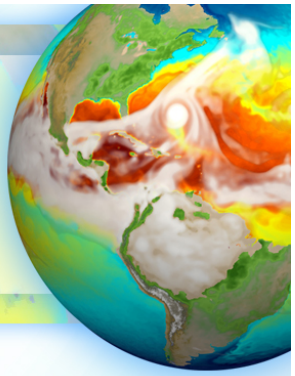
Performance Overview



Phil Jones, Sarat Sreepathi on behalf of Performance Group:

Oksana Guba, Noel Keen, Youngsung Kim, Jayesh Krishna, Azamat Mametjanov, Mark Taylor, Matt Turner, Pat Worley, Min Xu

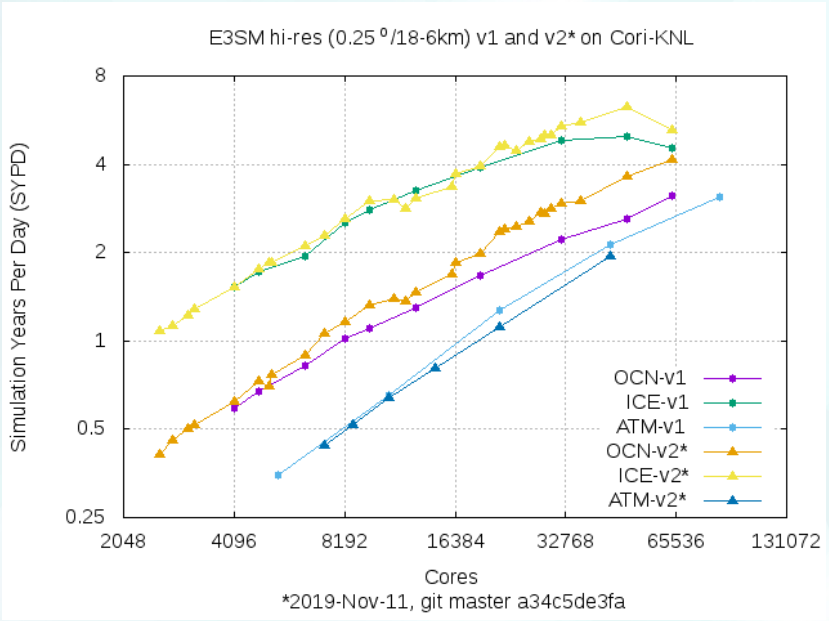
Ex Off: Nichols Romero, Xingqiu Yuan

Performance Group Goals

- Improve throughput on current sims/machines
  - E3SM/BER clusters
    - Anvil (Broadwell)
    - CompyMcNodeFace (Skylake)
    - Chrysalis (AMD Epyc)
  - Cori and Theta (Intel Phi)
  - Summit (IBM P9/Nvidia GPU)
- Prepare for new architectures
  - Perlmutter (EPYC/Nvidia GPU, no spinning disk)
  - A21 (Intel Xeon SP with Xe GPU, memory innov)
  - Frontier (AMD EPYC, AMD Radeon GPU)
  - Longer term – non-GPU?
- Measure and analyze performance
  - Standard benchmarks, profiling
  - Performance tools

Performance Metrics

- Throughput on standard hi-res benchmark cases
  - F-case (atm/sfc) 0.25 deg
  - G-case (ocean/ice), RRS 18-6
  - With/without I/O
- I/O benchmarks
- Incremental improvements in node performance (CPU)
  - Larger expected improvement for v2 final
  - Does not include GPU work
- Substantial improvement in I/O
  - SCORPIO
  - 10x and more



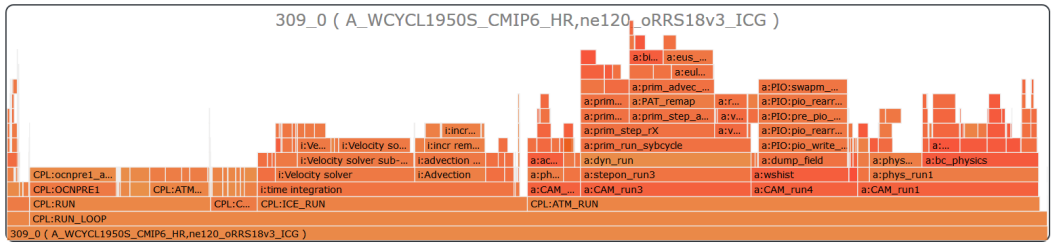
	PIO1/pnetcdf	PIO2/pnetcdf	PIO2/adios
High-res G-case Summit	12.36 GB/sec	27.89 GB/sec (BOX) 2.64 GB/sec (SUBSET)	129.58 GB/s
High-res G-case Cori	1.68 GB/sec	4.94 GB/sec (BOX) 2.86 GB/sec (SUBSET)	18.52 GB/sec
High-res F-case Summit	536.15 MB/sec	3.04 GB/sec (BOX) 1.31 GB/sec (SUBSET)	12.45 GB/sec
High-res F-case Cori	100.37 MB/sec	1031.03 MB/sec (BOX) 663.09 MB/sec (SUBSET)	4.86 GB/sec

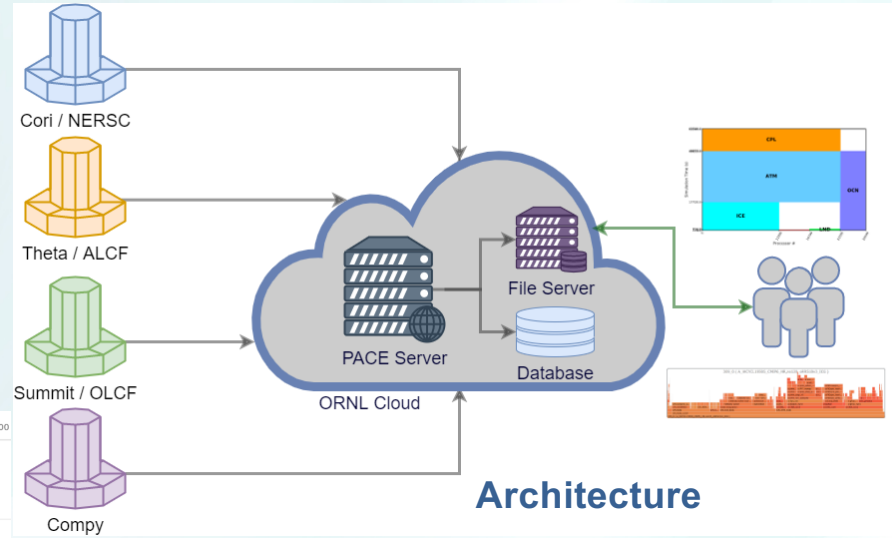
Performance Analytics for Computational Experiments

Infrastructure to provide executive summary of experiments performance.

- Collect performance of all simulations
- Central hub of performance data
- Interactively deep-dive as desired
- Track performance benchmarks of interest
- Facilitate performance research

Summary Performance Graphs



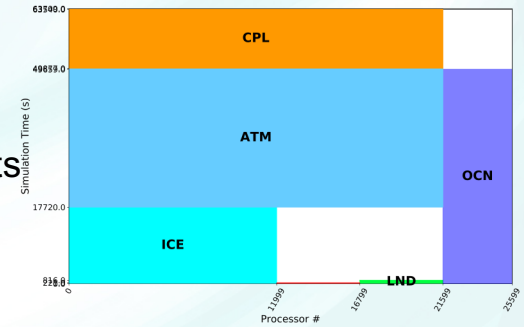


Architecture

Stats

- 184 users
- 10 platforms
- 36,683 experiments
- 1 million input files

Load Balancing

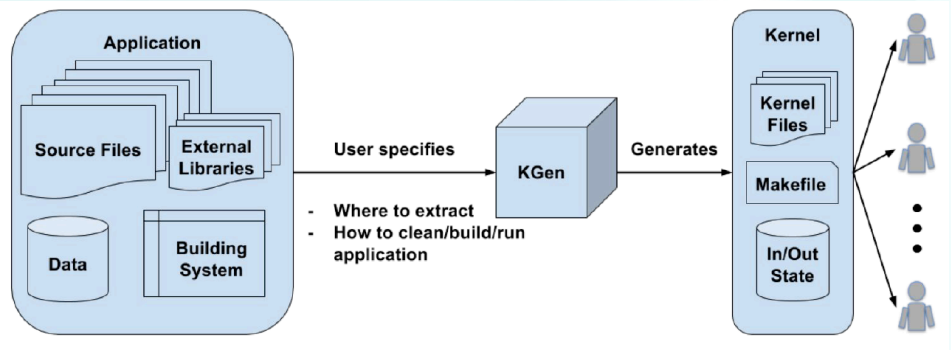


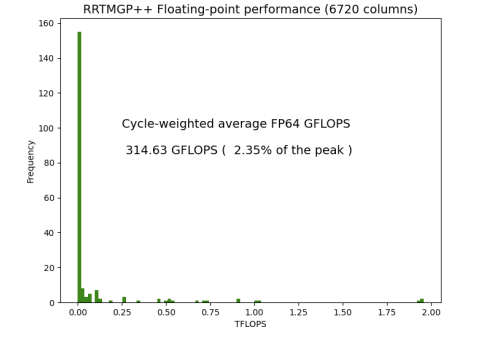
Sarat Sreepathi, ORNL

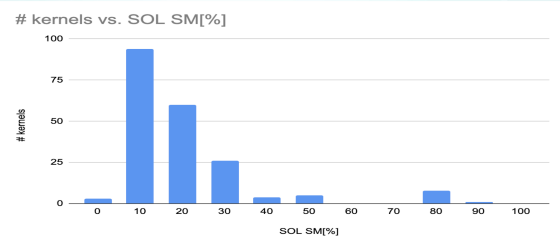
https://pace.ornl.gov

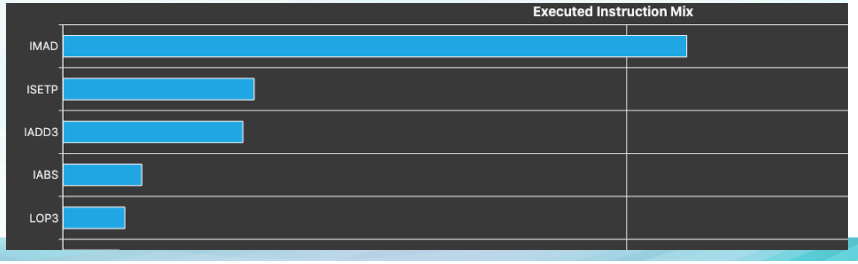
Kernel Extraction and Characterization

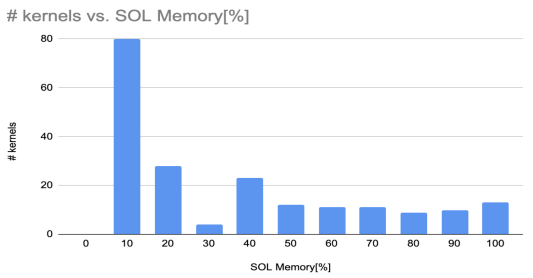
- Kgen tool
  - Extracts kernel and relevant data
  - Used for detailed characterization with realistic data
  - Vendor interactions
  - GitHub: E3SM-Project/KGen
- Detailed GPU characterization
  - Example: RRTMG++
  - Many kernels, diff behavior
  - Integer calculation (indexing)
  - Memory bandwidth
- Kernel classification





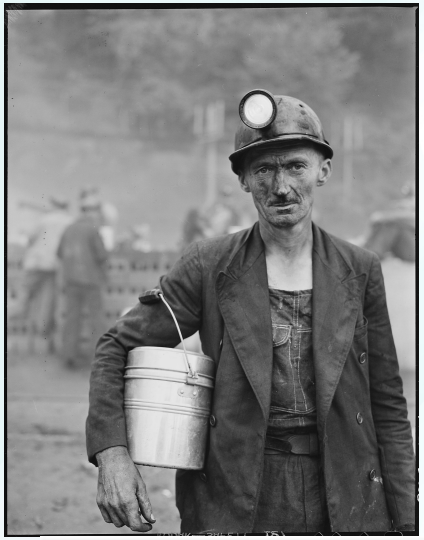






Day to Day

- Performance improvements and debugging
  - I/O
  - MPI collectives, MPI environment vars
  - Compiler bugs, compiler flags
  - Threading bugs, optimization
  - Non-deterministic behavior
  - Bad nodes
  - Initialization at scale
  - Atmospheric physics column configuration
    - Runtime vs compile-time
    - Reduced number of columns
    - Related chunk cost measurements
  - Memory use
    - Leaks, allocation errors
    - Scalability
- Prepare INCITE and other computer proposals

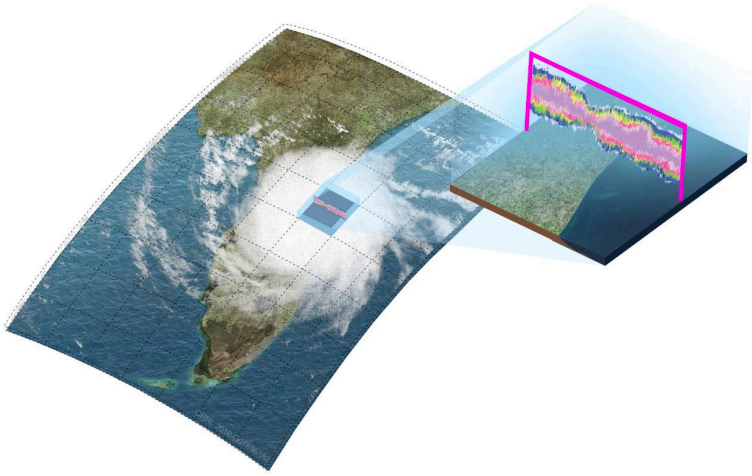


Coal loader image from Wikipedia

Trench image from 1914centenary.com

GPU Performance

- Porting and portability
  - Kokkos
    - Data model + Loop-level abstractions
    - Template metaprogramming
    - Must adopt C++
  - OpenACC/OpenMP
    - Directives for both data mgmt, loop execution
    - Only portable approach for Fortran
  - Issues
    - High occupancy still difficult
    - Both approaches require significant work for optimal implementation
    - Ultimate programming model still not clear, DSLs?
- Utilize GPU differently
  - Give every GPU 1000 subgrid models (CRM, LES)
  - Asynchronous execution to split work across devices



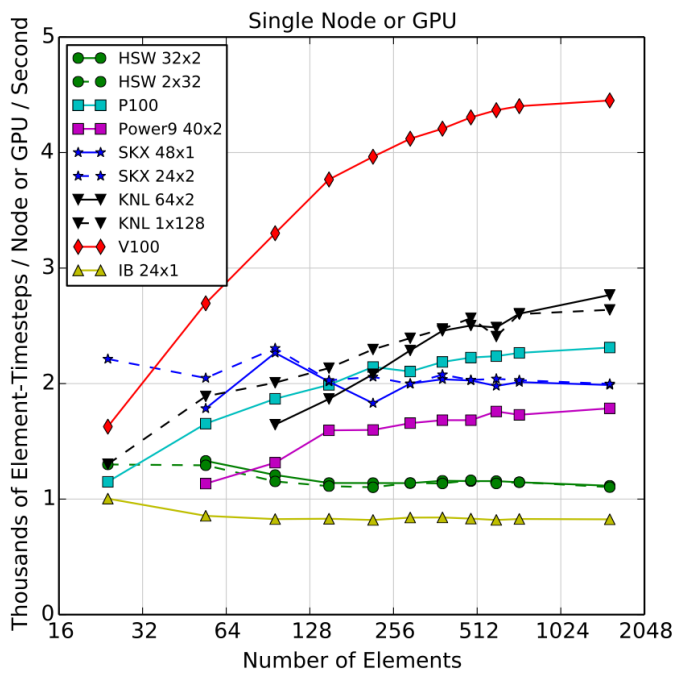
Leveraging for Advanced Architectures

- Much advanced work part of “ecosystem” efforts
- New programming models and GPU approaches
  - NGD: NH atm, SCREAM, Software and Algorithm
  - Exascale Computing Project (ECP): dycore, MMF/superparam, GPU ocean/ice
  - SciDAC CANGA: Exploration of Asynchronous Many-Task runtimes
- New algorithms
  - Software/Algorithm NGD: new ocean barotropic solver
  - SciDAC: new transport algorithms
  - SciDAC DEMSI: discrete-element sea-ice model
  - SciDAC CANGA: new coupling, remapping algorithms

Atmosphere

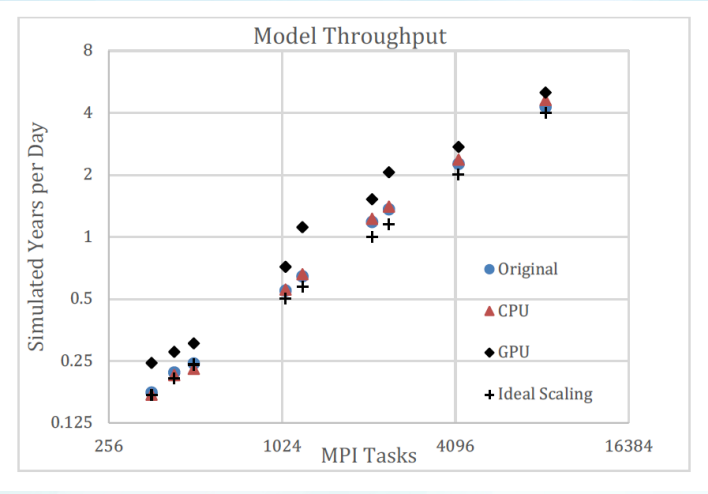
- HOMMEXX dycore
  - Kokkos
  - NGD + Prior work
- NH atmosphere
  - New physics in Kokkos (SCREAM/NGD)
  - IMEX, energy fixer/diagnostics
- Old v1 physics
  - Feasibility study porting MAM
  - OpenACC
  - Estimate of 2 man-years
  - Decided not to pursue
- Superparameterization
  - ECP project
  - Run cloud-resolving models on GPU
  - Initial results are promising

HOMEXX



Ocean/Ice

- Sea-ice
  - Threading/vector improvements
  - Initial GPU porting with OpenACC/OpenMP
    - Velocity solver
    - Comm dominated – difficult to achieve gains
  - SciDAC/ DEMSI
    - New discrete-element model
- Ocean
  - ECP: Ocean OpenACC port
  - Converted vel solver, tracer adv, eos, tridiag
  - Up to 30x speedup when min data txfr
  - Estimate 3-10x for final version, resident
- MPAS refactor, new ocean NGD
  - MPAS framework/style inhibiting progress
  - Embarking on significant refactor



MPAS-O throughput on Summit for GPU, CPU. Only 40% code converted.

MPI Tasks	Speedup for all GPU code	Baroclinic velocity (all)	Baroclinic velocity (excl. data transfers)	Tracer Advection	Equation of State	Vmix Solvers
8400	1.4	1.2	5.4	1.5	2.0	0.9
4200	1.7	1.7	9.3	1.9	1.7	1.0
2400	4.2	4.7	20.4	5.0	3.5	2.2
2100	2.1	2.3	14.0	2.4	1.8	1.1
1200	5.7	7.0	31.9	7.1	4.0	2.4
1050	2.5	2.8	17.9	2.8	1.8	1.3
504	2.8	3.1	18.9	3.2	1.8	1.3