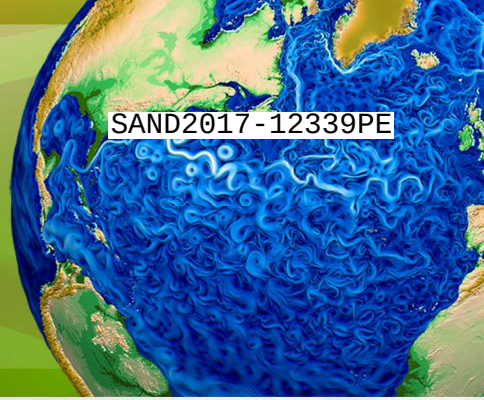
E3SM
Energy Exascale
Earth System Model



SAND2017-12339PE

Earth System Modeling on Upcoming Exascale Computers

Mark Taylor
mataylo@sandia.gov

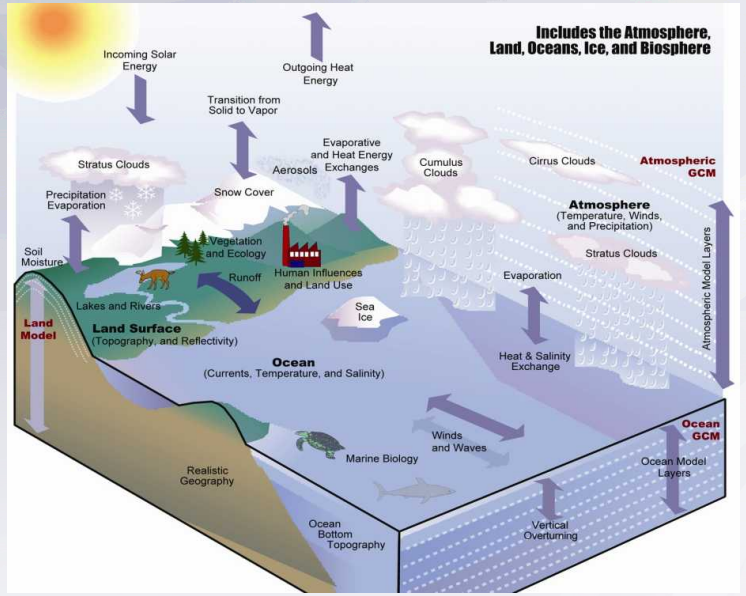
U.S. DEPARTMENT OF ENERGY

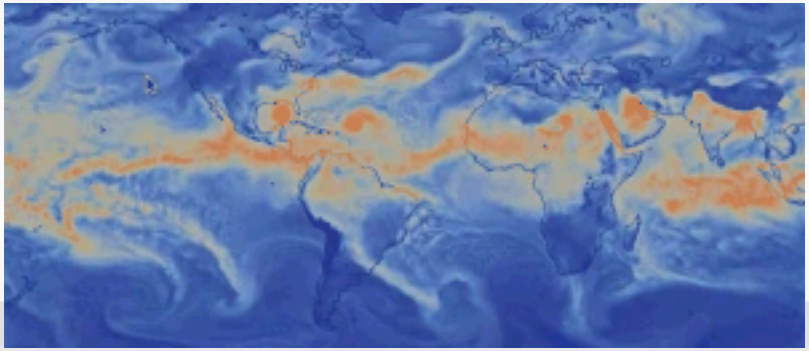
Outline

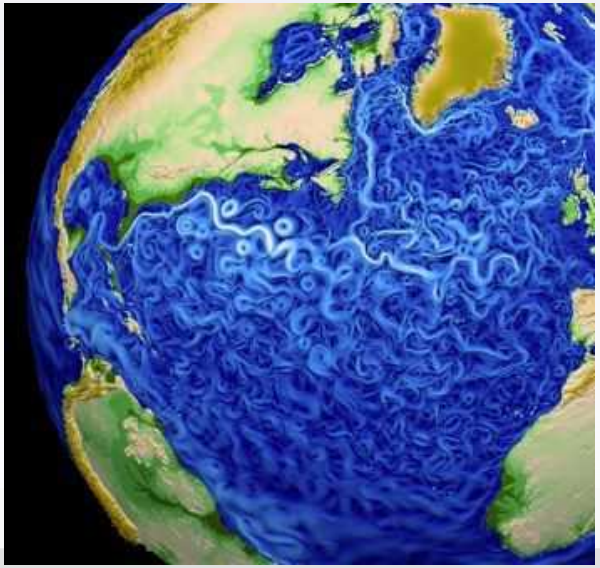
- Overview of E3SM (formerly ACME)
  - DOE's Coupled Atmosphere/Ocean/Sea Ice/Land Earth System Model
  - DOE Science Drivers
- Performance on existing LCF machines: Xeon Phi (KNL) and Xeon
- GPU Performance and Strategy
- E3SM-MMF
  - Multiscale "super-parameterization" configuration for GPU architectures

E3SM (formerly ACME)

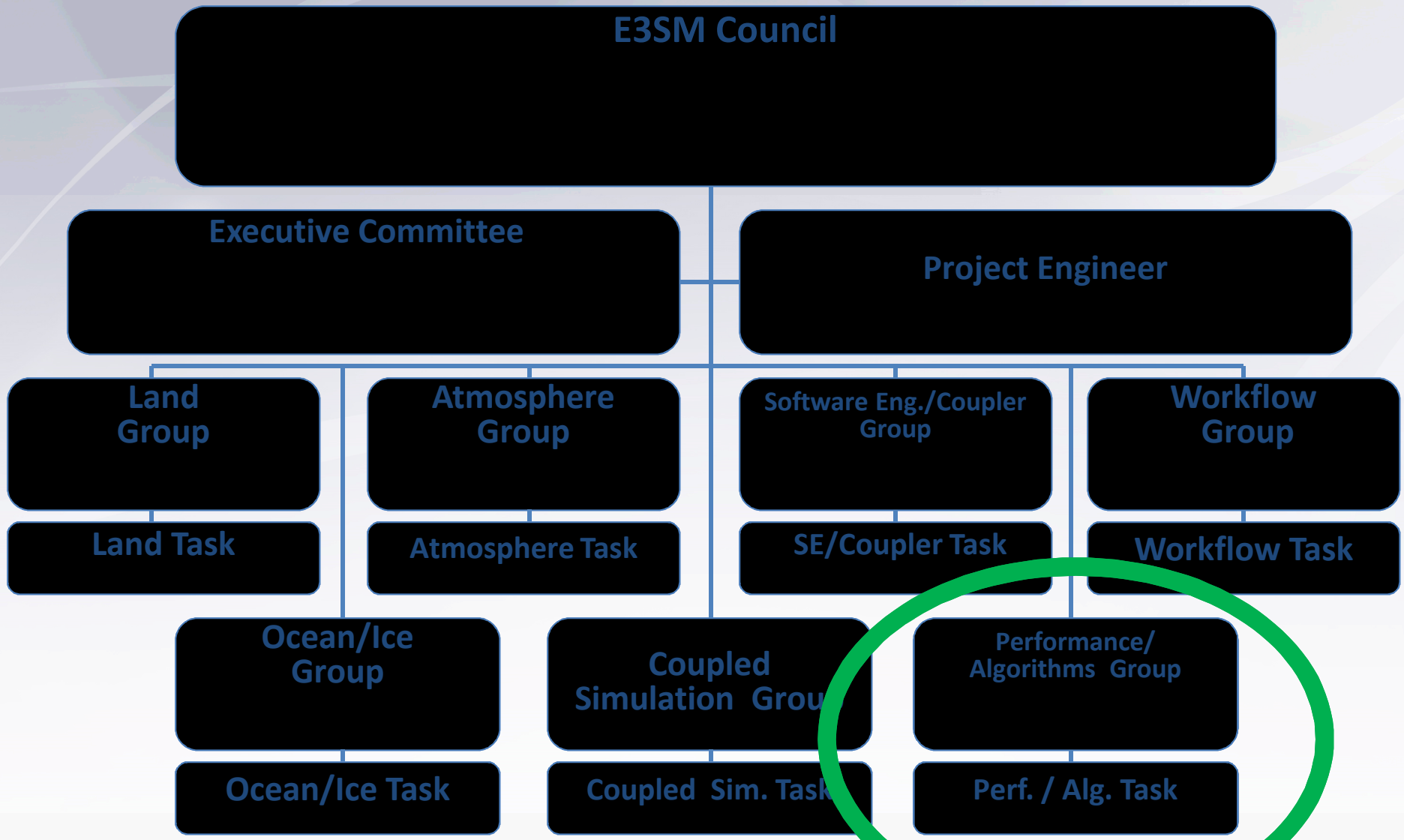
- Energy Exascale Earth System Model
- 8 DOE labs, NCAR, and universities. Total ~45 FTEs spread over 100 staff
- Atmosphere, Land, Ocean and Ice component models
- Development driven by DOE-SC mission interests: Energy/water issues looking out 40 years
- **Key computational goal: Ensure E3SM will run well on next generation DOE leadership computing facilities**
- E3SM is open source with first public release of code & simulations in early 2018







E3SM Structure



E3SM Performance Group

- Performance Group Leads:
  - Phil Jones, Pat Worley
- Researchers:
  - Az Mametjanov, Noel Keen, Matt Norman, Sarat Sreepathi, Ben Mayer
- New Additions:
  - Hongzhang Shan, Mathias Jacquelin, David Gunter

DOE computers

- Transition to new architectures will be disruptive
  - Comparable to the transition from vector to parallel supercomputers in the 2000's
  - DOE leads U.S. efforts to develop exascale computers
- By 2018, >95% of the computing power in DOE will be on multicore (Intel Phi) or NVIDIA GPU systems
  - Driven mostly by power considerations
- If we do nothing, today's codes will run *slower* on these systems than they run on today's systems
  - During the transition from vector to MPP: it took ~5 years before massively parallel supercomputers could could outperform parallel vector systems on climate applications

Earth Simulator, 2002

BG/L, 2008

Science Drivers

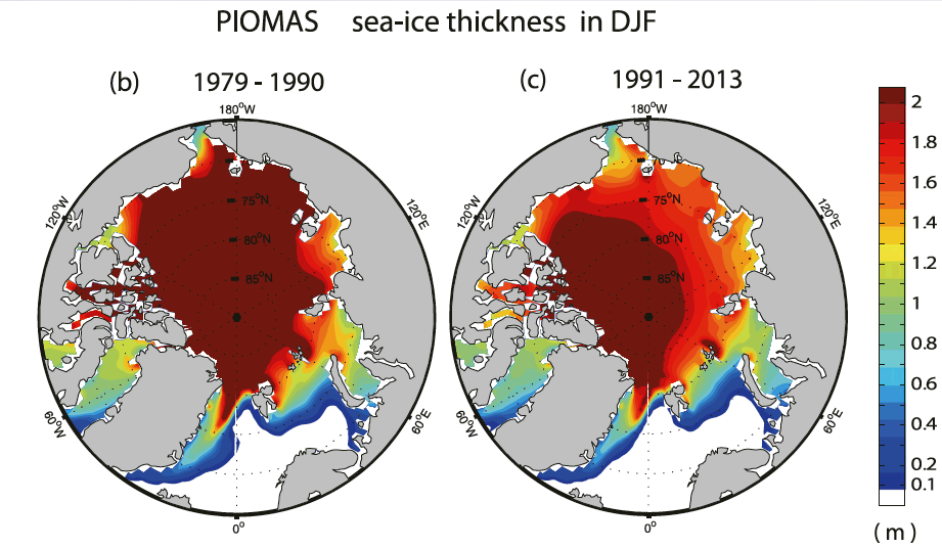
Climate Science Drivers

- *Water cycle:*
  - What are the processes and factors governing precipitation and the water cycle today and how will precipitation evolve over the next 40 years?
- *Biogeochemistry:*
  - What are the contributions and feedbacks from natural and managed systems to current greenhouse gas fluxes, and how will those factors and associated fluxes evolve in the future?
- *Cryosphere:*
  - What will be the long-term, committed Antarctic Ice Sheet contribution to sea level rise (SLR) from climate change during 1970–2050?

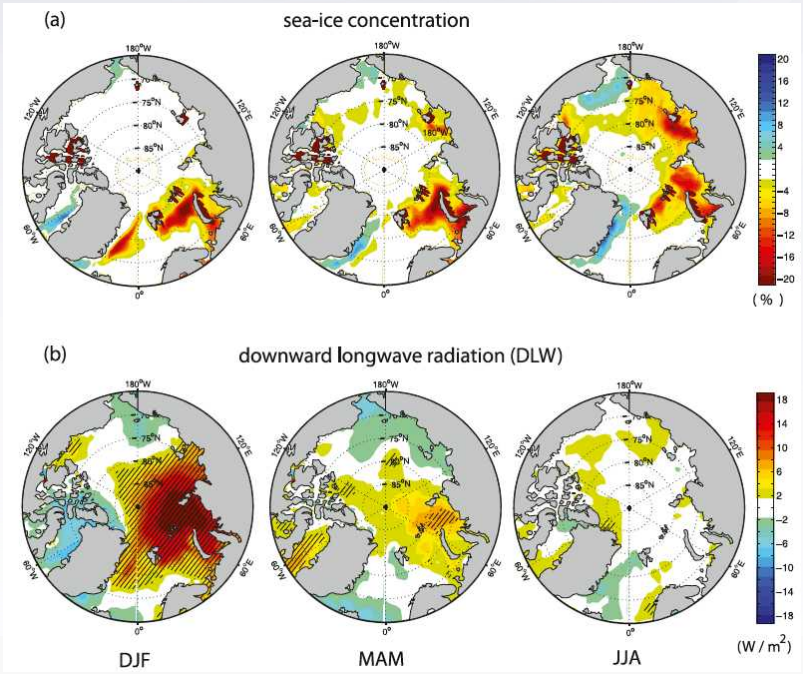
Water Cycle

- What are the impacts of poleward moisture transport on the seasonal melting of Arctic sea ice and how may they change in the future?

Significant sea ice loss between 1979 - 2013



Extreme moisture transport to the Arctic in winter/spring reduces sea ice in the summer – increasing trend of moisture transport explained half of the sea ice loss in the Atlantic sector between 1979-2011

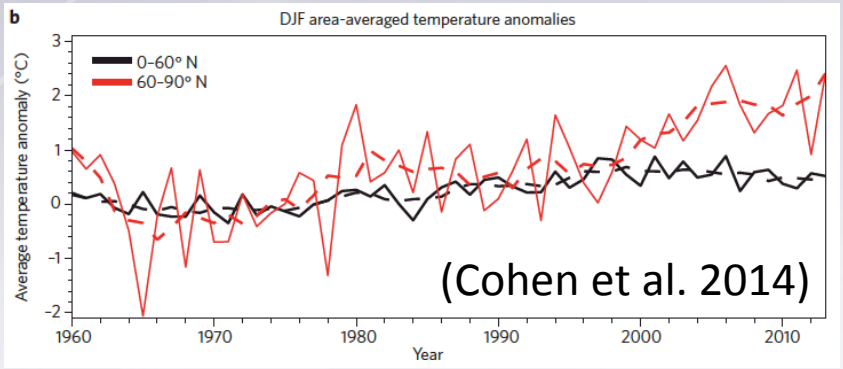


(Park et al. 2015a&b)

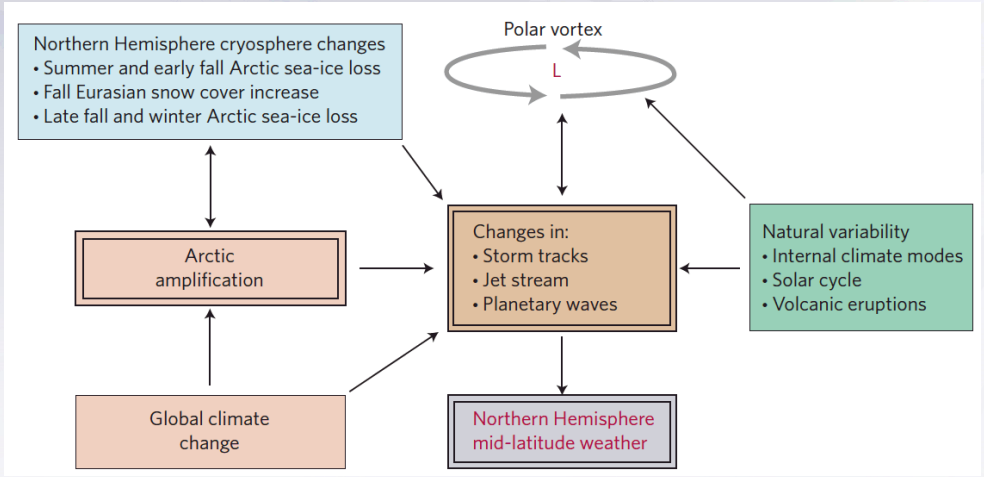
Water Cycle

- How may changes in the Arctic influence extreme events in the lower latitudes?

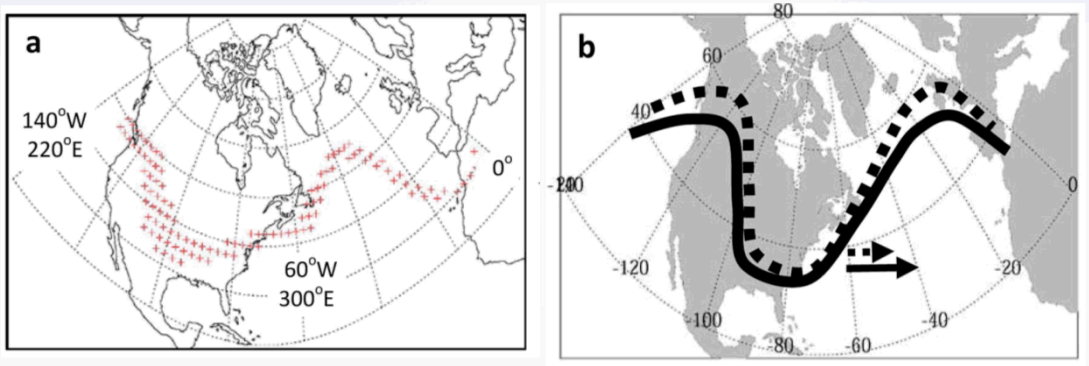
Polar amplification reduces meridional temperature gradients



Ways for polar amplification to influence mid-latitude weather



A slowed and more sinuous jet stream due to loss of sea ice is hypothesized to increase frequency of blocking and cold air outbreaks

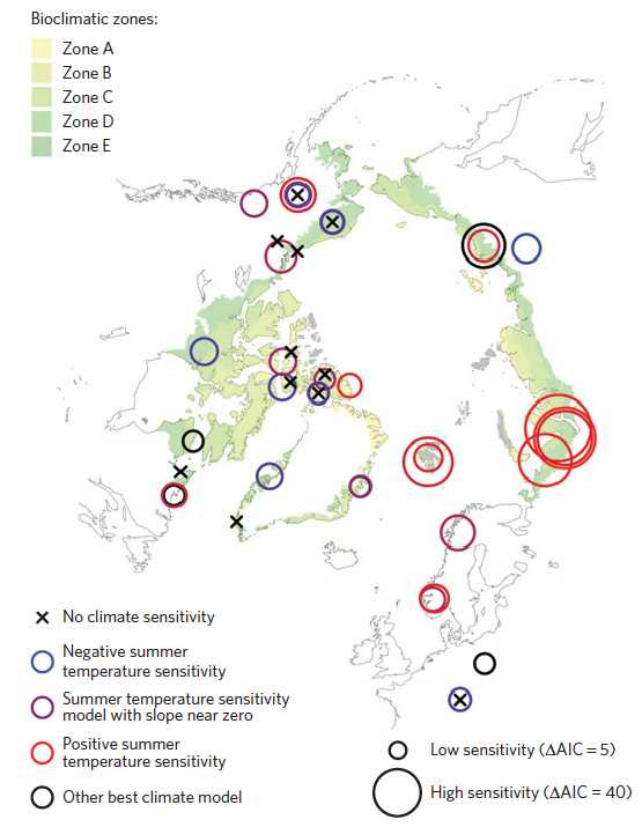


(Francis and Vavrus 2012)

Biogeochemical Cycle

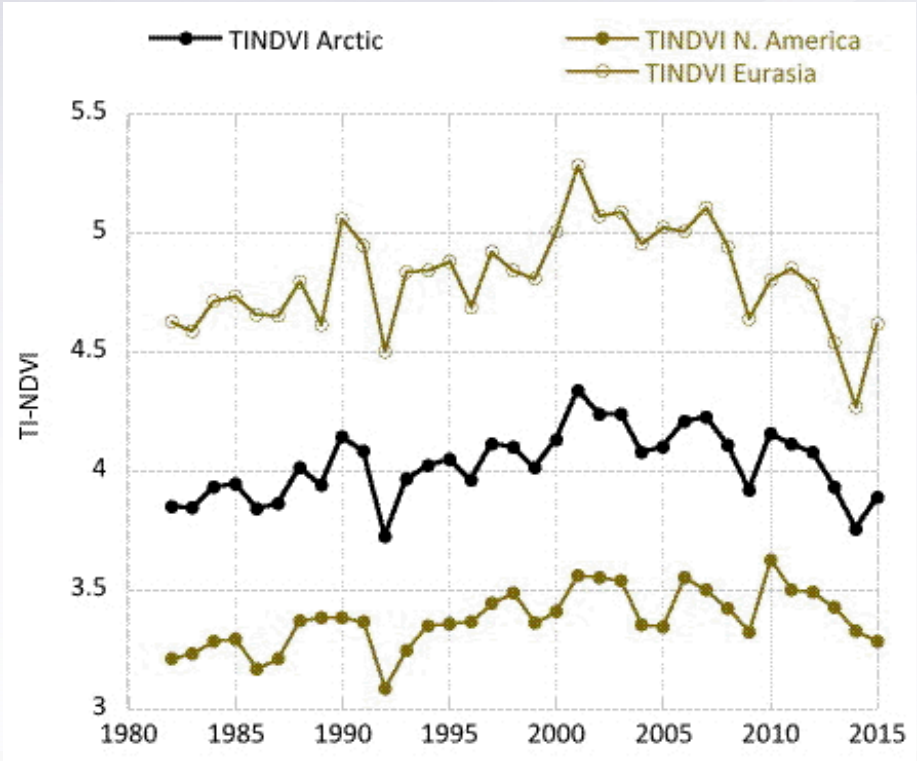
- How may vegetation respond to changes in the Arctic environments and perturb the biogeochemical cycle?

Diverse climate sensitivity across the tundra biome



(Myers-Smith et al. 2015)

Arctic greening and browning: large spatial and temporal variability

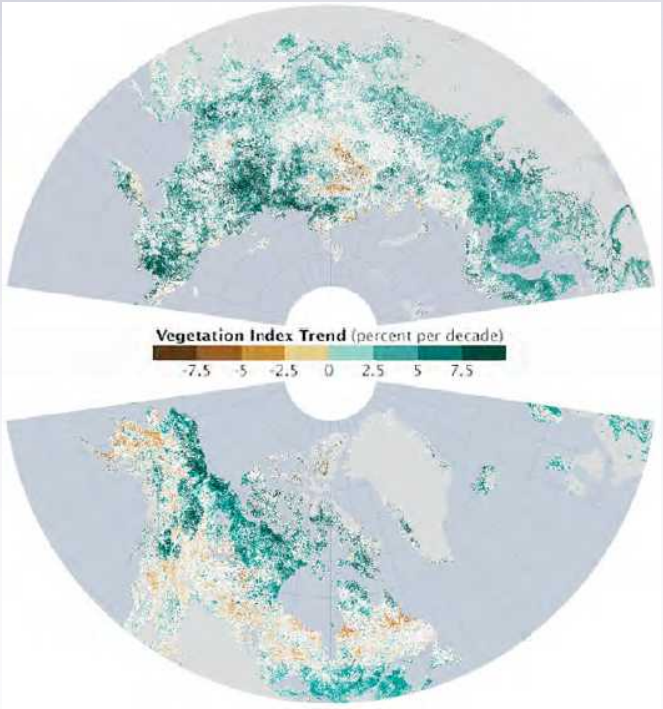


(Epstein et al. 2016)

Biogeochemical Cycle

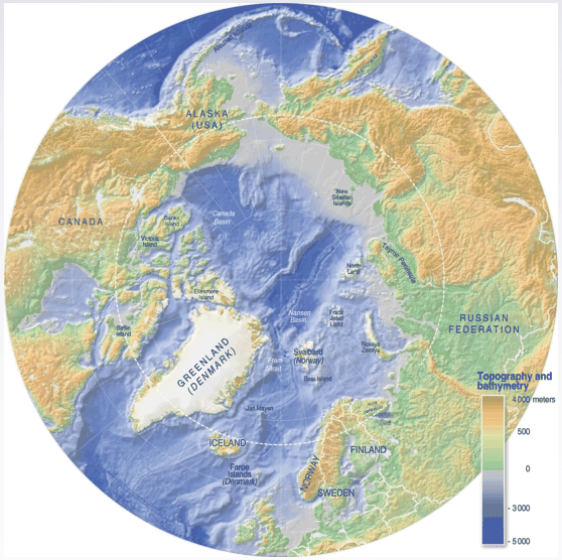
- What are the roles of surface heterogeneity in the Arctic and sub-polar region in modulating the Arctic ecosystem response and biogeochemical cycle changes?

Changes in vegetation (NDVI) between 1982 and 2011 showing greening in the tundra

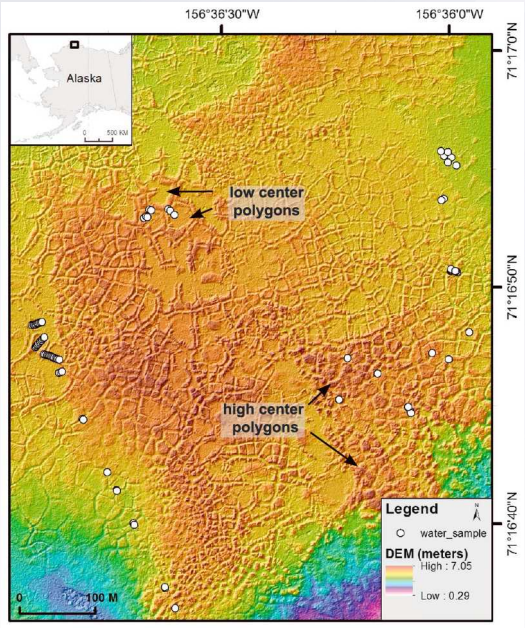


(NRC 2015)

Topographic variations



Polygon features and depths have important effects on biogeochemistry

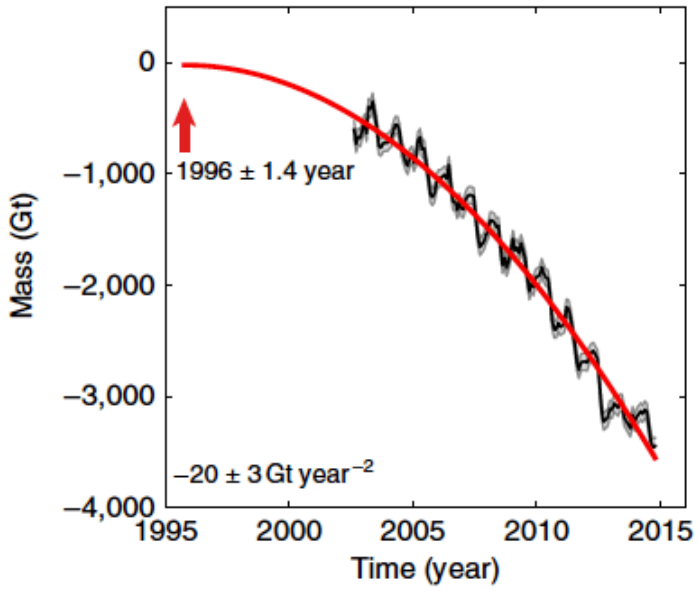


(Newman et al. 2015)

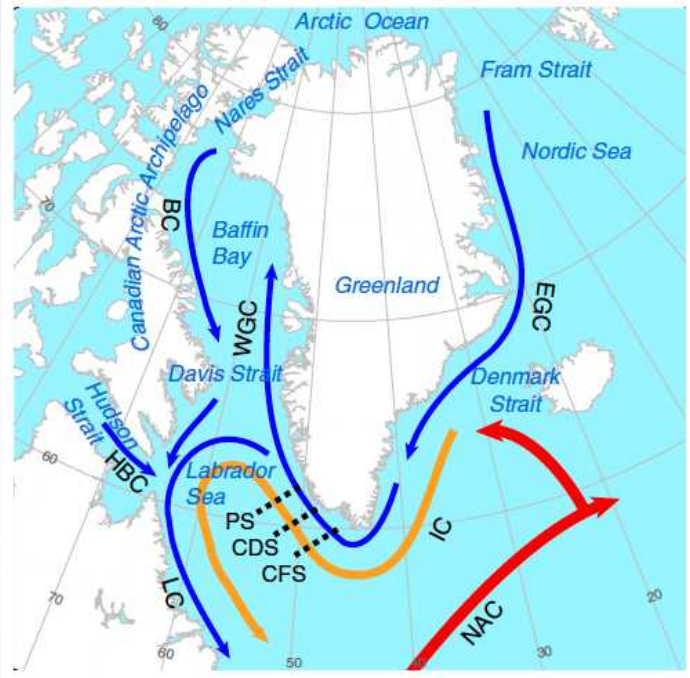
Cryosphere Systems

- How do changes in Arctic sea ice, Greenland ice sheet, net precipitation, river runoff, and ocean heat content impact the trends and variability in ocean water mass transformation in the Arctic Ocean and North Atlantic?

New estimates of Greenland freshwater flux and heat and salt flux from the North Atlantic into the Labrador Sea suggest a direct link of Labrador Sea Water formation to recent freshening, and a possible link to AMOC weakening



Melt water focusing in the Labrador Sea and in a short time period may have significant effect on AMOC

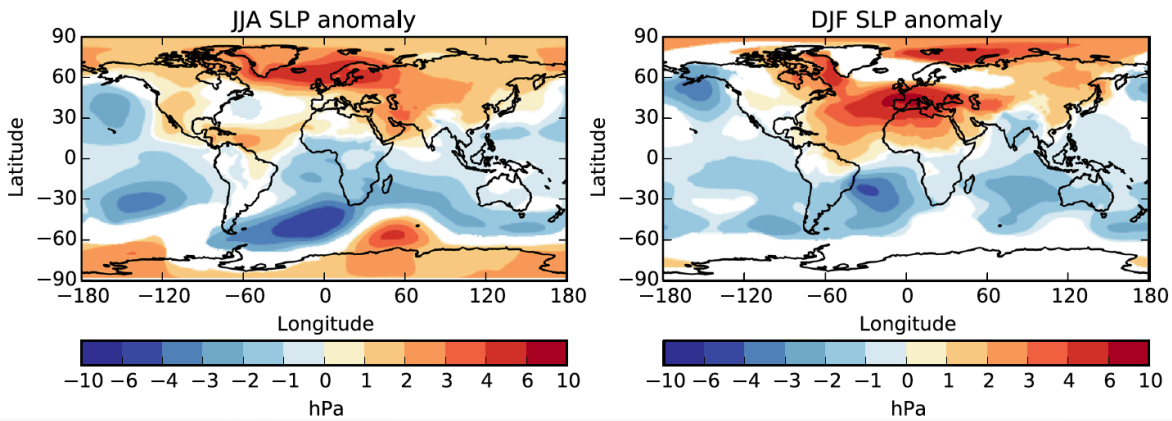


(Yang et al. 2016)

Cryosphere Systems

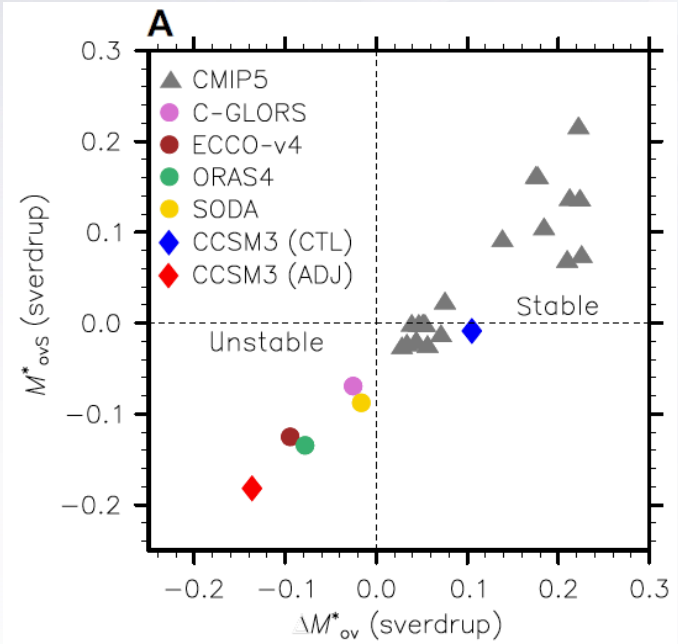
- What are the impacts of changes in the Arctic ocean circulation and the AMOC on global weather extremes?

Simulated changes in AMOC with freshwater input induce global changes in sea level pressure and weather patterns



(Jackson et al. 2015)

Models tend to simulate a more stable AMOC compared to observations

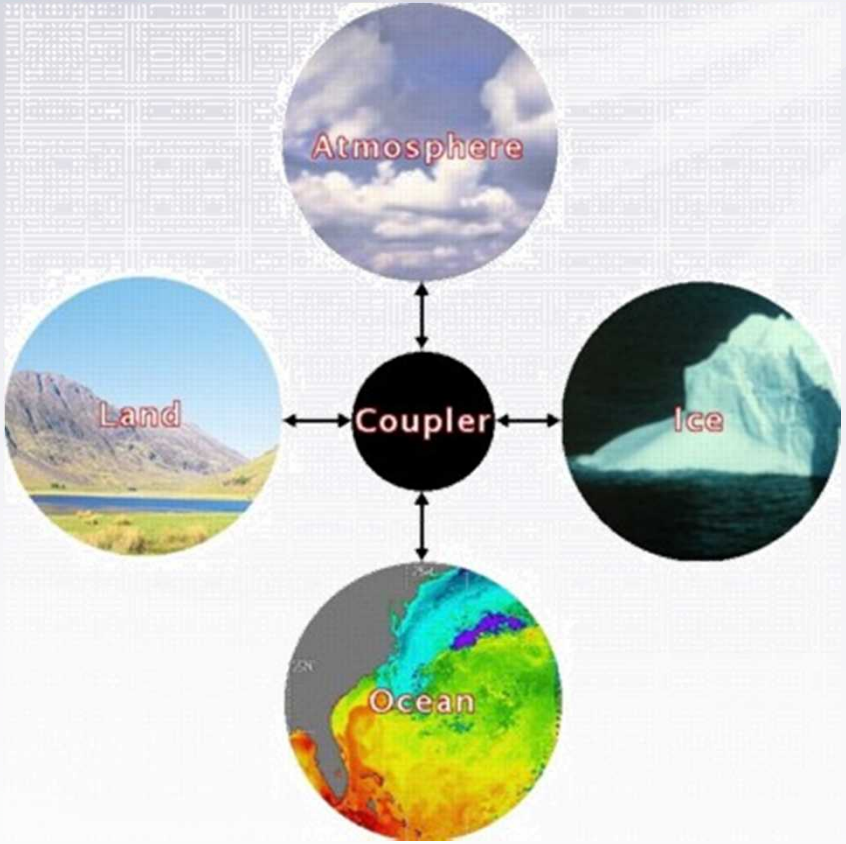


(Liu et al. 2017)

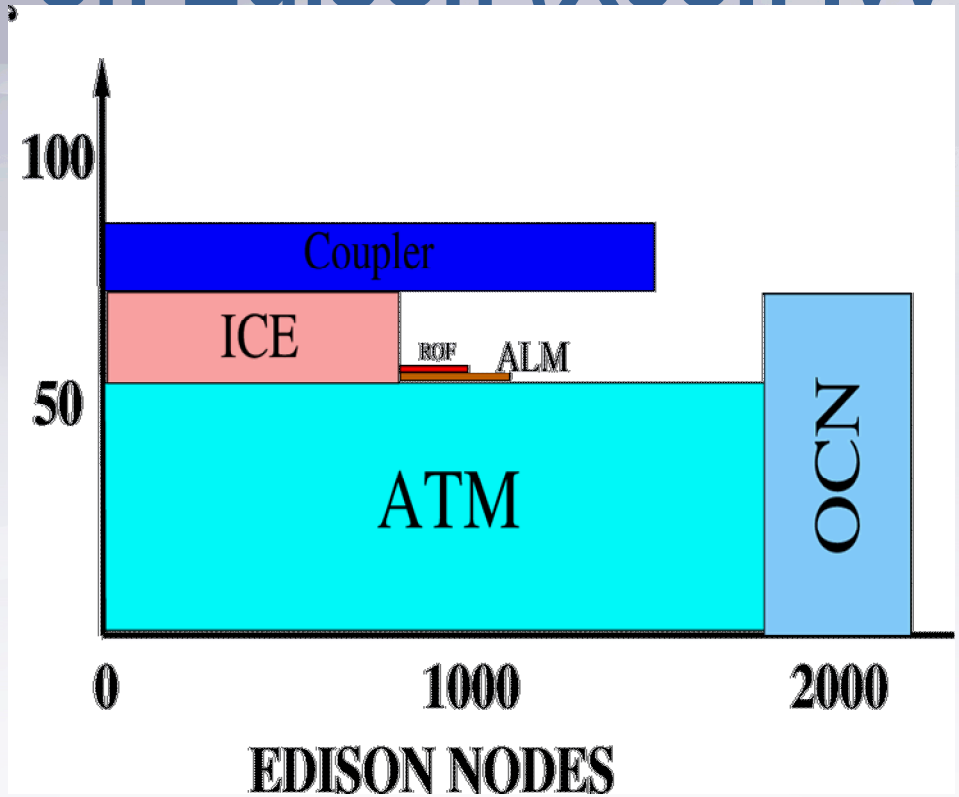
E3SM v1 Overview

E3SM Component Models

- Atmosphere: EAM
  - Branched from CESM's CAM
- Land: ELM
  - Brached from CESM's CLM4.5
- Ocean: MPAS-O
  - New model based on Model Prediction Across Scales (MPAS)
- Sea Ice: MPAS-SI
- Land Ice: MPAS-LI
  - Coupled land ice simulations in v2



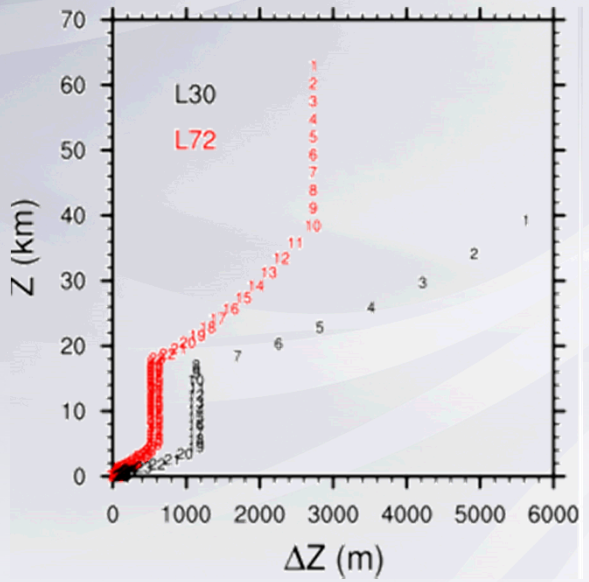
E3SM v1 on Edison (Xeon Ivy Bridge)

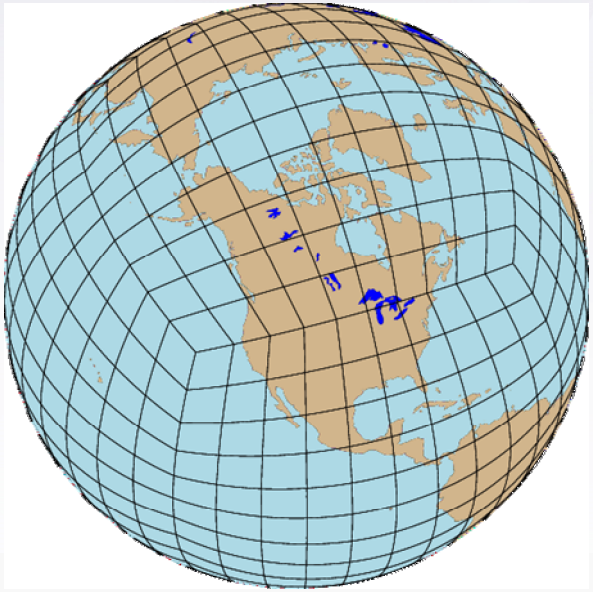


- Typical “load balancing” figure showing seconds per simulated day, concurrency and resource allocation by component.
- Atmosphere is most expensive component, followed by ocean and ice. Land and river runoff are negligible.

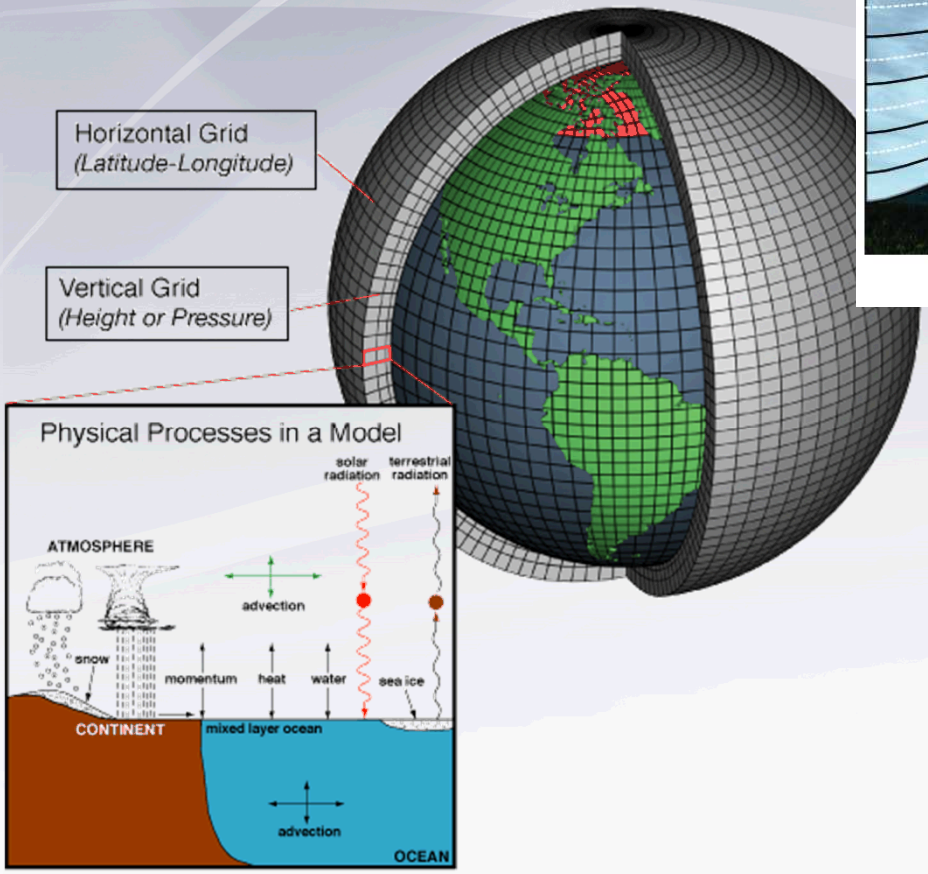
Atmosphere: EAM

- Updates from E3SM v0 (CAM5)
  - HOMME spectral element dynamical core
  - Increase vertical resolution from 30 to 72 Layers
  - Aerosols: MAM4, improvements to nucleation, resuspension, scavenging, convective transport, sea spray
  - Microphysics: MG2, ice nucleation
  - CLUBB shallow convection, ZM deep convection





Atmosphere Component



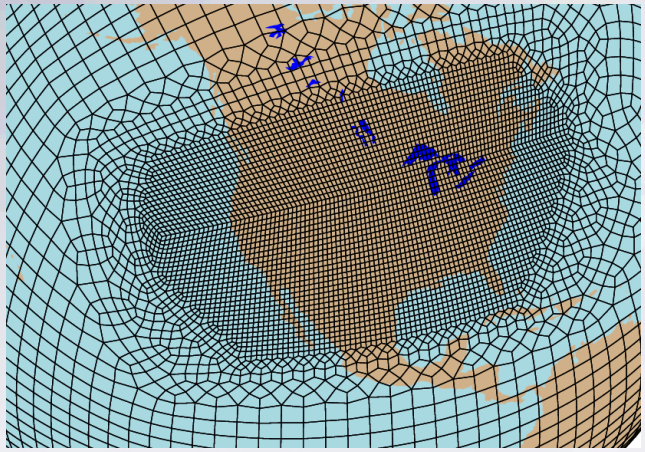
hydrostatic-pressure terrain-following coordinates

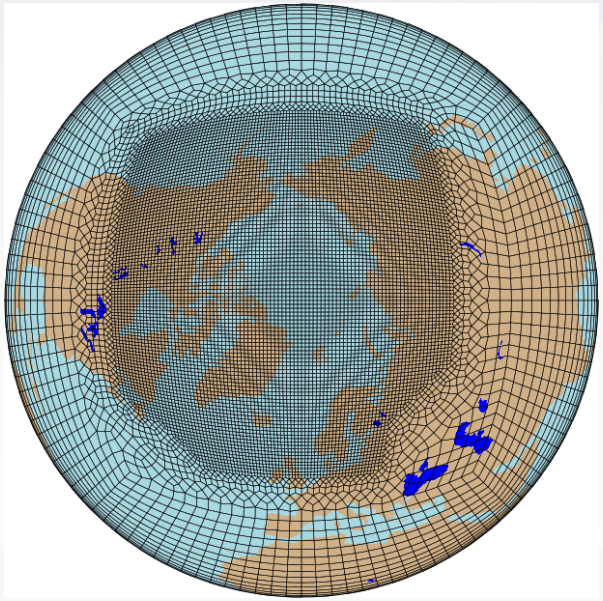
- Column Physics
  - Subgrid parametrizations: precipitation, radiative forcing, etc.
  - Embarrassingly parallel with 2D domain decomposition
- Dynamical Core
  - Solves the Atmospheric Primitive Equations
  - Linear transport of ~30 atmospheric species
  - Scalability bottleneck

Terrain following figure: D. Hall, CU Boulder
Source: http://celebrating200years.noaa.gov/breakthroughs/climate_model/welcome.html

E3SM Regionally Refined Model (RRM)

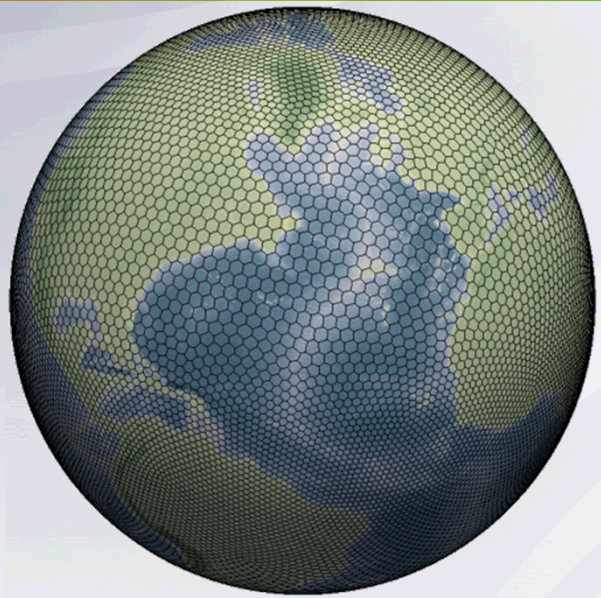
- Spectral Element discretization runs on fully unstructured finite element (quad) meshes
- Example grids:
  - 25km resolution over CONUS (top) and over the Arctic (bottom)
  - Transitioning to global 100km cubed-sphere grid
- E3SM supports “variable-resolution” grids in all components
  - Replaces nested grid approach for regional modeling and ultra-high resolution process studies.

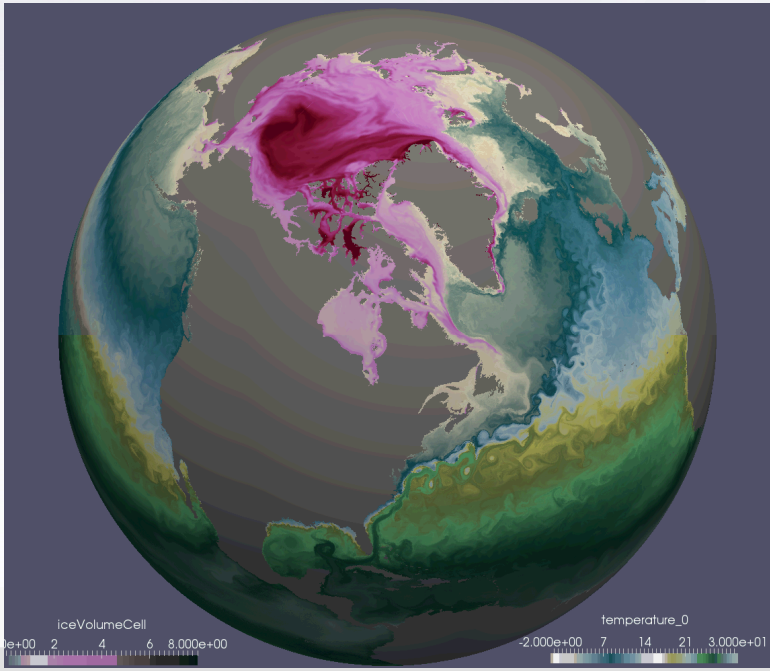




MPAS Ocean/Ice

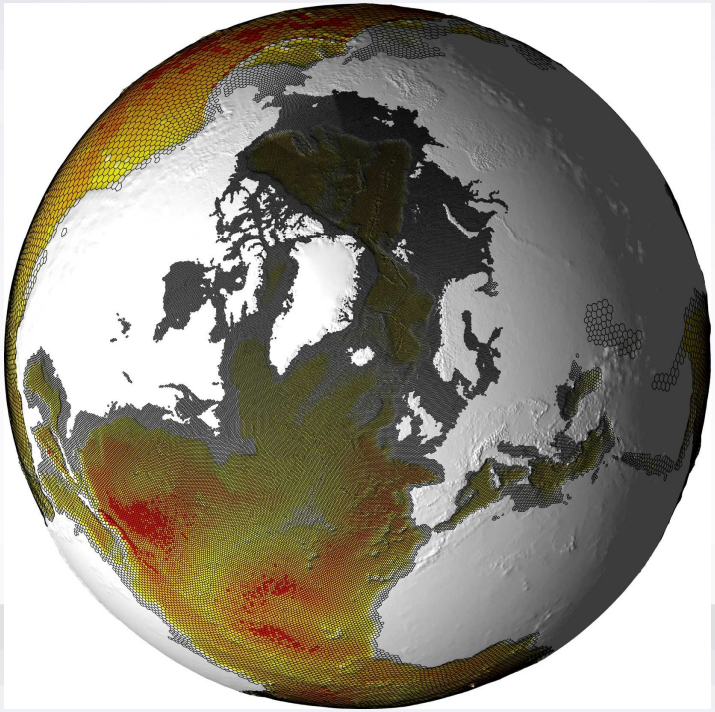
- New MPAS Ocean and Ice components (replaced POP/CICE)
  - Unstructured Voronoi mesh
  - Resolution adjusted to follow Rossby radius of deformation
  - Vertical resolution increased to 100L
  - Arbitrary Lagrangian-Eulerian vertical discretization
  - Split explicit time-stepping for a more scalable barotropic solve.
  - Incremental remap for ice transport

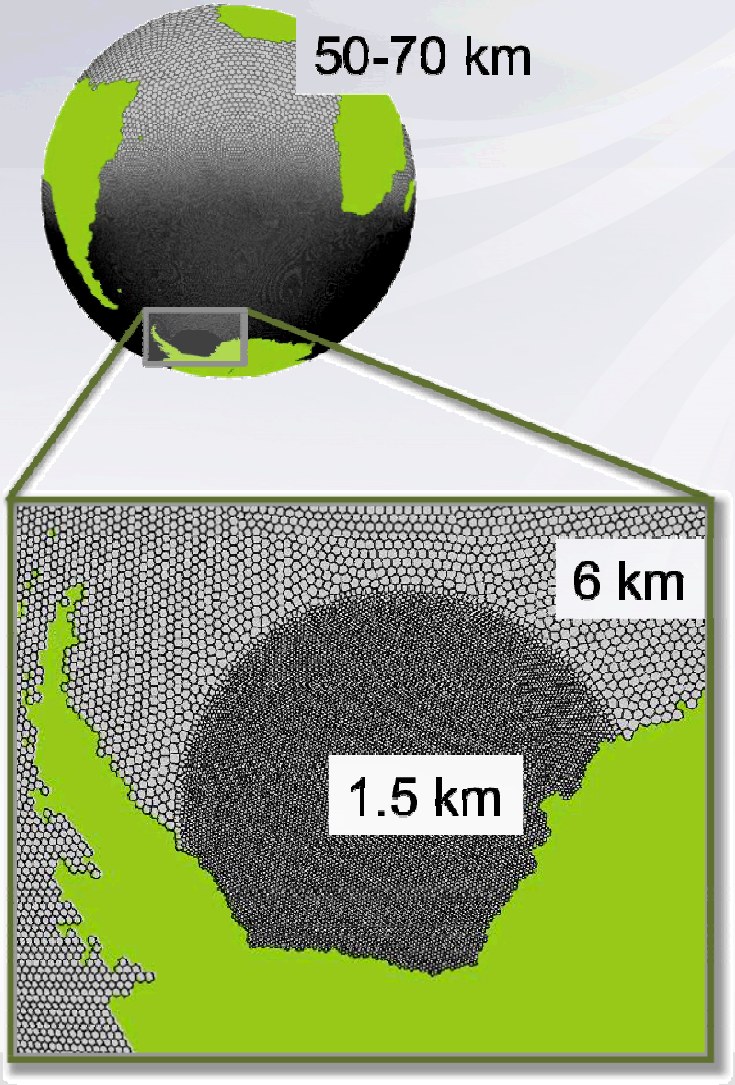




MPAS Ocean/Ice Grids

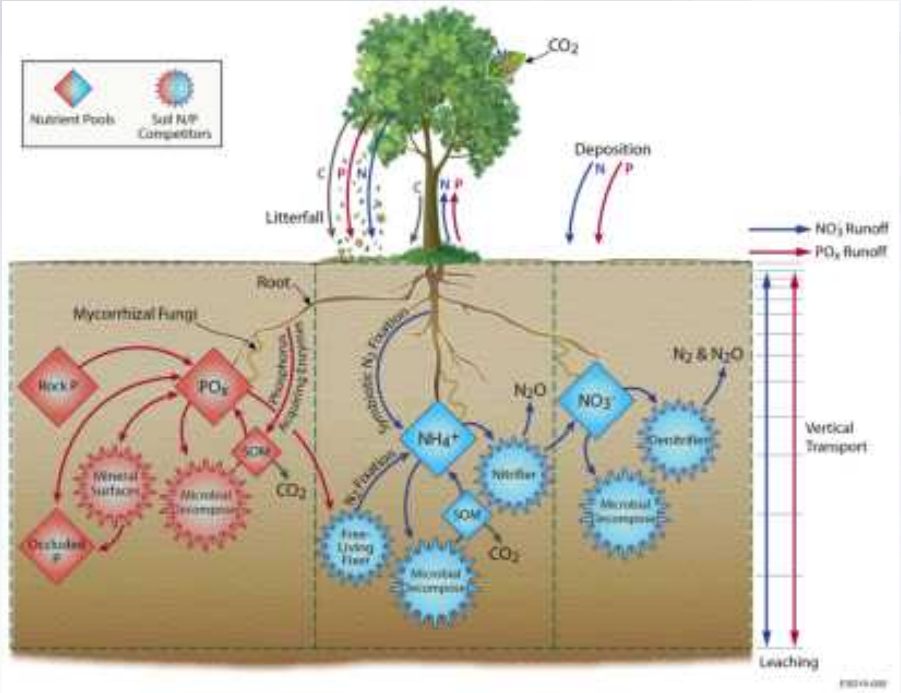
- MPAS components run on fully unstructured Voronoi meshes.
- Example variable resolution grids:
  - North Atlantic / Arctic ocean domain scales from 18 km at Equator to 6 km in the Arctic (4x computational savings)
  - 70km mesh transitioning down to 6km (Antarctic) and then 1.5 km to study ice shelf / ocean interactions.





E3SM Land Model (ELM)

- Hydrology
  - New river routing (MOSART)
  - New soil hydrology (VSFM)
- Vegatation
  - New crop model
  - Dynamic rooting distribution
  - Dynamic C:N:P stoichiometry (with ECA)
  - New allocation (from PiTS)
- Biogeochmistry
  - Coupled C-N-P model
  - New nutrient competition model (ECA)
  - New reactive transport code (BeTR)
  - Coupling to PFLOTRAN-BGC



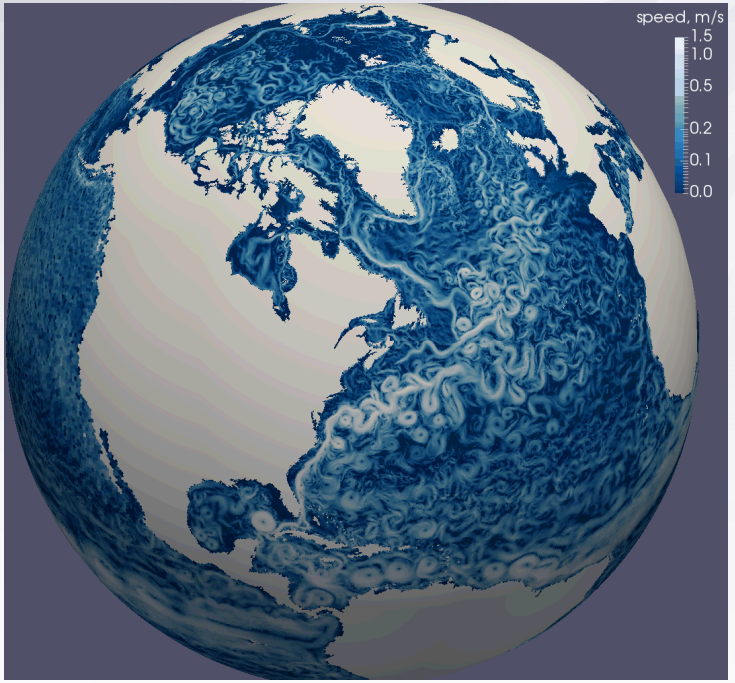
ACME v1 on Today's DOE LCF's O(10) petaflop

E3SM v1

- "Low-Resolution": 110km/72L atm/lnd, 60-30km/100L ocn/ice
  - Workhorse configuration - model development, CMIP/DECK type science campaigns, O(1000) years of simulation
- High-Resolution: 27km/72L atm/lnd, 18-6km/100L ocn/ice
  - E3SM v1: baseline simulations O(100) years
  - E3SM v2: Climate science campaigns on pre-exascale or exascale systems 2019-2023
- E3SM v1 is about 3.7x more expensive then E3SM v0
  - Increased degrees of freedom, increased physics, tighter coupling

E3SM v2/v3+

- RRM (Regionally Refined Model)
  - E3SM v1: All components capable of running on RRM meshes
  - E3SM v2: Affordable high resolution configurations for CONUS, Arctic
- Ultra-high resolution:
  - 1 km: Cloud resolving, Coastal modeling/inundation/ice shelves, ocean / ice shelf interaction
  - 100m resolution (LES regime, boundary layer mixing, sub-watershed resolution)
- E3SM-MMF:
  - Some aspects of cloud resolving, running at 5 SYPD
  - Via superparameterization and GPU acceleration



OLCF Titan

- 19K nodes, 8.2MW
  - 16 core AMD CPU + NVIDIA GPU
- Good machine for ACME:
  - ACME v0 high-res: 2 SYPD, 1.5M/year
  - ACME v1 high-res: 1.4 SYPD, 3.4M/year
- ACME v1 high resolution
  - 15% of our code (and growing) can make use of the GPU. Insufficient GPU utilization to be competitive in INCITE.
  - Exception: ACME-MMF

ALCF Mira

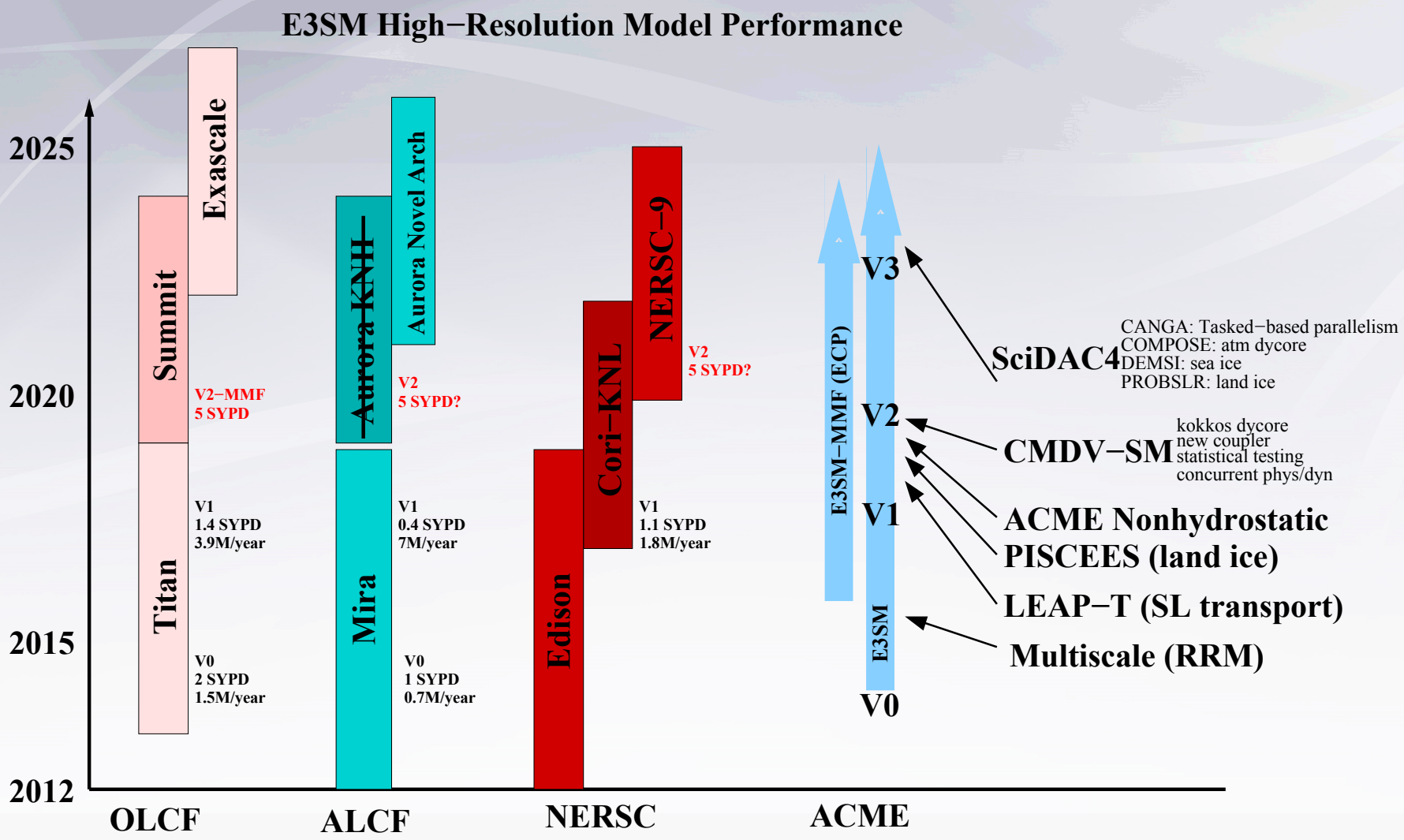
- Mira: 49K nodes (16 core BG/P) 3.9MW
- ACME v0 high-res: 1 SYPD, 0.7M/year
- ACME v1 high-res: .4 SYPD, 8M/year
- Great machine for ACME v0 high-res
  - 127 year pre-industrial control
  - 6x40 present day ensembles
- ACME v1 - too expensive
  - Performance work needed to fix this – but end-of-life machine so focusing on KNL architecture higher priority

KNL: Cori and Theta

- Intel KNL (64/68 cores per node)
- NERSC Cori-KNL 9145 nodes, 3.9MW
- ALCF Theta: 3624 nodes
- Most promising architecture for ACME v1 high-res:
  - 1.1 SYPD, 1.4M/year
- NERSC also has a traditional cluster, Edison
  - 5576 nodes, dual socket Xeon Ivy Bridge
  - 3.7MW

Timeline & Machine Roadmap

E3SM Timeline



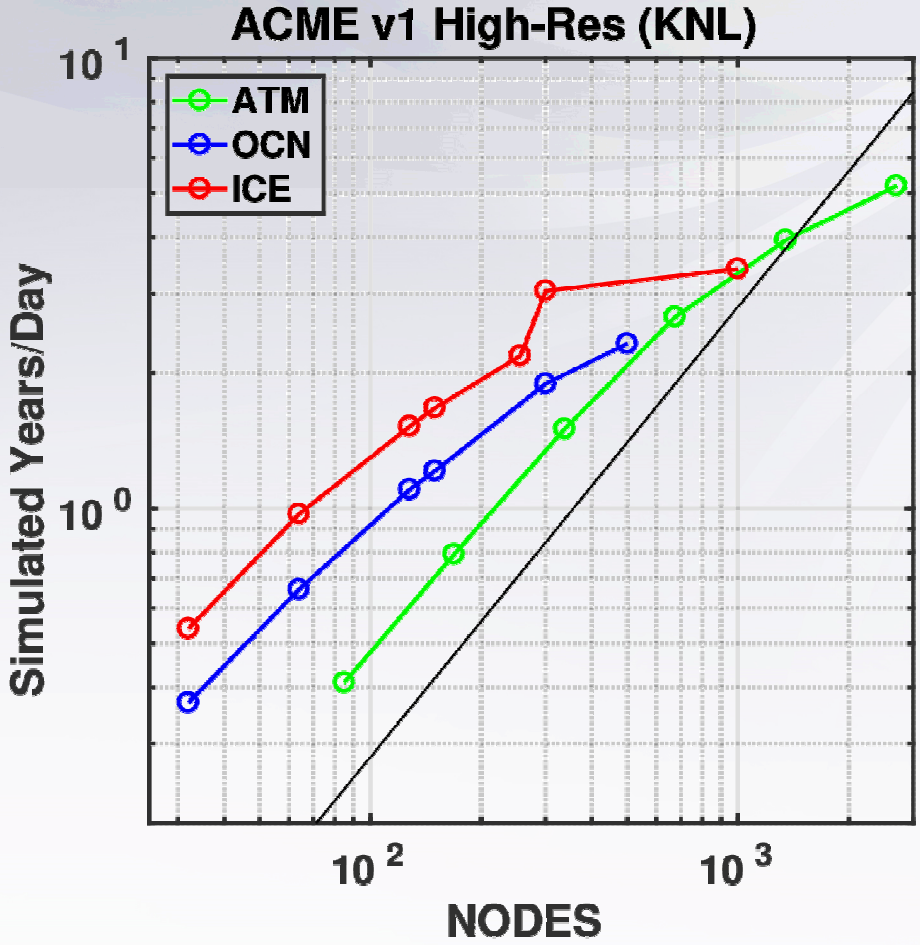
Comments

- E3SM high-res v1 model running on all DOE systems. We strong scale to very “thin” decompositions to maximize throughput
  - Throughput on today’s systems: insufficient for high-res v1 CMIP-style simulation campaigns
  - Throughput for high-res on next gen systems: should be sufficient
- Intel Xeon remains fastest general purpose architecture and the default choice for operational centers – but not suitable for Exascale (too much power)
- Intel Xeon Phi systems:
  - Aurora postponed, Xeon Phi architecture dropped. Does it have a future?
  - Good: New machine looks to be more promising then Xeon Phi
  - Bad: Where can we run high-res simulations in 2019-2020 timeframe?
- Summit (GPU system, 2019):
  - Power9 CPU should be competitive with Xeon
  - First machine to be able to run E3SM v1 high-res at 5 SYPD
  - E3SM has insufficient work per node to get large GPU speedups – allocations will go to other applications?
  - Targeted by E2SM-MMF (aka superparameterization) project spun off from E3SM
- NERSC 9 (2020): May be the first machine for large high-res science campaigns

Xeon Phi (KNL) Performance

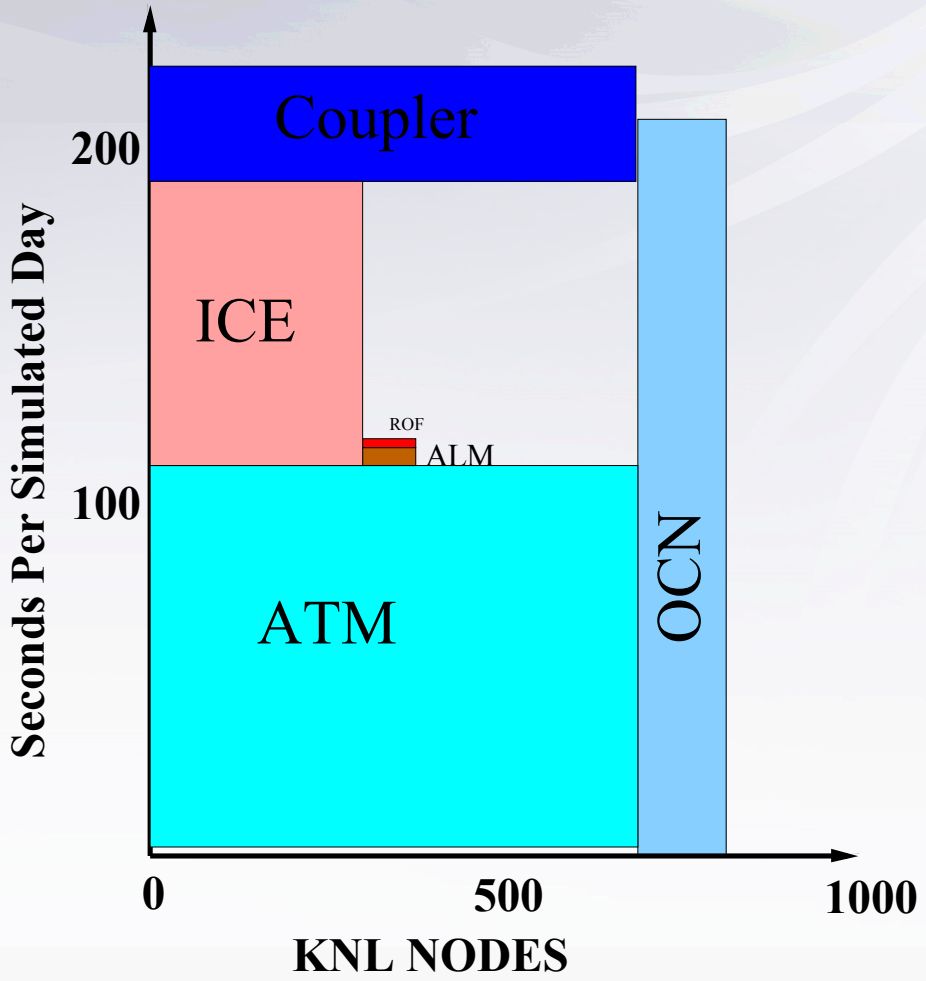
E3SM v1 High-Res Strong Scaling

- All components running 64x2 (MPI x threads) per node
- 64x4 sometimes faster, sometimes slower. Using less MPI and more threads usually slower
- Excellent scaling, ~172K threads in atmosphere, ~64K threads in ocean/ice
- MPAS components have dramatically improved scaling over POP/CICE, scaling reasonably well to 64K threads

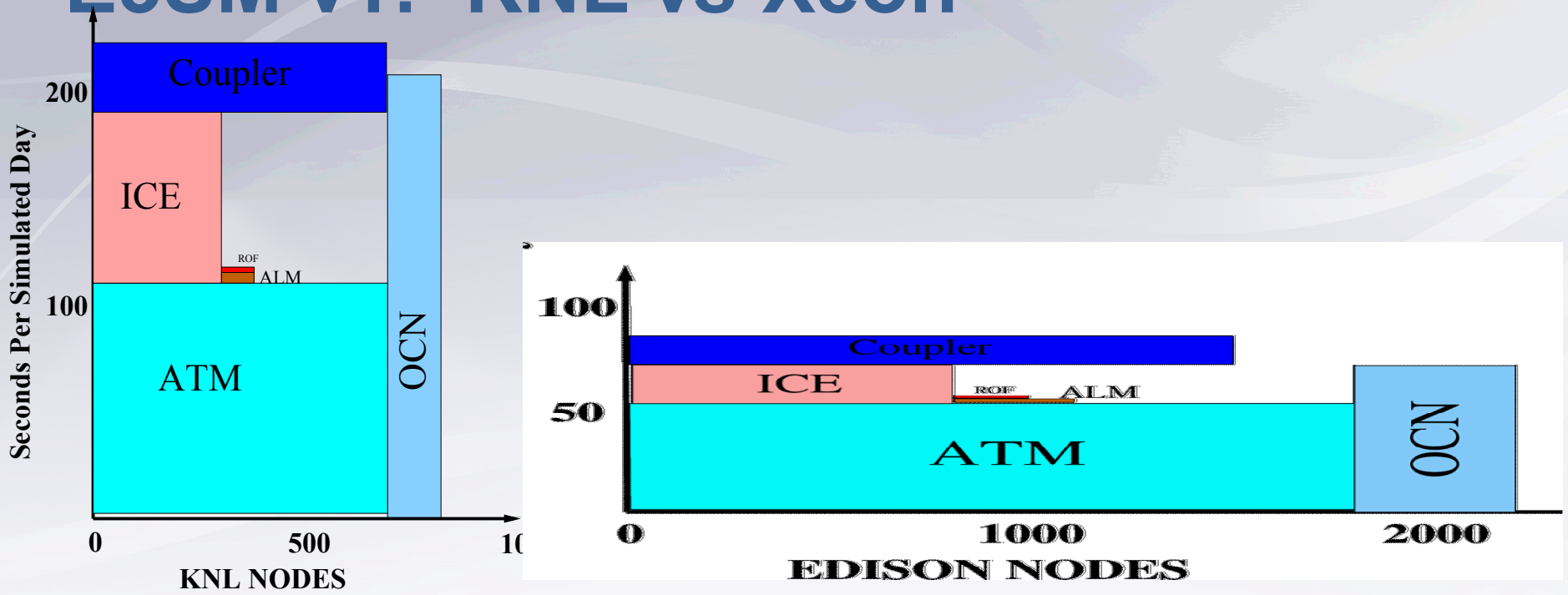


E3SM v1 High-Res Coupled System

- Example 825 node configuration for coupled model
- Ocean and Ice components are configured to run as fast as possible
- Atmosphere allocated sufficient processors so that ATM+ICE matches OCN performance
- 1.1 SYPD and costing 1.2M core-hours per simulated year (1.8M with NERSC charge factor)



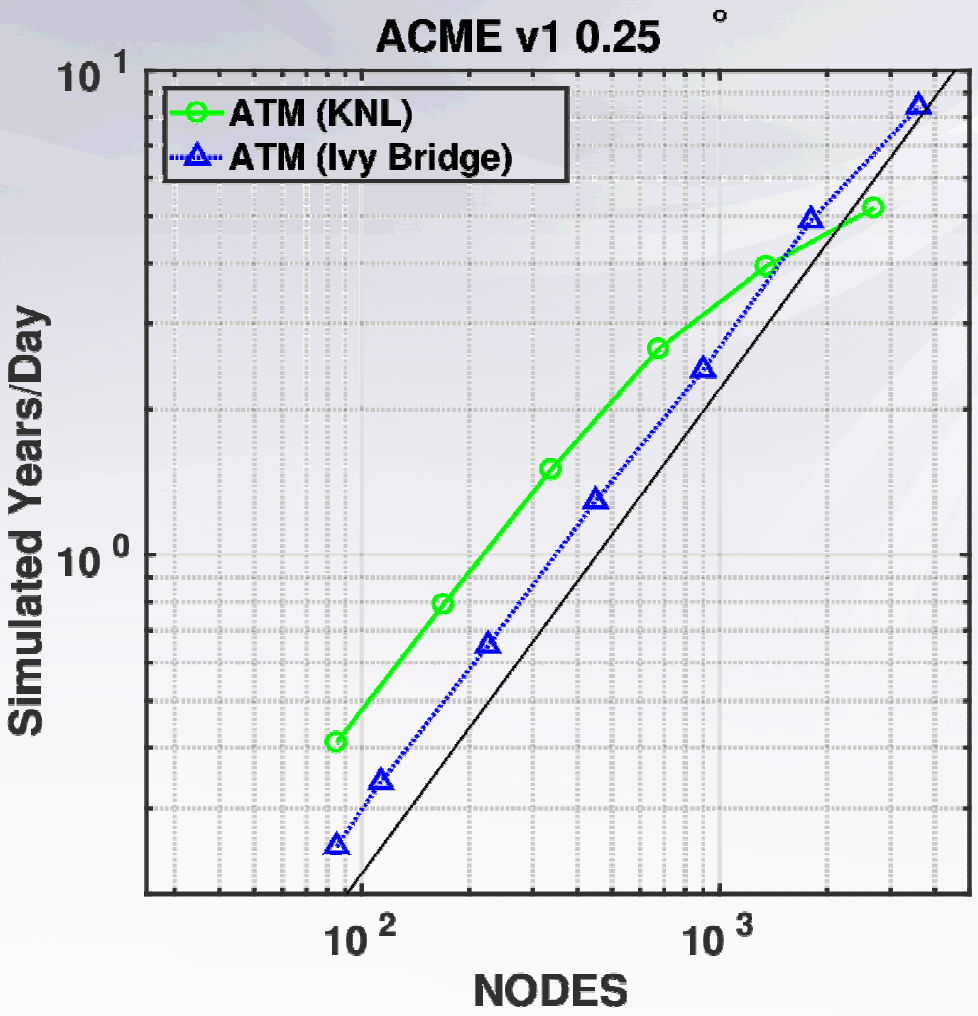
E3SM v1: KNL vs Xeon



- Edison (Xeon Ivy Bridge) more capable: 2.5x faster on 2.6x more nodes
- ATM: 2.4x faster on 2.7x more nodes
- OCN: 1.3x faster on 1.3x more nodes
- **ICE: 3.7x faster on 2.7 more nodes**

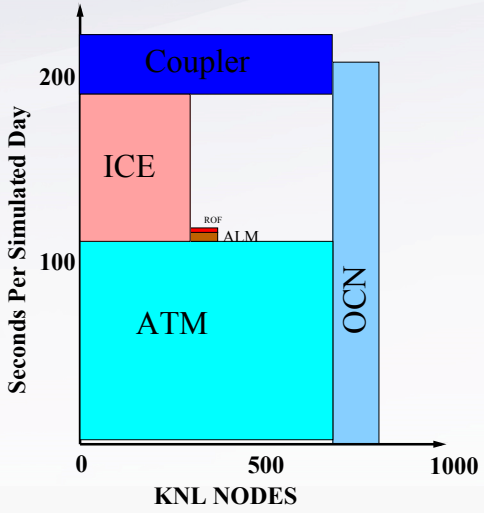
Atmosphere: Scalability

- Full ATM model
- KNL and Xeon Ivy Bridge are competitive up to O(600) columns per node
- Xeon continues to scale well beyond that and can achieve 2.5x better performance but at significantly higher cost
- KNL scaling beyond O(600) columns is poor (low-res data bottom right)
- NOTE: Comparing speed **per node.** If we compared per core, scaling would be the same, but KNL cores are slower than Xeon cores.

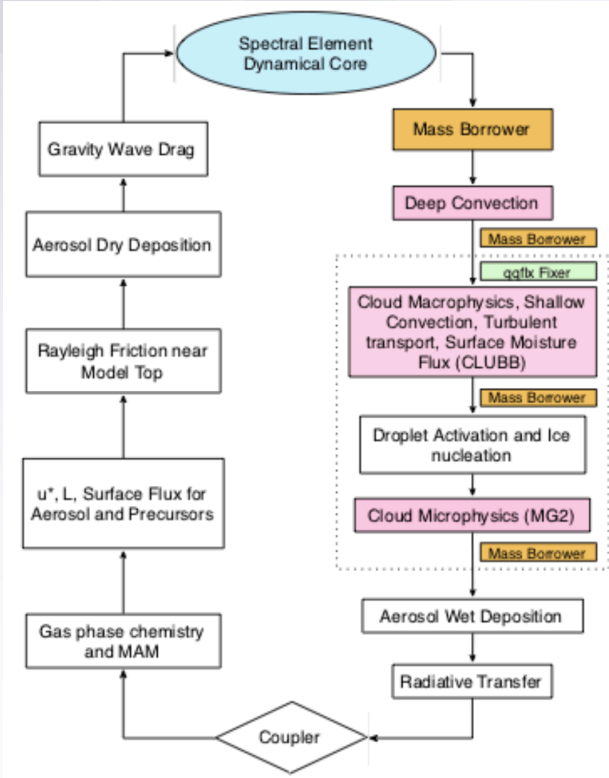


Atmosphere Performance

- Atmosphere is the most expensive component
- Physics (52%) is spread over dozens of parameterizations
- Dycore (transport + dynamics, 48% total) has the most work per lines of code and is usually our first porting/acceleration target

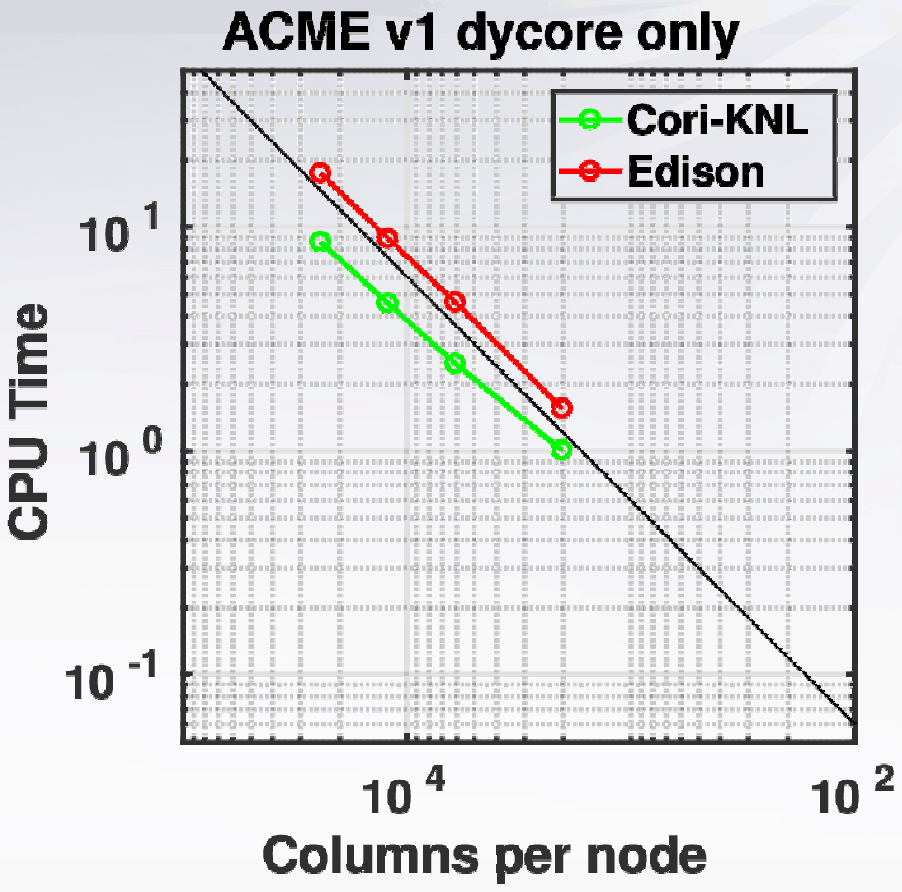


EAM Component	% time
Transport	0.35
Dynamics	0.13
Physics/Chem	0.52



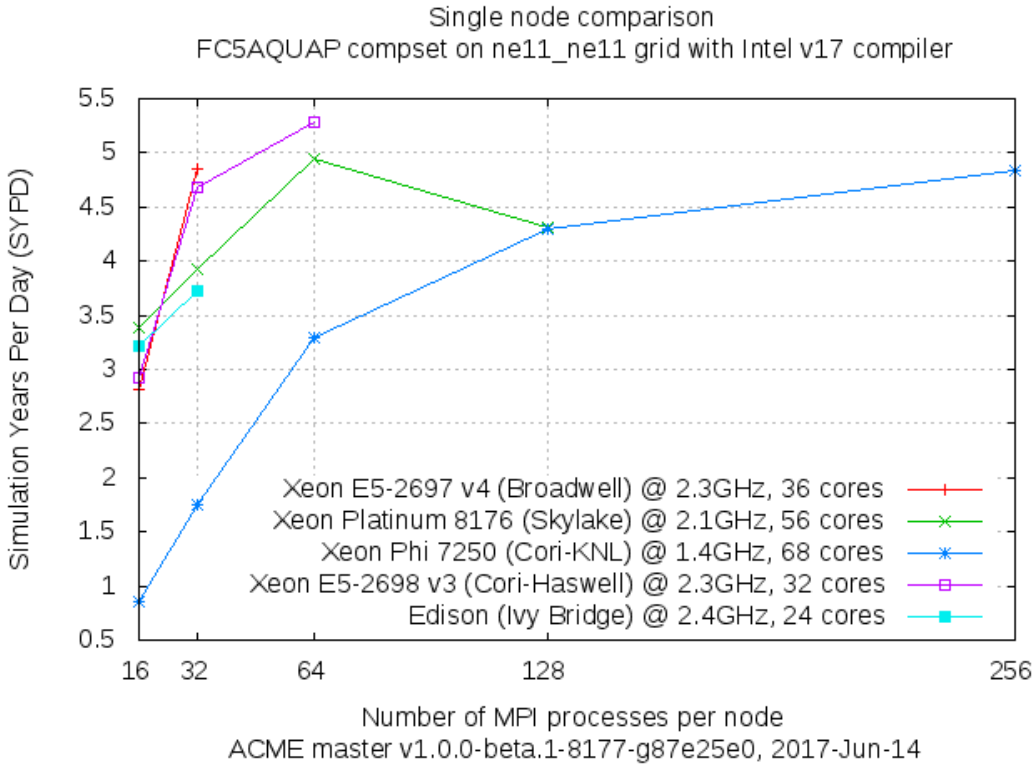
Dycore only: KNL vs Xeon

- E3SM uses the HOMME spectral element dycore
- HOMME does very well on KNL
- Spectral elements are well suited for KNL and has the biggest gains over Xeon
- 2x faster than Xeon Ivy Bridge when there is sufficient work per node O(7K) columns
- Dycore scales down to O(100) columns per node, where KNL and Ivey Bridge have similar performance
-



Full Atmosphere: Single Node

- Full ATM model – 7K columns per node
- With this much work per node, KNL node can outperform Ivy Bridge and is competitive with Haswell.
- But have to use 128 MPI tasks/threads vs 32 on Haswell.
- Need about 4x more work per node

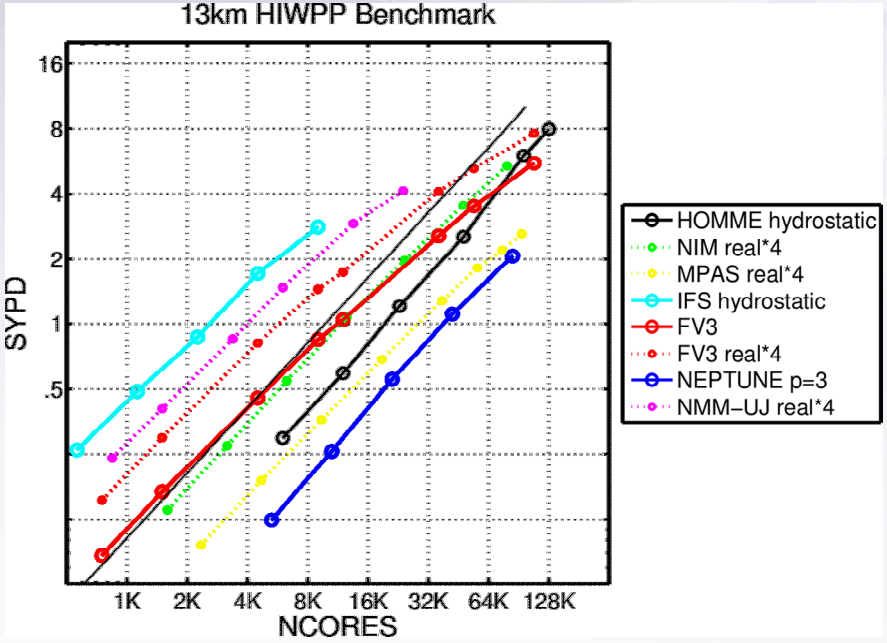


NGGPS Dynamical Core Evaluation

- https://www.weather.gov/sti/stimodeling_nggps_implementation_atmdynamics
- 13km and 3km benchmarks of several non-hydrostatic models on cubed-sphere, icosahedral and geodesic grids
- All models run on NERSC Edison (Xeon Ivy Bridge)
- Precise documentation, easy to reproduce with any model. Baroclinic instability flow with 10 tracers
- E3SM HOMME dycore results.
  - **Hydrostatic version**
  - 13km horizontal resolution
  - 128 vertical layers
  - 40s timestep
  - Monotone conservative transport
  - Report wall clock time for 2h simulation
  - No I/O

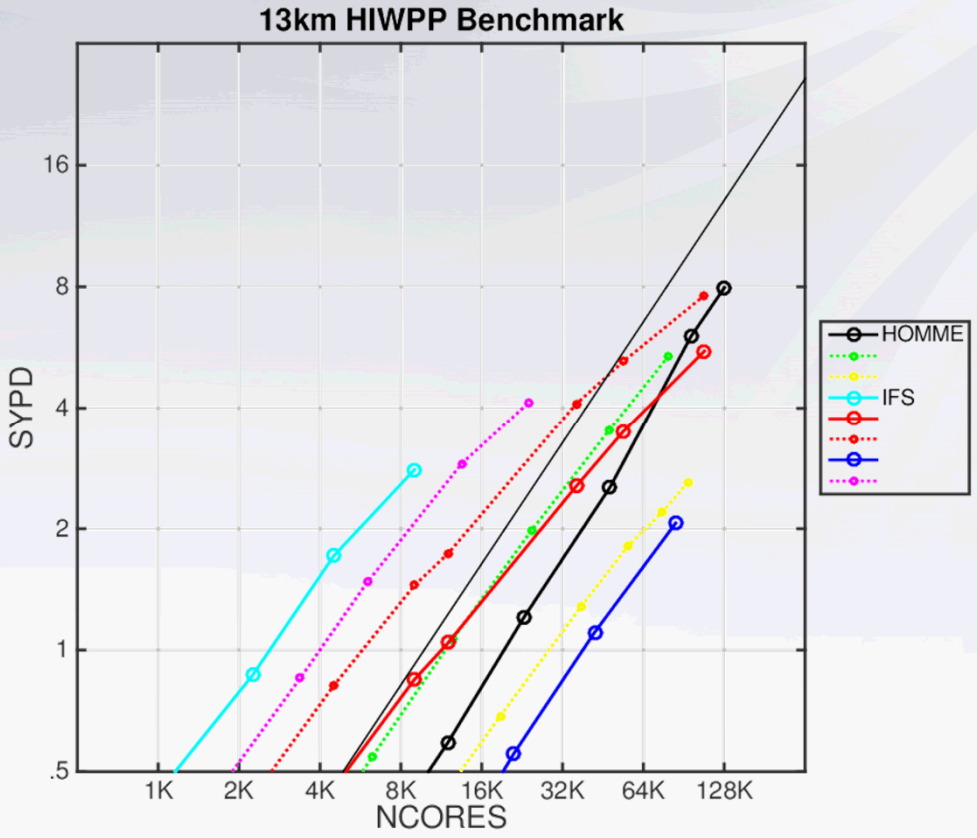
NGGPS 13km benchmark results

- Data replotted from AVEC report on log/log scale in terms of SYPD
- AVEC data from 2015
- Added HOMME data from 2016 (not part of original study)
- All models have been improved since then
- Most AVEC models are running real*4, FV3 reports both real*4 and real*8 results. HOMME is real*8
- HOMME (and IFS) are the only hydrostatic models



NGGPS 13km benchmark results

- NGGPS operational requirement: ~1 SYPD
- ACME climate model requirement: ~30 SYPD.
- Assuming 13km model strong scales as well as the 27km model:
  - HOMME expected to scale to 400K cores at 70% efficiency ~24 SYPD.
  - 3x larger than current Edison system = 12MW



GPU Performance

E3SM GPU Strategy

- CUDA Fortran
  - Switched to openACC ~2016, obtain competitive performance, easier maintenance.
- OpenACC
  - Atmosphere transport (done)
  - Atmosphere dynamics (nearly done)
  - Ocean/Ice dycore: in progress
- Kokkos
  - New effort to write Atmosphere dynamics in C++/kokkos
- CEED (DOE ECP project)
  - Atmoshere dycore could adopt the CEED high-order discretization infrastructure

E3SM GPU Results

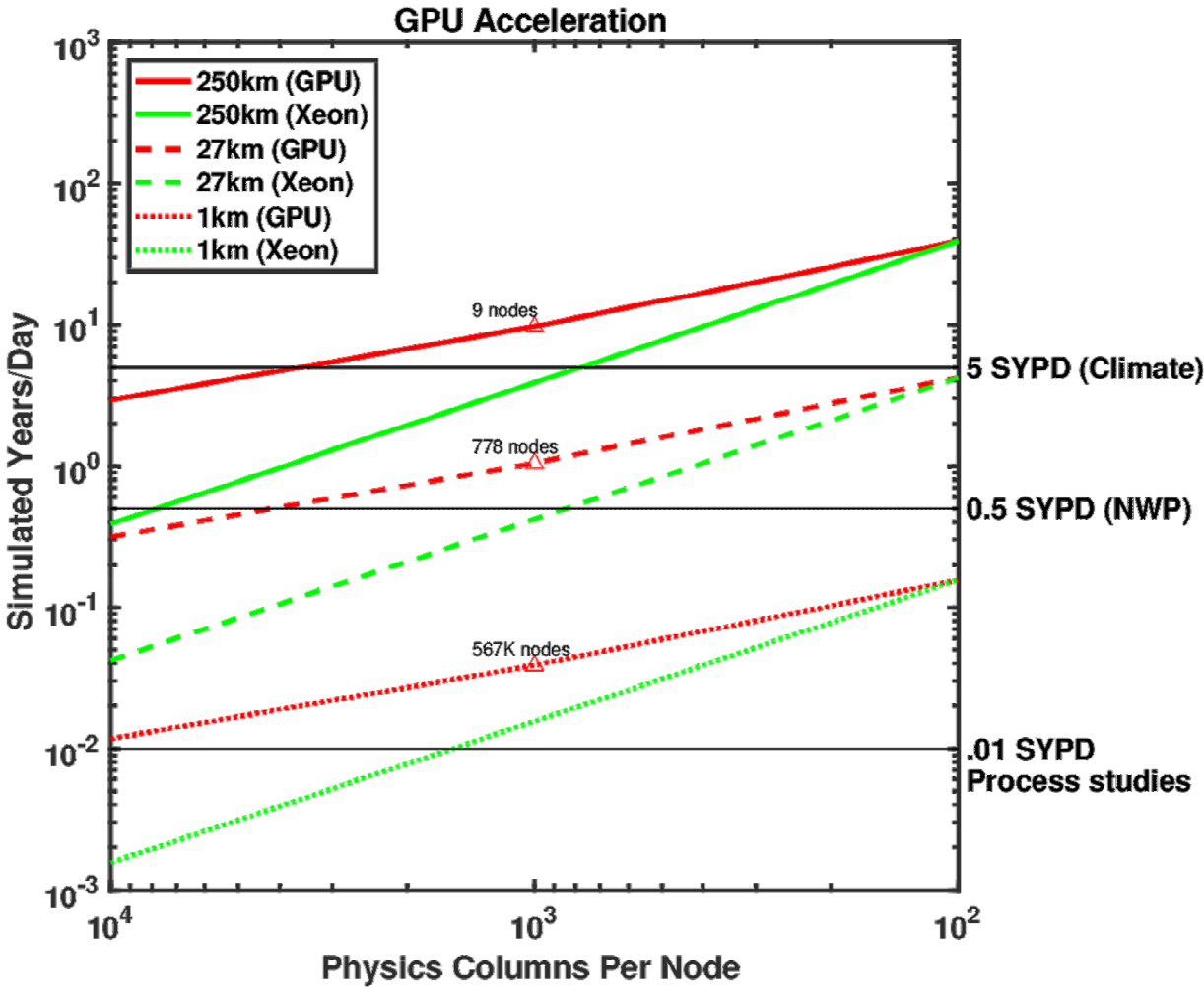
- OpenACC approach:
  - GPU speedup potential appears better than what we can get on KNL
  - But much more labor intensive than porting to KNL (openACC or C++/kokkos rewrite)
  - Difficult to hire staff for openACC development
  - With sufficient work per node can obtain speedups ~2.5x per GPU compared to single CPU.
    - Atmosphere transport: need 500 columns per node
    - Atmosphere dynamics: need 5000 column per node?
  - Currently only 20% of our code can benefit from the GPU, so the full model sees little acceleration

Issues on Summitdev

- Only PGI, Cray, and GNU support OpenACC
  - Cray has slowed development, and GNU has much to catch up on
- E3SM currently runs with OpenACC turned on on Titan with latest PGI compiler
- Summitdev node: 2 x “Power8+” CPU (160 threads, 20 cores) + 4 x P100 GPU
- E3SM cannot run on summitdev with PGI due to a modern-Fortran related bug
- Currently running E3SM with 20 MPI tasks per node on the CPU
- Eventual plan is to evenly divide MPI tasks over GPUs, tracer transport over GPUs, rest of model on CPUs
- Work in progress to try IBM OpenMP 4.x directives in place of OpenACC

GPU Potential

- Pre-Exascale syster
  With sufficient work
  node, assume 7.5x
  speedup over CPU.
- No speedup in the li
  strong scaling
- GPU Strengths: reg
  where throughput is
  issue:
  - large low-res ensemb
  - Ultra-high resolution
    studies
- E3SM's 25km resolu
  in a tough regime to
  use of GPU accelera



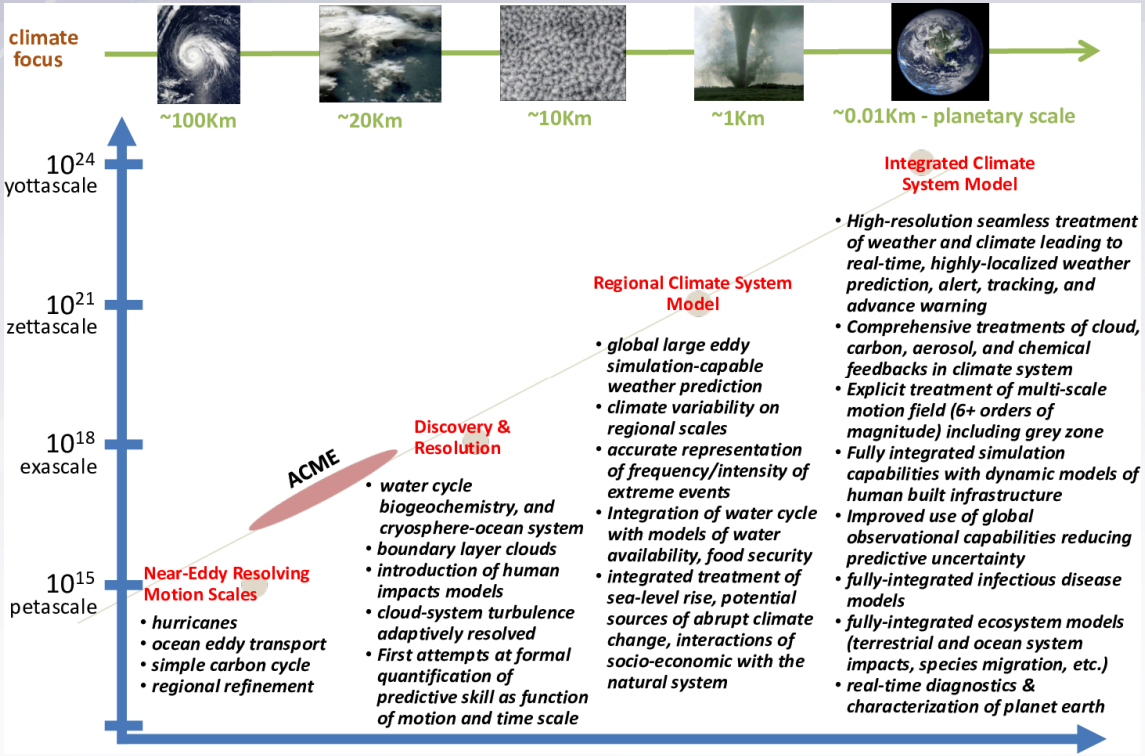
E3SM-MMF

E3SM-MMF Cloud Resolving Climate Model

- Develop capability to assess regional impacts of climate change on the water cycle that directly affect the US economy such as agriculture and energy production.
- A cloud resolving climate model is needed to reduce major systematic errors in climate simulations due to structural uncertainty in numerical treatments of convection – such as convective storm systems
- Challenge: Cloud resolving climate model using traditional approaches requires Zettascale resources.
- E3SM-MMF: Use a multiscale approach ideal for new architectures to achieve some aspects of cloud resolving convection on Exascale resources

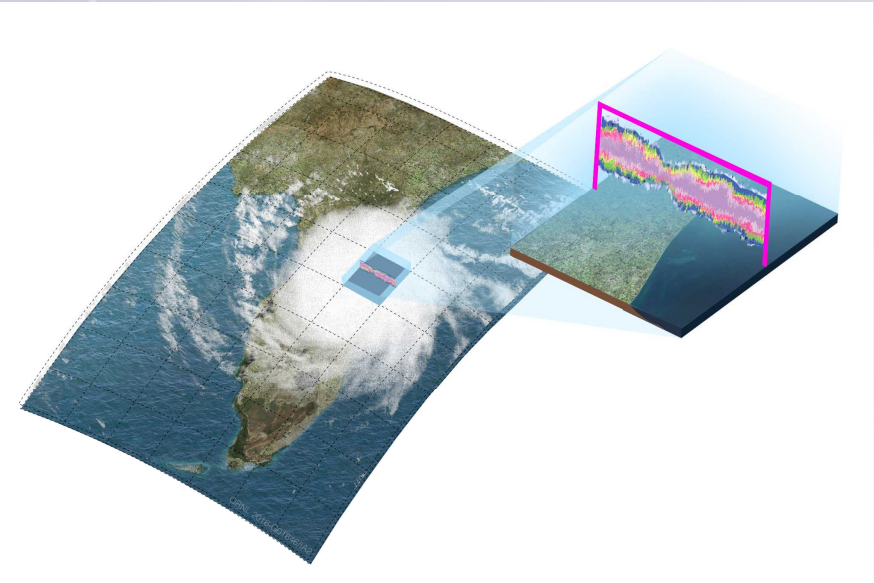
Convective storm system nearing the Chicago metropolitan area
http://www.spc.noaa.gov/misc/AbtDerechos/derechofacts.htm

E3SM-MMF Cloud Resolving Climate Model



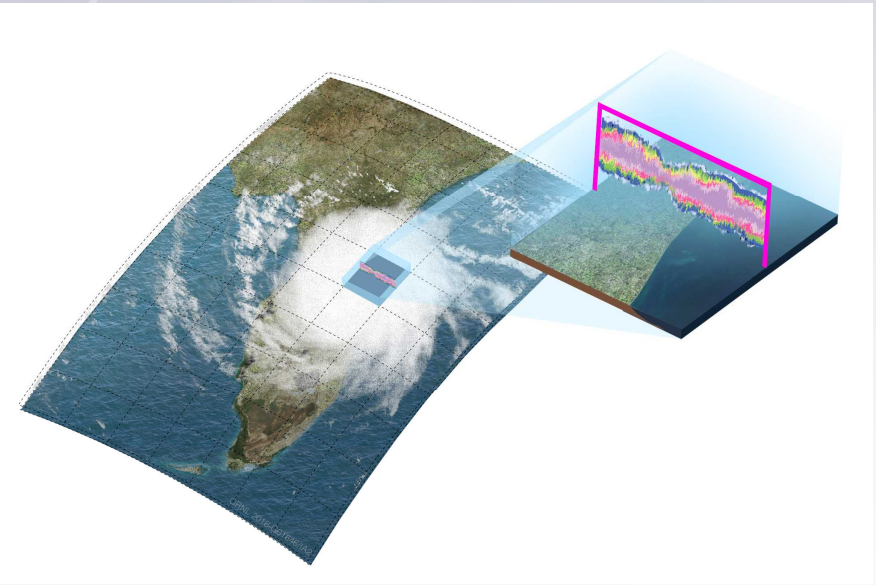
- Conventional approach may get us to 10km scales on Exascale machines (and 1km on Zetascale machines)
- E3SM-MMF approach gets us some aspects of 1km scales on Exascale machines

The Multiscale Modeling Framework (MMF)



- E3SM-MMF approach addresses structural uncertainty in cloud processes by replacing traditional parameterizations with cloud resolving “superparameterization” within each grid cell of global climate model
- Super-parameterization dramatically increases arithmetic intensity, making the MMF approach one of the few ways to achieve exascale performance on upcoming architectures.
- Exascale + MMF approach will make it possible to perform climate simulation campaigns with some aspects of cloud resolving resolutions.

The Multiscale Modeling Framework (MMF)



Conventional E3SM

EAM Component	% time
Transport	0.35
Dynamics	0.13
Physics/Chem	0.52

E3SM-MMF

EAM Component	% time
Transport	0.02
Dynamics	0.01
Physics/Chem	0.02
CRM	0.95

E3SM-MMF: Refactoring of the CRM

- ACME-MMF: Multiscale Modeling Framework
- Run a 2-D hi-res Cloud Resolving Model (CRM) at each physics column
- System for Atmospheric Modeling (SAM) used for the CRM
- Original code issues
  - rarely used modules
  - liberally used common blocks
  - runs each CRM serially inside a loop over GCM columns
- First refactor is to push loop over GCM columns into the CRM as the fastest-varying dimension for easy SIMD vectorization
  - Variable number of columns may be passed in for CPU / GPU accommodation
- Next, focus on two-moment microphysics by threading not only across GCM / CRM columns but within columns as well

E3SM-MMF: ECP Refactoring

- RRTMG radiation consumes as much time as 2-mom microphysics, but we are ultimately targeting RRTMGP
  - Also considering an RRTMG replacement using Deep Neural Network emulation of the LW and SW fluxes rather than by hand
- RRTMGP is already almost entirely ported via OpenACC, and is threading across and within columns and across spectral quadrature points
  - Kudos to Robert Pincus (CU) for an extremely well-written code
    - Threading already exposed in tightly nested loops
    - Modern FORTRAN outside the kernels, flat data arrays inside
- Transport & dynamics: small kernel times, large launch overheads
- However, they consume relatively little time; and we think CUDA 9 cooperative groups could possibly ameliorate this

Summary

Summary 1

- KNL vs Xeon for “conventional E3SM”
  - Full model: KNL performance per node is comparable
  - Atmosphere component is performing well – faster per node than Xeon (Ivy Bridge)
  - MPAS (ocean and ice) components slower per node on KNL, and they are the focus of current performance efforts
  - At strong scaling limit, Xeon is significantly faster but at significantly higher power

Summary 2

- GPU Systems
  - GPUs remain promising; can run some code significantly faster than the CPU - But only when there is sufficient work per node.
  - Porting to GPU significantly more disruptive than KNL
  - Much of our 1M lines of code is not yet GPU ready
- New approaches / new algorithms needed to take advantage of Exascale hardware
  - Increase arithmetic intensity
  - E.g. E3SM-MMF / super-parameterization, subcolumns, chemistry, ensembles

Speculation for 2021

- Exascale systems will be able to produce more simulated years per Watt in several simulation regimes:
  - Larger ensembles
  - Ultra-high resolution process studies (short simulations)
  - MMF (super-parameterization) and new approaches with high arithmetic intensity physics
- CMIP-style science campaigns at cloud resolving resolutions will remain impossible
- Xeon systems will remain superior terms of throughput-by-any-means-necessary
  - High throughput will require high power
  - Exascale performance not obtainable with DOE power budget (<< 100MW)

Thanks!

Mini Apps

Mini-apps and Kernels

- Collection of individual fortran subroutines + drivers
  - Used for debugging openACC
- Atmospheric transport mini-app
  - Linear transport with prescribed velocity, no physics
  - Spectral elements in spherical geometry
  - Fortran+MPI/openMP
- SIQK: New transport algorithm targetting E3SM v2
  - Multi-tracer efficient, Incremental remap
  - C++ with Kokkos for mesh intersections
- HOMME: Atmosphere dycore can be run standalone
  - 30K lines of code, Fortran+MPI/openMP, openACC
- Atmosphere dycore HOMME++
  - C++/Kokkos version, ready in 2018