Recent advances in the GFDL Flexible Modeling System

4th ENES HPC Workshop
Toulouse, FRANCE

V. Balaji and many others

NOAA/GFDL and Princeton University

6 April 2016

Outline

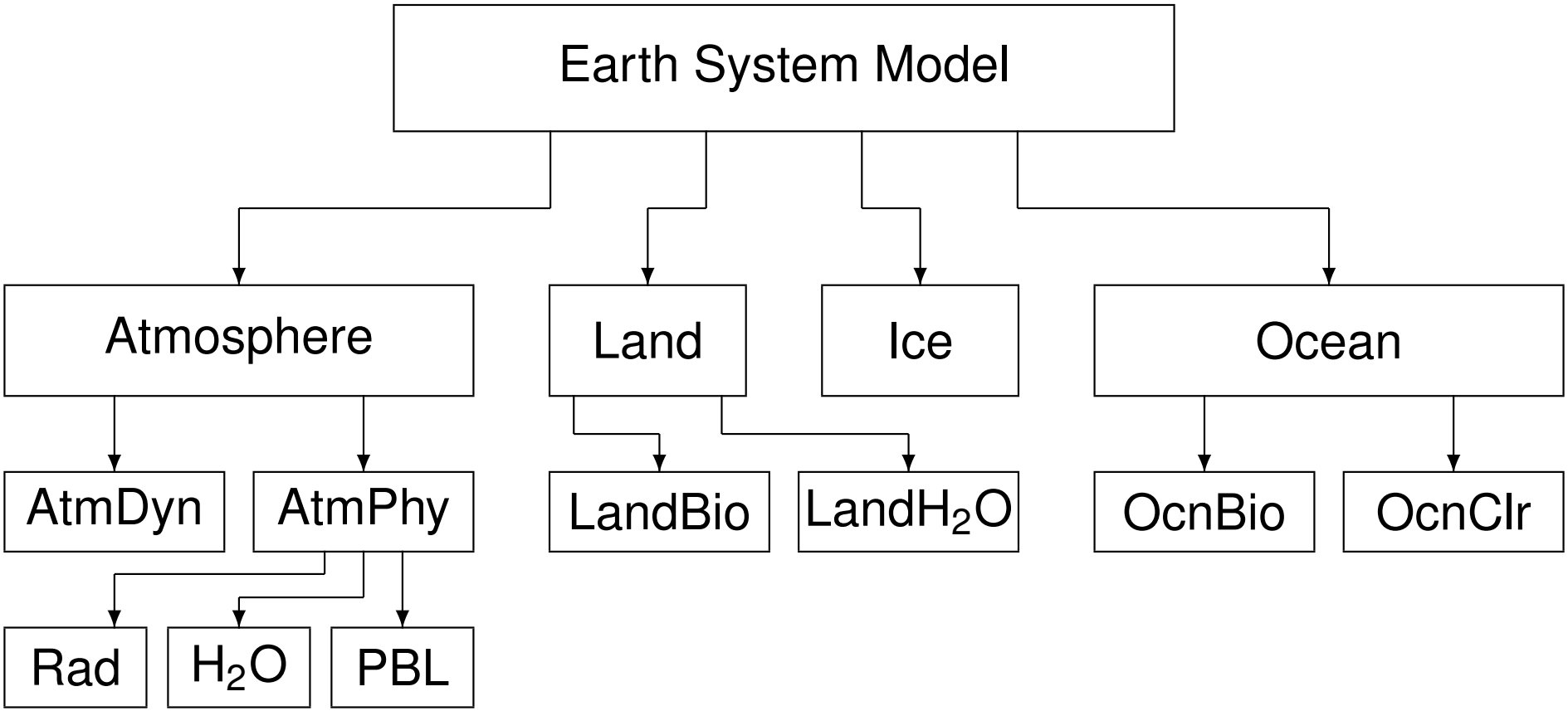
Outline

Climate modeling, a computational profile

- Intrinsic variability at all timescales from minutes to millennia; distinguishing natural from forced variability is a key challenge.
- coupled multi-scale multi-physics modeling;
- physics components have predictable data dependencies associated with grids;
- Adding processes and components improves scientific understanding;
- New physics and higher process fidelity at higher resolution;
- Ensemble methods to sample uncertainty (ICEs, PPEs, MMEs...)
- algorithms generally possess weak scalability.

In sum, climate modeling requires long-term integrations of weakly-scaling I/O and memory-bound models of enormous complexity.

Earth System Model Architecture



Complexity, resolution, UQ: components can have their own grids, timesteps, algorithms, multiple concurrent instances.

The hardware jungle

Upcoming hardware roadmap looks daunting! GPUs, MICs, DSPs, and many other TLAs. . .

- Intel straight line: IvyBridge/SandyBridge, Haswell/Broadwell: “traditional” systems with threading and vectors.
- Intel knight’s move: Knights Corner, Knights Landing: MICs, thread/vector again, wider in thread space.
- Hosted dual-socket systems with GPUs: SIMD co-processors.
- BG/Q: CPU only with hardware threads, thread and vector instructions. No followon planned.
- ARM-based systems coming. (e.g with DSPs).
- FPGAs? some inroads in finance.
- Specialized processors: Anton for molecular dynamics, GRAPE for astrophysics.

The software zoo

Exascale using nanosecond clocks implies billion-way concurrency!
It is unlikely that we will program codes with $10^6 - 10^9$ MPI ranks: it will be MPI+X. Solve for X . . .

- CUDA and CUDA-Fortran: proprietary for NVIDIA GPUs. Invasive and pervasive.
- OpenCL: proposed standard, not much penetration.
- ACC from Portland Group, now a new standard OpenACC.
- Potential OpenMP/OpenACC merging...?
- PGAS languages: Co-Array Fortran, UPC, a host of proprietary languages.
- Code generation:
  - Domain-specific languages (DSLs): e.g STELLA, Psy.
  - Source-to-source translators.

GFDL between jungle and zoo

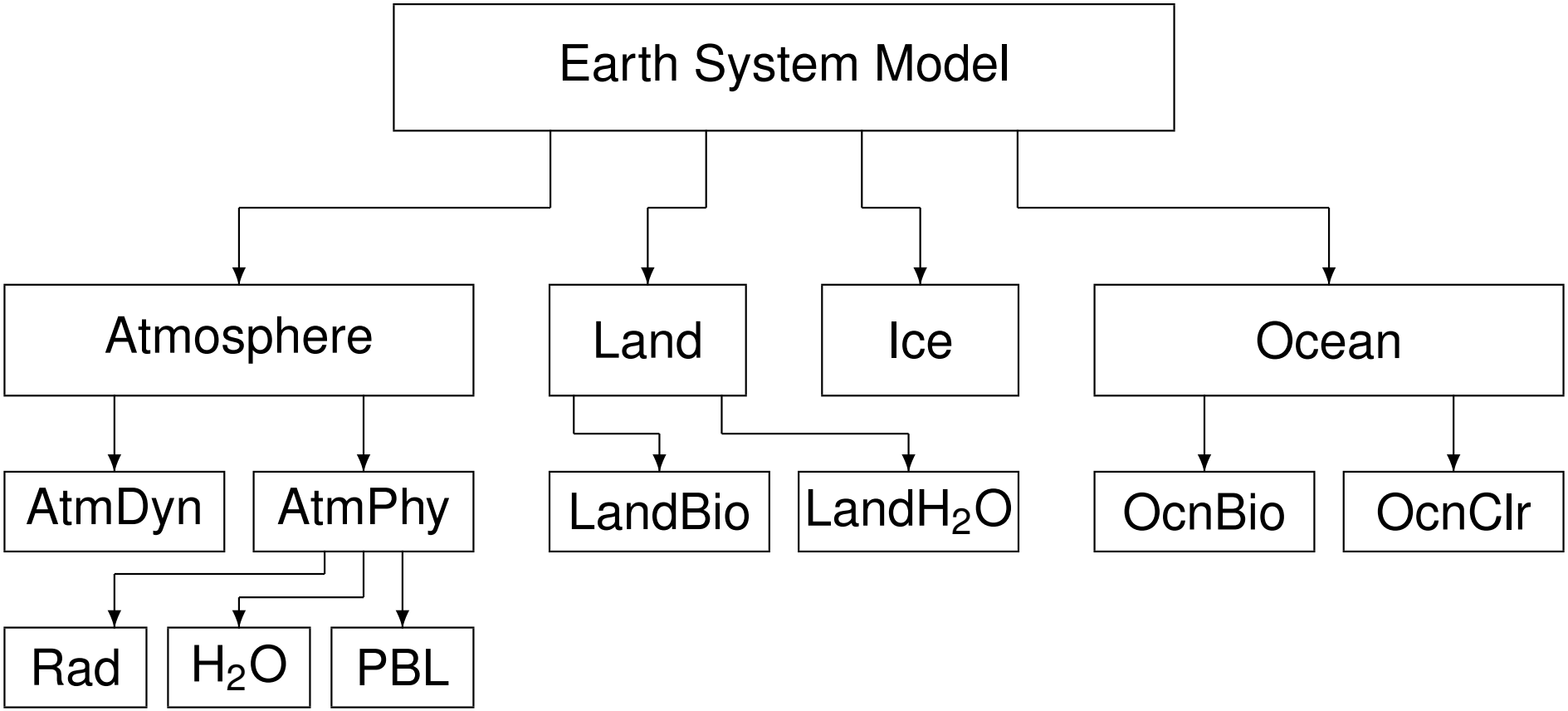
GFDL is taking a conservative approach:

- it looks like it will be a mix of MPI, threads, and vectors.
- Developing a three-level abstraction for parallelism: components, domains, blocks. Kernels work on blocks and must have vectorizing inner loops.
- Recommendation: sit tight, make sure MPI+OpenMP works well, write vector-friendly loops, reduce memory footprint, offload I/O.
- Other concerns:
  - Irreproducible computation
  - Tools for analyzing performance.
  - Debugging at scale.

Recent experience on Titan, Stampede and Mira reaffirm this approach.
This talk will focus on coarse-grained parallelism at the component level.

Outline

Earth System Model Architecture



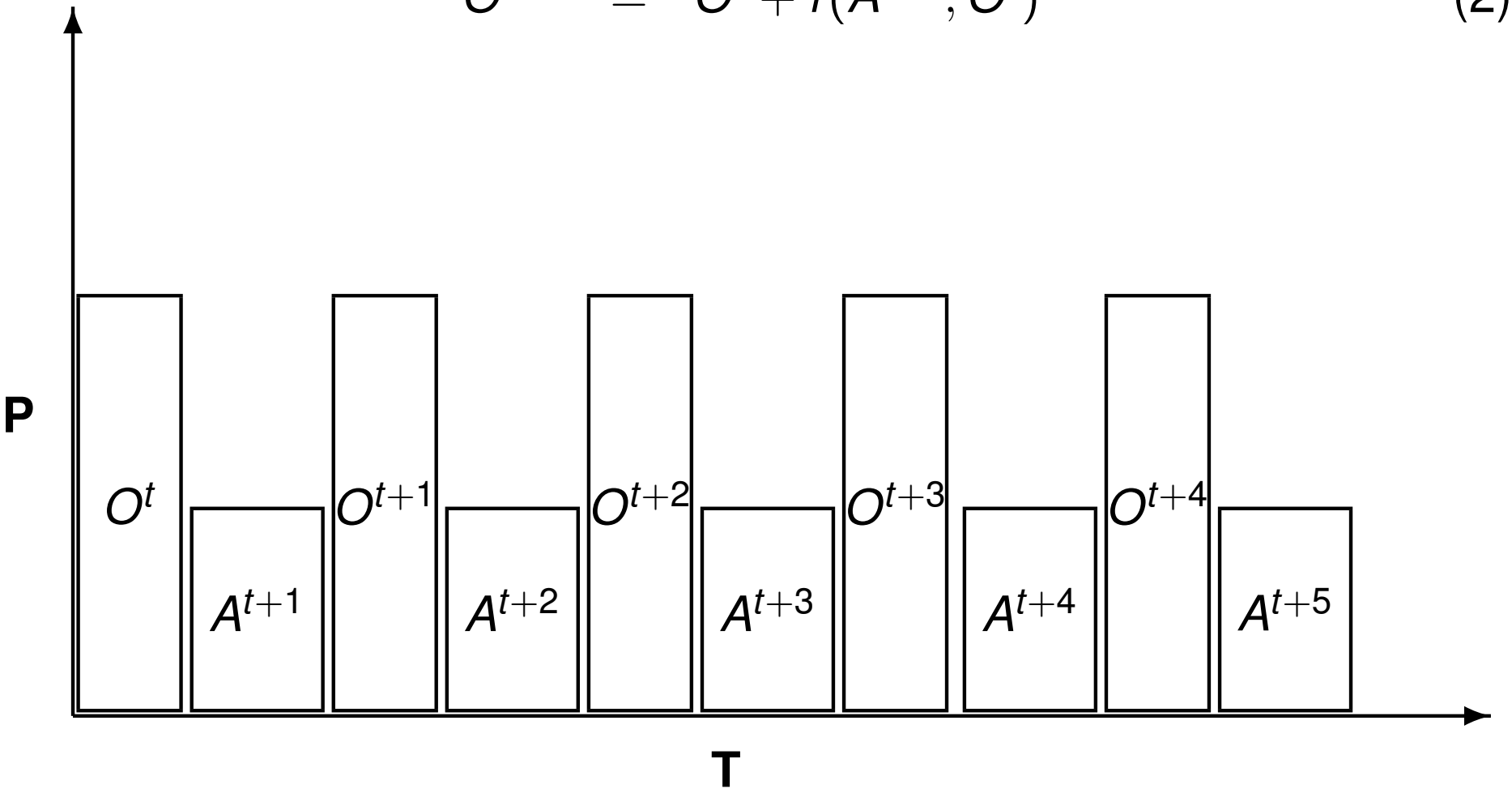
Extending component parallelism to $\mathcal{O}(10)$ requires a different physical architecture!

Serial coupling

Uses a forward-backward timestep for coupling.

$$A^{t+1} = A^t + f(A^t, O^t) \quad (1)$$

$$O^{t+1} = O^t + f(A^{t+1}, O^t) \quad (2)$$

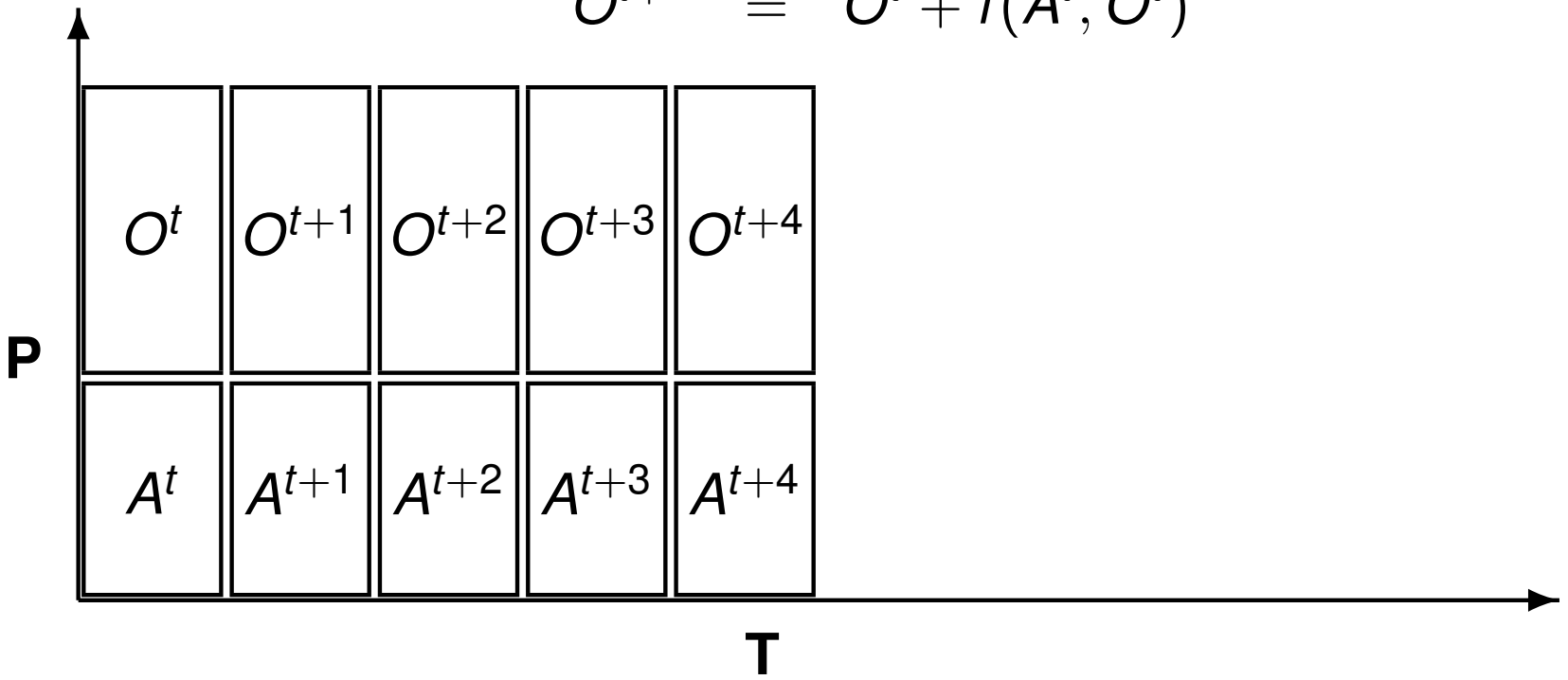


Concurrent coupling

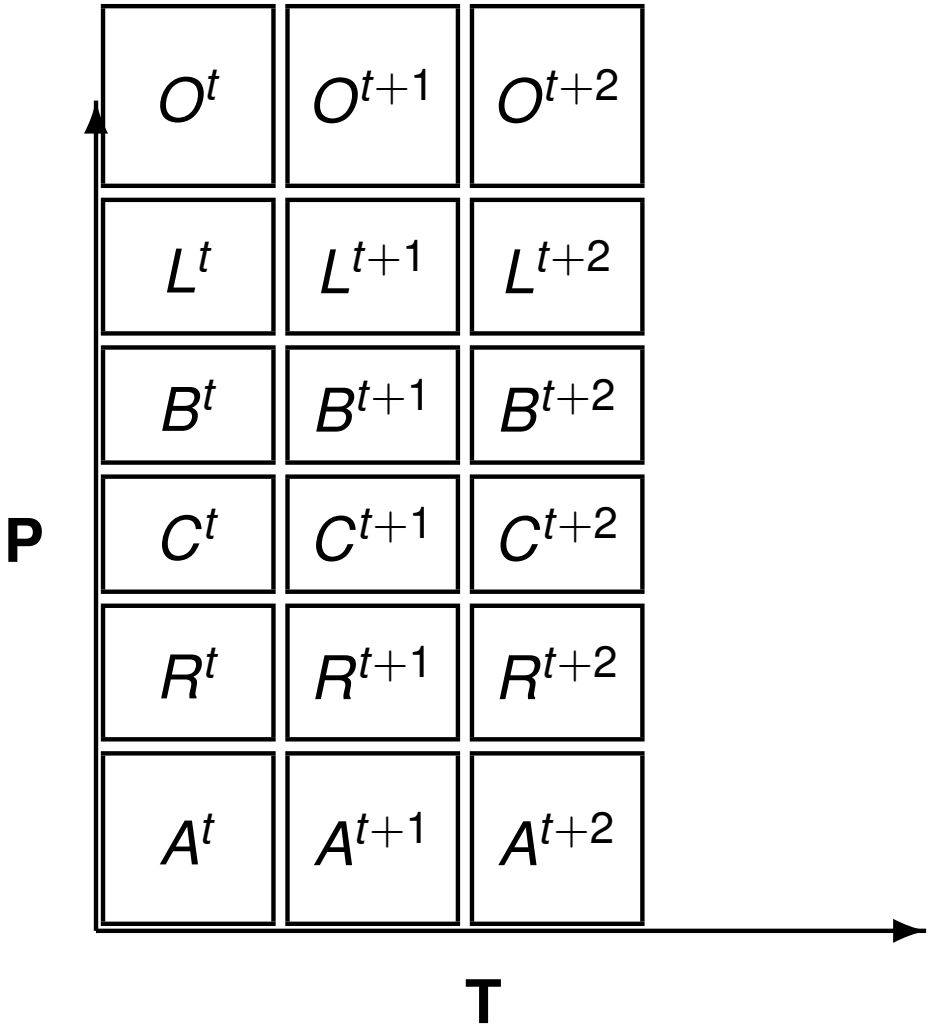
This uses a forward-only timestep for coupling. While formally this is unconditionally unstable, the system is strongly damped*. Answers vary with respect to serial coupling, as the ocean is now forced by atmospheric state from Δt ago.

$$A^{t+1} = A^t + f(A^t, O^t) \quad (3)$$

$$O^{t+1} = O^t + f(A^t, O^t) \quad (4)$$



Massively concurrent coupling



Components such as radiation, PBL, ocean biogeochemistry, each could run with its own grid, timestep, decomposition, even hardware. Coupler mediates state exchange.

Outline

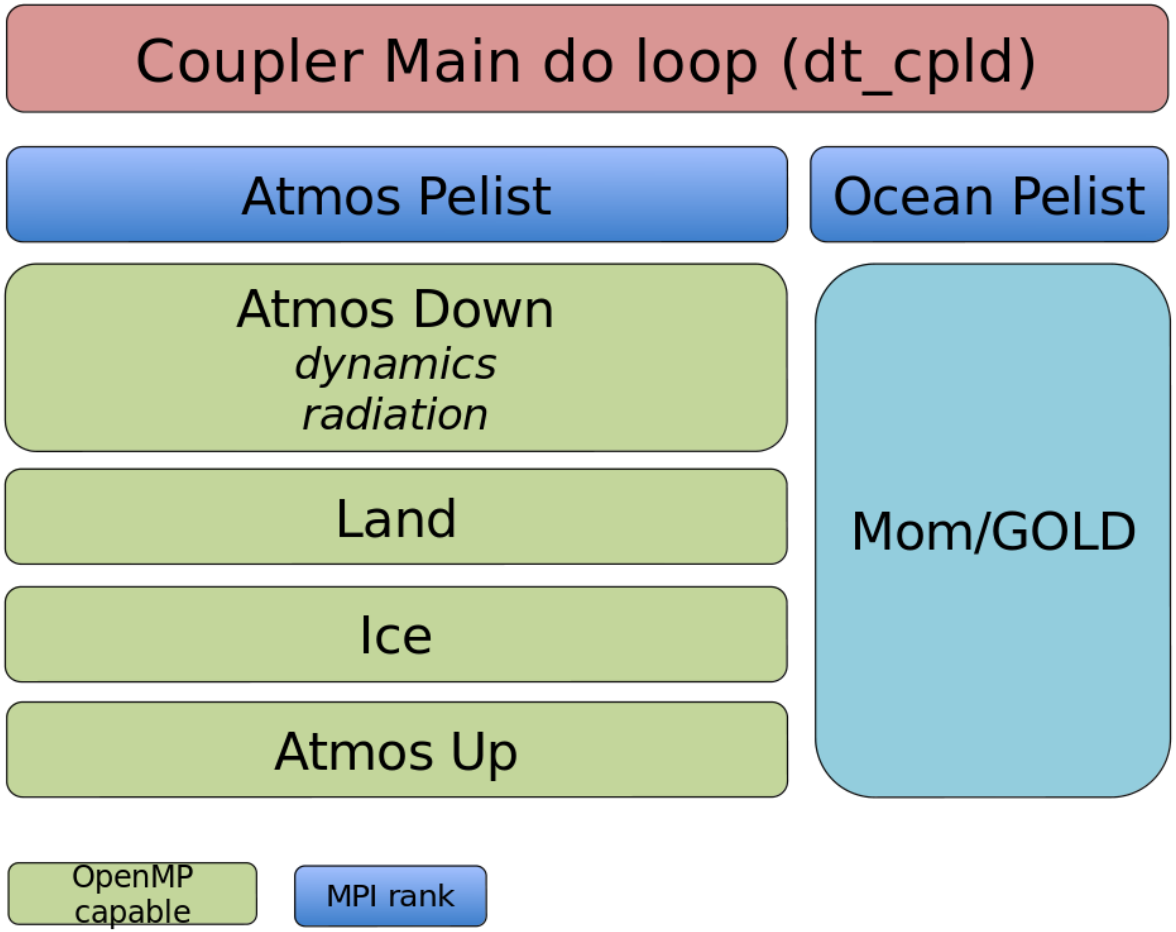
The radiation component

The atmospheric radiation component computes radiative transfer of incoming shortwave solar fluxes and outgoing longwave radiation as a function of all radiatively active species in the atmosphere (greenhouse gases, aerosols, particulates, clouds, ...).

- The physics of radiative transfer is relatively well-known, but a full Mie-scattering solution is computationally out of reach.
- Approximate methods (sampling the “line-by-line” calculation into “bands”) have been in use for decades, and “standard” packages like RRTM are available.
- They are still very expensive: typically $\Delta t_{rad} > \Delta t_{phy}$ (in the GFDL models typically 9X). The model is sensitive to this ratio.
- Other methods: stochastic sampling of bands (Pincus and Stevens 2013), neural nets (Krasnopolsky et al 2005)

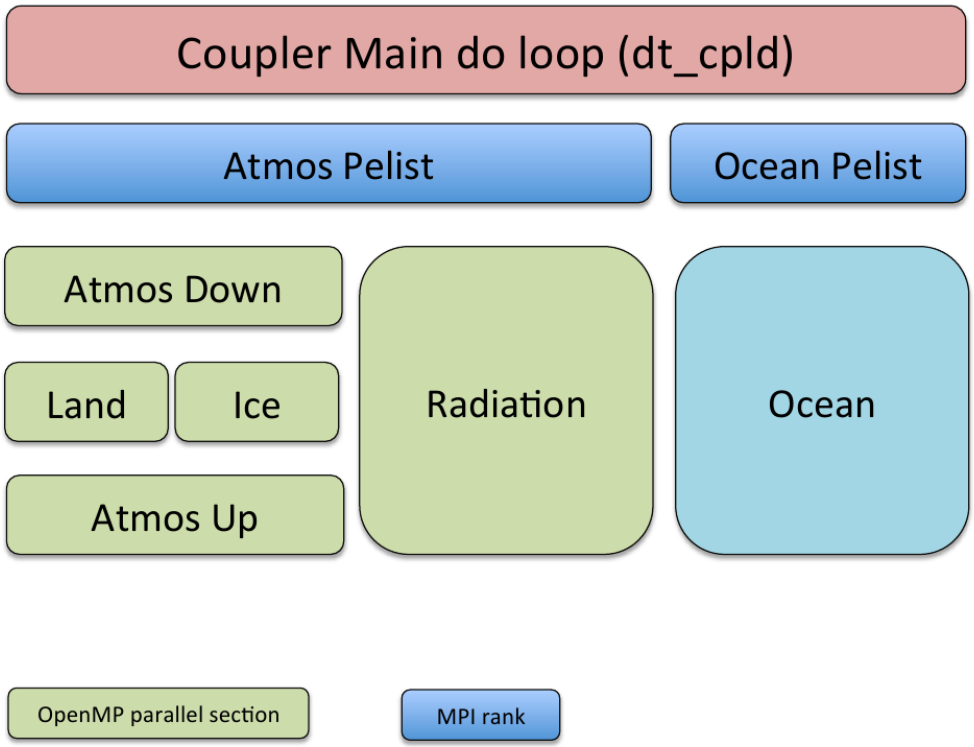
Challenge: can we exploit “cheap flops” to set $\Delta t_{rad} = \Delta t_{phy}$?

Traditional coupling sequence



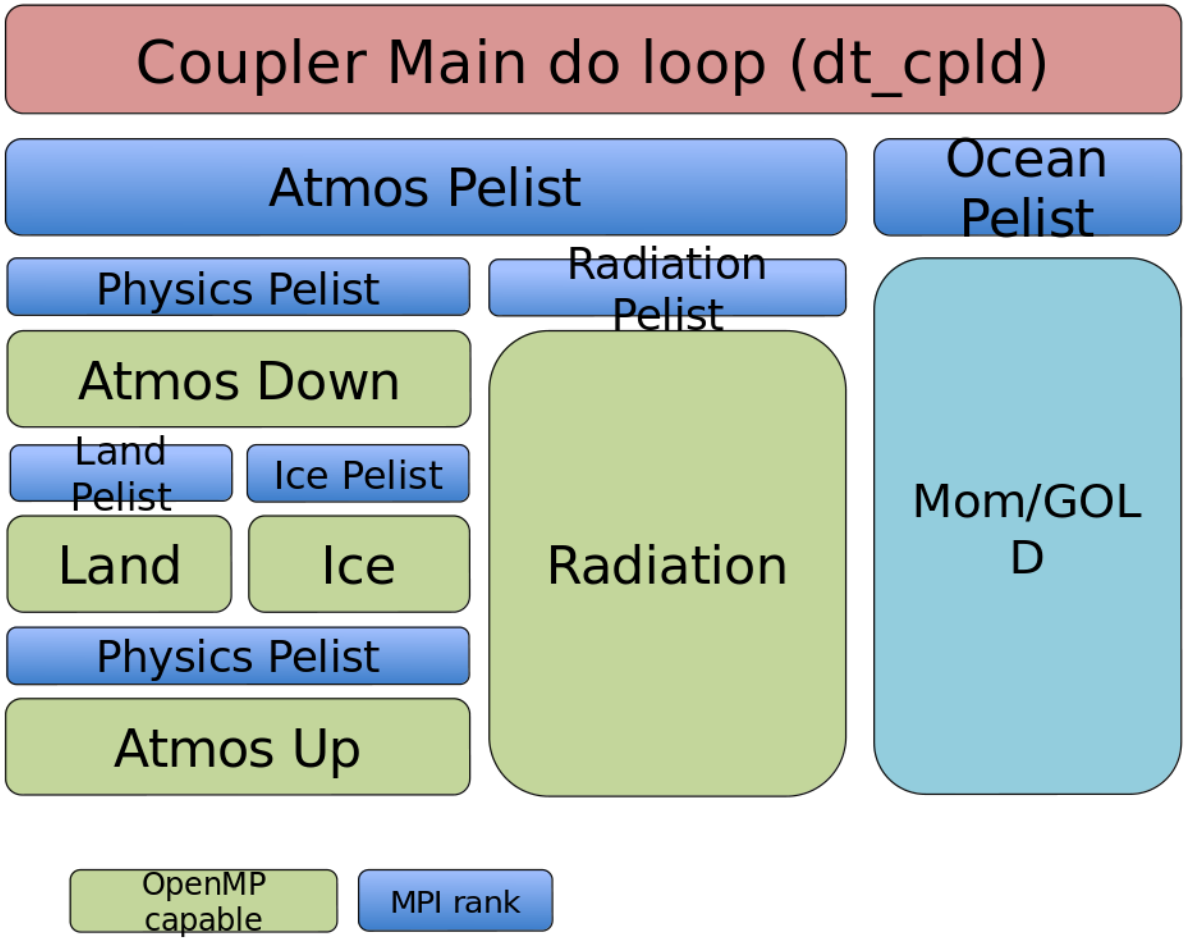
Radiation timestep much longer than physics timestep.
(Figure courtesy Rusty Benson, NOAA/GFDL).

Proposed coupling sequence



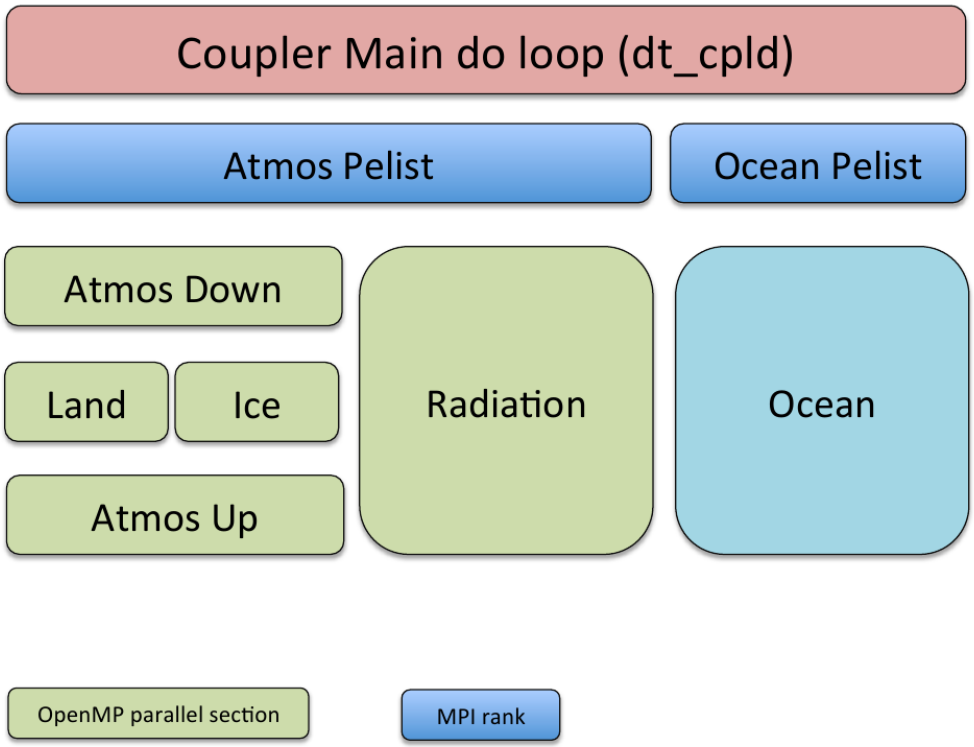
Radiation executes on physics timestep from lagged state.
(Figure courtesy Rusty Benson, NOAA/GFDL).

Proposed coupling sequence using pelists



Requires MPI communication between physics and radiation.
(Figure courtesy Rusty Benson, NOAA/GFDL).

Proposed coupling sequence: hybrid approach



Physics and radiation share memory.
(Figure courtesy Rusty Benson, NOAA/GFDL).

Results from climate run

20 year AMIP/SST climate runs have completed on Gaea (Cray XE6).

- Control: 9.25 sypd
  - $\Delta t_{rad} = 9\Delta t_{phy}$
  - 864 MPI-ranks / 2 OpenMP threads
- Serial Radiation: 5.28 sypd
  - $\Delta t_{rad} = \Delta t_{phy}$
  - 864 MPI-ranks / 2 OpenMP threads
- Concurrent Radiation: 5.90 sypd
  - $\Delta t_{rad} = \Delta t_{phy}$
  - 432 MPI-ranks / 4 OpenMP threads (2 atmos + 2 radiation)
  - Can get back to 9 sypd at about $\sim$2700 cores (roughly 1.6X).

Comparison of Concurrent Radiation to Control

- climate is similar
- TOA balance is off by $\sim 4 W/m^2$, mostly in the short wave, but easily retuned when ready to deploy

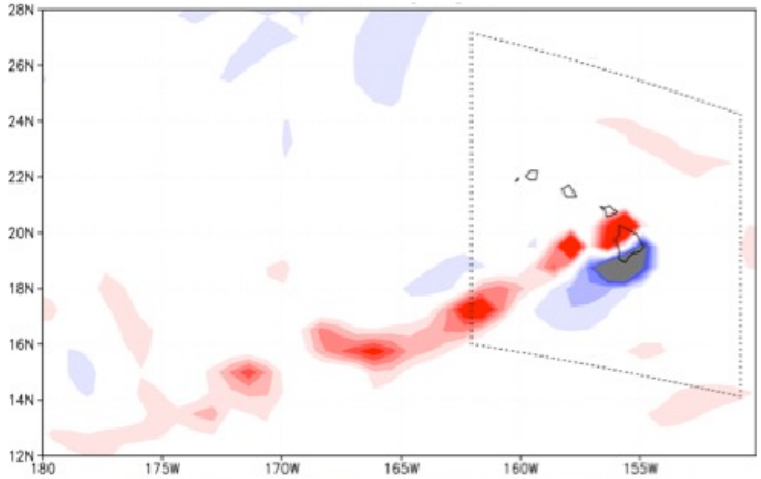
Results presented at AMS (Benson et al 2015). Article in the works for GMD special issue on coupling.

Outline

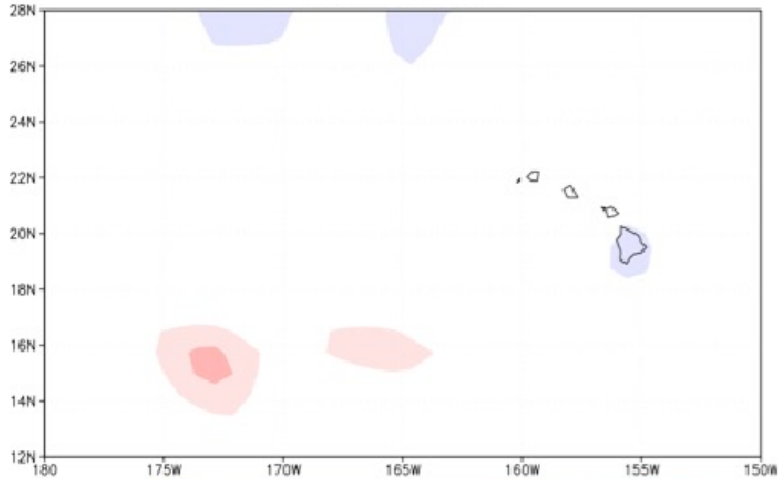
Lee vortices off Hawaii under two-way nesting

- 72 hr forecast from 1 Aug 2010 00Z with real topography
- Showing Vorticity × 10^5

c360 single grid



c120 single grid



c120 nested grid

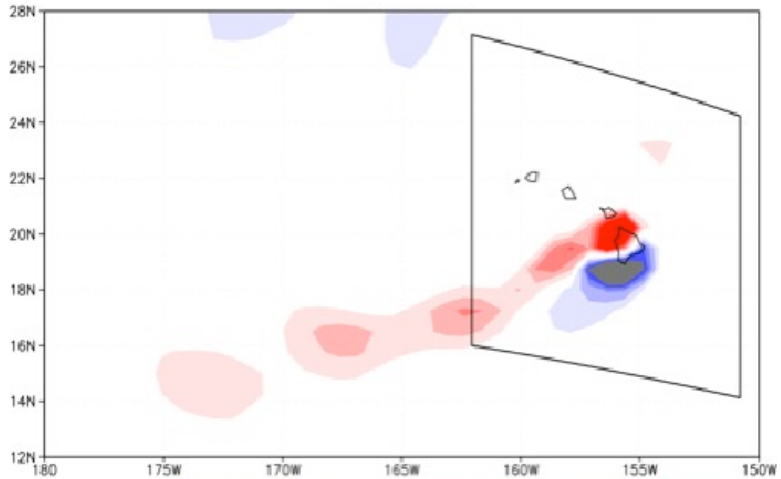
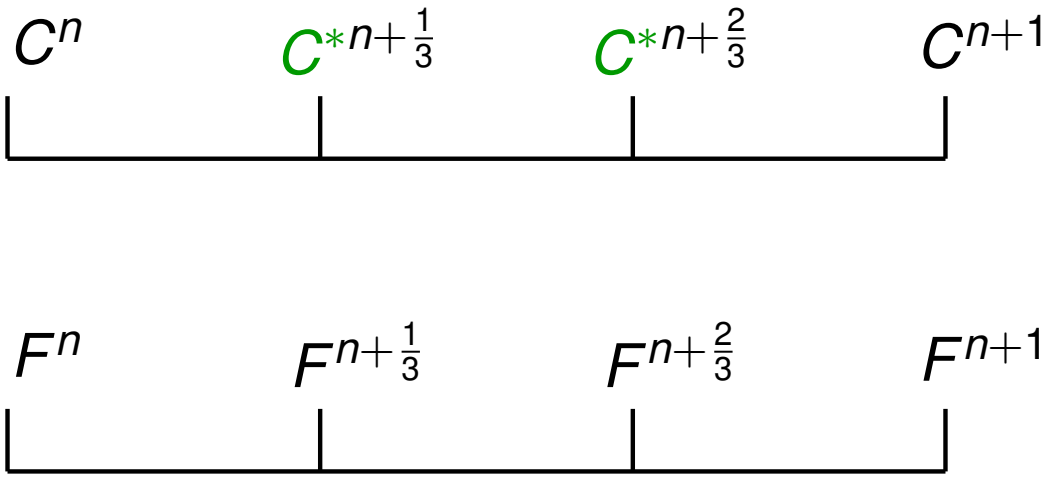


Figure courtesy Lucas Harris and S-J Lin, NOAA/GFDL.

Concurrent two-way nesting

Typical nesting protocols force serialization between fine and coarse grid timestepping, since the C^* are estimated by interpolating between C^n and C^{n+1}.



We enable concurrency by instead estimating the C^* by extrapolation from C^{n-1} and C^n, with an overhead of less than 10%. (See Harris and Lin 2012 for details.)

Sequential coupling

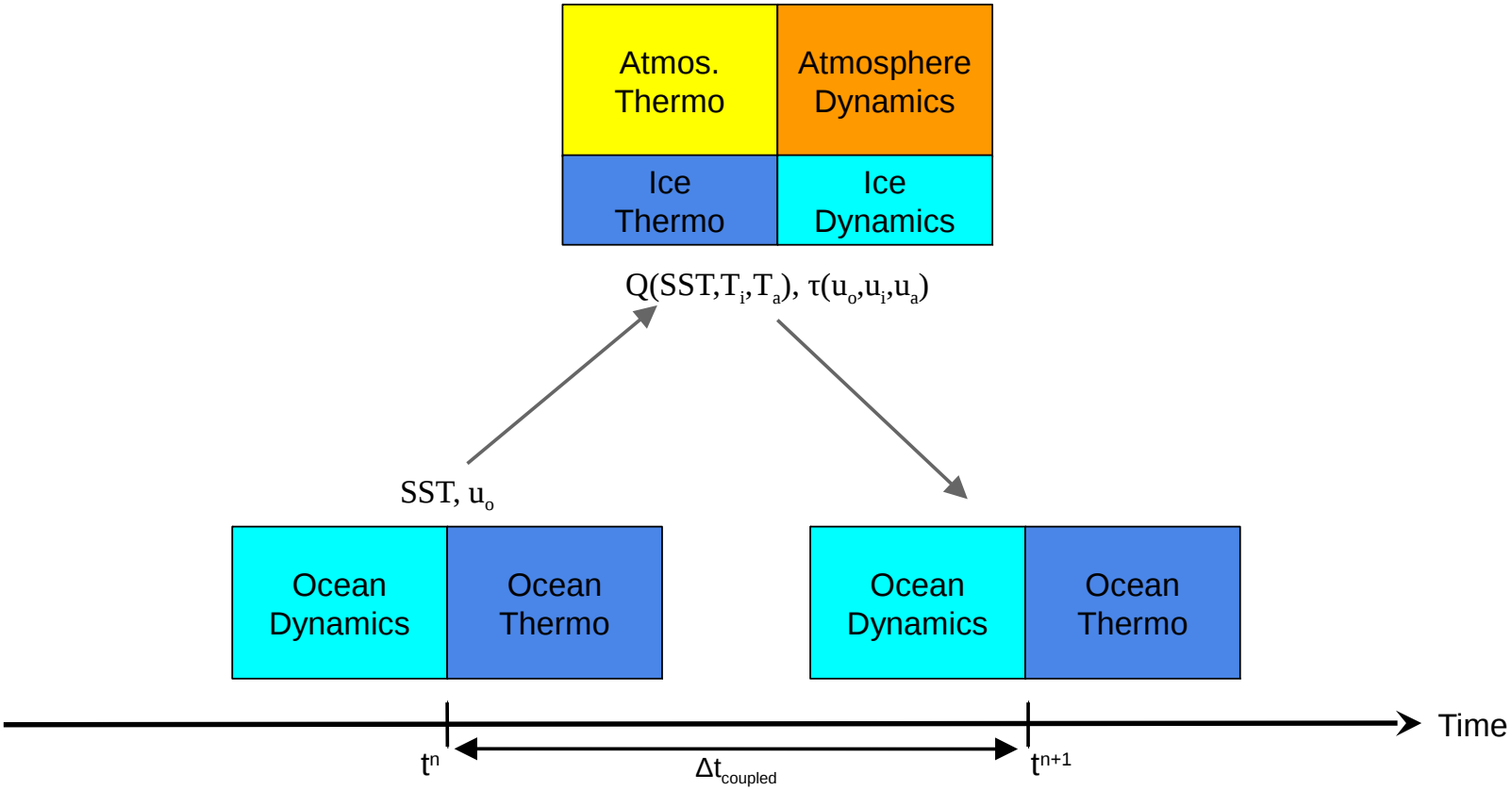


Figure courtesy Alistair Adcroft, Princeton University.

Concurrent coupling

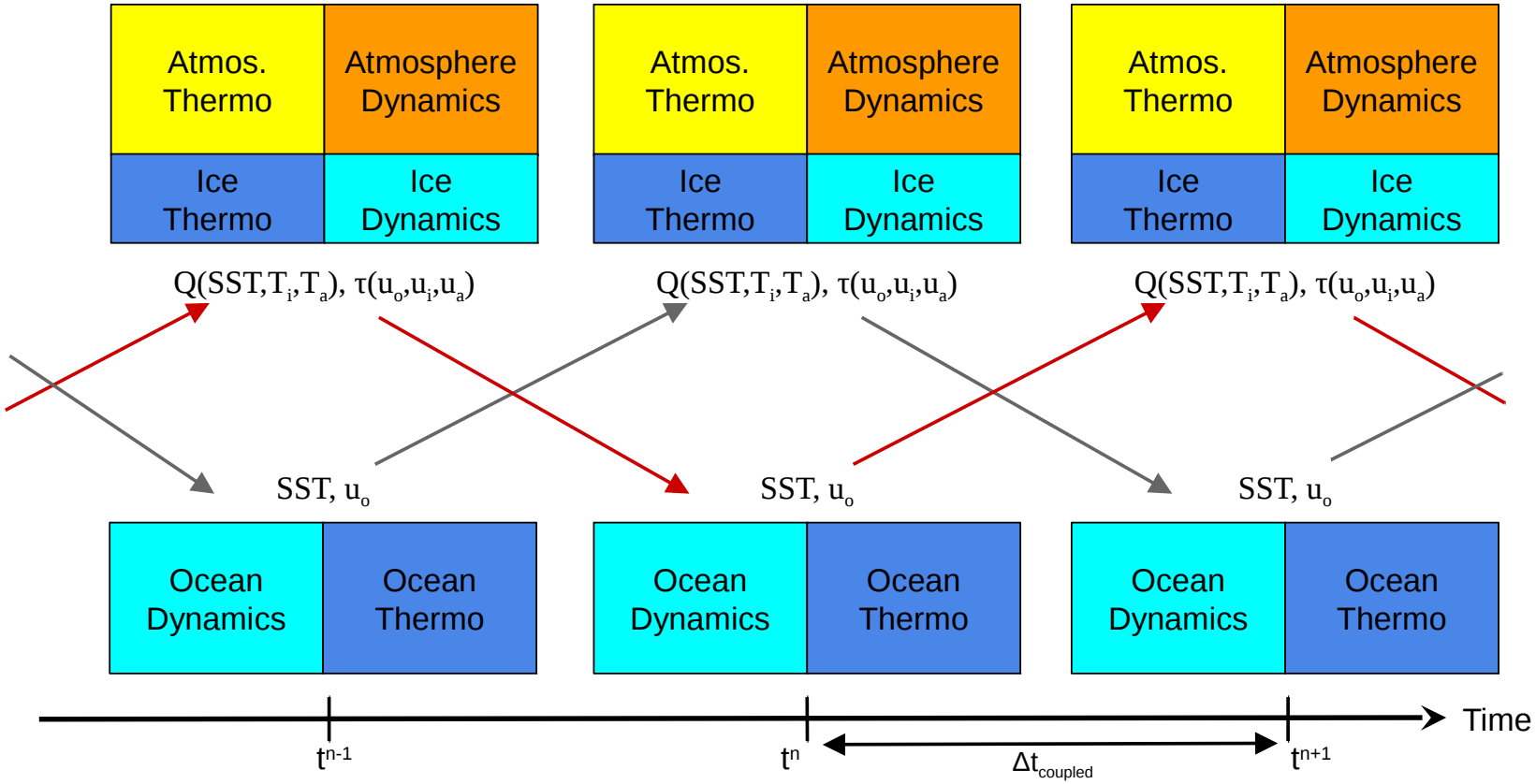


Figure courtesy Alistair Adcroft, Princeton University.

Staggered-concurrent coupling

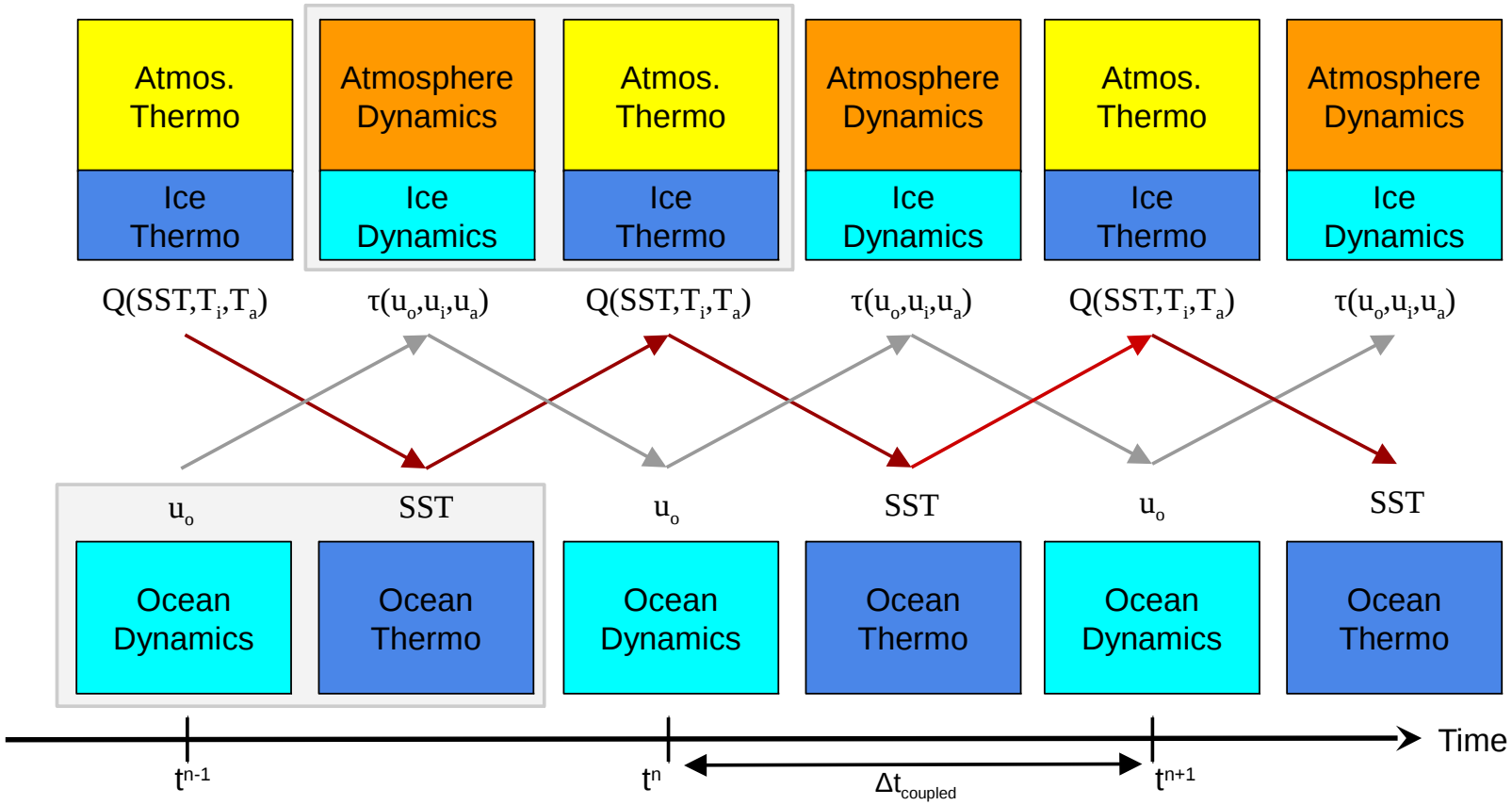


Figure courtesy Alistair Adcroft, Princeton University.

Sequential coupling + Adams-Bashforth

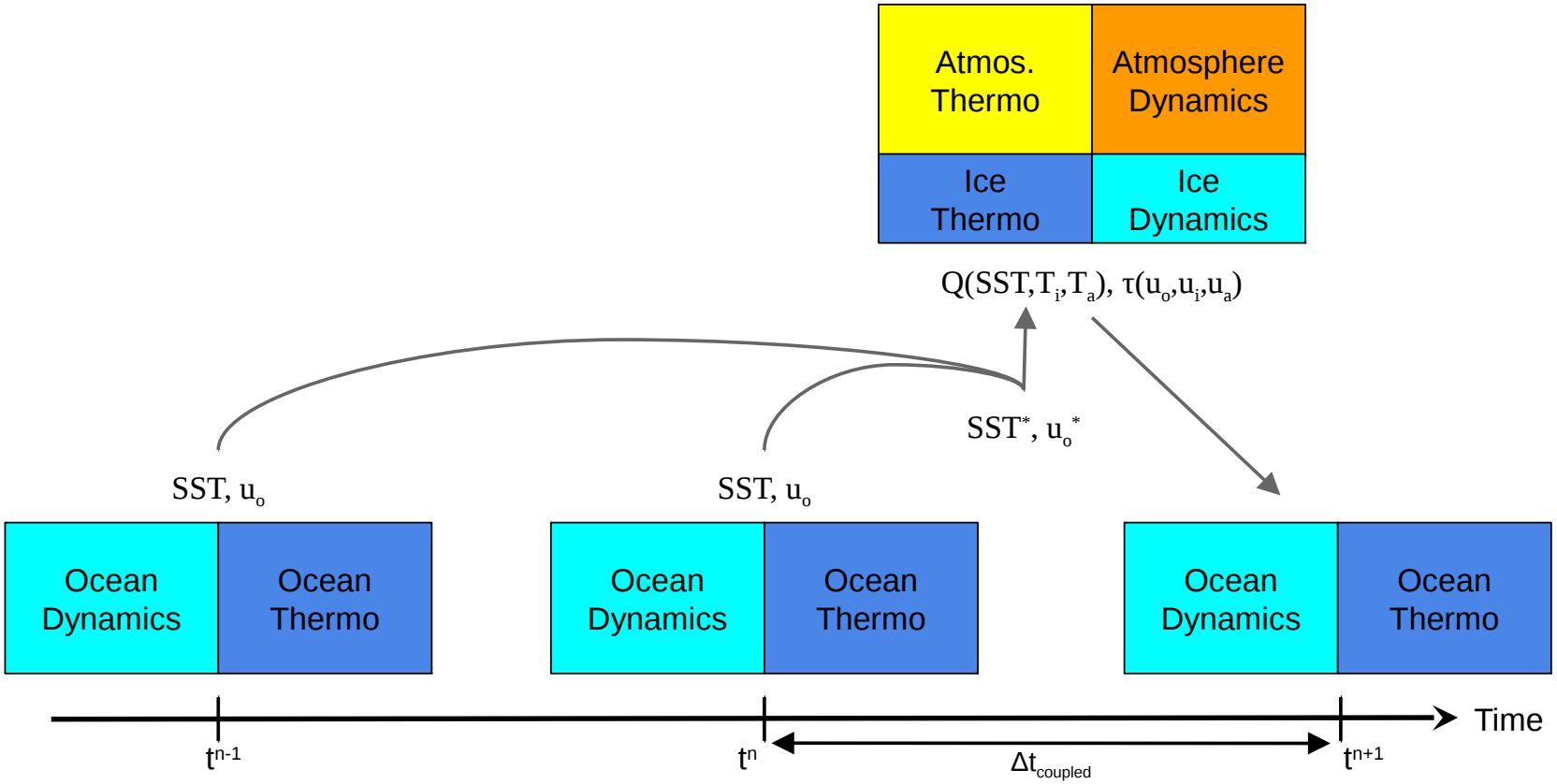


Figure courtesy Alistair Adcroft, Princeton University.

Concurrent coupling + AB

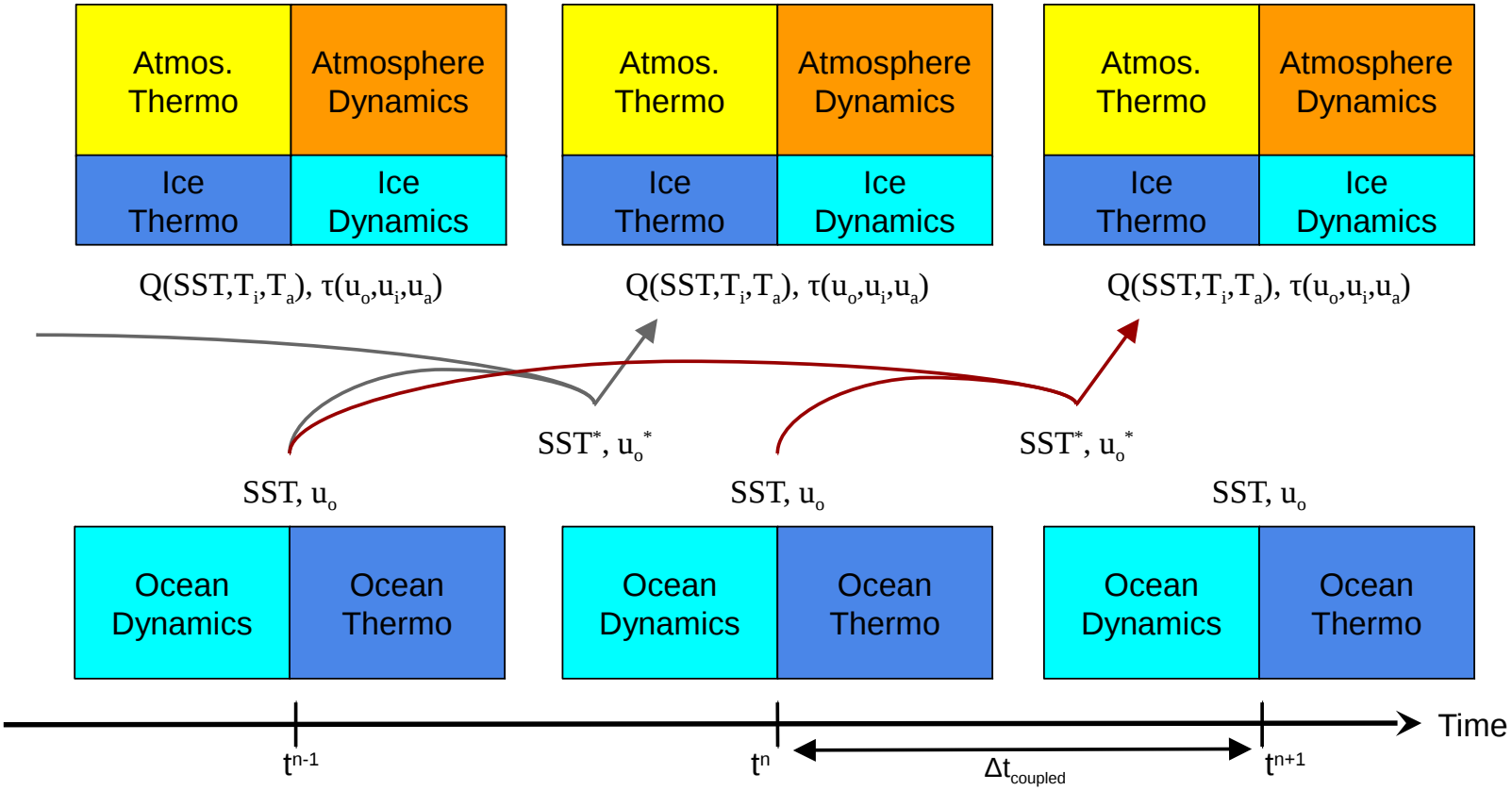


Figure courtesy Alistair Adcroft, Princeton University.

Staggered-concurrent coupling + AB

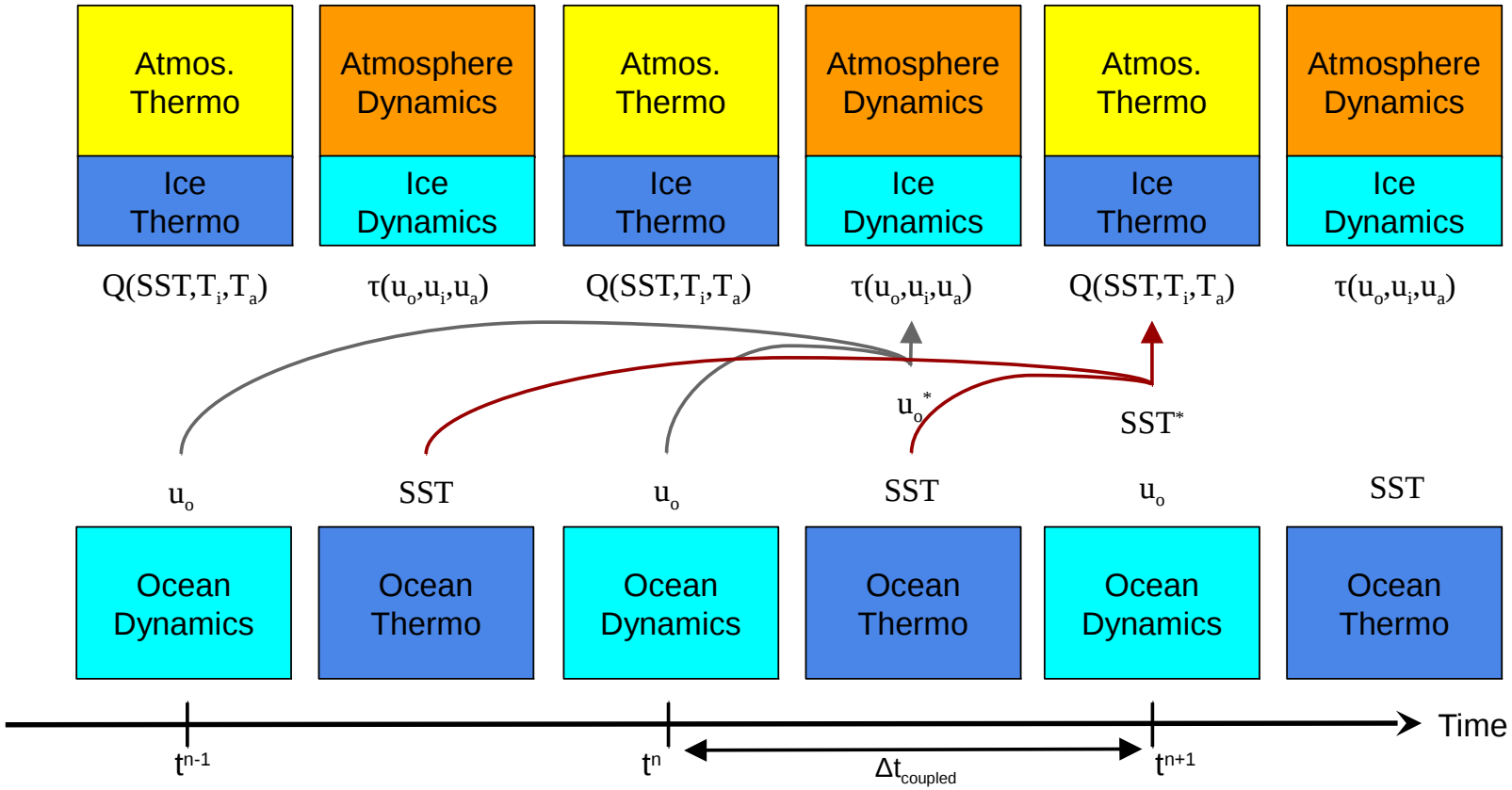


Figure courtesy Alistair Adcroft, Princeton University.

Outline

The NGGPS Effort

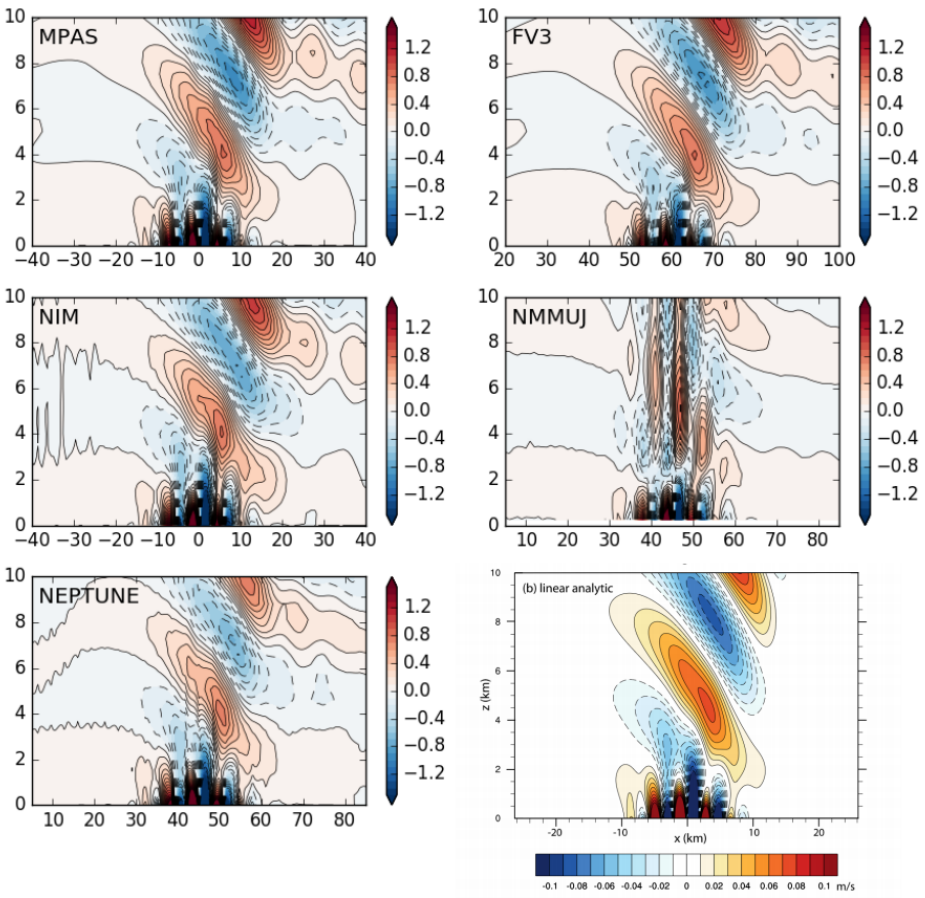
- NGGPS: Next-Generation Global Prediction System
- HIWPP: High-Intensity Weather Prediction Program

NGGPS and HIWPP launched a program to select a dynamical core for the next-generation forecast model (target: 3 km non-hydrostatic in 10 years). Selected dycores will undergo a substantial re-engineering effort for novel architectures.

- Scaling tests
- Idealized baroclinic wave test with embedded fronts (DCMIP 4.1)
- non-hydrostatic orographic mountain waves on reduced-radius sphere, no rotation
- idealized supercell thunderstorm on reduced-radius sphere, no rotation

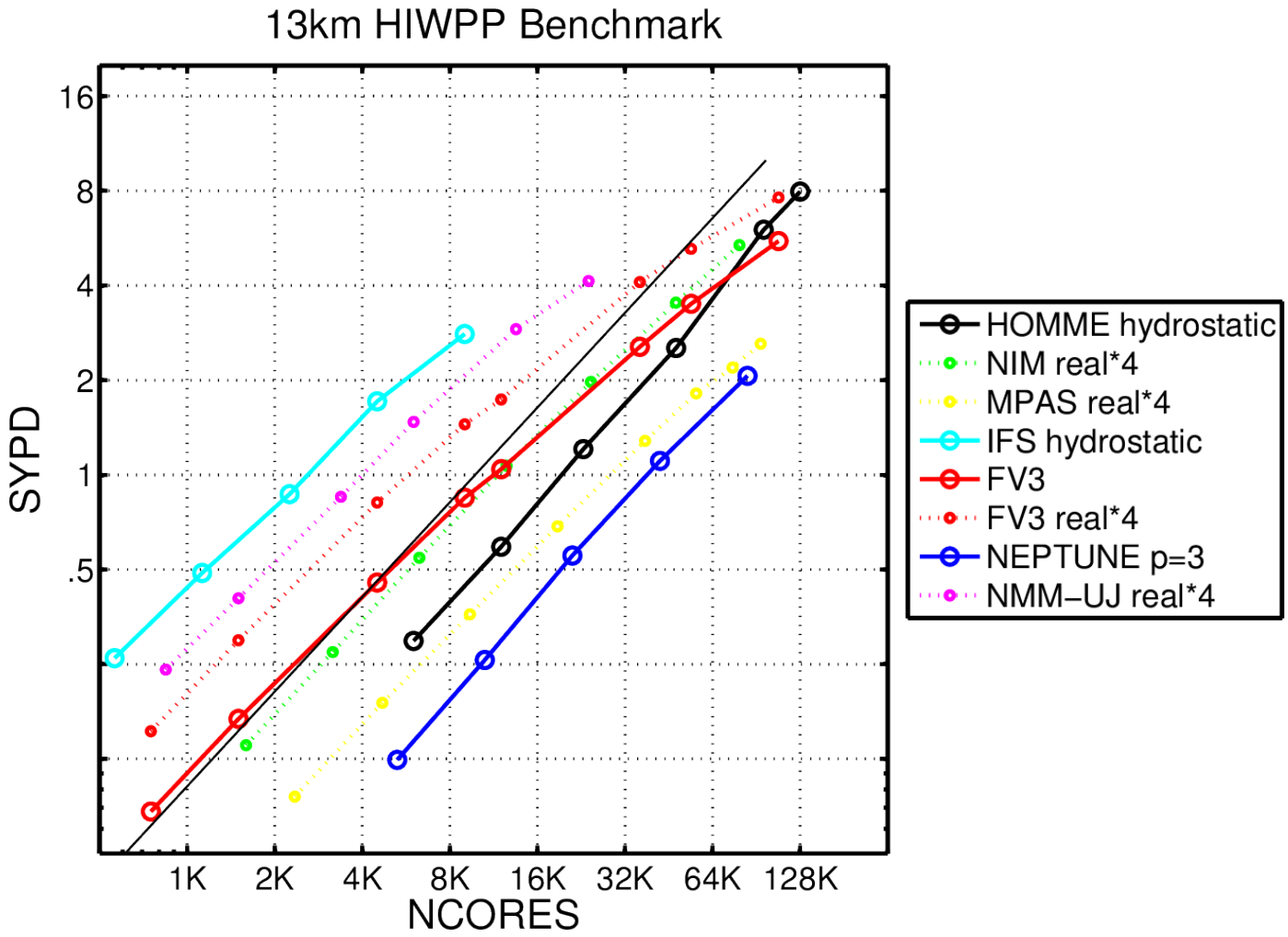
http://www.nws.noaa.gov/ost/nggps/dycoretesting.html

NGGPS Mountain Wave test case



http://www.nws.noaa.gov/ost/nggps/dycoretesting.html

NGGPS Scaling Study



Next steps: two models selected for Phase II: well-known NWP test cases with common physics (GFS) at 13 km. Results later this year.

SENA: Software Engineering for Novel Architectures

SENA is a multi-year NOAA effort to prepare key NOAA models for novel architectures (GPUs and MICs).

- SENA metrics of success: fraction of identified models successfully run on novel architectures.

GFDL projects under SENA:

- Close collaboration with compiler groups: PGI (Nvidia) and CCE (Cray) targeted at OpenACC.
- All porting efforts based on well-defined scientific benchmarks (viz. NGGPS, radiation).
- Likely to bear fruit when hardware memory architectures improve: expected in Volta (GPU-CPU with shared memory) and Knights Landing (MCDRAM).

Outline

Summary

- Moore's law has taken us from the von Neumann model to the "sea of functional units" (Kathy Yelick). Not easy to understand, predict or program performance.
- ... but the "free lunch" decades are over, they've come to take away your plates.
- Coarse-grained parallelism is an area in the current effort to reclaim performance from the encroaching "sea".
- The "component" abstraction still may let us extract some benefits out of the machines of this era:
  - sharing of the wide thread space.
  - distribute components among heterogeneous hardware?
  - concerns about stability, conservation, and accuracy.
- Presented at AMS, CW2015, GMD paper near submission: "Coarse-grained component concurrency in Earth System modeling", Balaji et al 2016.
- NGGPS and SENA efforts: driven by scientific benchmarks.