Performance Portability and Optimization for Strong Scaling: Examples from the Community Earth System Model (CESM)

Patrick Worley
Oak Ridge National Laboratory
(*representing the greater CESM development and software engineering community*)

SciDAC 2010
The Chattanoogan
Chattanooga, TN
July 12, 2010

Acknowledgements

- Research sponsored by the Atmospheric and Climate Research Division within the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research, both in the Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.
- This research used resources (Cray XT5) of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725, and resources (IBM BG/P) of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357.

- These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Announcement

4/01/10: CCSM4 (Community Climate System Model, version 4) released

6/25/10: CESM1 (Community Earth System Model, version 1) released

- New!
  - First public releases in 6 years!
- Improved!
  - New components!
  - New numerical algorithms!
  - New grids and grid resolutions!
  - New and improved physical process representations!
  - Improved computer performance and performance scalability!
- Easier to use!
  - Subversion repository access to both model and data sets!
  - Single executable!

What's Special about Climate?*

*(*viewpoint of a computer scientist working with the CESM)*

- Strong scaling (in practice)
  - **Serial-in-time, time-accurate forward simulation models** with timestep lengths decreasing with increase in spatial resolution
    - Many timestep lengths and timestepping algorithms: some explicit, some semi-implicit, some operator-split, some time-split, subcyling, etc.
    - Long(er) timestep and time-parallel methods are being investigated.
  - **Long simulations**: 10, 100, 1000, 10000 simulated years, depending on science
  - **Ensemble simulations**: 3 to O(10) proposed for IPCC Fifth Assessment (AR5)
  - **Multiple scenarios** to quantify uncertainties, assess impacts, evaluate policies, …
  - **Practical upper bounds on execution time** of single experiment and **need to use the allocation efficiently** (not necessarily the processor efficiently)
  - **Each configuration (grid, grid size, enabled processes, …) requires tuning and validation**. Traditional weak scaling (increase problem size => improve science) is problematic because of significant overheads in changing problem size.

Result: Execution speed requirements, limits on allocation, and cost of validation constrain grid resolution and enabled physical processes.

What's Special about the CESM?*

*(*viewpoint of a computer scientist working with the CESM)*

- Software Engineering
  - **Community model** that must run on many different platforms, ranging from University departmental systems to the largest DOD/DOE/NSF systems.
  - **Coupled model** (atmosphere, ocean, land, sea ice, and land ice components), each component with its own grid and exchanging boundary data via a coupler component.
  - **Component models provided by multiple sources**, each with its own (often distributed) core development team, its own coding style and history, and its own CESM working group deciding on research directions. Mechanisms exist for evaluating and importing externally developed contributions as well.
  - **Continually evolving model**, but with component model updates scheduled via a gatekeeper and rigorous testing required for inclusion in a validated CESM configuration.
  - **Continually ported to or evaluated on new architectures**, but must remain "backward compatible" with current production systems.
  - **Large code base**: Over 1 million lines of Fortran (developed over 20+ years).

What's Special about the CESM?*

*(*viewpoint of a computer scientist working with the CESM)*

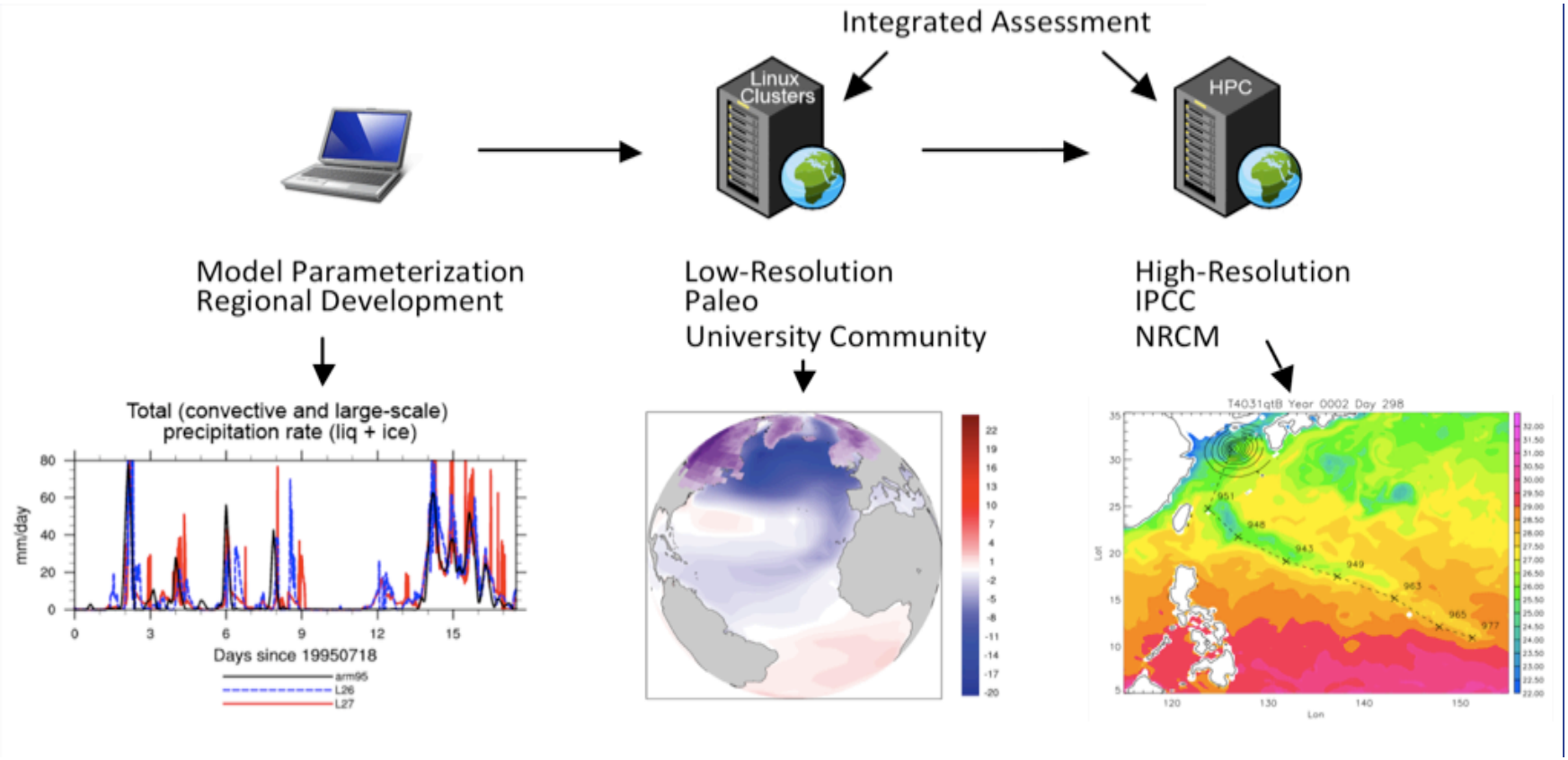
- Numerous configurations (and science options)
  - From single active component model with data models for all other components to all active component models.
  - Production horizontal grid-cell spacings of 500 km to 20 km in atmosphere/land and 400 km to 10 km in ocean/ice. (Cloud resolving atmosphere models will require 3 km grid-cell spacing.)
  - Production vertical resolutions of 26, 30, and 66 levels for the atmosphere and 42 and 60 levels for the ocean.
  - Four different atmospheric dynamics solvers, with at least two more solvers in development.
  - Many subgrid-scale parameterization and physics options: radiation scheme, boundary layer scheme, cloud physics scheme, chemistry packages, with or without biogeochemistry, dynamic vegetation, ecosystem model, …
  - CESM used for many aspects of computational climate science, including further development of simulation capabilities.

Model Component Complexity

Atmosphere Component	CAM DATM (WRF)
CAM Modes: Multiple Dycores, Multiple Chemistry Options, WACCM, single column	
Data-ATM: Multiple Forcing/Physics Modes	
Land Component	CLM DLND (VIC)
CLM Modes: no BGC, BGC, Dynamic-Vegetation, BGC-DV, Prescribed-Veg, Urban	
Data-LND: Multiple Forcing/Physics Modes	
Ice Component	CICE DICE
CICE Modes: Fully Prognostic, Prescribed	
Data-ICE : Multiple Forcing/Physics Modes	
Ocean Component	POP DOCN(SOM/DOM) (ROMS)
POP Modes: Ecosystem, Fully-coupled, Ocean-only, Multiple Physics Options	
Data-OCN : Multiple Forcing/Physics Modes (SOM/DOM)	
New Land Ice Component	
Coupler **Regridding, Merging, Calculation of ATM/OCN fluxes, Conservation diagnostics**	

Slide courtesy of M. Vertenstein (NCAR).

End-to-End Cycle of Model Development, Integration and Prediction with One Unified Model Code Base



Slide courtesy of M. Vertenstein (NCAR).

What's Special about the CESM?*

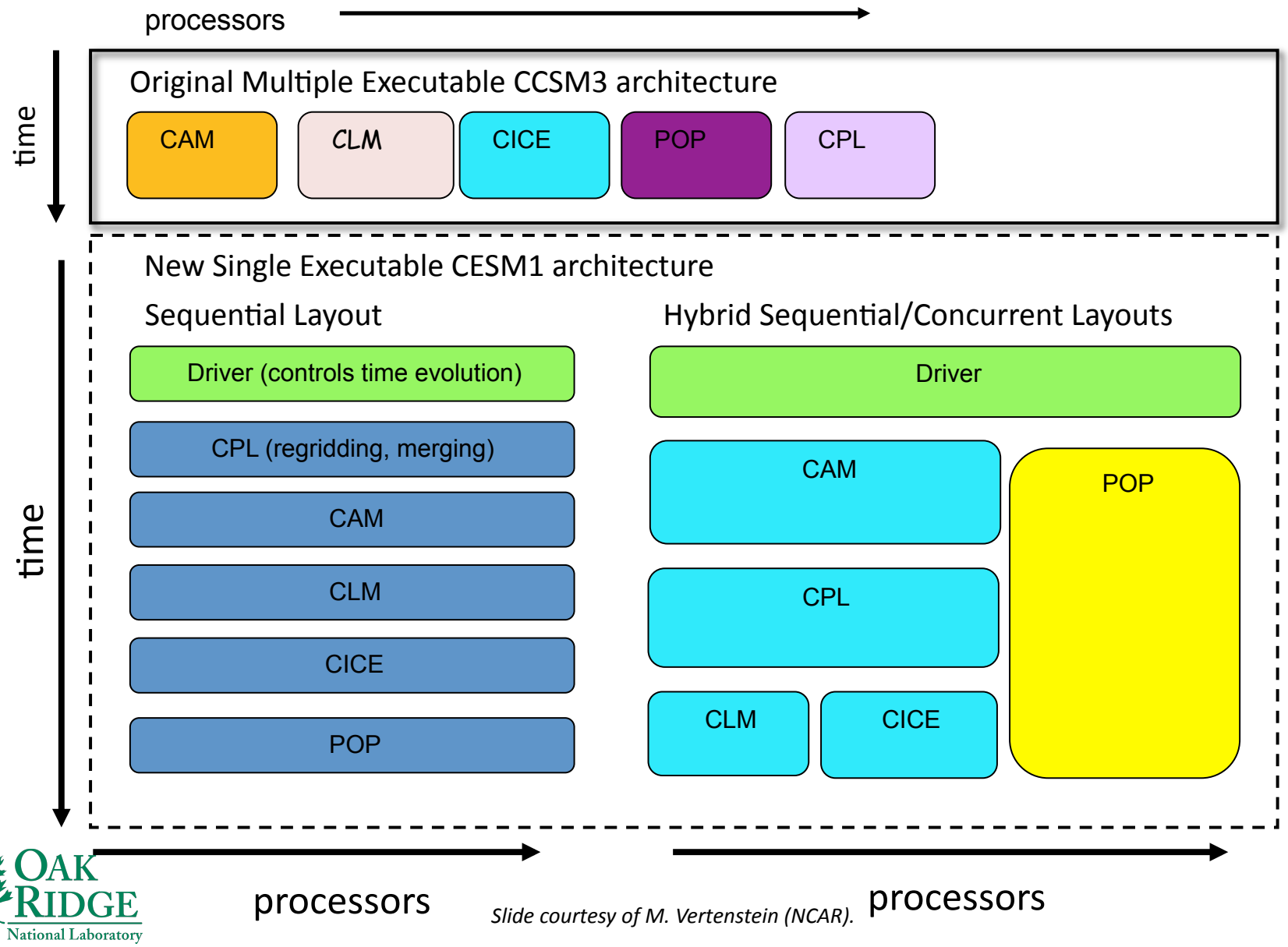
*(*viewpoint of a computer scientist working with the CESM)*

- Numerous parallel algorithm and implementation options
  - **CESM**: allocation and assignment of processor cores to component models
  - **All Components**: MPI and/or OpenMP; flow-control collective communication options; number and location of I/O processes and choice of parallel I/O library
  - **Atmosphere**: per phase degree of parallelism; per phase domain decomposition and assignment to processes (thus load balancing); per phase communication protocols; alternative parallel algorithms for certain phases; subblock size
  - **Ocean**: domain decomposition and assignment to processes; subblock size
  - **Sea Ice**: domain decomposition and assignment to processes, including weights (e.g., climatological) for load balancing; subblock size
  - **Land**: subblock size

most of which are new since the CCSM3 release.

*Result: Enhanced functionality, performance and performance portability **AND** an unwieldy mess confusing to the users and a search space difficult and expensive to optimize over.*

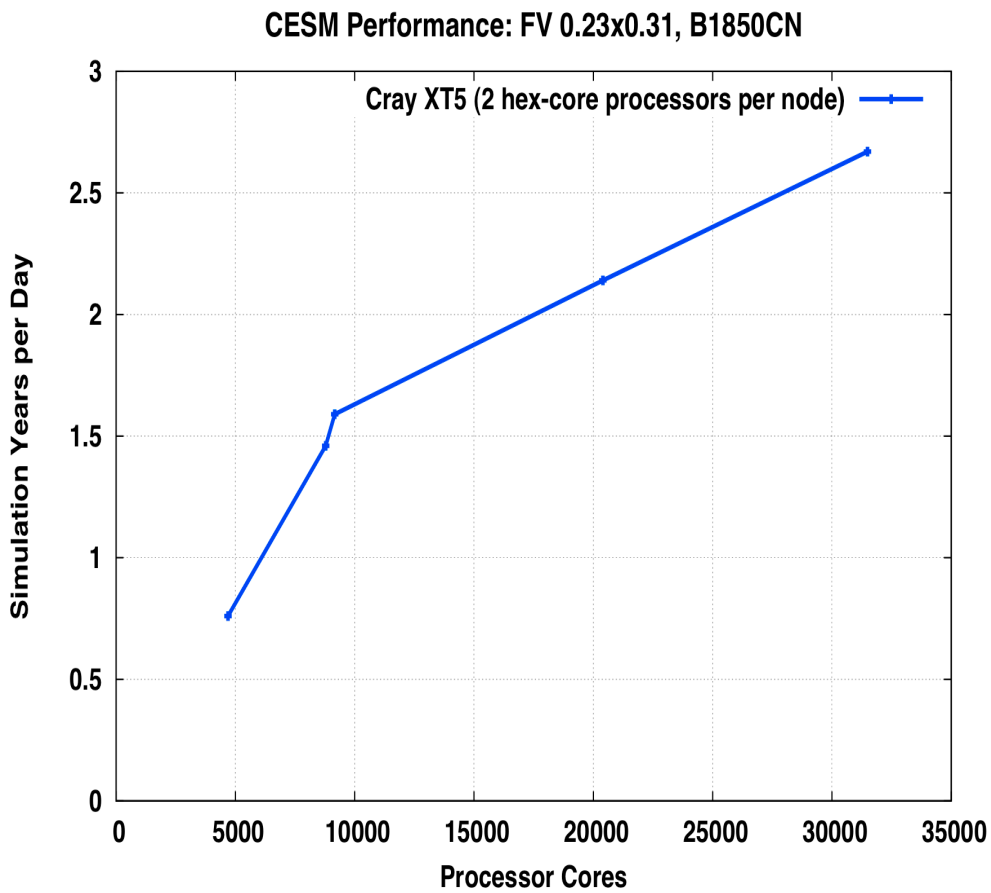
New CESM Architecture
(NCAR CSEG / SciDAC SA)

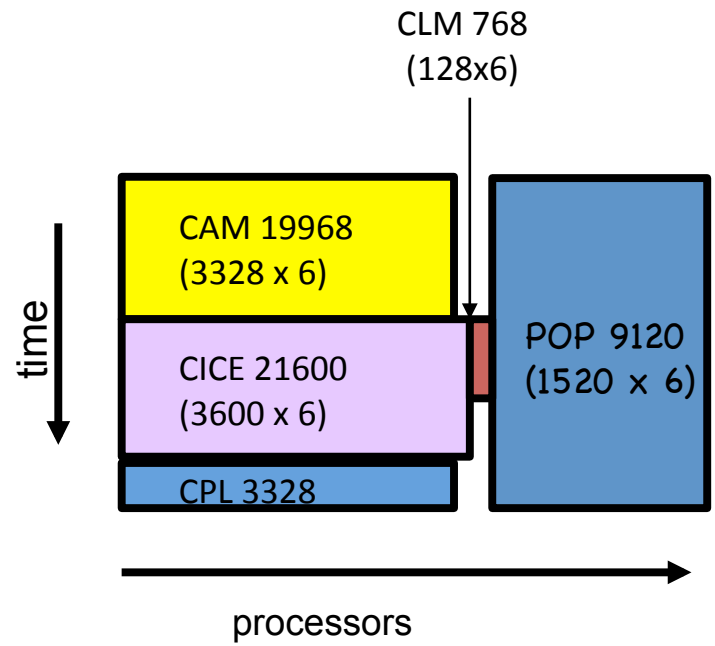


Slide courtesy of M. Vertenstein (NCAR).

CESM1 Cray XT Scalability

35 km atm grid / 10 km ocn grid





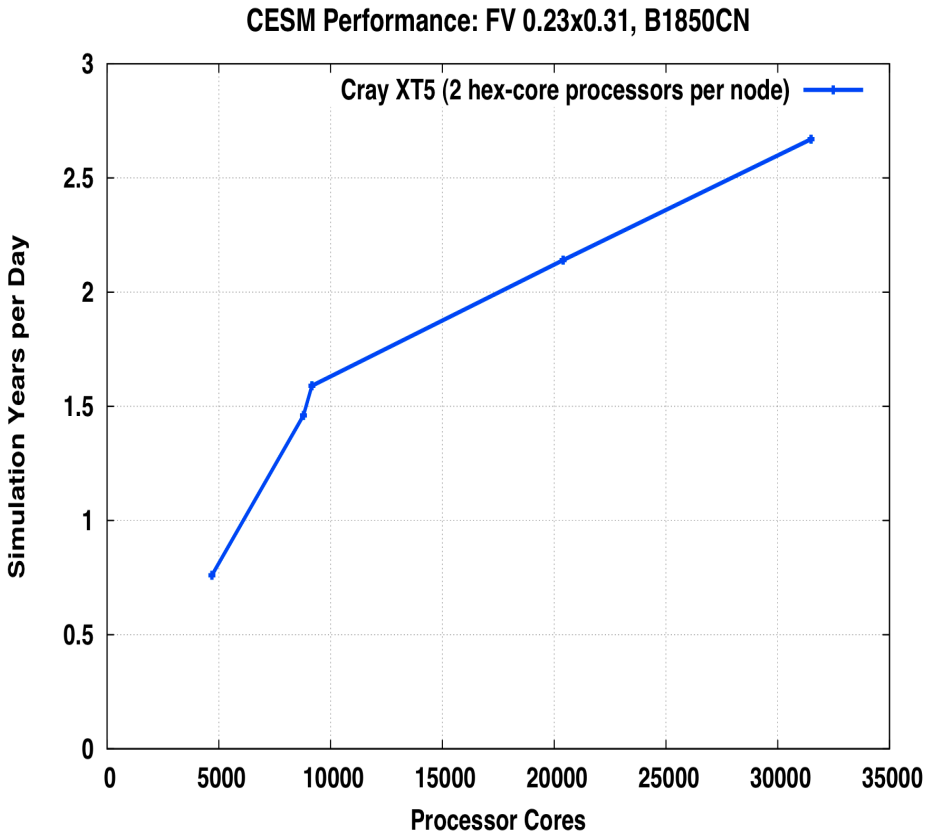
2.7 SYPD on 31,488 cores
(P processes x T threads)

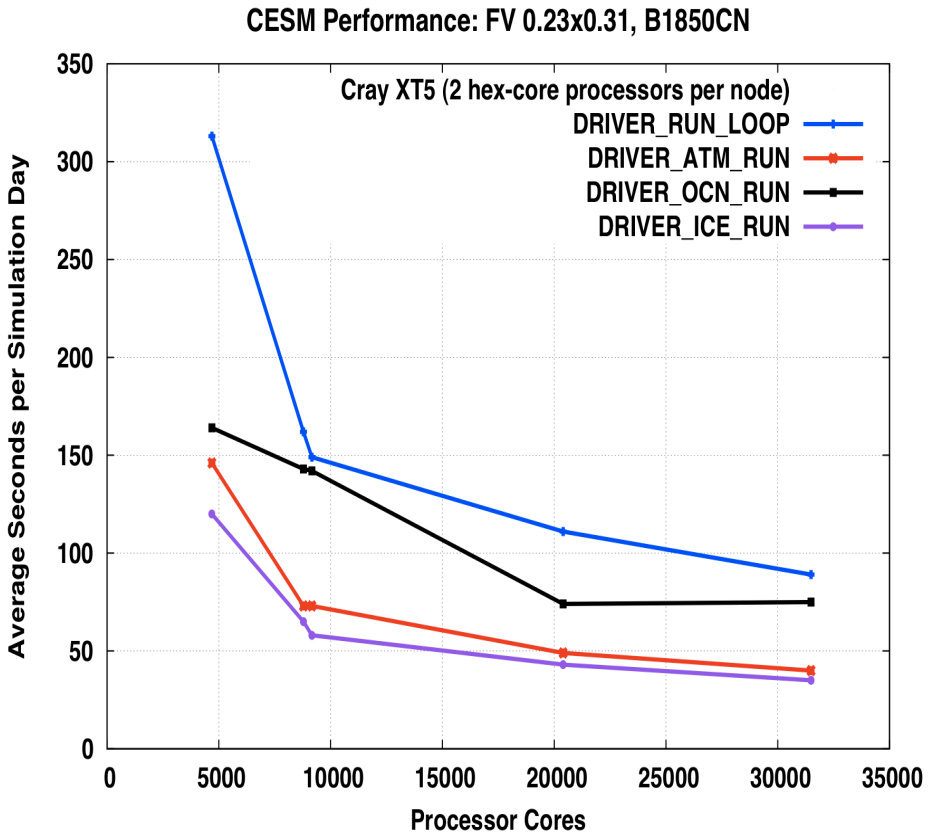
CAM + CICE + CPL is the performance limiter

Preliminary results. Collected June 18-25, 2010.

CESM1 Cray XT Scalability

35 km atm grid / 10 km ocn grid



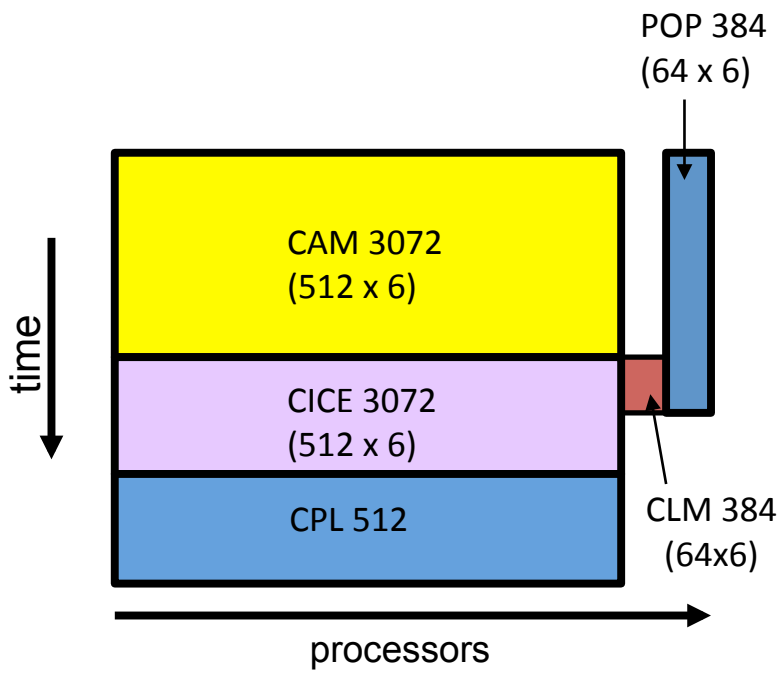


Balancing act: increasing processor core count/decreasing runtime for one component may or may not improve CESM performance without changes in other components.

CESM1 Cray XT Configurations

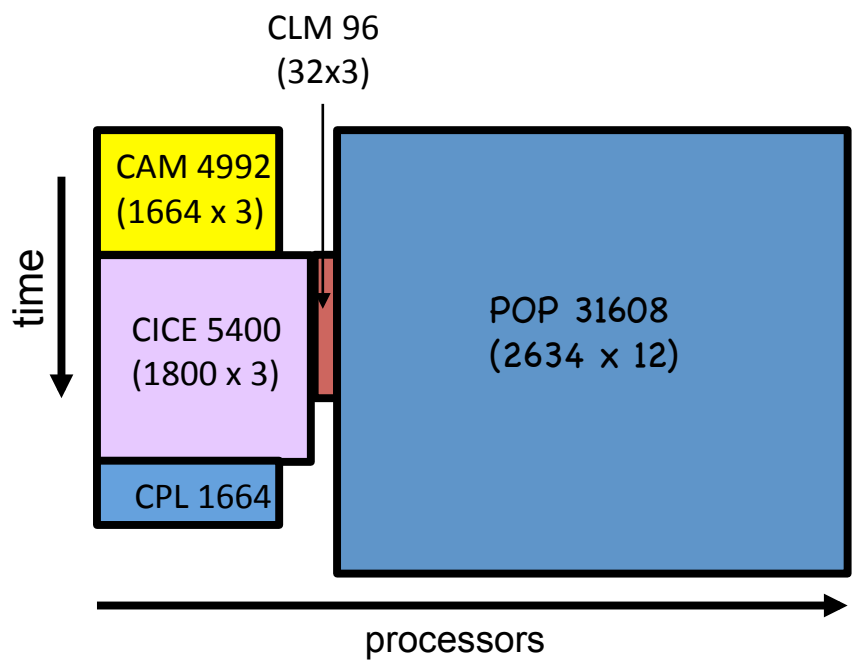
Other Examples (maximum observed performance)

140 km ATM/LND, CCSM4 physics
100 km OCN/ICE



32.7 SYPD on 3,844 cores
(P processes x T threads)

70 km ATM/LND, CCSM4 physics
10 km OCN/ICE



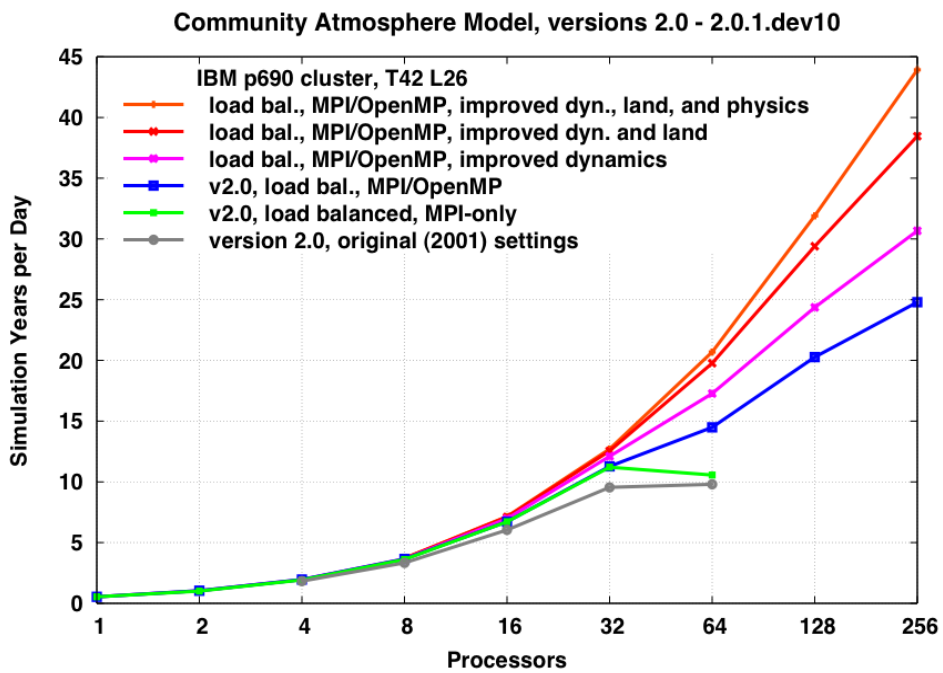
3.9 SYPD on 37,104 cores
(P processes x T threads)

Performance Limiters: Left is CAM; Right is POP.

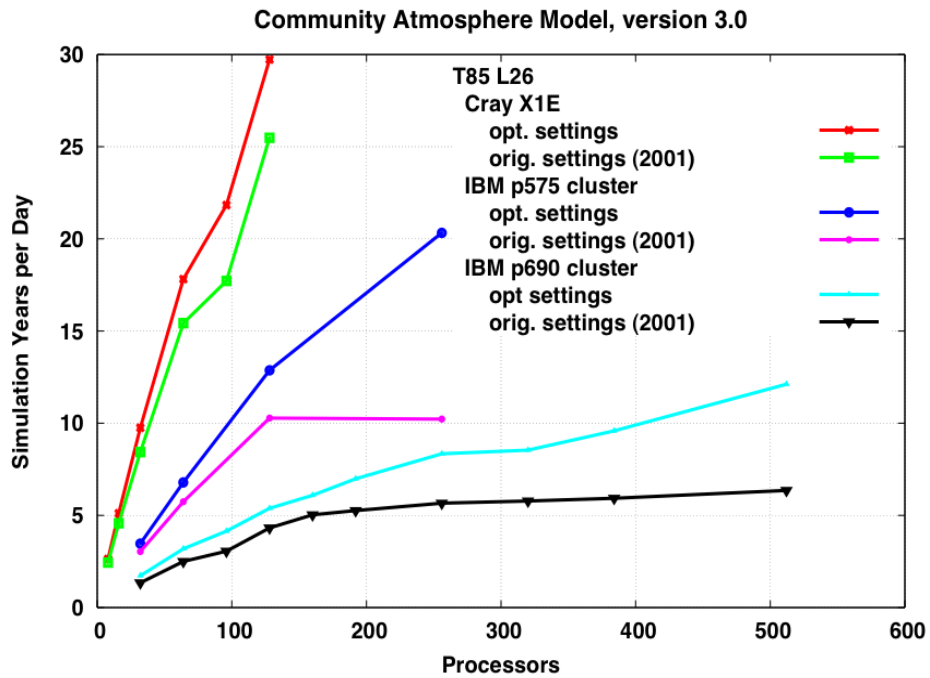
Preliminary results. Collected June 18-25, 2010.

CAM Performance Optimization

2001-2004 (SciDAC SA and ISIC)



300 km ATM/LND grid, CAM3 physics

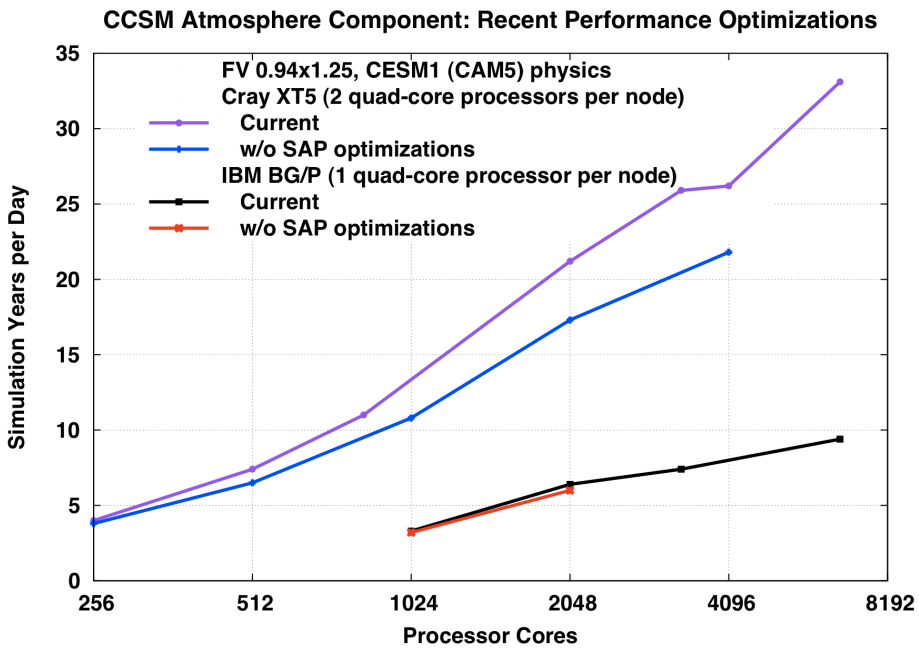


150 km ATM/LND grid, CAM3 physics

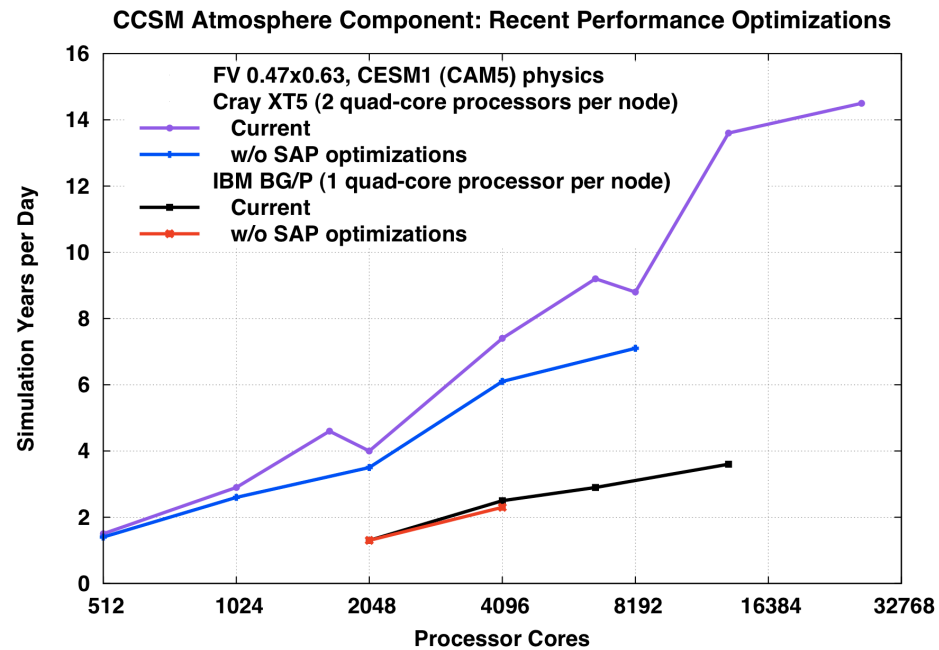
Introduced hybrid MPI/OpenMP parallelism and load balancing, identified and eliminated scalability bottlenecks, and added tuning options to communication algorithms and to data structures. Right figure shows impact of optimizing over new tuning options only.

CAM Performance Optimization

2007-2009 (SciDAC SAP)



140 km ATM/LND grid, CAM5 physics

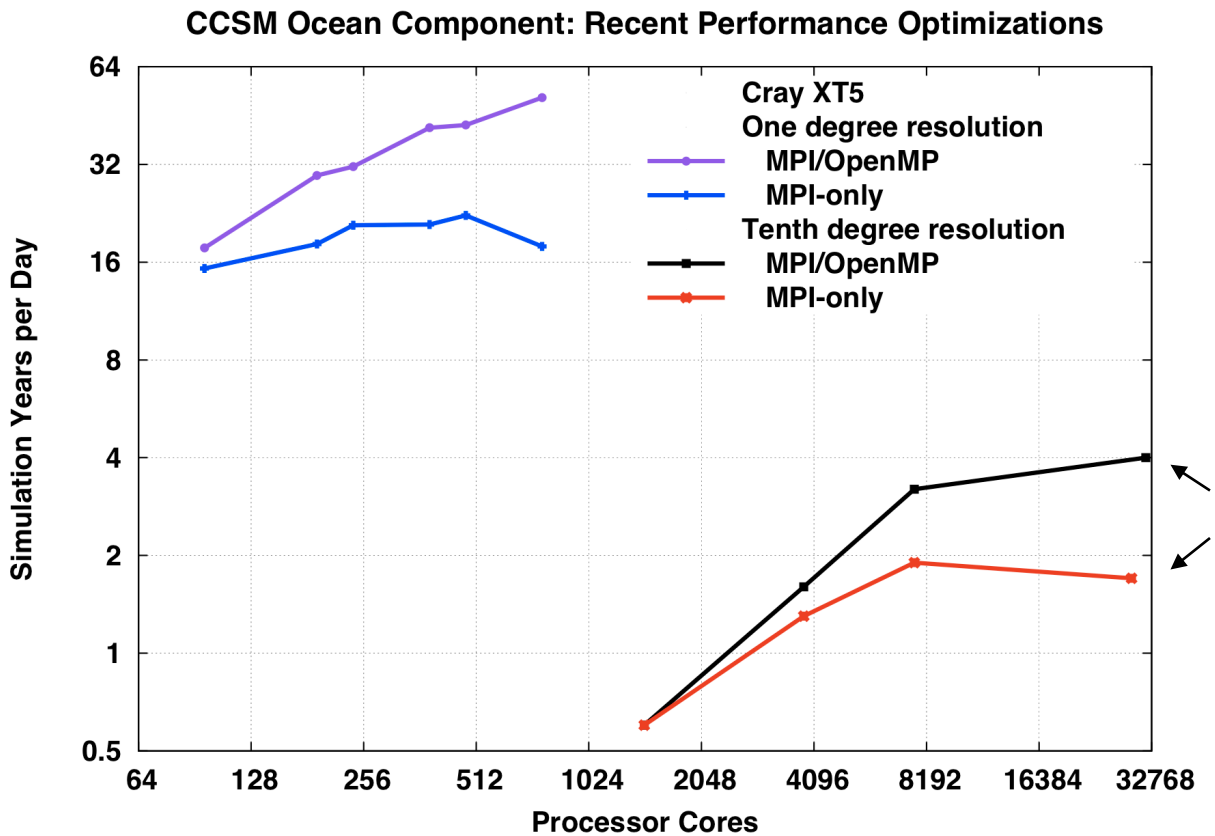


70 km ATM/LND grid, CAM5 physics

As the code evolves monitoring, maintaining and improving performance are critical tasks. Recent optimizations came from identifying new opportunities for parallelism, and continuing to identify and eliminate scalability bottlenecks and optimize communication algorithms.

POP Performance Optimization

2009-2010 (SciDAC SAP/NCAR CSEG)



Resolutions

100 km grid

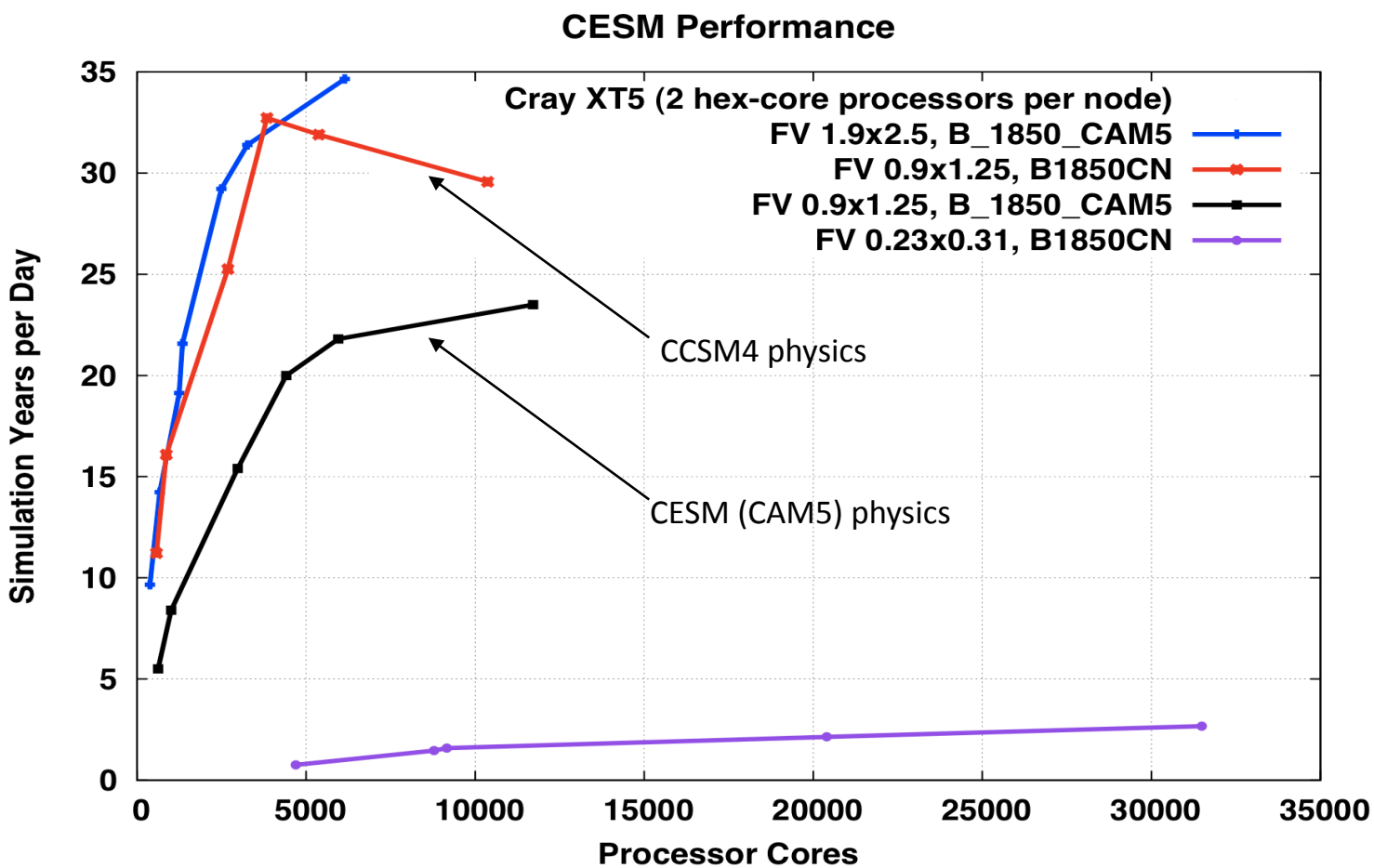
10 km grid

Preliminary 30,000 core results (collected 7/11/10)

POP has an equally long performance optimization history, including optimizing communication algorithms and subblocking and space-filling-curve decompositions to minimize static load imbalance and communication overhead. Most recently, OpenMP parallelism was re-enabled and further optimized.

CESM1 Cray XT Performance

CCSM4 vs. CESM1 (CAM5) physics



Resolutions

280 km ATM/LND
100 km OCN/ICE

140 km ATM/LND
100 km OCN/ICE

35 km ATM/LND
10 km OCN/ICE

Fixed size problems have limited scalability, and performance curves "roll over" eventually. However, this represents an opportunity for more expensive (local) physics. Using CESM1 (CAM5) physics was expected to be significantly more than twice as expensive as CCSM4 physics.

Preliminary results. Collected June 18-25, 2010.

What's Special about (CESM) Strong Scaling?

- Improved performance and performance scalability are critical for accelerating the science (even more so than for weak scaling).
- Performance improvements require attention to detail: performance optimization of all components is important, nothing scales forever, and constants matter.
- "Headroom" exists for significantly more expensive (local) physics.

Next Steps:

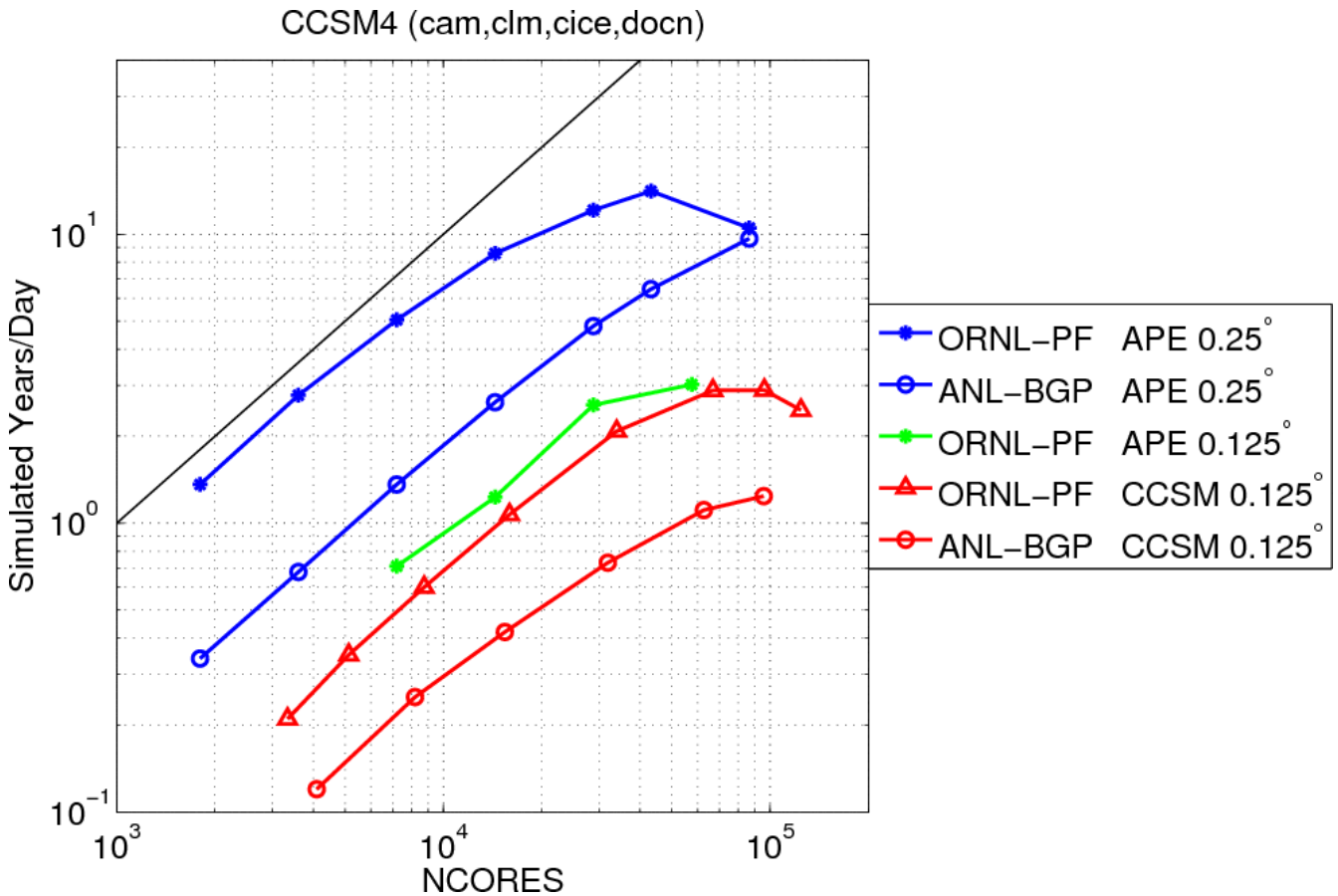
New, more scalable, numerical methods are being fast tracked for the atmosphere and ocean:

- *Spectral element on cubed sphere grid (HOMME)*
- *Finite volume on cubed sphere grid (FVCUBED)*
- *Finite volume on hexagonal grid (MPAS, for both atmosphere and ocean)*

Investigations have begun for sea ice as well.

CCSM/HOMME Scalability

18 and 35 km atm grid / 10 km ocn grid (DOCN)



Scalability demonstrated out to 86,000 cores (IBM BG/P) and achieve 3 SYPD (Cray XT5). *(SciDAC SA, NCAR CISL, NCAR CSEG, ..)*

Figure courtesy of M. Taylor (SNL).

What's Special about tuning the CESM?

Existing performance tuning options …

- Were each introduced for a reason, and have been very effective.
- Provide performance portability when changing platforms, when changing processor core count, and when changing problem specifications.
- Are not tested as often as they should be, due to the number of options.
- Are often not intuitive. In particular, determining a good choice for subblock size, domain decomposition, and process assignments in ocean and sea ice is a laborious manual task currently. This can make load balancing components in the CESM difficult.

An opportunity:

Tuning would be much easier to use if augmented with predictive modeling capability, auto-tuning infrastructure, and runtime auto-fault detection and mitigation software.

What's Special about Running at Scale?

- Communication algorithms that worked well for smaller processor counts may not perform well, or at all, at scale. In particular, any M-to-N collective in which M >> N will have problems on the Cray XT system, and performance issues have arisen on the IBM BG/P and other platforms in similar situations.
- Increased performance variability (on some systems)

And the usual issues

- I/O
- Memory scaling (application and system)
- Algorithms with O(process count), O(problem size), … computational complexities (anything that does not scale, but was not important enough to fix or was not obvious at smaller processor core counts)
- Faults
- …

Robust Communication Algorithms

- The arrival of large numbers of "unexpected messages" can cause performance problems or application failure.
  - Prepost all receive requests for a given communication collective?
- Preposting all receive requests can exhaust system memory and increase the cost of matching requests with incoming message.
- Preposting can not guarantee that requests are posted before message arrival unless additional synchronization in invoked.
- When enabled, CESM flow control communication preposts (some) receive requests, specifies message ordering and caps the maximize number of outstanding requests, and/or uses "4-byte" handshaking messages to guarantee that receives are requested before messages arrive. Specified per communication algorithm at compile- or runtime.
- Flow control in communication algorithms is critical on the Cray XT systems
  - Without flow control enabled, cannot run 30,000 processor 0.23 ATM/LND, tx0.1v2 OCN/ICE B1850CN experiment. Run out of node memory otherwise.

CESM Performance Engineering

What's Next?

- Just Getting Started: resolutions and other configuration options
  - 0.47x0.63 FV with 0.1 degree OCN/ICE, both CAM4 and CAM5
  - Trop_mozart, WACCM, …
- Just Getting Started: platforms
  - IBM BG/P (Intrepid), IBM Power6 (BlueFire), …
- Just Getting Started: ATM dycores
  - Spectral Eulerian (EUL), Spectral Element (HOMME), …
- Just Getting Started: performance optimization
  - Component load balancing
  - High process count decompositions for OCN and ICE
  - Detailed component performance analysis and further optimization
  - I/O

CESM Performance and Software Engineering Improvements Contributors

(including, but not limited to, …)

- CESM Software Engineering Group (NCAR)
  - David Bailey, Tony Craig, Brian Eaton, Chris Fischer, Diane Feddema, Brian Kauffman, Erik Kluzek, Andy Mai, Nancy Norton, Mathew Rothstein, Mariana Vertenstein, Francis Vitt, Jon Wolfe
- CESM Software Engineering Working Group (DOE/SciDAC, NCAR/CISL, ESMF)
  - John Dennis, John Drake, Jim Edwards, Nathan Hearn, Forrest Hoffman, Rob Jacob, Phil Jones, Fei Liu, Sheri Mickelson, Art Mirin, Ray Loy, Mark Taylor, John Truesdale, Pat Worley, ESMF Core Team
- CESM Scientists

For this talk, special thanks to my collaborators Art Mirin (LLNL) and Mariana Vertenstein (NCAR).