High-Resolution Weather Forecasting on Linux Superclusters

Daniel B. Weber, Ph.D.
Research Scientist
and
Kelvin Droegemeier, Ph.D.
Director
CAPS/University of Oklahoma
Outline

- Application background (ARPS)
- Forecast challenges and examples
- Engineering issues
  - optimization issues (scalar, parallel)
  - file system issues
  - timing statistics (including Los Lobos)
- Future of ARPS on clusters
  - 3km CONUS forecasts, Fact or Fiction?
- Bottom Line - We want to do science!!!!
The Advanced Regional Prediction System (ARPS)

Center for Analysis and Prediction of Storms

University of Oklahoma
ARPS Model

General features

- Fully self-contained meso- and storm-scale prediction model
- Can be run at any resolution (10m to 60km)
- Fortran 77 with Fortran 90 extensions
  - Fortran 90 version (5.0) due out in late 2000
- Multiple I/O formats: binary, HDF, NetCDF, GrADS, GRIB, AVS, Vis5D, ASCII, packed binary
- Extensive in-code documentation and user’s guide (online)
- Email-based user support system + online FAQ
- Code available online (http://www.caps.ou.edu/ARPS)
- Entire ARPS system is in the public domain - no restrictions
- Designed for all architectures
  - MPI for shared and distributed memory parallel computers
  - UNIX workstations
  - Linux and Windows PCs
ARPS Details

Dynamics and Numerics
- Non-hydrostatic and fully compressible
- Generalized terrain-following vertical coordinate
- Arakawa C-grid
- User-defined vertical stretching
- Polar stereographic, Lambert conformal, Mercator projections
- 1-D, 2-D, 3-D geometry
- Split-explicit solution with vertically implicit option
  - 2nd and 4th order quadratically conservative centered differences
  - Zalesak multi-dimensional FCT
  - Multi-dimensional positive definite centered difference (MPDCD)
- Initialization
  - Horizontally homogeneous (analytic, sounding)
  - 3-D inhomogeneous
ARPS Model

Physics

- Moist Processes
  - Kuo or Kain-Fritsch cumulus parameterization
  - Kessler, Lin-Tao, Schultz NEM grid-scale microphysics (all highly optimized)

- Surface and PBL
  - Convective PBL scheme based on TKE formulation
  - Stability-dependent bulk aerodynamic drag for surface heat, momentum, and moisture fluxes
  - 2-layer diffusive soil model with surface energy budget (multiple soil types in 1 grid cell; API initialization)
  - Full long- and short-wave radiation (NASA code) including cloud interactions, cloud shadowing, and terrain gradient effects
  - 1 km resolution (over US) USDA sfc characteristics data base and pre-processing software; 30 second global terrain database; 3 second for US plus pre-processing software
Numerical Weather Prediction
Run the Computer Model

- Over the course of a single forecast, the computer model solves billions of equations
- Requires the fastest supercomputers in the world -- capable of performing billions to trillions of calculations each second
Small-Scale Weather is LOCAL!

- Severe Thunderstorms
- Fog
- Rain and Snow
- Intense Turbulence
- Snow and Freezing Rain
- Severe Thunderstorms
Grids: Achieving the Goal

20 km CONUS Ensembles

10 km

3 km

1 km
Numerical Weather Prediction
Run the Computer Model

- Break the forecast into grid boxes (finite grid)
- Solve complicated equations within each grid box to account for
  - wind speed and direction
  - temperature
  - humidity
  - sun heating the ground
  - surface vegetation
  - lakes and oceans
  - clouds, rain, hail, snow
  - terrain
  - turbulence
Numerical Weather Prediction

1. Make Observations
2. Collect and Process Data
3. Run Forecast Model on Supercomputer
4. Create Products
5. Dissemination to End Users
ARPS Components

Data Acquisition & Analysis
ARPS Data Analysis System (ADAS)
- Ingest
- Quality control
- Objective analysis
- Archival

Parameter Retrieval and 4DDA
Single-Doppler Velocity Retrieval (SDVR)
- Variational Velocity Adjustment
- Data Assimilation
- Thermo-dynamic Retrieval

Forecast Generation
ARPS Numerical Model
- Multi-scale non-hydrostatic prediction model with comprehensive physics

Product Generation and Data Support System
ARPSPLT and ARPSVIEW
- Plots and images
- Animations
- Diagnostics and statistics
- Forecast evaluation

Incoming data
Lateral boundary conditions from large-scale models
Grided first guess
Mobile Mesonet
Rawinsondes
ACARS
CLASS
SAO
Satellite
Profilers
ASOS/AWOS
Oklahoma Mesonet
WSR-88D Wideband
But...Fine Grids Require Fine-Scale Observations: NEXRAD Doppler Radar Data
NEXRAD Doppler Radar Data

KTLX
Oklahoma City, OK
Tue 18 Jul 2000
15:35 UTC

Reflectivity
Tilt 1 Elev 0.5
Precip Mode
Max: 46.0 dBZ
ADAS Cloud Analysis Scheme

GOES Visible Image at 1745 UTC on 07 May 1995
ADAS Cloud Analysis Scheme

Vertical E/W Cross-Section: METAR + GOES IR + WSR-88D
<table>
<thead>
<tr>
<th>Model</th>
<th>Rain Water (g/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d16a03.bin</td>
<td>y-z wind x-dist=19.500 km</td>
</tr>
<tr>
<td>f16a54.bin</td>
<td>y-z wind x-dist=19.500 km</td>
</tr>
</tbody>
</table>

**Dual-Doppler**

**SDVR-Retrieved**
Impact of SDVR
Center for Analysis and Prediction of Storms (CAPS)
Spring 1999 Special Operational Test

- Advanced Regional Prediction System (ARPS)
- Assimilation of NEXRAD base data + profilers, MDCRS, OK Mesonet, METARS, GOES
- 3 km grid spacing
- Full model physics
- Run on 256-node massively parallel SGI Origin supercomputer at NCSA
- Real time output on web
The 3 May 1999 Oklahoma Tornado Outbreak
Largest Tornado Outbreak Ever In Oklahoma

- 60+ tornadoes statewide (55 in Norman CWFA)
- Previous state record 26 tornadoes in one day
- First F5 since 1982
- First F5 ever for Oklahoma City

Courtesy Dave Andra, Oklahoma City Area National Weather Service Forecast Office
The Forecasts and Warnings

- Severe thunderstorms in 4:30 AM forecast
- Thunderstorm outlooks mention tornadoes
- First warnings issued upstream 4:15 PM
- Short term forecast at 5:40 PM mentions tornadic storms moving into metro by 7:00 PM
- Numerous warnings and detailed statements tracked tornado into and through metro area

Courtesy Dave Andra, Oklahoma City Area National Weather Service Forecast Office
May 3 Tornado Damage
May 3, 1999 Forecast Configuration

- Full single-Doppler velocity retrieval using base (Level II) data from the Oklahoma City (KTLX) WSR-88D radar
- Other observations
  - Oklahoma Mesonet, GOES Satellite, Wind Profiler, Surface METARS, MDCRS Commercial Aircraft
- Forecast time: 2200 - 0200 UTC
- Cold start (no dynamic data assimilation)
  - Simplest possible configuration
  - More sophisticated tests now underway
- Full model physics including radiation, ice microphysics, terrain, surface energy budget
- 3 km spatial resolution
CAPS Numerical Forecasts of the May 3 Tornadic Storms

5:00 pm - Model Initialization Time

ARPS Prediction Model
(0 hour forecast)

NEXRAD Radar Observations

Storm Beyond Velocity Range of NEXRAD
CAPS Numerical Forecasts of the May 3 Tornadic Storms

5:30 pm - 30 min Forecast

NEXRAD Radar Observations

ARPS Prediction Model
(1/2 hour forecast)
CAPS Numerical Forecasts of the May 3 Tornadic Storms

6:00 pm - 1 hour Forecast

ARPS Prediction Model
(1 hour forecast)

NEXRAD Radar Observations
CAPS Numerical Forecasts of the May 3 Tornadic Storms

6:30 pm - 1.5 hour Forecast

ARPS Prediction Model
(1 1/2 hour forecast)

NEXRAD Radar Observations
Forecasts **With and Without** NEXRAD Data

**WITH**

Moore, OK
Tornadic Storm

2-Hour CAPS Computer Forecast Down to the Scale of Counties

**WITHOUT**

NEXRAD Radar Observations
Science on Roadrunner

- CAPS conducted thunderstorm turbulence research in a collaborative effort with Honeywell.
- The object was to characterize turbulence at commercial jet cruise altitudes (64 processors/5 days).
- Successful in generating turbulence features inside and outside the storm.
- Provided a conduit for further development of the ARPS microphysics.
Engineering Issues

- Optimization Summary
  - scalar
  - parallel (mpi, pvm, hpf)
  - file system

- Benchmarks
  - ARPS
  - ARPI 3D (instrumented research code)
Optimization Tools

Poorman’s approach
- use system calls to a timer inside the program around the loop/code of interest (etime, second works for some systems, but you must be careful)

Commercial Software
- Totalview, Perfview, Perfx, Speedshop Deep, Vampir etc...
Scalar Performance

- Optimize loops (ARPS is 95% loops)
  - remove redundant calculations (21% less work)
  - combine work, eliminate calls to operator type subroutines - not good for readability
  - bigger loops are better, less instruction overhead and memory access
  - remove divides!, very slow on any processor
  - preprocess constants, saves cycles, but increases memory footprint (13 3-D arrays)
Scalar Performance

- Current optimization gain of 20-33% depending on the platform
- Results indicate that ARPS is memory bound
- Rewrite ARPS in a cache friendlier manner?
  - >100k lines of Fortran
Cache Issues

- Samex Domain runs (23x23x43) aided by larger cache (4mb O2000 vs 1mb O200)
- 15% reduction of execution time due solely to larger cache
- Majority of the computations are performed on the small time step.
- Small time step requires 42 3-D arrays
- 44 3-D arrays can fit into 4mb cache
- Result: Need to fit ARPS into cache
ARPS Optimization Costs

- Larger memory footprint (13 additional 3-d arrays) but memory is not a big issue with mpi runs
- Increased granularity - more work done in upper levels routines
- Code is less modular, more complex loops
- In general, the data will not fit into cache!
Parallel Performance

- MPI, PVM outperform HPF and OpenMP
- Distributed computing (MPI)
  - perform benchmark tests
  - assess mpi requirements, including a code review of the message passing design, reduce the number of processor to processor communications
  - combine data exchanges, fewer calls with larger amounts of data sent is more efficient than many smaller exchanges
Benchmarks

- Results from a number of different platforms, including clusters ARPS and ARPI3D
- Approach ==> fixed domain size on individual processors, so the overall size grows with increasing nodes
- Tests the network without changing the local cache behavior
ARPS Parallel Performance on 1328 Node Cray T3E-900

Operational ARPS Timings on Cray T3E

Fixed Local Problem Size
(19x19x35 zones)
Parallel Performance on SGI/Cray Origin 2000

Operational ARPS Timings on Origin 2000

- Fixed Local Problem Size (28x28 zones)
- Fixed Global Problem Size (403x403 zones)
ARPS 3km Grid Timings

- linux cluster
  2proc/node
- Origin 2000
  R10K
- IBM Winter Hawk I
- IBM Winter Hawk II
- Compaq Alpha EV-67
- linux cluster
  1proc/node
ARPI3D 2 Processor/Node Benchmark
Los Lobos

Wall Clock Time (seconds)

Total time
File Writes only
Calculations + MPI
Calculations Only

Processors
Benchmark Summary

- MPI overhead is 45% for 480 processors (non-dedicated mode)
- Computations are slower than Alpha’s and Winter Hawk II, similar to SGI.
- 2 Processor/node tests (for 128 tests) is 50% slower than the 1 processor/node case. There appears to be a bottleneck with the memory bus/network port.
Benchmark Summary

- 20% increase in compute time for 2proc/node configuration
- File system is a weak link
  - 5.5mb/sec sustained for 480 2proc/node tests
  - passing through linux file server not r6000
Can we use the current Cluster technology (UNM) for High-Resolution ARPS CONUS simulations?

Need to estimate the following:
- memory
- cpu time
- disk space/network capacity
- scheduling
Memory

- Generally not an issue.
- Since memory access during the computations is the bottleneck. (I.e. memory technology is behind the processor state)
- For a 47x65x47 grid defined on each processor, the memory requirement is approximately 75mb
CPU Time

Function of the grid:

- A 47x65x47 per processor grid size (3km x 3km x 500m) runs approximately 1.4 times faster than wall clock on a single LL processor (4200km x 3000km x 22km)
- 1.1 times slower than real time at 64 processors. (includes file writes)

MPI overhead? --> no estimate yet

Load balance will be an issue with isolated precipitation
Current status of ARPS requires binary file writes for mpi jobs

History dumps are performed by each processor 8mb for a 47x65x47 grid

Total storage for a 12hr 512 processor simulation storing every hour is 50gb.

At the present file write speed, it would require 3 hours to write the data to disk

Alternative --> write to local disk
Other Requirements

- Rapid access to **observational data**, especially Doppler radars
- **Visualize** the output interactively and perhaps even immersively
- Keep data files distributed?
- Scheduling the job on the machine