
- advance_uv: 7.3%
- horizontal_press: 1.9%
- memcpy_d_h: 1.2%
- zero_ten: 2.1%
- calc_p_rho: 1.8%
- memcpy_h_d: 2.0%

Other: 62.3%
Cost of data transfer- P100 GPU + Haswell CPU

Avoid data transfers

GPU speedup vs 1-core CPU on Pure GPU and Hybrid CPU-GPU modes

- 128x128-Pure
- 256x256-Pure
- 512x512-Pure
- 128x128-hybrid
- 256x256-hybrid
- 512x512-hybrid

Vertical levels:
- 20
- 40
- 60
- 80
- 120
- 160

Speedup:
- 0
- 1
- 2
- 3
CONUS results - Elapsed seconds / timestep
Results: GPU WRF Strong Scaling for CONUS 2.5 km

~5x Speedup Full Model: 4 x P100 vs. 4 x HSW
(1 x node) vs. (2 x nodes)

- Higher is Better

Performance [MM grid points / sec]

- ~5x Speedup Full Model: 4 x P100 vs. 4 x HSW
- (Haswell vs. P100)

CONUS 2.5 km Case on PSG Cluster - 4 nodes
Source: TQI – Abdi; Apr 18

- Based on WRF 3.8.1 trunk
- 1501 x 1201 grid x 35 levels
- Total 60 time steps, SP run
- Physics option modified:
  - WSM6
  - Radiation *off*
  - 5-layer TDS

- All WRF runs single precision
- PSG cluster node configuration:
  - 2 CPUs, 16 cores each
  - 4 x P100 GPUs
  - Or 4 x V100 GPUs
  - CPU-only 1 MPI task each core
  - CPU+GPU 1 MPI task per GPU

CONUS 2.5km Source: [http://www2.mmm.ucar.edu/wrf/bench/benchdata_v3911.html](http://www2.mmm.ucar.edu/wrf/bench/benchdata_v3911.html) (Note “Physics options modified” in side bar)
~7x Speedup Full Model: 4 x V100 vs. 4 x HSW

(1 x node)  (2 x nodes)

4 x V100 Result:
- 35% Speedup vs. P100
- 690% Speedup vs. HSW

Higher is Better

CONUS 2.5 km Case on PSG Cluster - 4 nodes
Source: TQI – Abdi; Apr 18

- Based on WRF 3.8.1 trunk
- 1501 x 1201 grid x 35 levels
- Total 60 time steps, SP run
- Physics option modified:
  - WSM6
  - Radiation *off*
  - 5-layer TDS
- All WRF runs single precision
- PSG cluster node configuration:
  - 2 CPUs, 16 cores each
  - 4 x P100 GPUs
  - Or 4 x V100 GPUs
  - CPU-only 1 MPI task each core
  - CPU+GPU 1 MPI task per GPU

CONUS 2.5km Source: http://www2.mmm.ucar.edu/wrf/bench/benchdata_v3911.html (Note “Physics options modified” in side bar)
Results: GPU WRF Strong Scaling for EM_LES

~5x Speedup: 4 x P100 vs. 4 x HSW

EM_LES case too small for GPU efficient scaling

Results for EM_LES Case on PSG - 4 nodes
Source: TQI – Abdi; Dec 18

- Based on WRF 3.8.1 trunk
- 1024 x 1024 grid x 60 levels
- Physics options:
  - Kessler
  - Mostly dycore time
- PSG cluster nodes:
  - 2 CPUs, 16 cores each
  - 4 x P100 GPUs
  - CPU-only MPI task each core
  - CPU+GPU MPI task per GPU
Other customer namelist speedup results

- ROKAF
  - Volta: 6.6x faster
  - Pascal: 4.74x faster
  - K80: 2.83x faster
- Weatherbell
  - Volta: 7x faster
- Agriculture
  - Pascal: 5x faster
TempoQuest Systems Architecture

**Cloud Hardware**
- Multi-node GPU (Pascal/Volta)
- NVLINK
- Shared Memory
- Storage

**Software**
- AceCAST WRF CUDA
- Weather Workflow and System Interface
- O/S Visualization
- Deep Learning/Analytics

**Data Sources**
- Gov't
- Weather Satellites
- Weather Radar
- Commercial
- Other (Academic, NGO)

**Weather Service Providers**
- Transportation
- Energy
- Agriculture
- Wild Fires
Conclusions

- TQI is a micro-weather prediction company with the goal of accelerating WRF by up to 10x using NVIDIA GPUs.
- We had a breakthrough with acceleration of end-to-end WRF runs by 5x to 7x.
- We deliver on-premise or software-as-service on the cloud.
- Future goal: we feel the need for more speed...