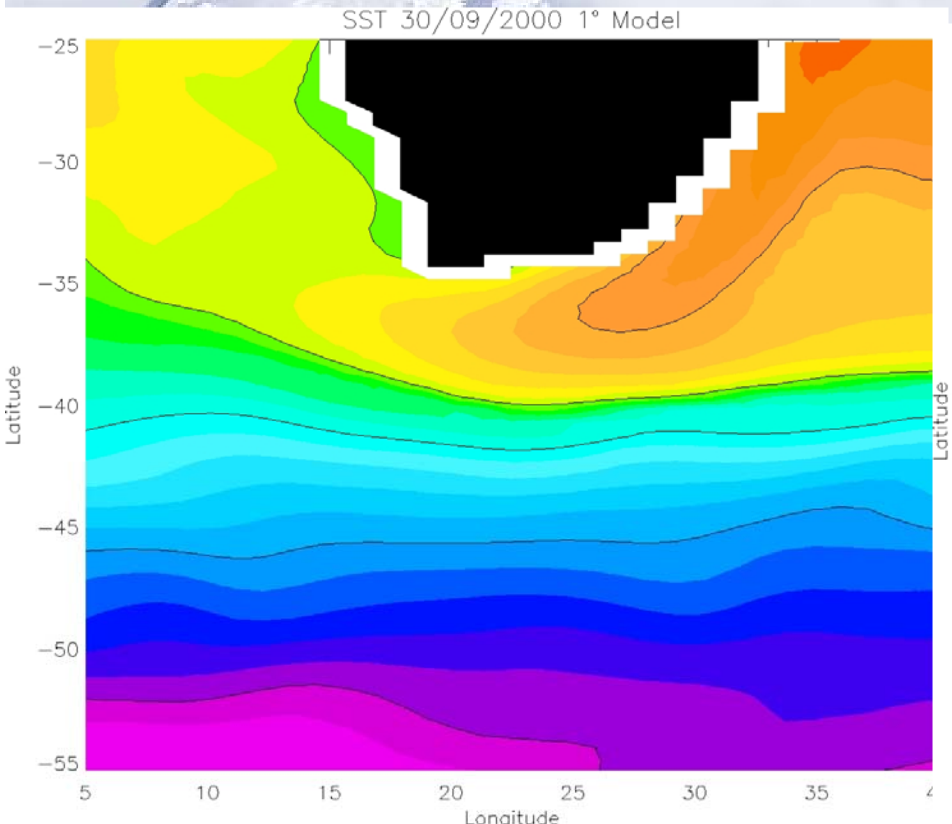
Optimizing High-Resolution Climate Variability Experiments on Cray XT4 and Cray XT5 Systems at NICS and NERSC

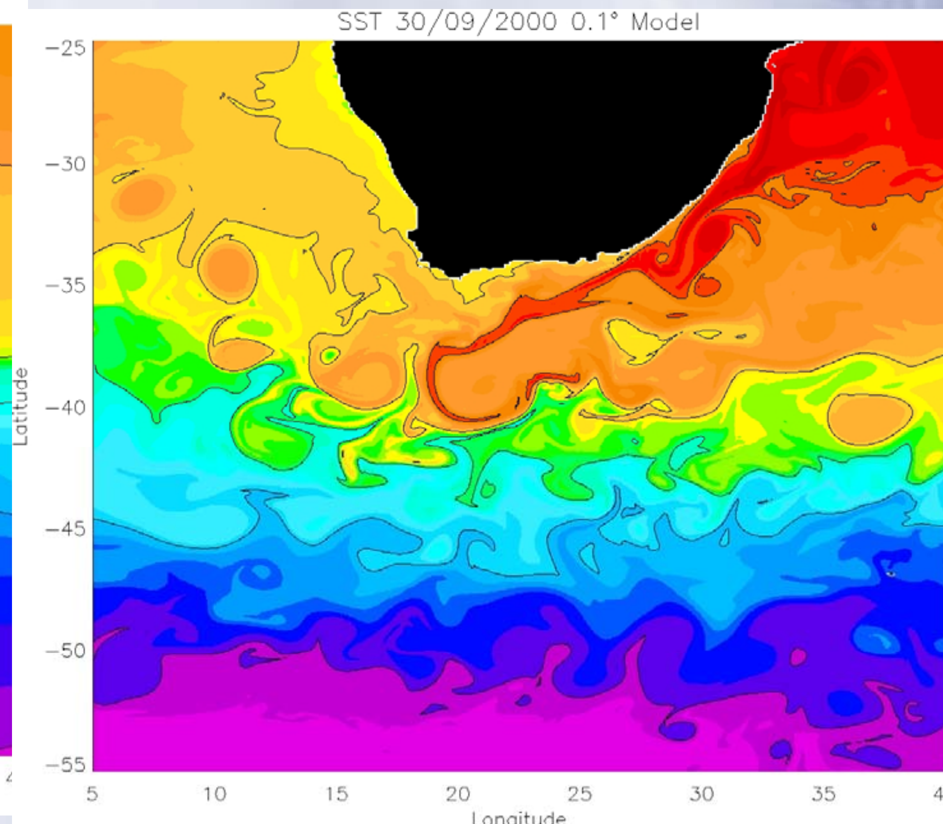
John Dennis and Richard Loft
National Center for Atmospheric Research
Boulder, Colorado

Outline

- Science Motivation
- Compute systems used
- CCSM Coupled System
- Scaling and Performance
  - Benchmark results
  - I/O variability
- Conclusions

Why High Resolution? Resolving Ocean Mesoscale Eddies



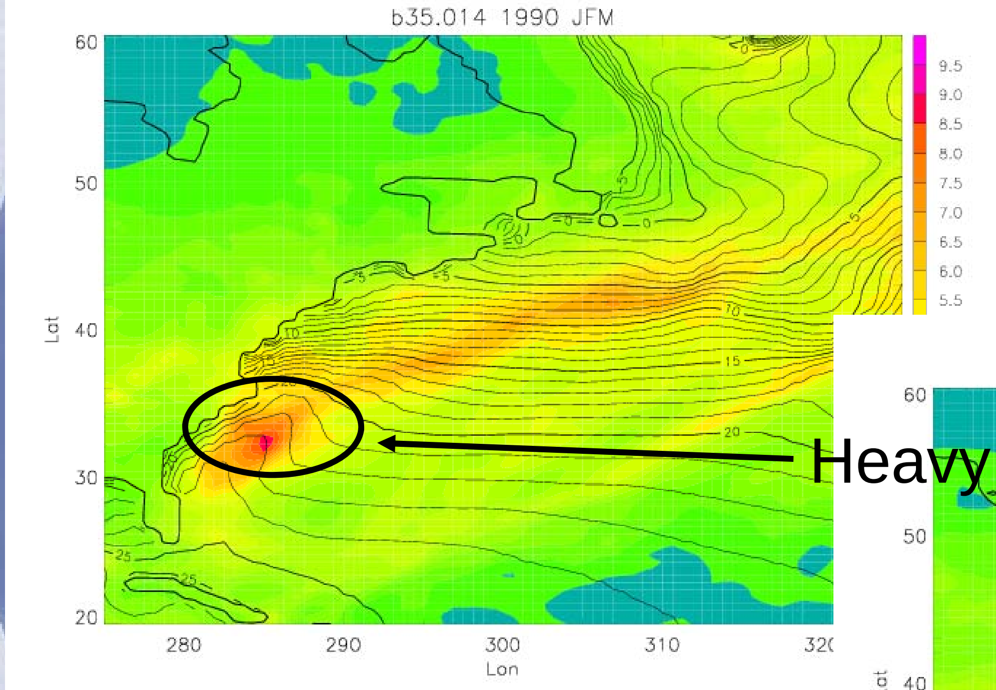


Ocean component of CCSM (Collins et al, 2006)

Eddy-resolving POP (Maltrud & McClean,2005)

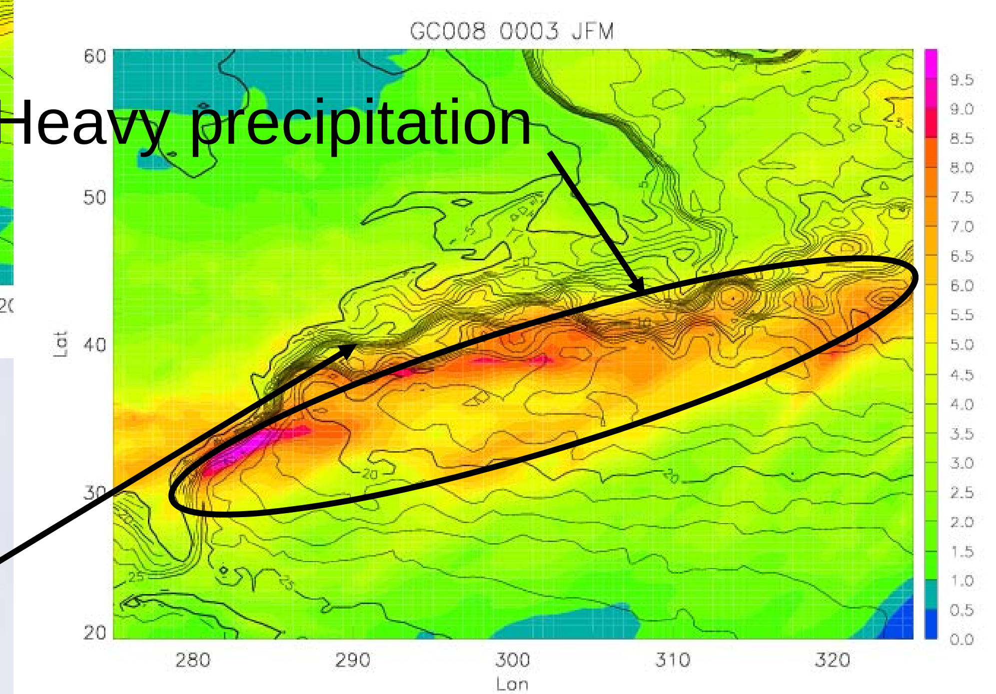
NCAR

Ocean-Atmosphere Interactions: North Atlantic Winter storm track



0.5° atm + 0.1° ocn

Heavy precipitation



0.5° atm + 1° ocn

Strong SST gradient

PetaApps: Interactive Ensembles

- Interactive ensembles
  - Multiple instances of component models
  - Explore the role of weather noise in climate
    - Test hypothesis that noise is “reddened” and influences low-frequency components of climate system
- 35M CPU hours TeraGrid [2nd largest]
  - 6000 core job: 7 months non-stop

PetaApps: Interactive Ensembles (con't)

- PetaApps project members:
  - J. Kinter, C. Stan (COLA)
  - B. Kirtman (U of Miami)
  - C. Bitz (U of Washington)
  - W. Collins, K. Yelick (U of California)
  - F. Bryan, J. Dennis, R. Loft, M. Vertenstein (NCAR)

Funding Sources

- Department of Energy: CCPP Program Grants
  - DE-FC03-97ER62402 [SciDAC]
  - DE-PS02-07ER07-06 [SciDAC]
- National Science Foundation:
  - OCI-0749206 [PetaApps]
  - OCE-0825754
  - Cooperative Grant NSF01

Outline

- Science Motivation
- Compute systems used
- CCSM Coupled System
- Scaling and Performance
  - Benchmark results
  - I/O variability
- Conclusions

Franklin Cray XT4 at NERSC

Courtesy NERSC

Kraken XT-5 at NICS

Courtesy of Pat Kovatch, NICS

Compute platforms

System	Franklin	Kraken	Atlas
processor	AMD Opteron Quad core 2.3Ghz	AMD Opteron Quad core 2.3 Ghz	AMD Opteron Dual core 2.6 Ghz
Memory/core	2 GB	2 GB/ 1GB	1.5 GB
Sockets/node	1	2	4
network	Cray Seastar	Cray Seastar	IB
total nodes	9,660	8,256	1,200
Total cores	38,640	66,048	9,600

Outline

- Science Motivation
- Compute systems used
- CCSM Coupled System
- Scaling and Performance
  - Benchmark results
  - I/O variability
- Conclusions

Community Climate System Model (CCSM)

- Multiple component models on different grids
- Flux and state between components [CPL]
- Large code base: >1M lines
  - Developed over 20+ years
  - 200-300K lines are critically important --> no comp kernels, need good compilers
- Demanding on networks:
  - need good message latency + bandwidth

NCAR

CCSM Coupling and Execution Flow

CLM – Land Model
(land/atm fluxes and land albedos)

CICE – Sea Ice Model
(ice/atm & ice/ocn fluxes and ice albedos)

CPL - Coupler
(ocean/atm fluxes and ocean albedos)

CAM - Atmosphere
(downward SW and bottom level states)

POP - Ocean
(sea surface temperature)

Coupling of CAM/CLM/CICE/AOF is every CAM timestep (30 min)

Coupling of POP is every 24 hours (1º) and every 6 hours (.1º)
CAM/CLM/CICE/AOF fluxes averaged over coupling interval

Courtesy Vertenstein, CGD/NCAR

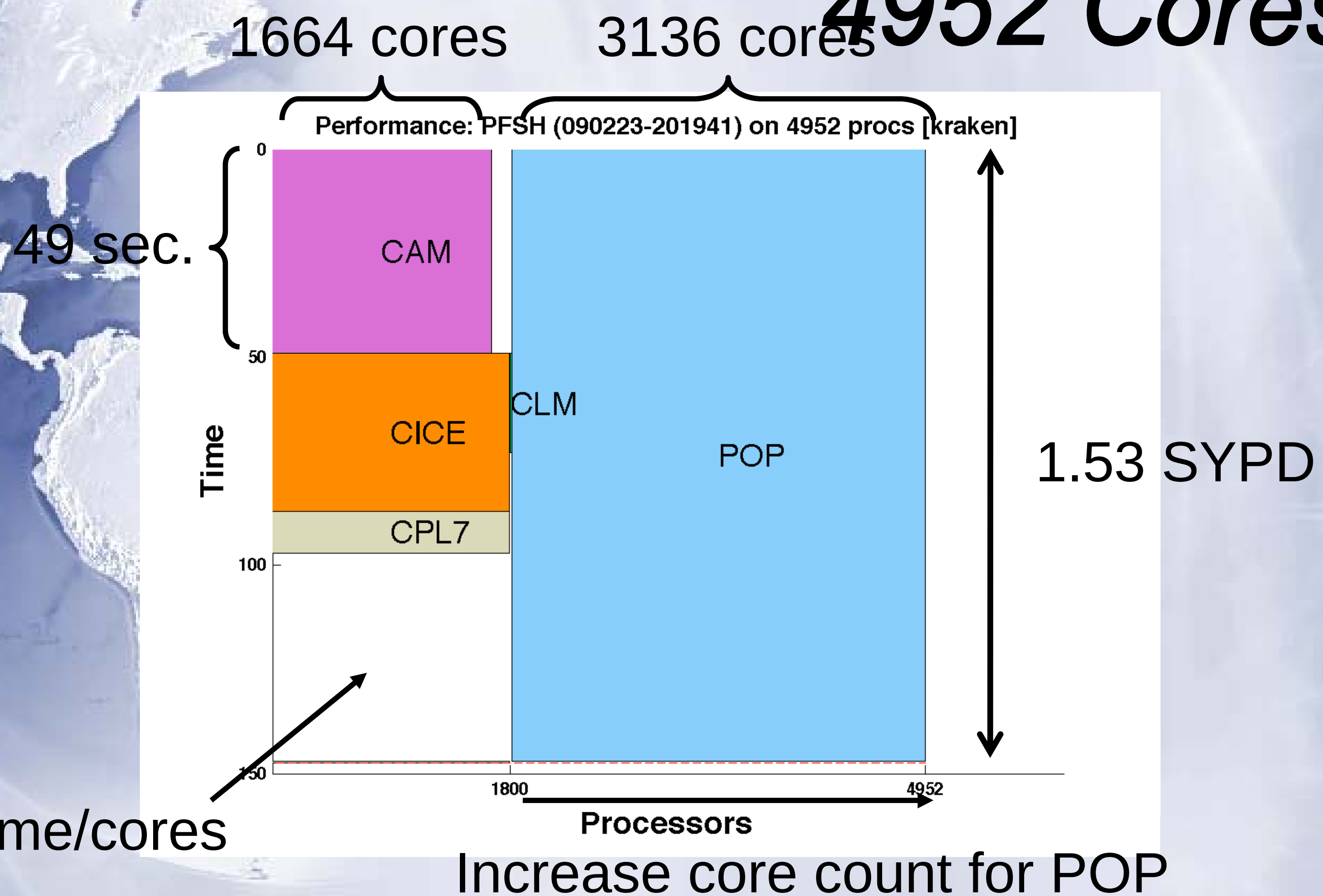
Outline

- Science Motivation
- Compute systems used
- CCSM Coupled System
- Scaling and Performance
  - Benchmark results
  - I/O variability
- Conclusions

CCSM4_alpha Benchmark Configurations

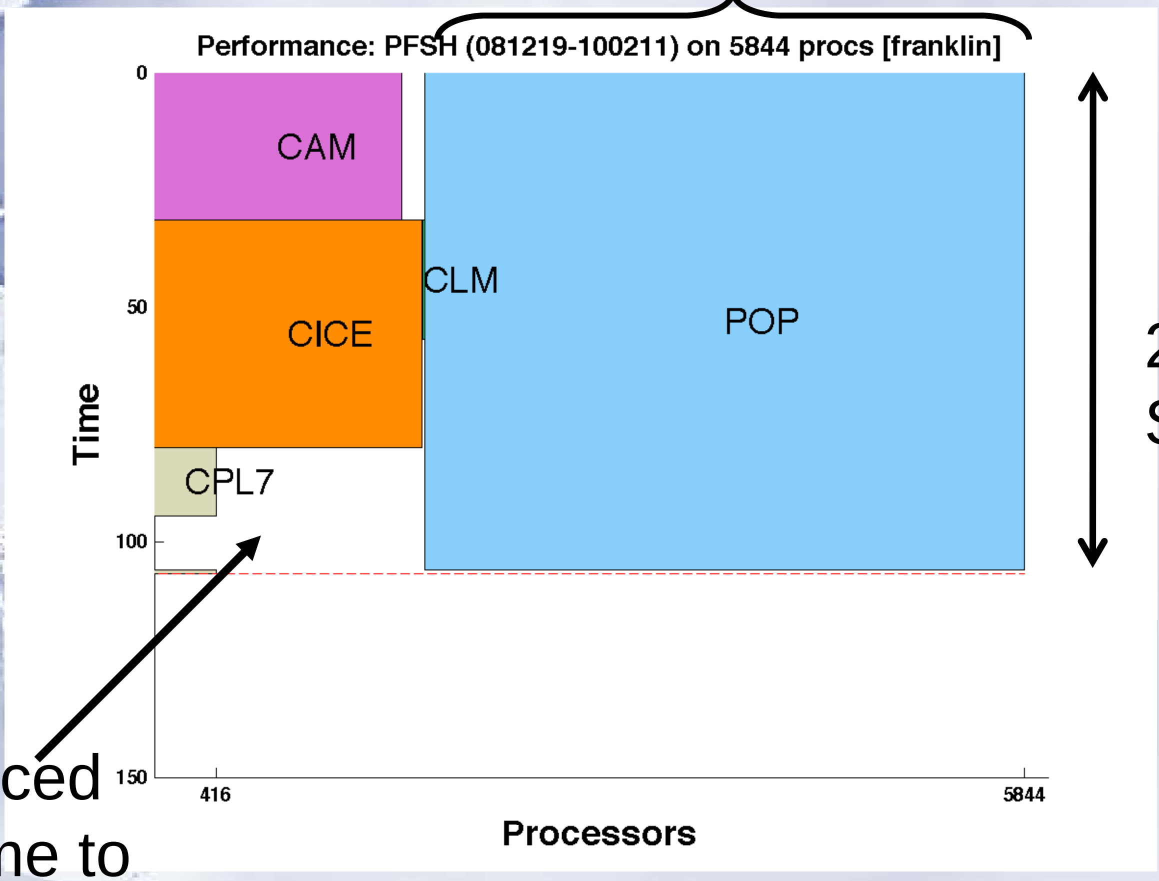
- 0.50° ATM [576 x 384 x 26]
- 0.50° LND [576 x 384 x 17]
- 0.1° OCN [3600 x 2400 x 42]
- 0.1° ICE [3600 x 2400 x 20]
- 5 days/ no writing to disk
- 5 processor configurations:
  - XS: 480 cores
  - S: 1024 cores
  - M: 1712-1865 cores
  - L: 3488-3658 cores
  - XL: 4952-6380 cores

CCSM4_alpha on 4952 Cores



CCSM4_alpha on 5844 Cores

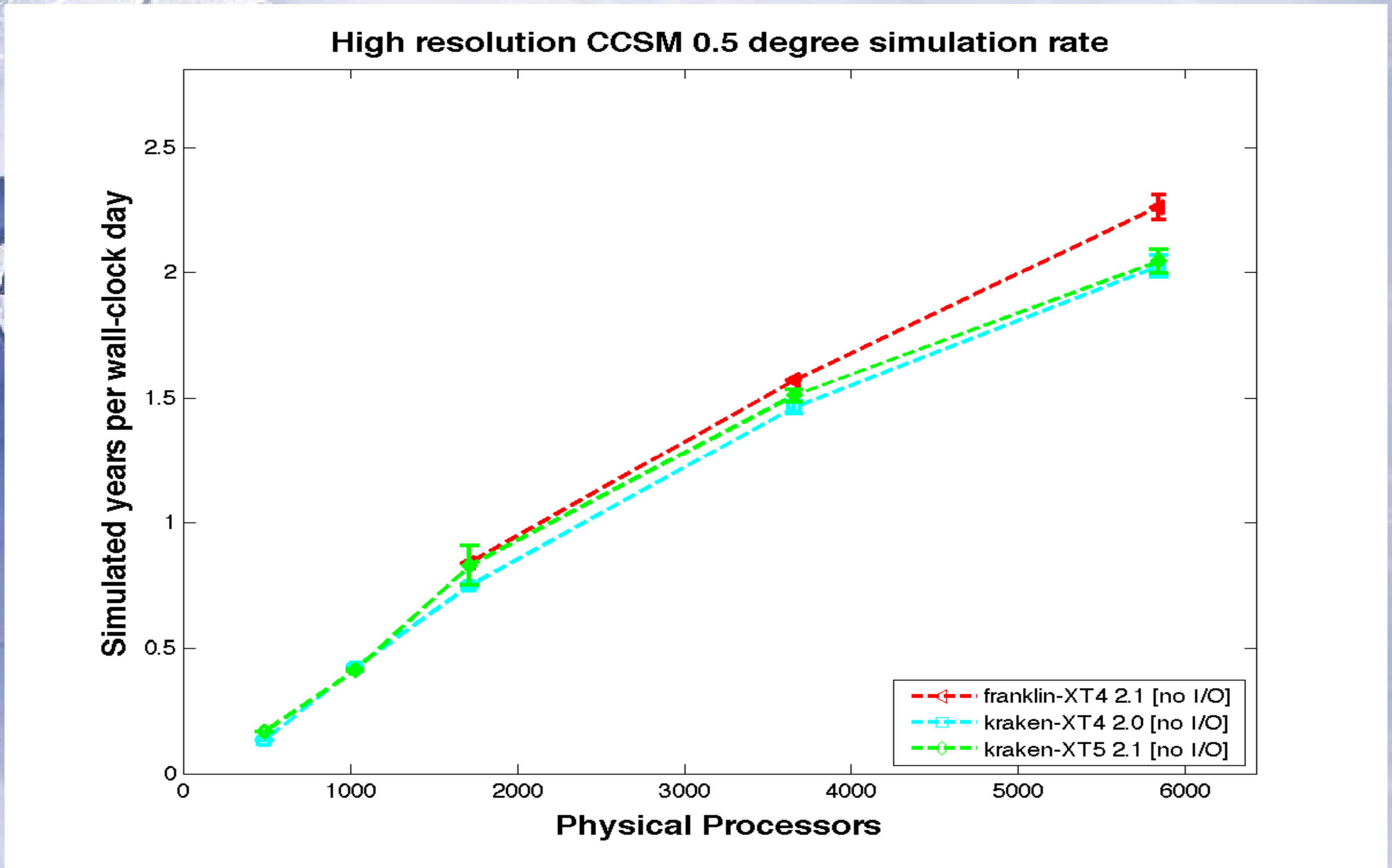
4028 cores



2.23 SYPD

Reduced Idle time to 8%

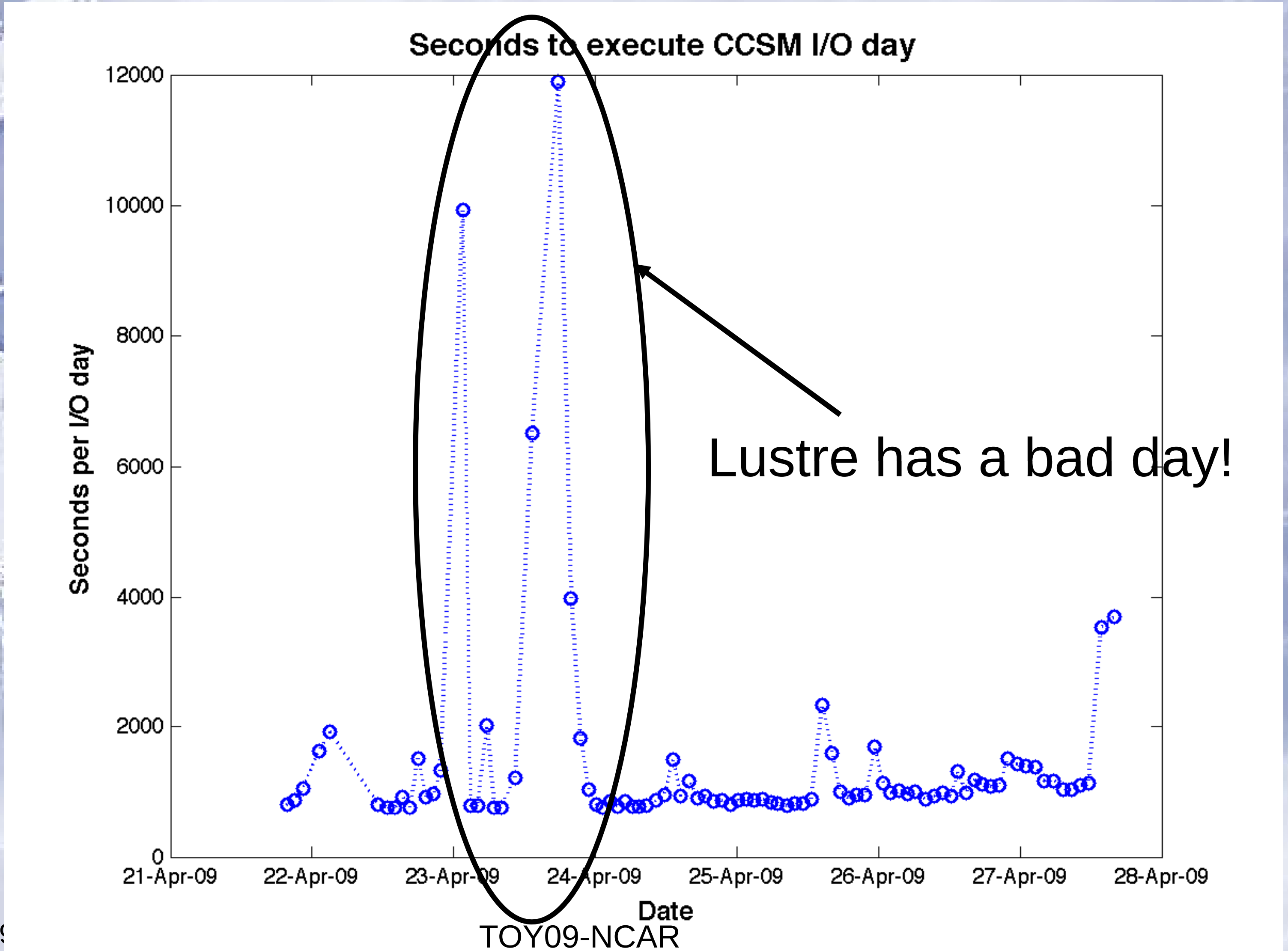
CCSM4_alpha Cray XT Scalability (no I/O)



Outline

- Science Motivation
- Compute systems used
- CCSM Coupled System
- Scaling and Performance
  - Benchmark results
  - I/O variability
- Conclusions

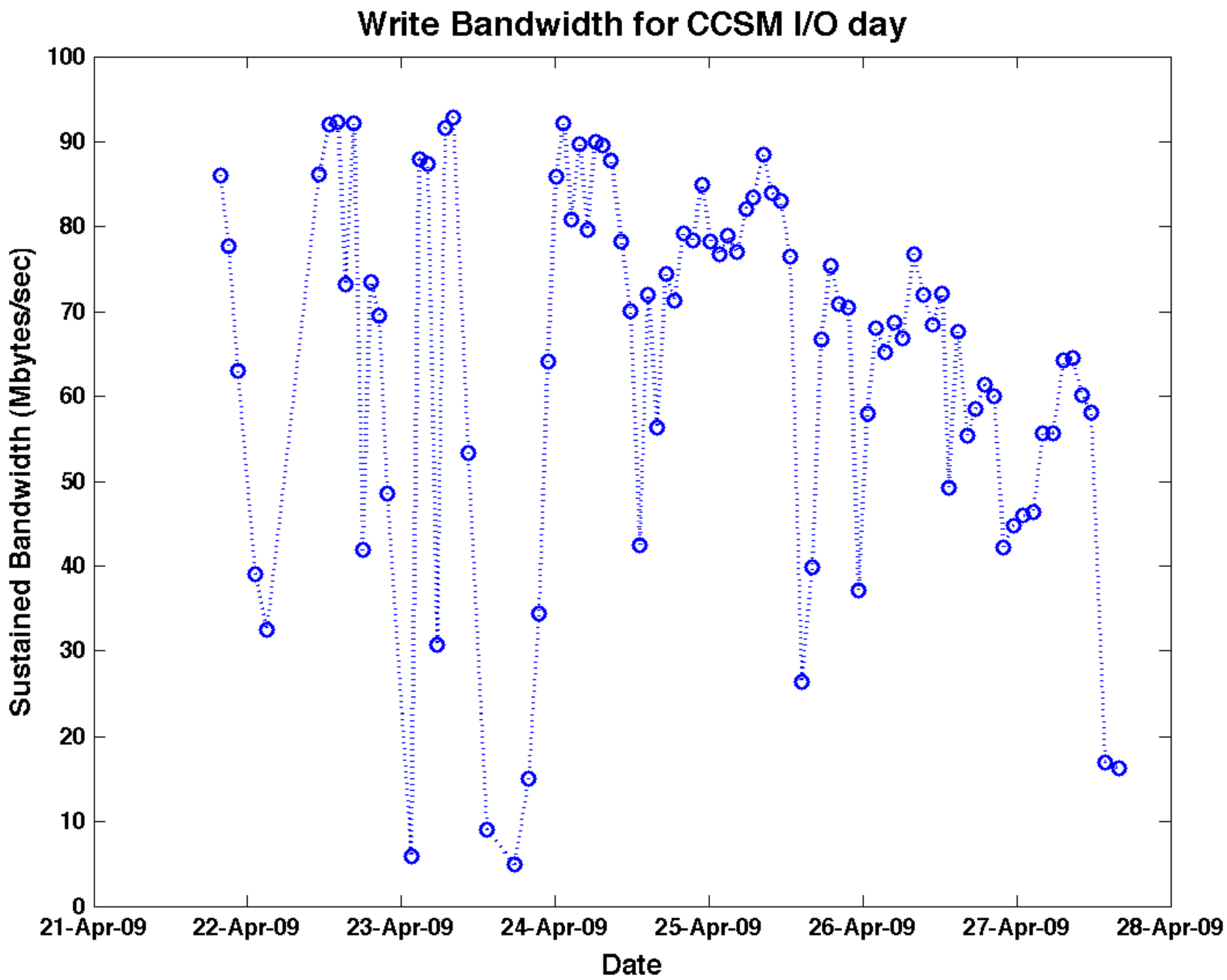
Variability of CCSM File Write Times on Kraken



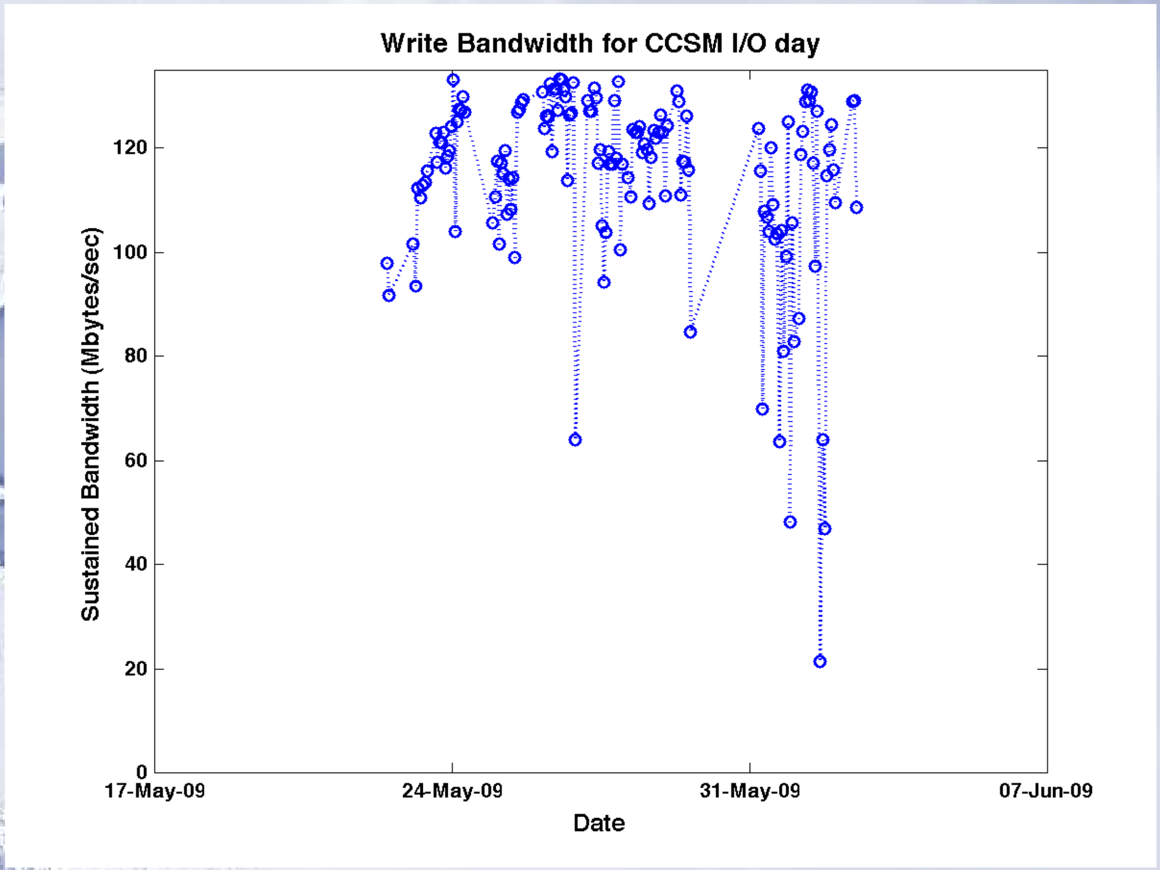
CCSM Sustained Output Bandwidth on Kraken [using big-endian]

High = 92 MB/sec

Low = 5 MB/sec



Lustre variability doubled cost of disk I/O from 11% to 23% of total



Write bandwidth for I/O day for high resolution CCSM on Kraken [little-endian]

Eliminated page size [4 Kb] I/O ops for POP restart

Conclusions

- Preliminary ultra-high-resolution science runs
  - Mesoscale processes: Atlantic storm track
- Control run in production @ NICS (Teragrid)
  - 84+ years complete
  - Generating 2.5 TB of data per week!
- Future work:
  - Improve disk I/O performance [10 - 25% of time]
  - Improve memory footprint scalability
  - OS jitter investigation

Acknowledgements and Questions?

- NCAR:
  - D. Bailey
  - F. Bryan
  - B. Eaton
  - N. Hearn
  - K. Lindsay
  - N. Norton
  - M. Vertenstein
- COLA:
  - J. Kinter
  - C. Stan
- U. Miami
  - B. Kirtman
- U.C. Berkeley
  - W. Collins
  - K. Yelick (NERSC)
- U. Washington
  - C. Bitz

- NICS:
  - M. Fahey
  - P. Kovatch
- ANL:
  - R. Jacob
  - R. Loy
- LANL:
  - E. Hunke
  - P. Jones
  - M. Maltrud
- LLNL
  - D. Bader
  - D. Ivanova
  - J. McClean (Scripps)
  - A. Mirin
- ORNL:
  - P. Worley

and many more…

- Grant Support:
  - DOE
    - DE-FC03-97ER62402 [SciDAC]
    - DE-PS02-07ER07-06 [SciDAC]
  - NSF
    - Cooperative Grant NSF01
    - OCI-0749206 [PetaApps]
    - CNS-0421498
    - CNS-0420873
    - CNS-0420985
- Computer Allocations:
  - TeraGrid TRAC @ NICS
  - DOE INCITE @ NERSC
  - LLNL Grand Challenge
- Thanks for Assistance:
  - Cray, NICS, and NERSC

Scaling High-Resolution Climate to the Next Level

John Dennis dennis@ucar.edu

Motivation

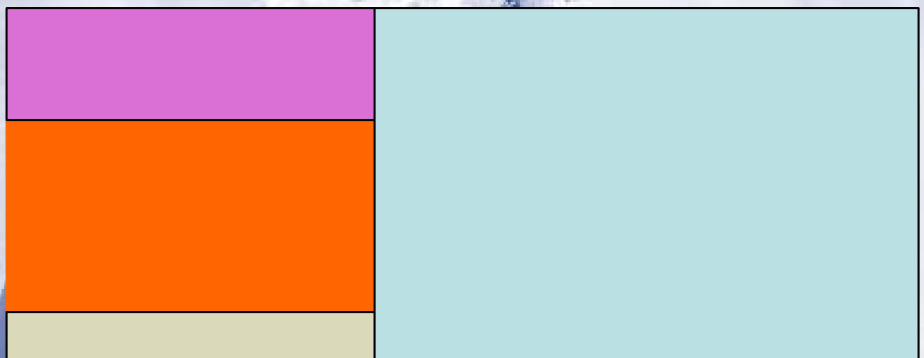
- Current simulation is rather slow…
- Why is simulation rate important?
  - Heroic effort to generate necessary time series [100-200 years]
  - Example: 108 years -> 63 days of non-stop computing
- Simulated years per wall-clock day [SYPD]

NCAR

Motivation [con't]

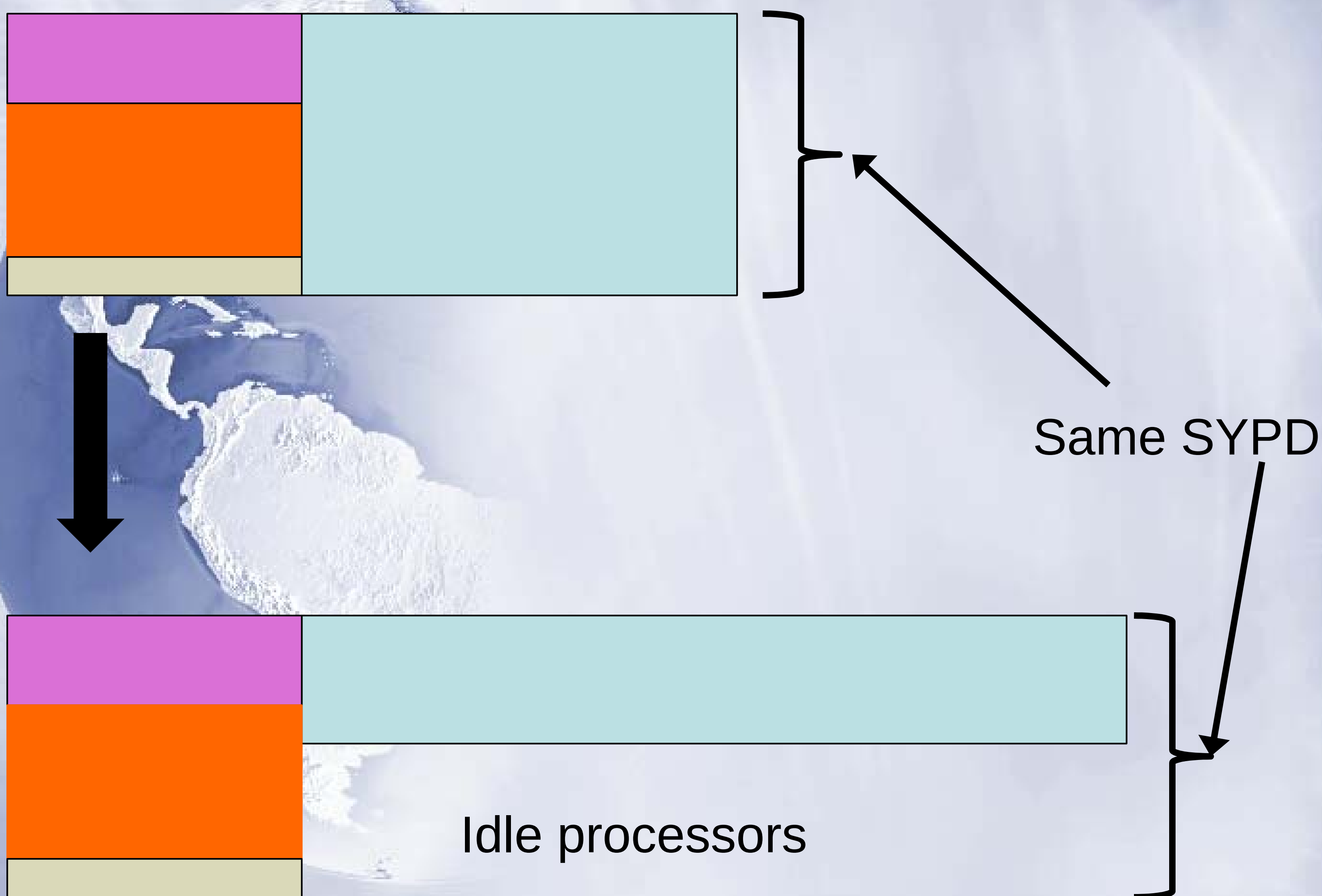
- Typical climate rate: 5 SYPD
- Currently ~1.7 SYPD
- Single thread speed is not increasing
- Need more parallelism!

Our Goal



Increase core count
Increase simulation rate

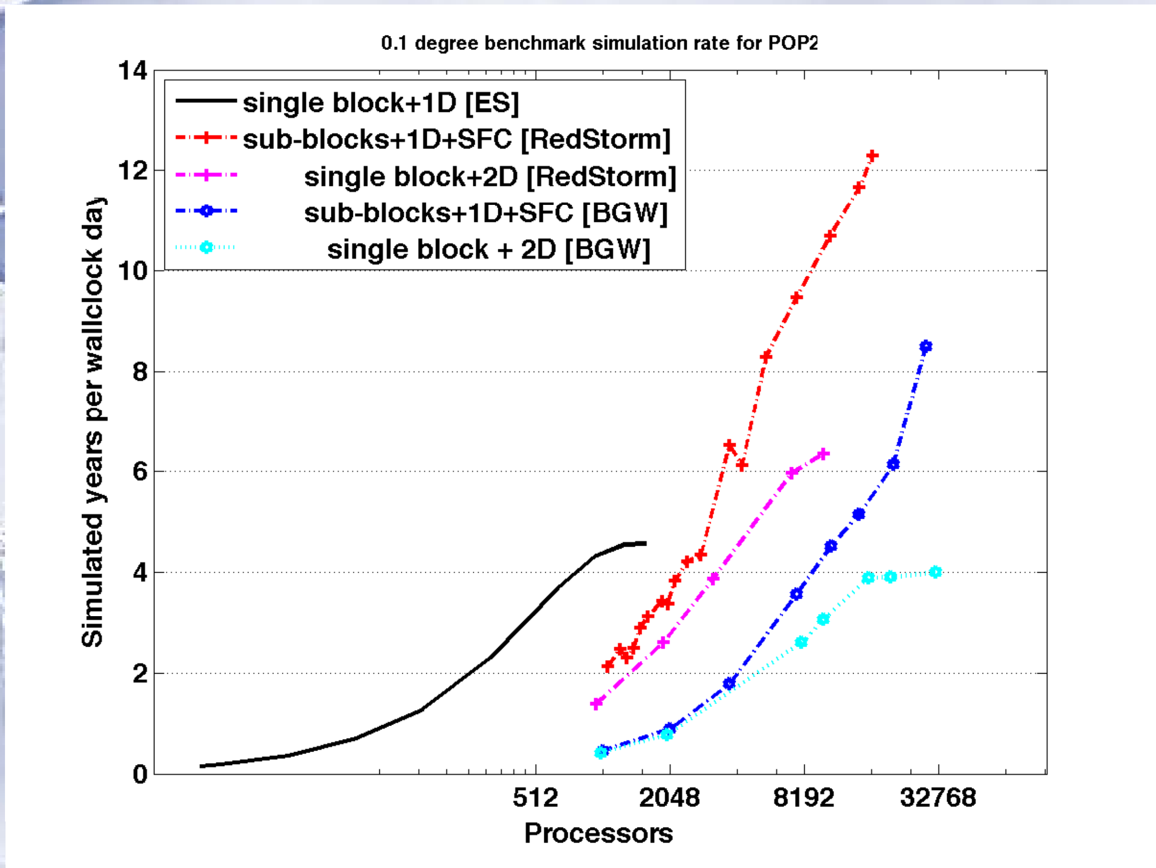
Need to scale all components!



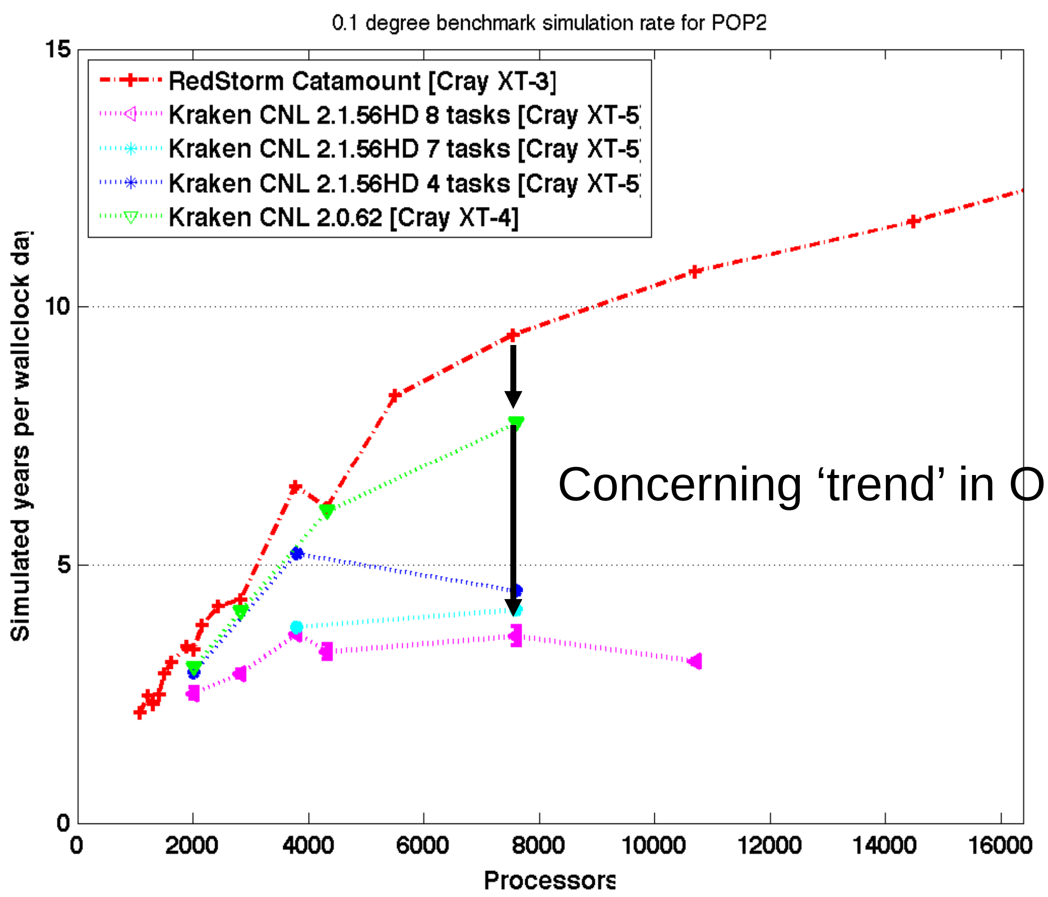
Outline

- Motivation
- Increasing scalability
  - POP
  - CICE
  - CAM
  - CPL
- Conclusions

POP scalability [3 years ago]



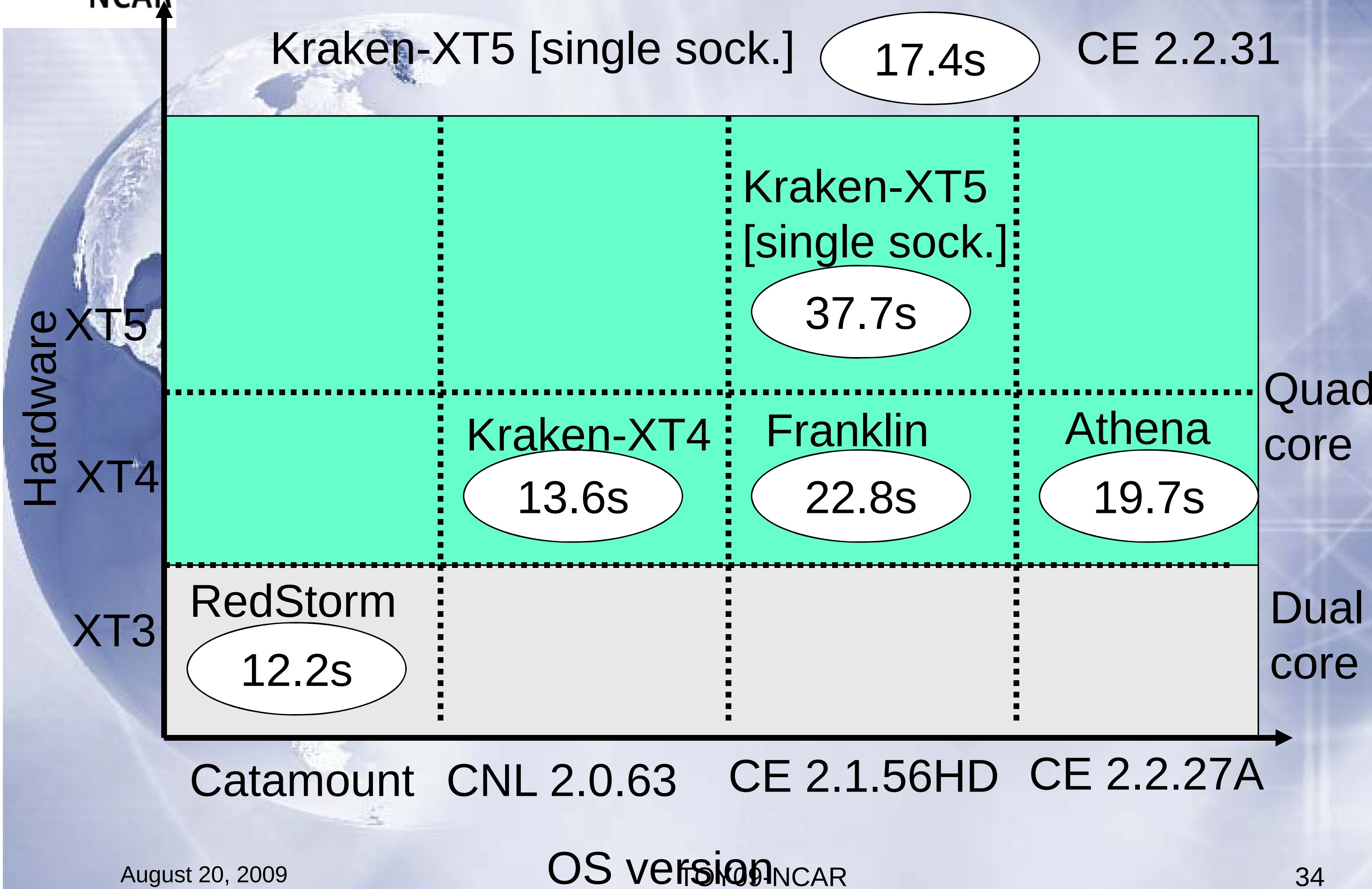
POP scalability [Several months ago]



Concerning 'trend' in OS/HW evolution?

NCAR

Execution time of Barotropic section of POP on 7600 cores



Issues with POP scalability on XT5

- POP amplifies OS interference/jitter
  - Past: Jitter kills reduction performance
  - Current: Contention at network interface
- Solutions:
  - Fix OS [Cray]
  - Modify partitioning algorithm [Dennis]
  - Overlap comm/comp [Worley]
  - OpenMP/MPI ?

Outline

- Motivation
- Increasing simulation rate
  - POP
  - CICE
  - CAM
  - CPL
- Conclusions

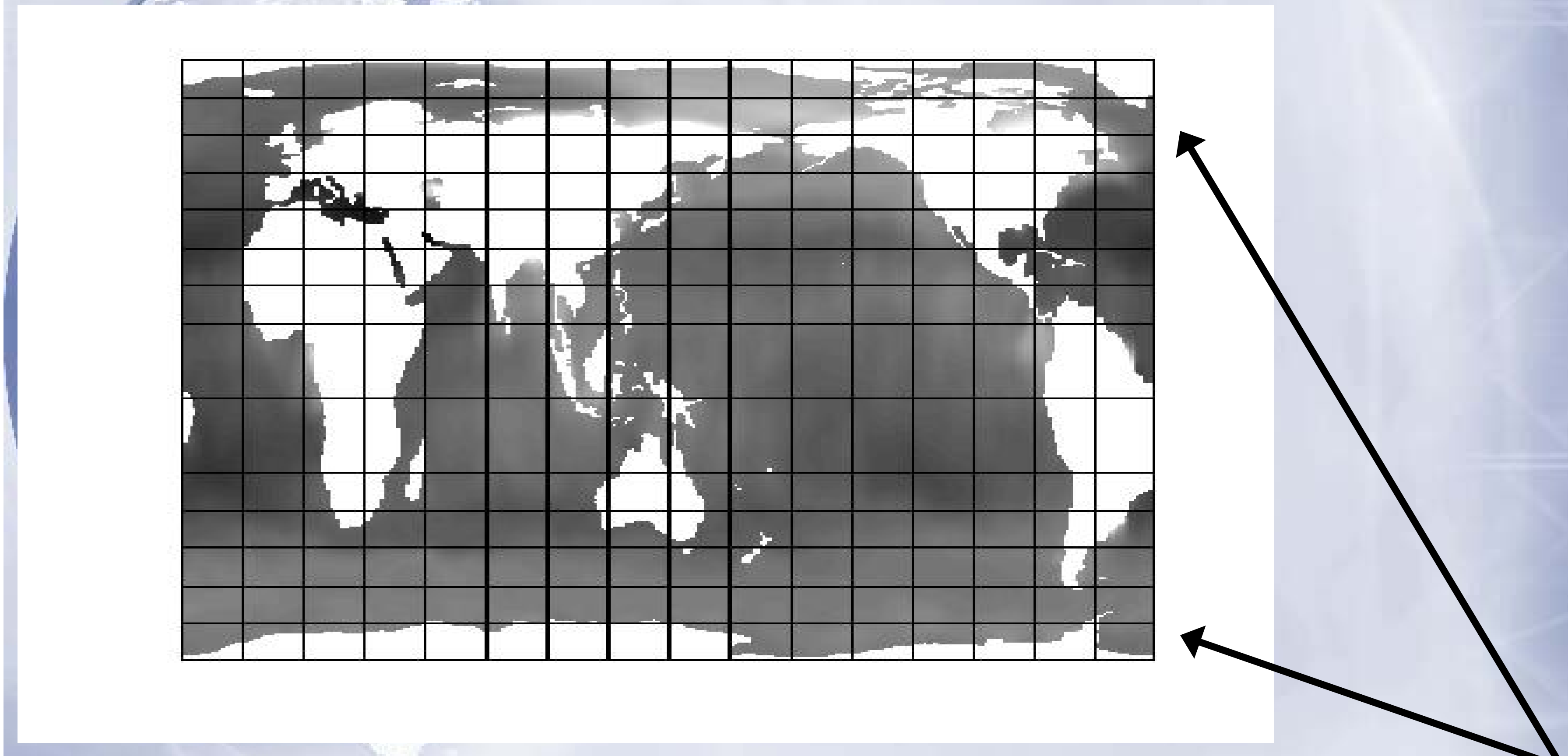
CICE: Sea-ice Model

- Developed at LANL
- Shares grid and infrastructure with POP
- Unique feature: computational load can disappear
- CICE 4.0
  - Sub-block data structures (POP2)
  - Reuse techniques from POP2 [Dennis]
  - Partitioning grid using weighted Space-filling curves:
    - 40% reduction in execution time at 0.1° on 1800 cores

Partitioning with Weighted Space-filling curves

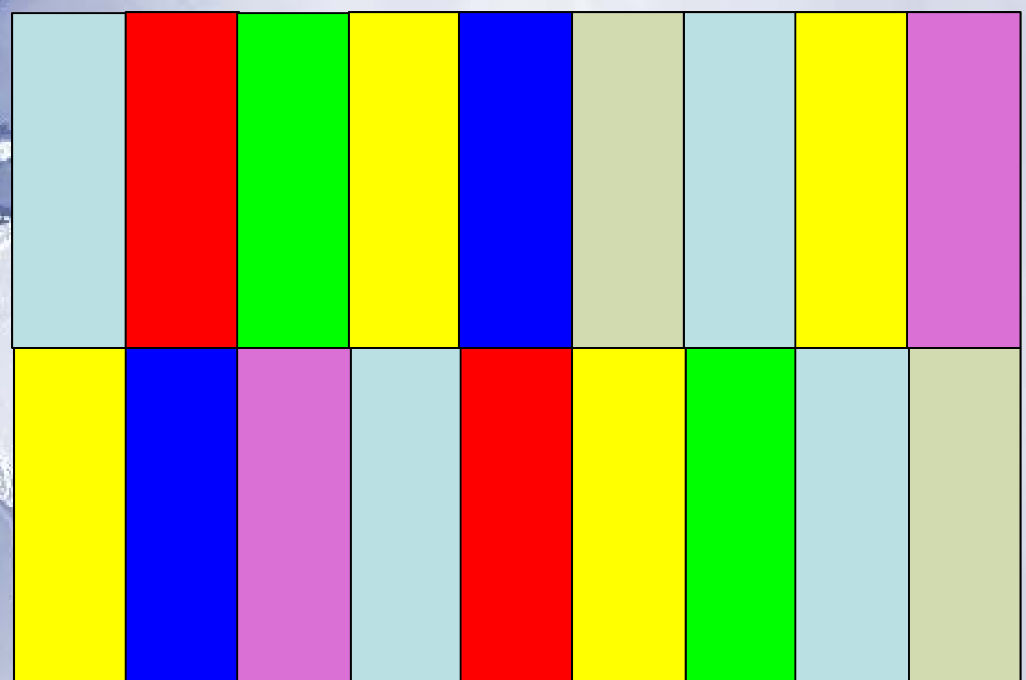
- Weight space-filling curve (wSFC)
  - Estimate work based on Probability function
  - Partition for equal amounts of work
- Probability block contains Sea-ice
  - Depends on climate scenario
    - Control-run
    - Paelo
    - CO_2 doubling
  - Estimate of probability
  - Bad estimate -> Slower simulation rate

CICE: computational grid 1°

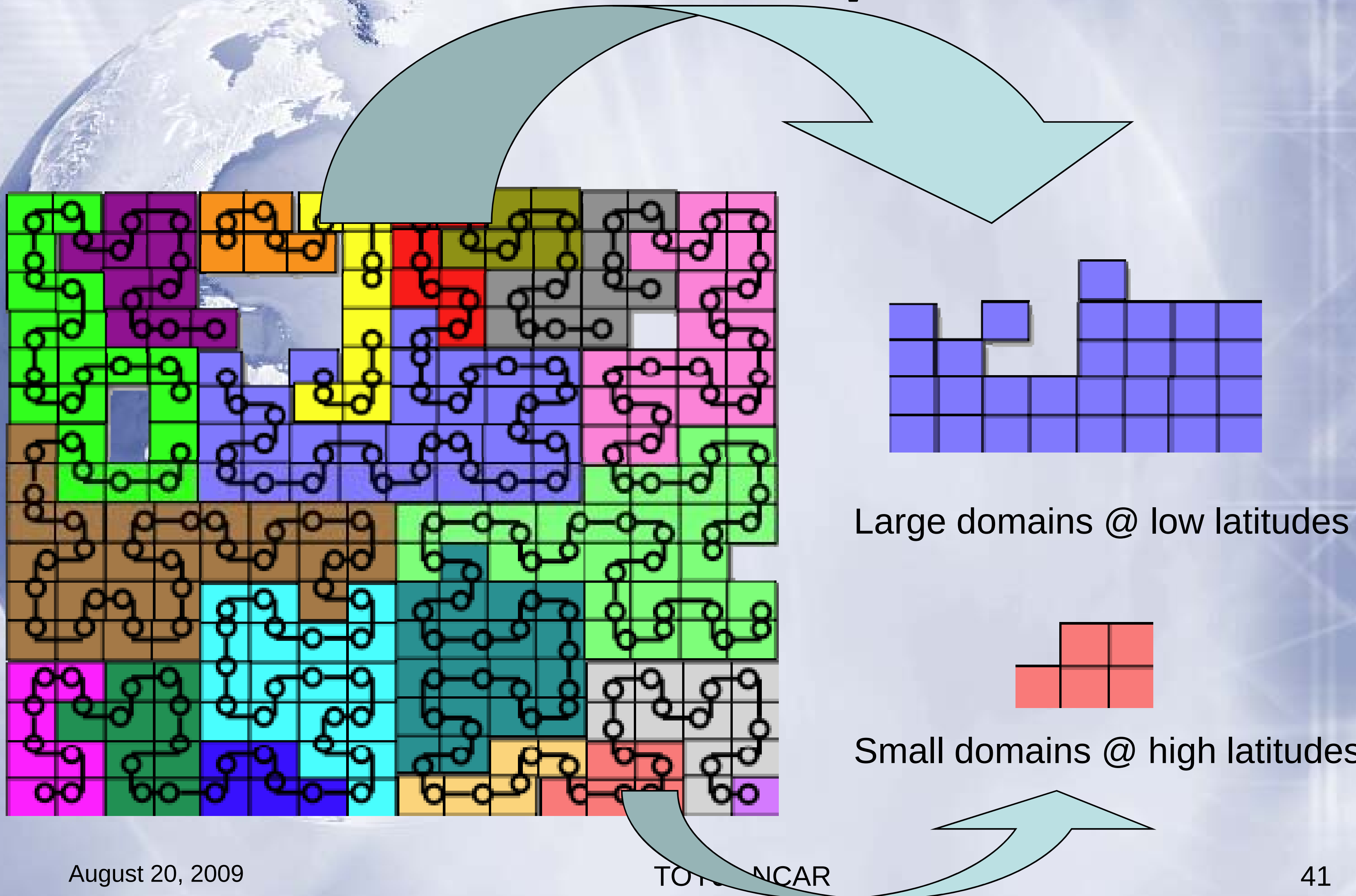


Sea ice located at high latitude

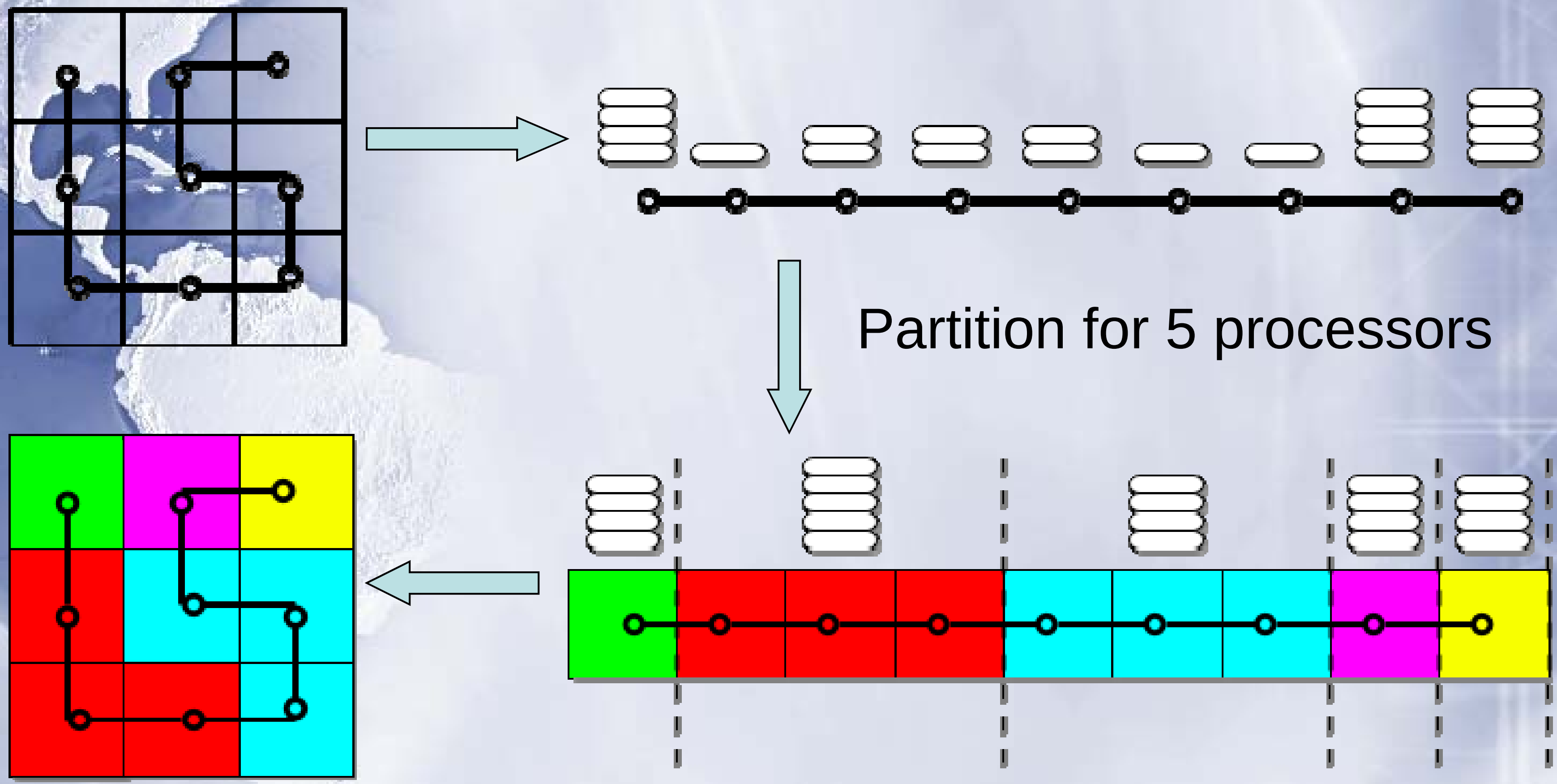
Long-skinny Cartesian partitioning



1° CICE4 on 20 processors



Partitioning with w-SFC



Weighted Space-filling curves

Predict time for grid block i

$t^p_i = c^1 * nocn_i + c^2 * nice_i + c^3 * (bsx + bsy) + c^4 * fnice_i * (bsx * bsy)$

where:

$nocn_j$: ocean points in block j

$nice_i$: sea ice points in block i where $P_i > 0.05$

$fnice_i = \{1$ if $nice_i > 0, 0$

P_i: probability sea ice present

bsx,bsy: size of grid block

Performance model: c = [c1 c2 c3 c4]

Weighted Space-filling curves (cont')

Linear system:

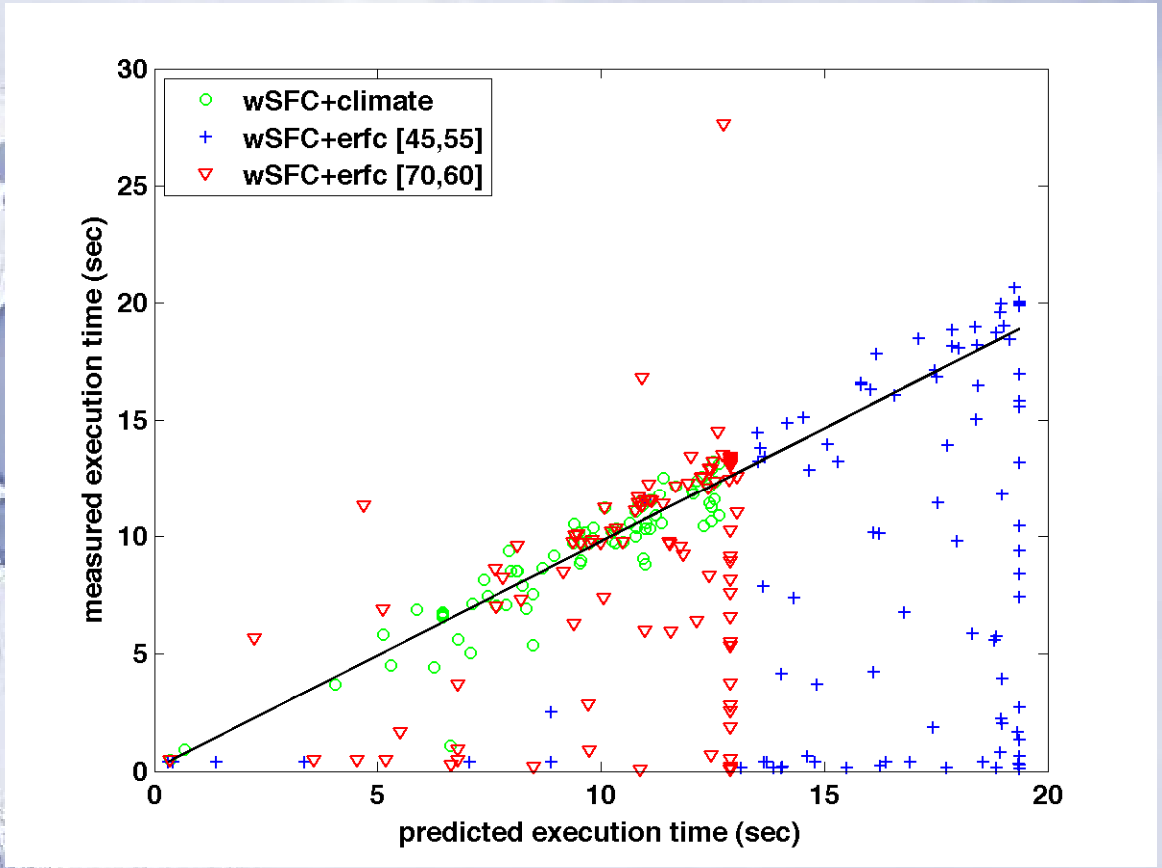
$$Ac = t^m$$

where t^m is measured execution time

Solve for performance model: c

Prediction execution time Qt^p for partition Q:

$$QAc = Qt^p$$



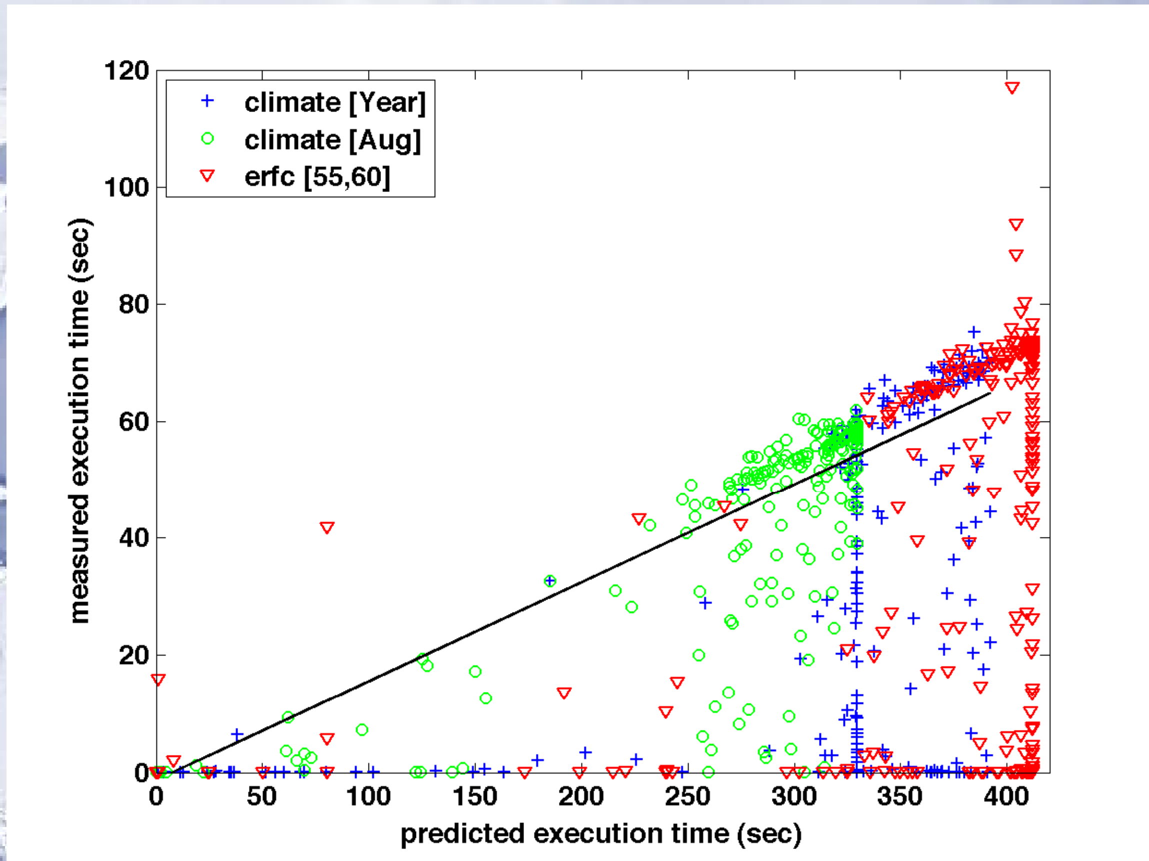
Execution time for dynamics sub-cycling 1° CICE on 128 cores of BGL

Impact of error in Probability on Performance

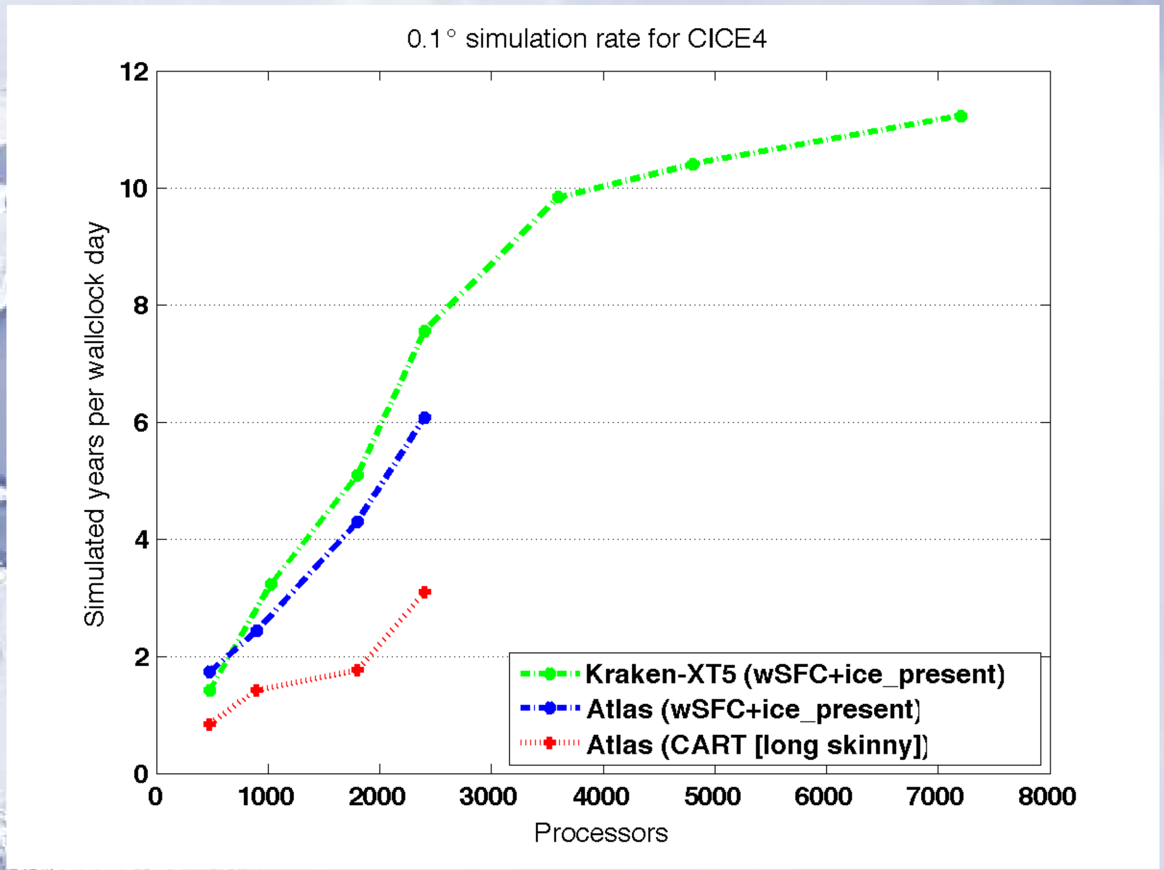
- Perfect a priori knowledge of sea-ice not possible
- What impact error in P_i?
- Sea ice extent:
  - Overestimate: erfc[45,55]
  - Underestimate: erfc[70x60]
- Best to overestimate sea ice extent

More potential error

- Single performance model ?
  - One per compute platform?
  - One per resolution?
  - For performance model: c
    - Derived on BGL at low-resolution
    - Test on ATLAS at high-resolution, high processor count



Execution time for dynamics sub-cycling 0.1° CICE on 480 cores of ATLAS



Simulation rate for CICE 0.1° on Cray XT5 & ATLAS

Outline

- Motivation
- Increasing simulation rate
  - POP
  - CICE
  - CAM
  - CPL
- Conclusions

CAM scalability

- Not always the dominate cost
- Incremental Solution:
  - Parallelizing tracers [Mirin,Worley]
  - OpenMP/MPI
- Radical Solutions:
  - New scalable dynamical core/method

Outline

- Motivation
- Increasing simulation rate
  - POP
  - CICE
  - CAM
  - CPL
- Conclusions

CPL scalability

- Total redesign versus previous generation
  - 10-20x increase in core count
- Minor impact at 1800 cores
- Tested at ~12,000 cores on BGL

Conclusions

- Possible to increase scalability of high-resolution CCSM4
- Load balance is critical
  - POP imbalance causes contention for network
  - CICE imbalance of computational/communication costs
- Improve parallel I/O performance
- Number of subtle issues

Acknowledgements and Questions?

- NCAR:
  - D. Bailey
  - F. Bryan
  - B. Eaton
  - N. Hearn
  - K. Lindsay
  - N. Norton
  - M. Vertenstein
- COLA:
  - J. Kinter
  - C. Stan
- U. Miami
  - B. Kirtman
- U.C. Berkeley
  - W. Collins
  - K. Yelick (NERSC)
- U. Washington
  - C. Bitz

- NICS:
  - M. Fahey
  - P. Kovatch
- ANL:
  - R. Jacob
  - R. Loy
- LANL:
  - E. Hunke
  - P. Jones
  - M. Maltrud
- LLNL
  - D. Bader
  - D. Ivanova
  - J. McClean (Scripps)
  - A. Mirin
- ORNL:
  - P. Worley

and many more…

- Grant Support:
  - DOE
    - DE-FC03-97ER62402 [SciDAC]
    - DE-PS02-07ER07-06 [SciDAC]
  - NSF
    - Cooperative Grant NSF01
    - OCI-0749206 [PetaApps]
    - CNS-0421498
    - CNS-0420873
    - CNS-0420985
- Computer Allocations:
  - TeraGrid TRAC @ NICS
  - DOE INCITE @ NERSC
  - LLNL Grand Challenge
- Thanks for Assistance:
  - Cray, NICS, and NERSC