MPAS on GPUs

Dr. Raghu Raj Kumar
Project Scientist I & Group Head
Special Technical Projects (STP) Group
National Center for Atmospheric Research

September 2017

Overview

- Introduction
  - Team
  - System & Software Spec
- The Plan
  - Analyzing MPAS
  - Approach
- Results
  - Obtaining portable code
  - GPU Performance comparison
- MPAS Physics Plan
- Questions

Developers Team

- **NCAR**
  - Dr. Raghu Raj Kumar, Project Scientist, STP
  - Michael Duda, Software Engineer, MMM
  - Negin Sobhani, Post Doc Researcher, STP
- **NVIDIA/PGI**
  - Dr. Carl Ponder, Senior Applications Engineer
  - Brent Leback, PGI Compiler Engineering Manager
- **University of Wyoming**
  - Supreeth Suresh, Pranay Reddy, Sumathi Lakshmiranganathan, Cena Miller, Bradley Riotto- GRAs
- **Korean Institute of Science and Technology Information**
  - Jae Youp Kim, GRA
- **IBM/TWC**
  - Constantinos Evangelinos, IBM Researcher

System Spec

- NVIDIA's PSG Cluster
  - Dual socket Haswell (32 cores per node) with 4 P100 per node, 16 nodes
  - Minsky- Dual socket Power8 (20 cores per node) with 4 P100 per node, 2 nodes
- NCAR's Cheyenne
  - Dual socket Broadwell (36 cores per node), 4,032 nodes
- TACC's Stampede
  - Single socket KNL (68 cores per node), 500+ nodes
- IBM's SummitDev
  - Minsky- Dual socket Power8 (20 cores per node) with 4 P100 per node, nodes
  - Witherspoon- Power9 + Volta GPUs

Software Spec: DyCore only MPAS

- **Software**
  - MPAS Release (or MPAS 5.0)
  - Intel Compiler 17.0, PGI Compiler 17.1
- **Baroclinic Instability Test**
  - Dry dynamics test-case produces baroclinic storms from analytic initial conditions
  - Split Dynamics: 2 sub-steps, 3 split steps
  - Current work: 60km resolution (163k grid points, dt=300s), 120km (40k grid points, dt=600s)
  - Number of levels = 56
  - Double precision
  - Execution time for simulating 1 day
    - 144 timesteps (600 sec) for 120 km
    - 288 timesteps (300 sec) for 60 km

Software Spec: MPAS on GPUs

- **Software**
  - MPAS 5.2
  - Compilers- PGI 17.7, IBM XL, Intel 17.0
- **Configuration**
  - Grell Freitas Convection, WSM 6 Microphysics, Noah Land surface, YSU Boundary Layer, Monin-Obhukov Surface layer, RRTMG radiation, Xu Randall Cloud Fraction
  - Split Dynamics: 2 sub-steps, 3 split steps
  - 120km (40k grid points, dt=720s)
  - Number of levels = 56
  - Radiation interval: 30 minutes
  - Single Precision

Analyzing MPAS

DyCore only MPAS Kernel-wise Execution Time for Intel Compiled 40k (120km) Dataset with 32 MPI Ranks

Execution time-

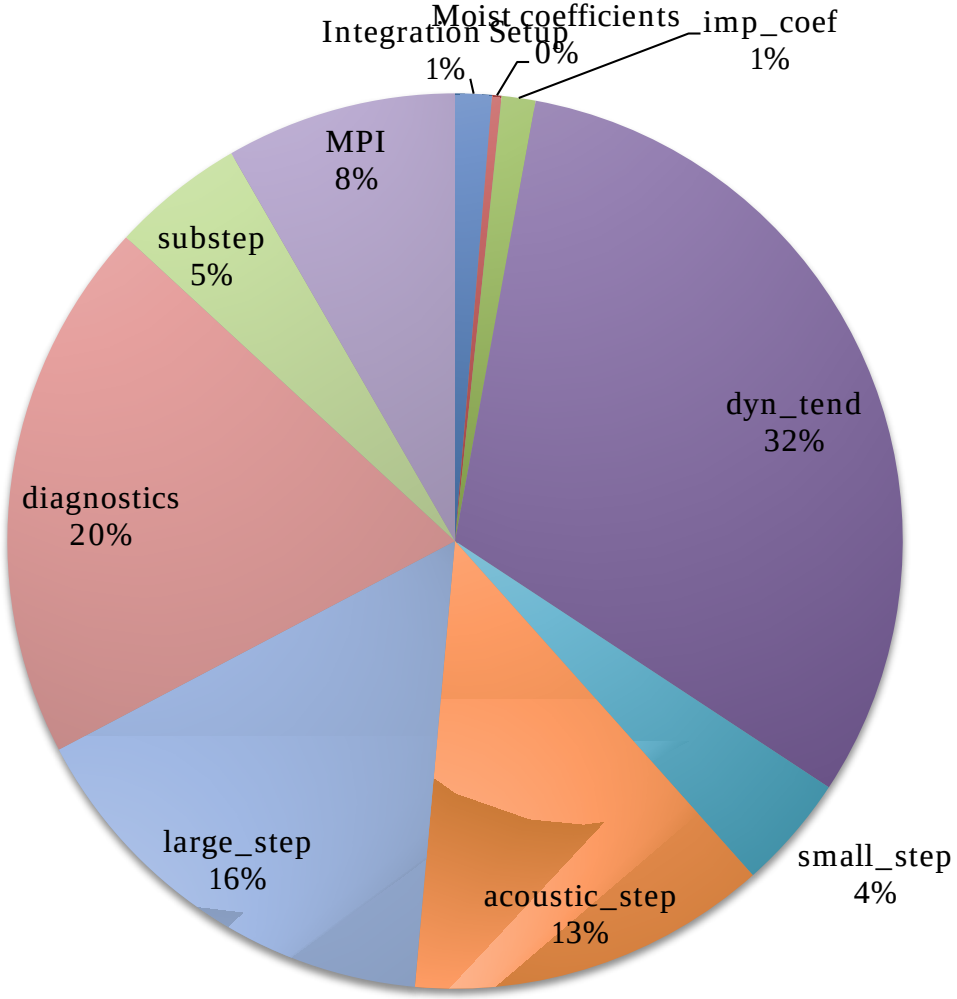
Physics: 45%

DyCore: 55%

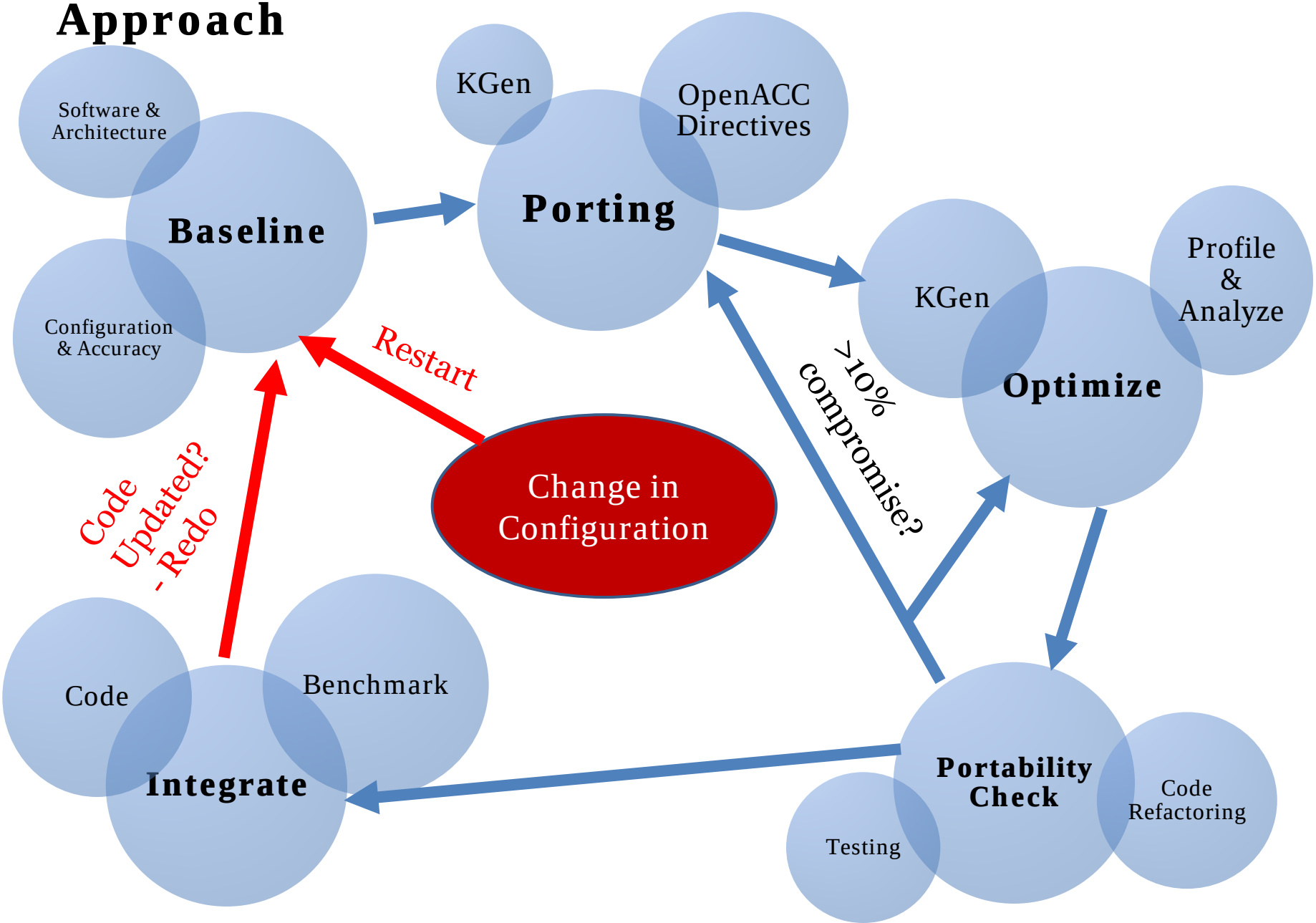
Lines of Code-

Physics: 110,000

DyCore: 10,000



Approach


Software & Architecture
Baseline
Configuration & Accuracy
KGen
OpenACC Directives
Porting
KGen
Profile & Analyze
Optimize
Restart
Change in Configuration
Code Updated? - Redo
>10% compromise?
Code
Benchmark
Integrate
Portability Check
Testing
Code Refactoring

Approach

- **Porting**
  - F77 to F90 conversion
  - Use OpenACC parallel directives
  - About 5% of total code lines
    - 500 lines for DyCore
- **Optimization**
  - Code: Standard techniques
    - 32 is good, branches are bad, etc.
  - Data: Create when created on CPU, map to use
- **Portability Check**
  - Revert, re-transform or retain

Obtaining Portable Code

- **Broadwell**
  - Single node, OpenMP enabled and optimized, Fully subscribed
- **Intel Xeon Phi**
  - Single node, 64 MPI ranks per node, 4 OMP threads per rank
- **Timers & Precision**
  - MPAS GPTL timers, double precision

Resolution	Time in seconds					
	Broadwell			KNL		
	Before Portability Phase	After Portability Phase	Improvement	Before Portability Phase	After Portability Phase	Improvement
120 Km	0.92	0.85	8%	0.65	0.59	10%
60 Km	3.87	3.49	11%	2.01	1.75	15%

The code obtained after portability phase is 2-15% better than MPAS Release

Measuring Performance on GPUs

- **Timing includes**
  - Data transfers (Host<->Device)
  - Data mapping
  - Memory allocations and initializations

Resolution	Time in seconds			Speed Up	
	Broadwell	KNL	P100	Broadwell	KNL
120 Km	0.85	0.59	0.37	2.3	1.6
60 Km	3.49	1.75	1.26	2.8	1.4

Understanding Performance on GPUs

- **Portability Check is not Optimizing**
  - GPU optimized code checked for portability
  - Portability check does not involve Xeon optimization
- **Cache pressure**
  - Example: 3o+ 2D or 3D variables in one routine, double precision
  - Broadwell performance better for single precision
- **GPU-GPU MPI**
  - MPI has linked lists with member linked lists
  - MPI via CPU involves data transfers

MPAS Physics: Current Status and Future Plan

- **Porting**
  - Convection, Microphysics, Boundary layer, Surface layer, Cloud fraction on GPU
  - Radiation, Land surface on CPU
- **Development**
  - MPI rewrite using arrays
  - Lagged radiation
  - Adopting Sion library instead of PIO
- **Optimization**
  - Includes Power architecture optimization for Radiation

Thank you! Questions?

Strong Scaling Curve for 30 Km Resolution Optimized MPAS Release Code