

*INCITE:
Ultra-high resolution CCSM on Blue Gene*

John M. Dennis: dennis@ucar.edu

Mariana Vertenstein: mvertens@ucar.edu

December 05, 2007

Motivation

- ✧ Can CCSM utilize Petascale systems?
- ✧ Use Blue Gene as a prototype system
 - ✧ 'Easy' access to large systems
 - ✧ Demanding platform [limited memory]
 - ✧ Similar set of techniques for Blue Gene/XT
- ✧ Have demonstrated scalability of each component
- ✧ Preparation for ACLF INCITE proposal

Target Configuration: 0.47x0.63_tx0.1v2

- ✧ 5x 50 year runs [1980-2030]
- ✧ Configuration:
 - ✧ FV-CAM (0.47° x 0.63° , L26 or L31)
 - ✧ CLM (0.47° x 0.63°)
 - ✧ POP @ 0.1° *
 - ✧ CICE4 @ 0.1° *
 - * tripole grid with Caspian sea + PBC
- ✧ CPL7 (Sequential/Hybrid Coupler)
- ✧ Goal:
 - ✧ 4K - 32K BG/P processors
 - ✧ Runs begin in April 2008
 - ✧ **Modifications a part of CCSM4 (Stock code)**

Very large systems exist!

- ✧ Increasing common access to large systems
 - ✧ LLNL Appro AMD: 9K processors [today]
 - ✧ TJ Watson IBM Blue Gene/L: 40K processors [today]
 - ✧ ORNL Cray XT3/4 :
 - ✧ 22K processors [today]
 - ✧ 44K processors [Jan 2008]
 - ✧ **BNL/SUNY IBM Blue Gene/L: 38K processors [today]**
 - ✧ NERSC Cray XT4: 19K processors [today]
 - ✧ TACC Sun: 55K processors [Jan 2008]
 - ✧ ANL IBM Blue Gene/P:
 - ✧ 32K processors [Jan 2008]
 - ✧ **160K processors [Spring 2008]**
- Debug system
- Production system

People Involved

- ✧ T. Craig (NCAR)
- ✧ J. Dennis (NCAR)
- ✧ B. Eaton (NCAR)
- ✧ J. Edwards (IBM)
- ✧ R. Jacob (ANL)
- ✧ R. Loy (ANL)
- ✧ A. Mirin (LLNL)
- ✧ P. Worley (ORNL)
- ✧ M. Vertenstein (NCAR)

PIO Telecom Group

Funding Sources

✧ Department of Energy: CCPP Program Grants

✧ DOE-BER

✧ DE-FC03-97ER62402

✧ DE-PS02-07ER07-06

✧ DE-FC02-07ER64340

✧ B&R KP1206000

✧ DOE-ASCR

✧ B&R KJ0101030

✧ National Science Foundation:

✧ Cooperative Grant NSF01

Outline:

- ✧ Motivation
- ✧ **The puzzle pieces**
 - ✧ POP
 - ✧ CICE
 - ✧ CLM
 - ✧ CAM
 - ✧ PIO
 - ✧ CPL7
- ✧ The whole puzzle
 - ✧ Assembly
 - ✧ Status
- ✧ Conclusions

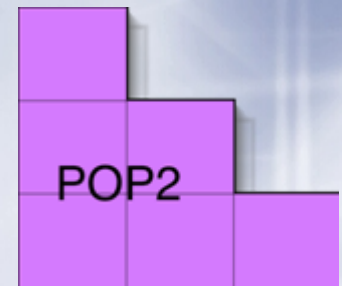


POP



Status of POP

- ✧ Developed at LANL
- ✧ Two lines of code development
 - ✧ CCSM-POP (current CCSM3.5 ocean model)
 - ✧ LANL-POP
- ✧ Simulation rate of offline LANL-POP2 benchmark
 - ✧ Blue Gene +50%
 - ✧ Cray XT4 +33%
 - ✧ Does not include MPI_reduce fixes [P. Worley]
- ✧ Won BGW cycle allocation
 - ✧ 110 Rack Days/ 5.4M CPU hours
 - ✧ Completed 9.5 year of spinup [7600 processors]
- ✧ Prototype version of PIO [binary]
 - ✧ restart: CCSM-POP & LANL-POP + my mods
 - ✧ tavg: LANL-POP + my mods



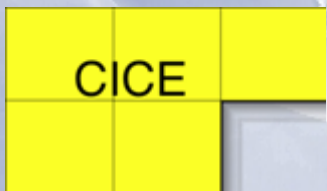
POP: Issues

- ✧ Large cost for maintaining two [or three] versions of POP
 - ✧ CCSM-POP
 - ✧ LANL-POP
 - ✧ Hacked up POP2 for scalability [BGW runs]
 - ✧ Space-filling curves
 - ✧ Parallel-I/O
 - ✧ Need a unified version
- ✧ Improved support for 0.1° in CCSM POP
- ✧ Performance issues for tripole grid
- ✧ Full support for PIO [netcdf]



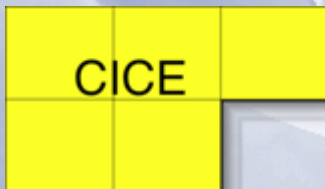


CICE



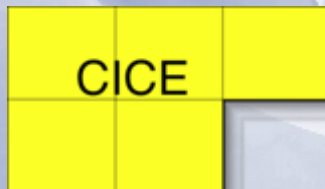
CICE4

- ✧ Developed at LANL (current CCSM3.5 sea-ice model)
- ✧ Shares grid and infrastructure with POP
 - ✧ Reuse techniques from POP work
- ✧ Computational load-imbalance for CICE4 creates challenges:
 - ✧ ~15% of grid has sea-ice
 - ✧ Use *weighted* Space-filling curves?
- ✧ Evaluate offline CICE4 @ 0.1° (computational grid [3600 x 2400 x 20]) using benchmark:
 - ✧ 1 day/ Initial run / 30 minute timestep/ no Forcing
 - ✧ 10K Cray XT3 processors
 - ✧ 40K Blue Gene/L processors



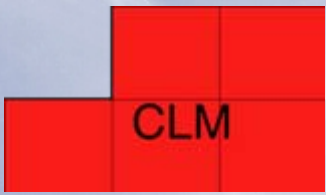
CICE4: Issues

- ✧ Implementation of OpenMP directives
- ✧ Implementation of parallel I/O
 - ✧ Based on POP2 prototype

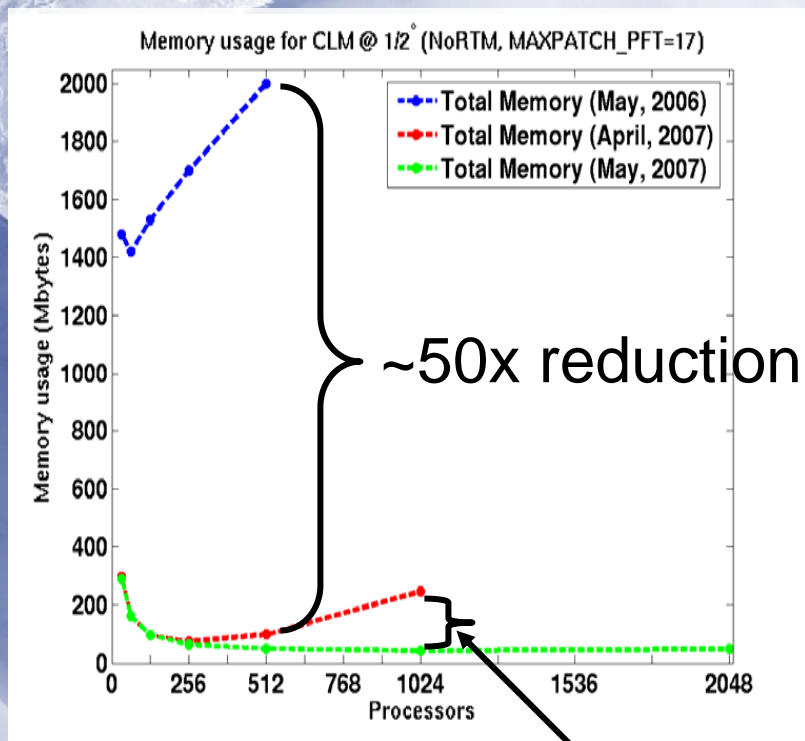




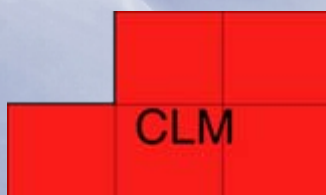
CLM



Status of CLM

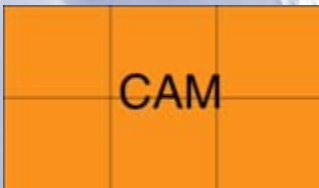


- ✧ CLM is inherently embarrassingly parallel
- ✧ Accomplished (in CCSM3.5)
 - ✧ **Elimination of global memory** and reworking of decomposition algorithms
 - ✧ Separation of CAM/CLM grids
 - ✧ Work of Tony Craig
- ✧ Current Work
 - ✧ Implementation of parallel I/O using PIO
 - ✧ Investigation of scalability at 1/6° & 1/10°



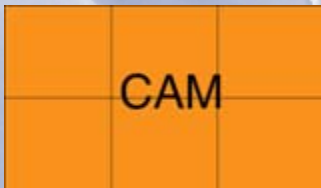



CAM



CAM@ 0.47°x0.63°

- ✧ Scalability of FV-CAM [Mirin & Worley 2007]
 - ✧ Enable execution of physics on more processors than dynamics
 - ✧ Improvements in performance of MPI reduction
 - ✧ Reductions in memory usage now enables execution of FV-CAM @ 0.47° x 0.63° on BGL
- ✧ Memory issue in cam_history.F90 located
 - ✧ The 'names' of history fields, {'T','PS','U','V', ...}
 - ✧ 72 Mbytes/processor
 - ✧ 72 Gbytes total
 - ✧ 15% of all memory
 - ✧ Elimination will enable use of all processors on Blue Gene/P

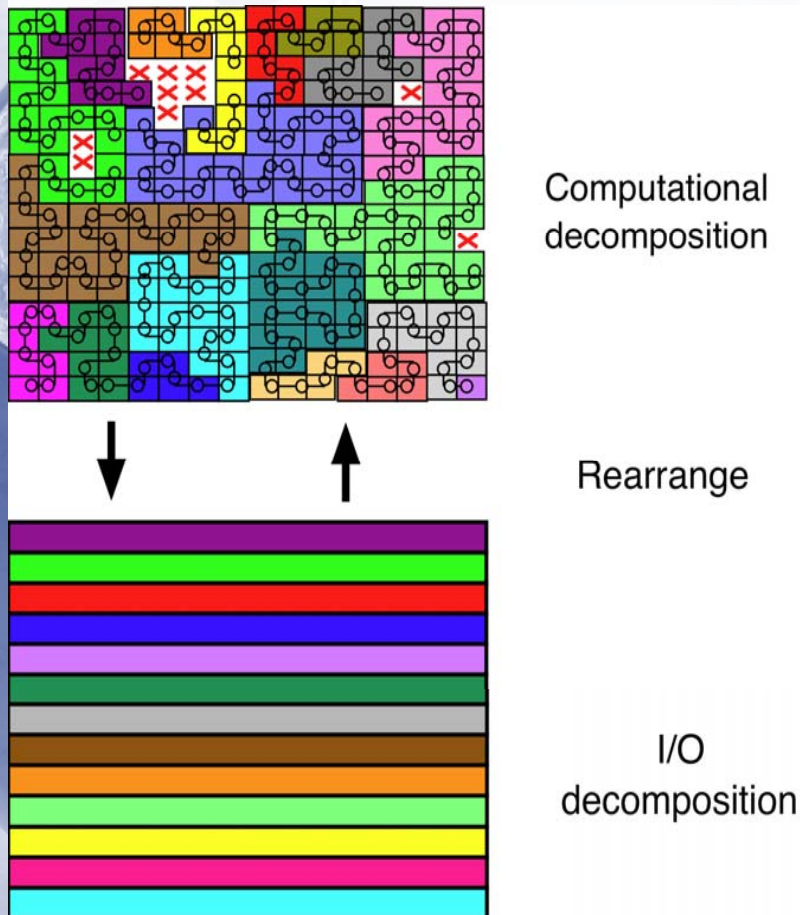




Parallel I/O (PIO)

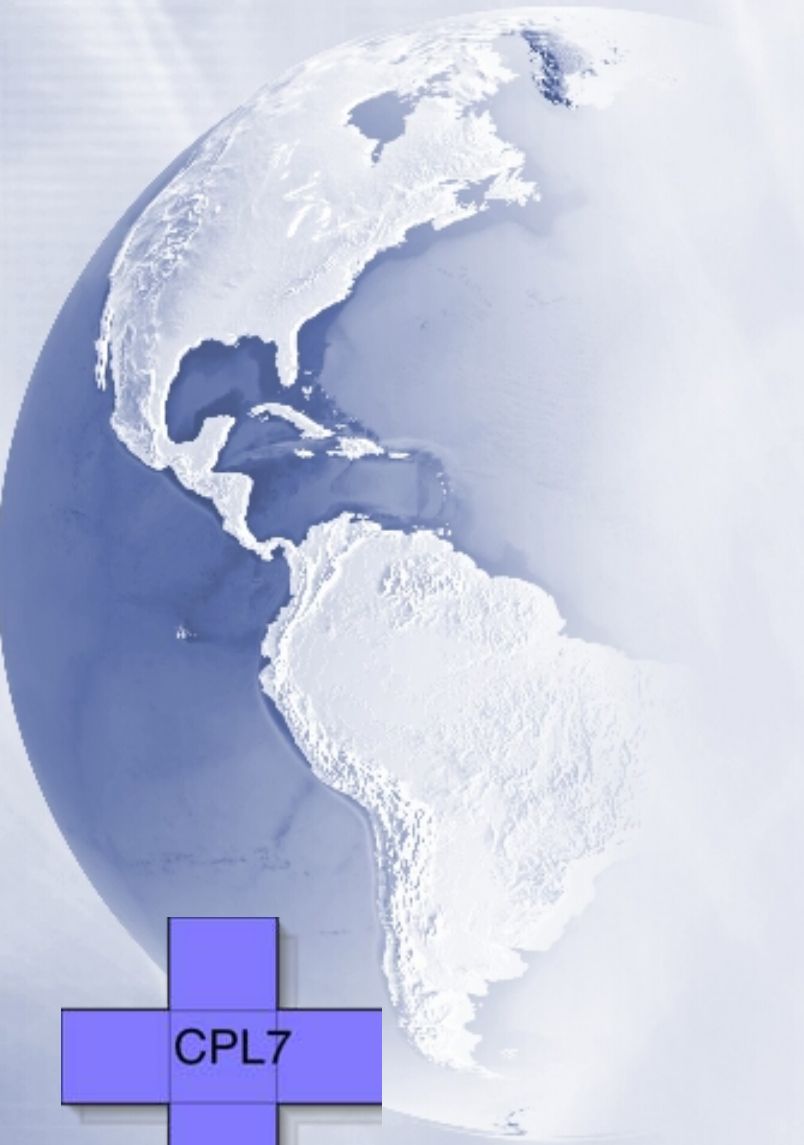
PIO

PIO: Parallel I/O library

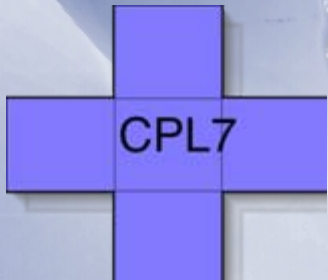


- ✧ Work of J. Dennis, R. Loy, & J. Edwards
- ✧ All component models need parallel I/O
 - ✧ Serial I/O blows memory
- ✧ Supports: binary & netCDF
- ✧ Flexibility to adapt to I/O system
- ✧ Prototype implementations
 - ✧ POP2 (CCSM & LANL)
 - ✧ CLM
 - ✧ CAM
- ✧ Rearrangement:
 - ✧ MCT
 - ✧ ESMF [work in progress]
- ✧ Critical for high {resolution, processor counts, low memory machines}





CPL7



CCSM4 (CPL7) Goals

- ✧ Target broader global resolutions
 - ✧ 3° -> .1° degree ocean/ice
 - ✧ 4° degree -> .25° (atm/land)
- ✧ Target broader range of platforms
 - ✧ Single pe (linux laptop)
 - ✧ Linux clusters
 - ✧ IBM AIX
 - ✧ Cray XT4, IBM BG (1K-> 30K) pes
 - ✧ Next Generation petascale architectures

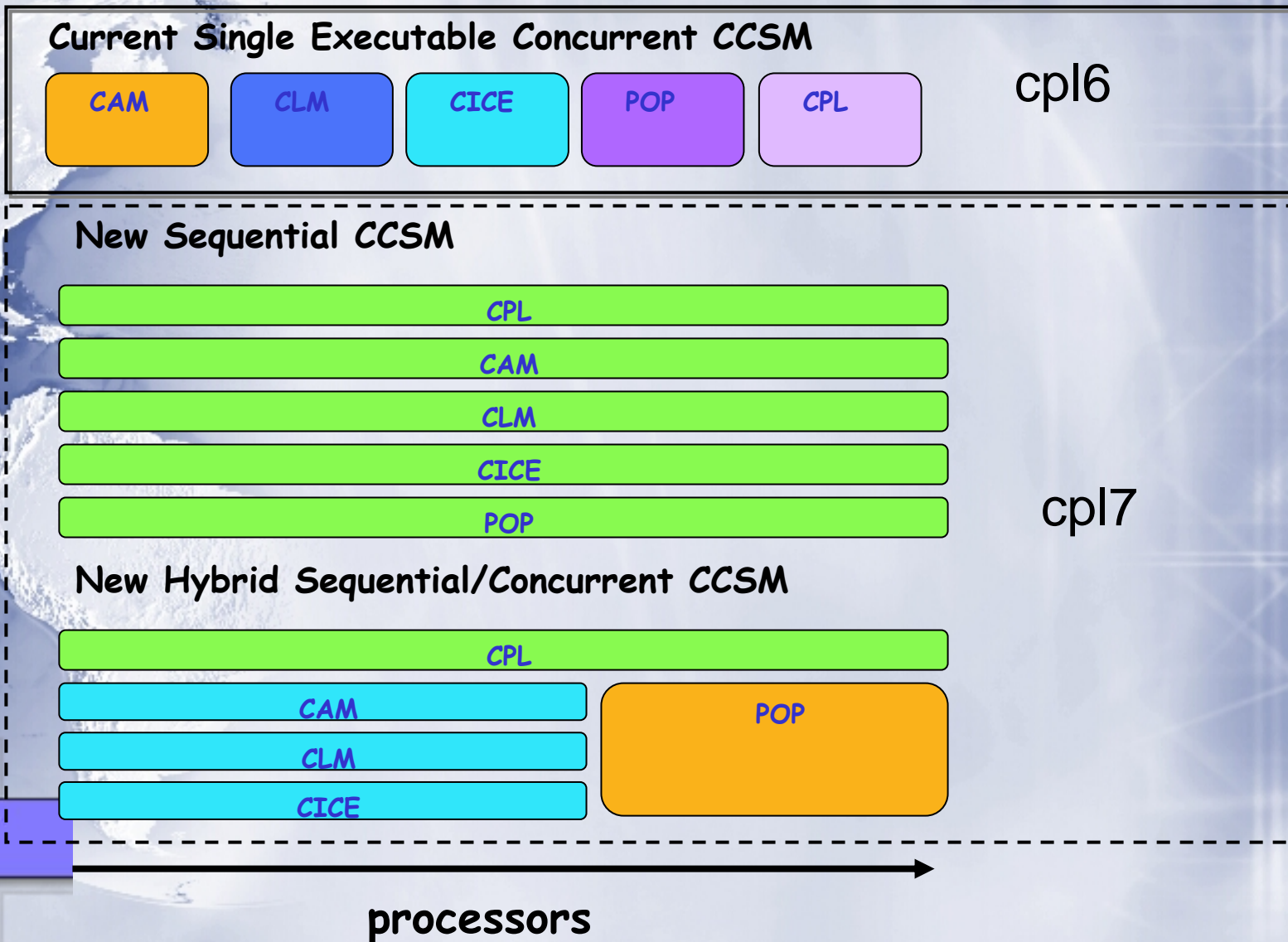


CCSM4 (CPL7) Goals (cont)

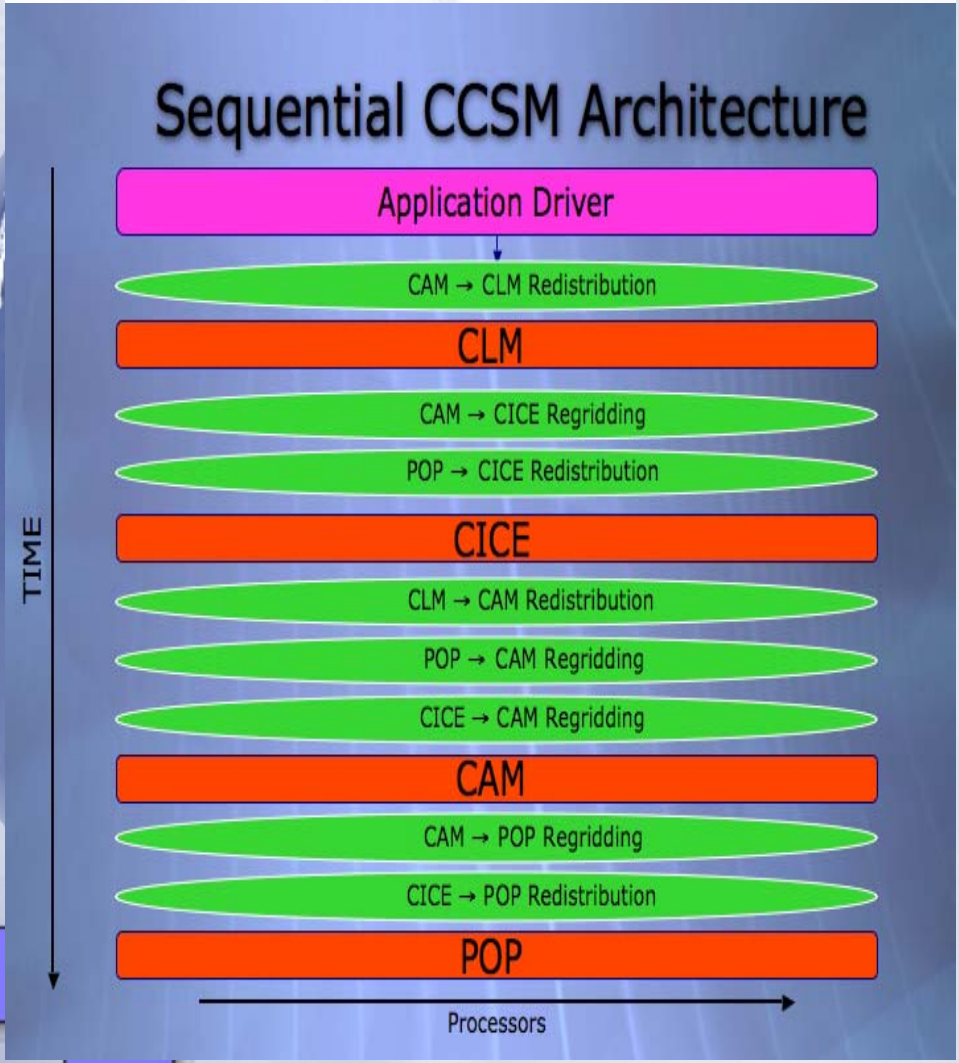
- ✧ Target new flexibility in run time configuration of system
 - ✧ Mixed sequential/concurrent component sequencing
 - ✧ Make it easier to add new components vs CCSM3
- ✧ Target unification of CCSM3 stand-alone models
 - ✧ Use CCSM4 architecture to replace former stand-alone component code
- ✧ Target only single executable mode
- ✧ Work of M. Vertenstein, T. Craig, R. Jacob, E. Kluzek, Dennis



CPL6 -> CPL7 Design



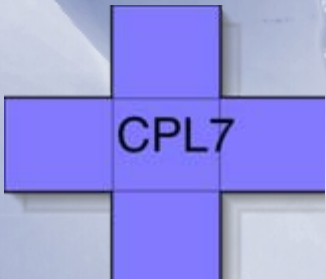
CPL7: More Detailed View

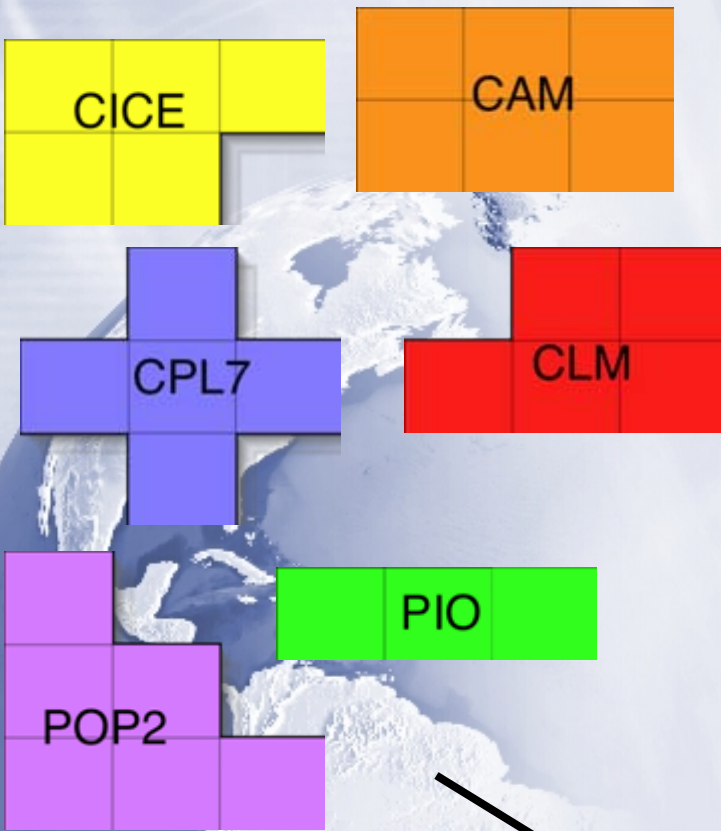


- ✧ Simple elegant design containing
 1. Application Driver
 2. Mappers (green)
 3. Mergers (not shown)
 4. ESMF Clocks (not shown)
- ✧ Eliminates two stage communication
- ✧ Permits possible quasi-local communications
- ✧ Places premium on highly scalable components with low memory usage

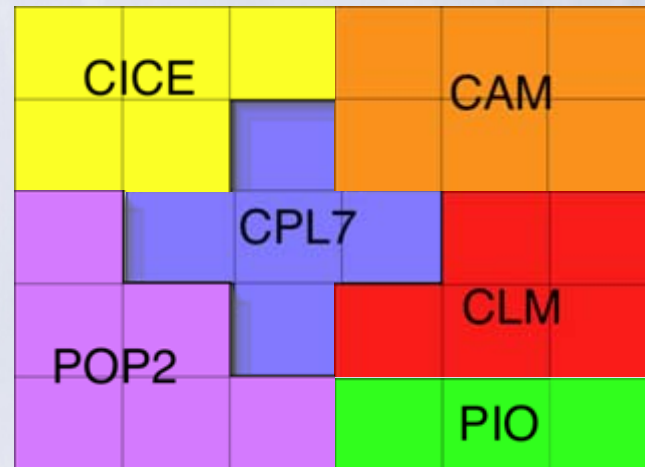
Issues: Coupling Frequency

- ✧ Currently, CAM/CLM/CICE components communicate hourly (and daily with POP)
- ✧ New CAM physics will require coupling on CAM physics time step for CAM/CLM/CICE
- ✧ Higher resolution CAM will also be impacted:
 - ✧ Appears that $.5^\circ$ FV CAM can still couple hourly
 - ✧ Expect that $.25^\circ$ FV CAM will require at least 30 min coupling
- ✧ Increased coupling frequency results in significant cost in cpl6
 - ✧ Migrate to cpl7 system after validation
 - ✧ Use new CICE modifications underway to not invoke dynamics less frequently
- ✧ Albedo computation subtleties are currently being resolved



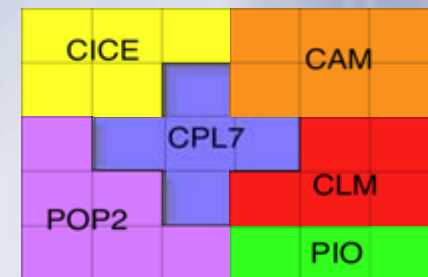


The whole puzzle



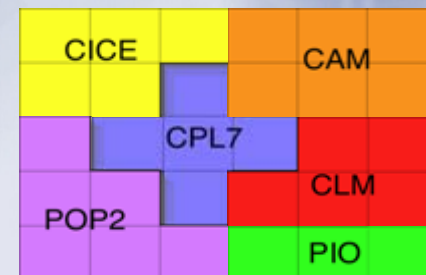
Puzzle Assembly

- ✧ Using CCSM4 alpha tags [B_PRESENT_DAY]
- ✧ Res: 1.9x2.5_gx1v5 on Frost [3 hours]
 - ✧ NCAR system
- ✧ Res: 1.9x2.5_gx1v5 on NYBlue [2 days]
 - ✧ BNL/SUNY system
 - ✧ Just released to users
- ✧ Res: 0.47x0.63_tx0.1v2 on NYBlue [2 weeks]
 - ✧ Generation of new datasets
 - ✧ Rediscovery of 'old' bugs
 - ✧ Some difficulty with access to 480 & 960 processors



Status of 0.47x0.63_tx0.1v2

- ✧ run01:
 - ✧ POP dt_count = 1920, 1 day --> wrote restart
- ✧ run02:
 - ✧ spunup_start from run01 restartfile
 - ✧ POP dt_count = 800, 1 day -> wrote restart
- ✧ run03:
 - ✧ spunup_start from run02 restartfile
 - ✧ POP dt_count = 400, 27 days [vertical CFL violation]
 - ✧ Zero'd POP output
- ✧ run04:
 - ✧ Spunup_start from run02 restartfile
 - ✧ POP dt_count = 480, 10.8 days [SEGV in CICE]

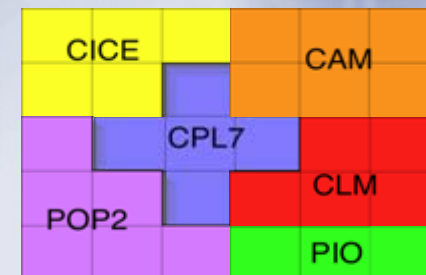


Changes `alpha08' TAG

✧ CCSM POP2

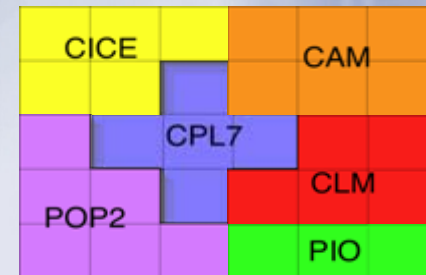
✧ Source code changes

- ✧ No velocity diagnostics
- ✧ Altered CCSM output file-naming conventions
- ✧ Changes to region ids
- ✧ No BSF
- ✧ No history files



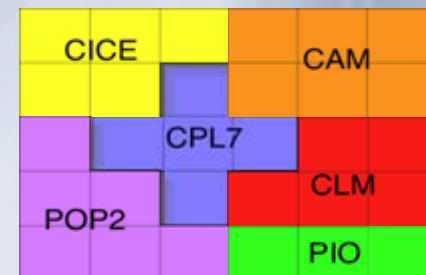
Changes to `alpha08' TAG (con't)

- ✧ CCSM POP2 (con't)
 - ✧ Namelist/input parameter changes
 - ✧ uses PBC
 - ✧ h_momentum_choice='del4', + parameters
 - ✧ ldiag_global_tracer_budgets=.false.
 - ✧ diag_gm_bolus = .false.
 - ✧ qsw_diurnal_cycle = .false./coupled_freq=6 hours
 - ✧ sw_absorption_type = 'top-layer'
 - ✧ ms_balance = .false.
 - ✧ deleted BSF from tavg_contents



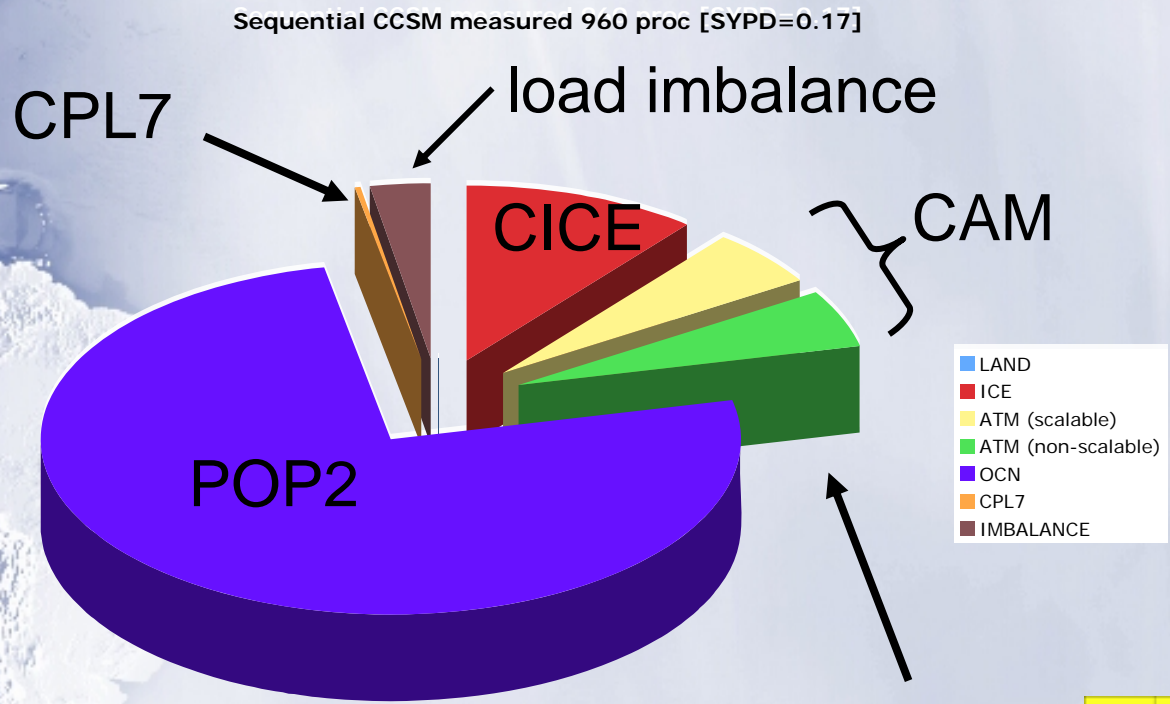
Changes `alpha08' TAG (con't)

- ✧ CAM
 - ✧ pfls=1,000
 - ✧ No history or restart
 - ✧ atm_cpl_dt=1800
- ✧ CLM
 - ✧ None
- ✧ CICE
 - ✧ No history or restart files
 - ✧ ice_cpl_dt=1800
- ✧ CPL7
 - ✧ No river runoff
 - ✧ No history or restart files

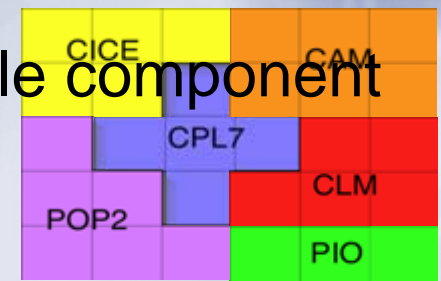




Measured Sequential CCSM on 960 processors [SYPD=0.17]



Non-scalable component

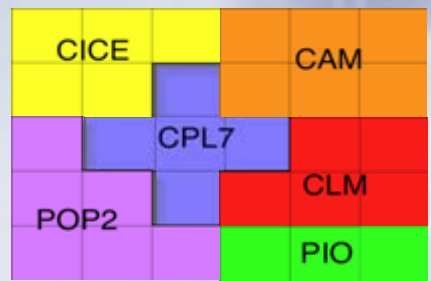
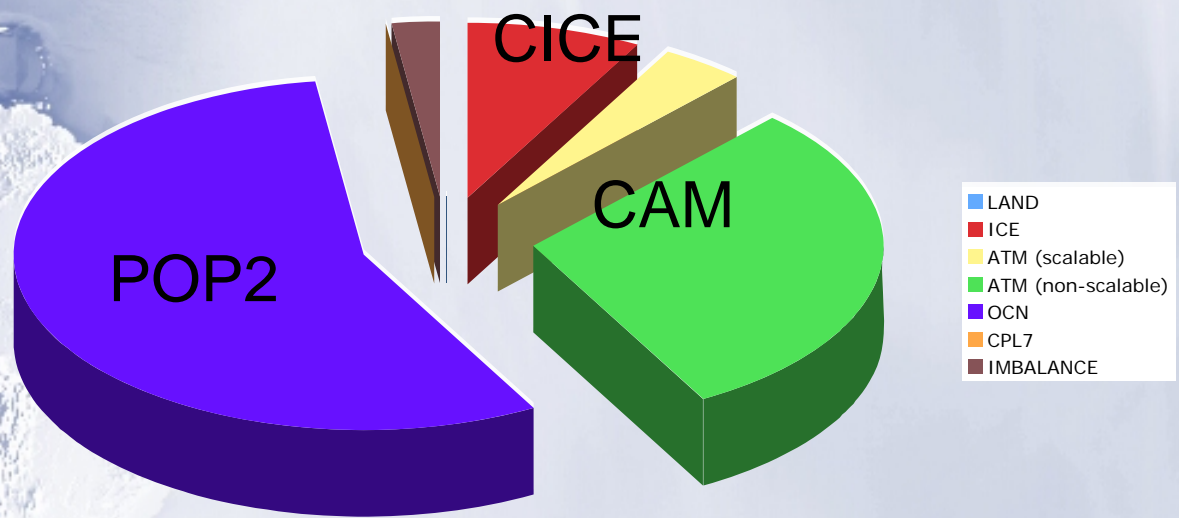


Performance estimations for coupled system

- ✧ Current performance #'s do not include
 - ✧ Space-filling curves
 - ✧ All improvements in CAM
 - ✧ 1D data structure in solver
 - ✧ I/O costs
- ✧ Assume:
 - ✧ MPI only
 - ✧ 90% scalability
- ✧ Most optimistic expectations for coupled system!

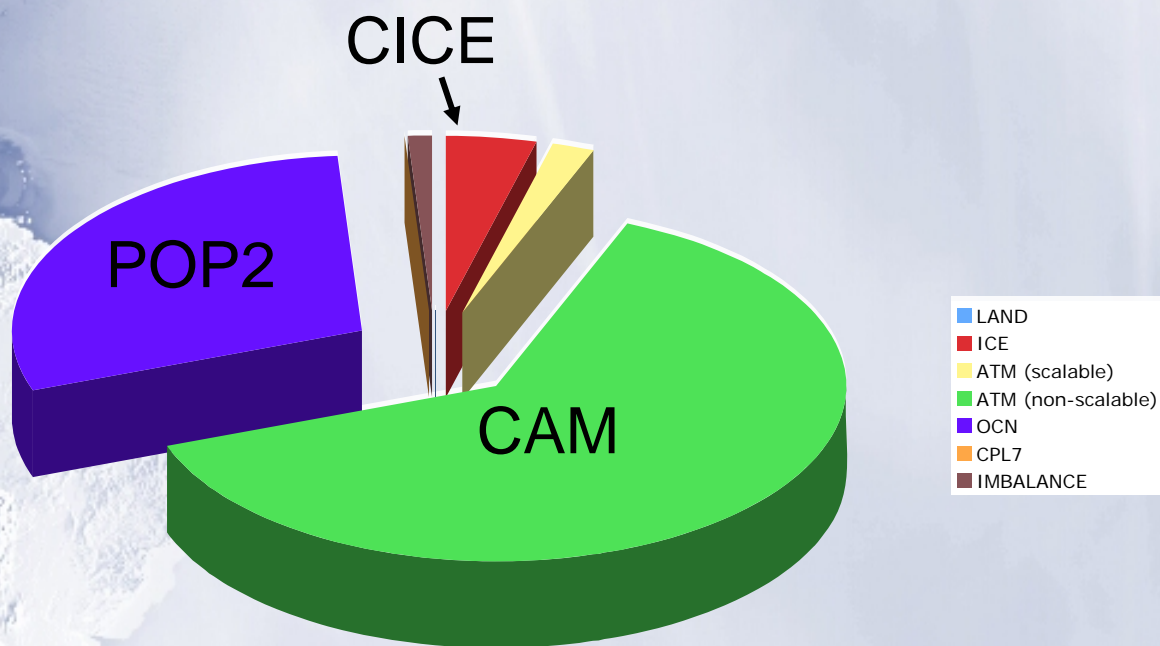
Estimate Sequential CCSM on 8192 processors [SYPD=0.72]

SEQUENTIAL CCSM estimated 8192 processors [SYPD=0.72]

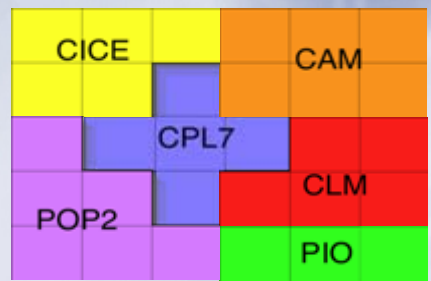


Estimate Sequential CCSM on 32K processors [SYPD=1.54]

Sequential CCSM 32K Estimate [SYPD=1.54]



Non-scalable CAM dynamical core

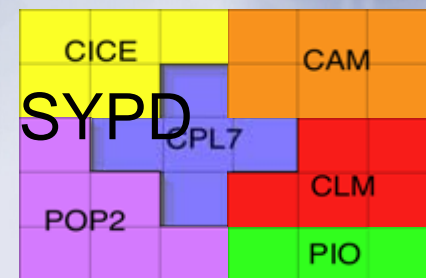


SYPD for 0.47x0.63_tx0.1v2 on Blue Gene

Processors	Sequential	Hybrid/Concurrent
960	0.17	
8192	0.72*	
16384	1.11*	1.15*
32768	1.54*	1.81*

*Will never exceed estimates

XT4/5 gives 4x Blue Gene simulation rate --> 7.24 SYPD



Conclusions

- ✧ Ultra-high resolution coupled system runs on Blue Gene! [10-27 days]
 - ✧ Significant changes to CCSM3 & CCSM3.5 components
 - ✧ Minor modifications to CCSM4
 - ✧ 480 & 960 processors
 - ✧ Only needs 512 Mbytes/processor --> Use all Blue Gene/P processors
 - ✧ POP generates restart files otherwise no I/O
- ✧ Expected simulation rate:
 - ✧ Blue Gene: ~1 SYPD
 - ✧ XT4+: ~5 SYPD
 - ✧ Limited by non-scalable CAM dynamical core

Conclusions (con't)

✧ Parallel I/O (PIO)

- ✧ For INCITE work, parallel I/O will need to be added to each component
- ✧ Prototype implementations for CAM, CLM and POP2

Conclusions (con't)

- ✧ Issue: OpenMP
 - ✧ Limitations of CRAY XT4(?) and BG/P system will require all components to run either in pure MPI or hybrid mode (no heterogeneous capability).
 - ✧ OpenMP already in CAM/CLM
 - ✧ OpenMP needs to be put in CICE for future performance flexibility
 - ✧ OpenMP needs to be tested in POP2

Conclusions (con't)

- ✧ Issues: Regridding Maps
 - ✧ Need to create runoff mapping file for $.5^\circ$ land \rightarrow $.1^\circ$ ocean.
 - ✧ Experienced other difficulties with creating SCRIP mapping weights as part of this process.

Acknowledgements/Questions?

✧ Thanks to:

D. Bailey (NCAR)
F. Bryan (NCAR)
T. Craig (NCAR)
J. Edwards (IBM)
B. Fox-Kemper (MIT,CU)
E. Hunke (LANL)
B. Kadlec (CU)
D. Ivanova (LLNL)
E. Jedlicka (ANL)
E. Jessup (CU)
R. Jacob (ANL)
P. Jones (LANL)
S. Peacock (NCAR)
K. Lindsay (NCAR)
W. Lipscomb (LANL)
R. Loy (ANL)
A. Mirin (LLNL)
M. Maltrud (LANL)
J. McClean (LLNL)
T. Qian (NCAR)
M. Taylor (SNL)
H. Tufo (NCAR)
P. Worley (ORNL)
M. Zhang (SUNY)

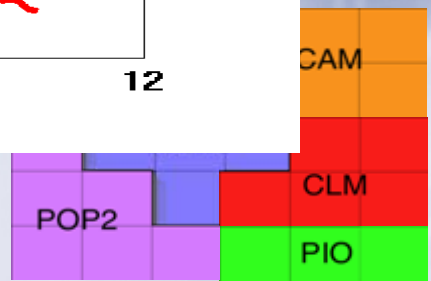
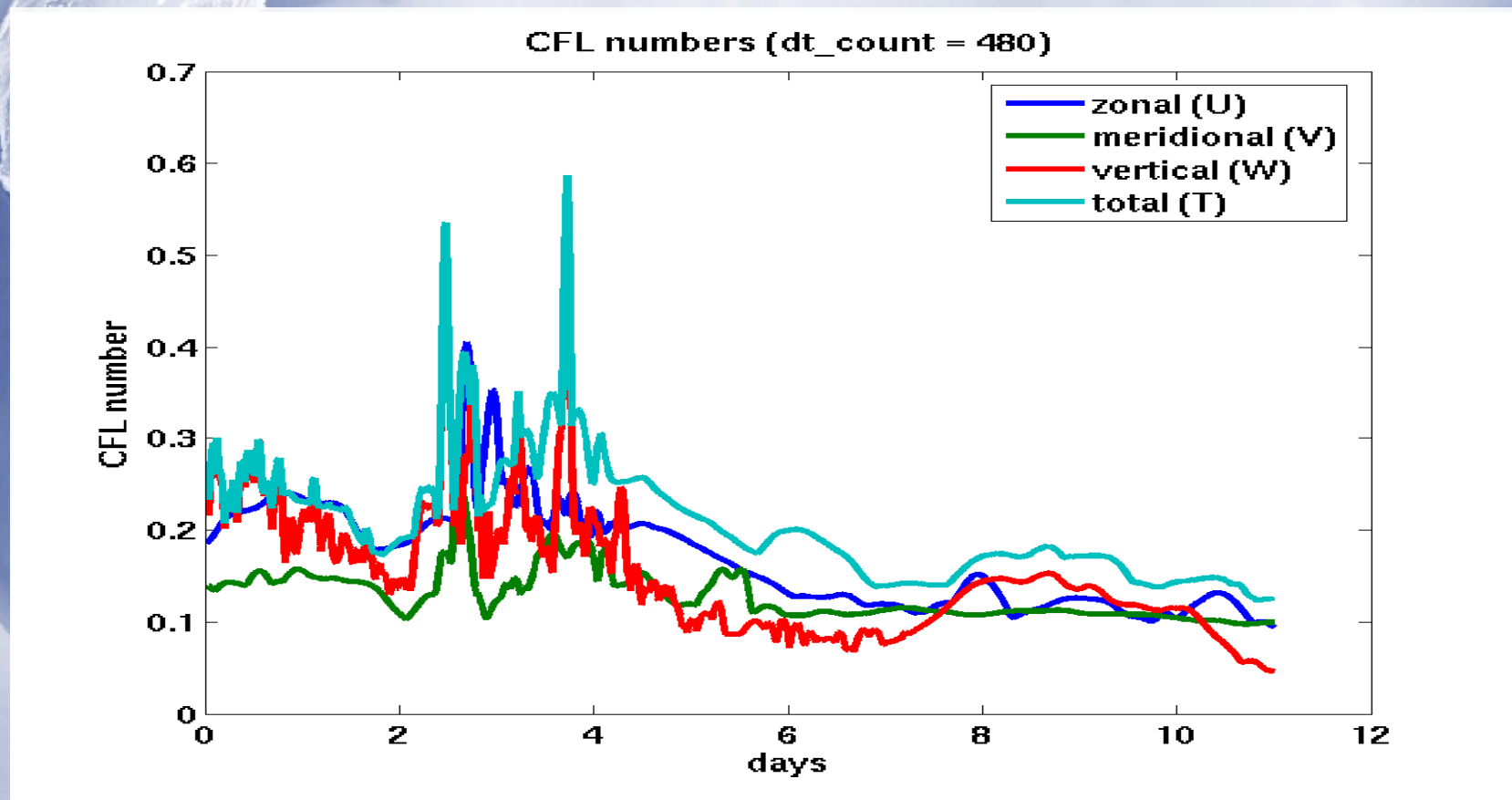
✧ Funding:

- ✧ DOE-BER CCPP Program Grant
 - ✧ DE-FC03-97ER62402
 - ✧ DE-PS02-07ER07-06
 - ✧ DE-FC02-07ER64340
 - ✧ B&R KP1206000
- ✧ DOE-ASCR
 - ✧ B&R KJ0101030
- ✧ National Science Foundation Cooperative Grant NSF01

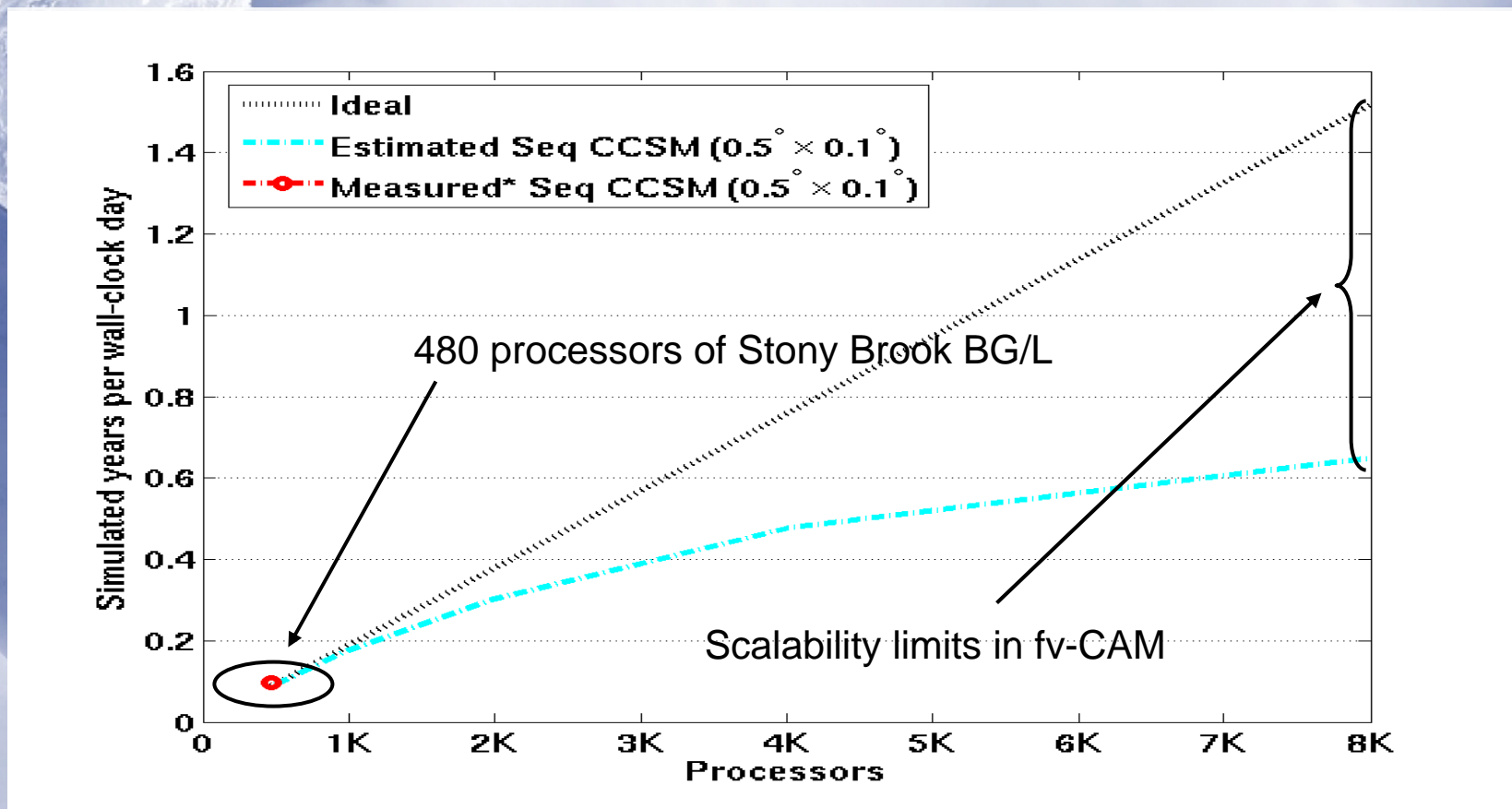
✧ Computer Time:

- ✧ Blue Gene/L time:
 - NSF MRI Grant
 - NCAR
 - University of Colorado
 - IBM (SUR) program
 - BGW Consortium Days
 - IBM research (Watson)
 - LLNL
 - BNL/SUNY
- ✧ CRAY XT3/4 time:
 - ORNL
 - Sandia

POP CFL numbers for 0.47x0.63_tx0.1v2 (dt=480)



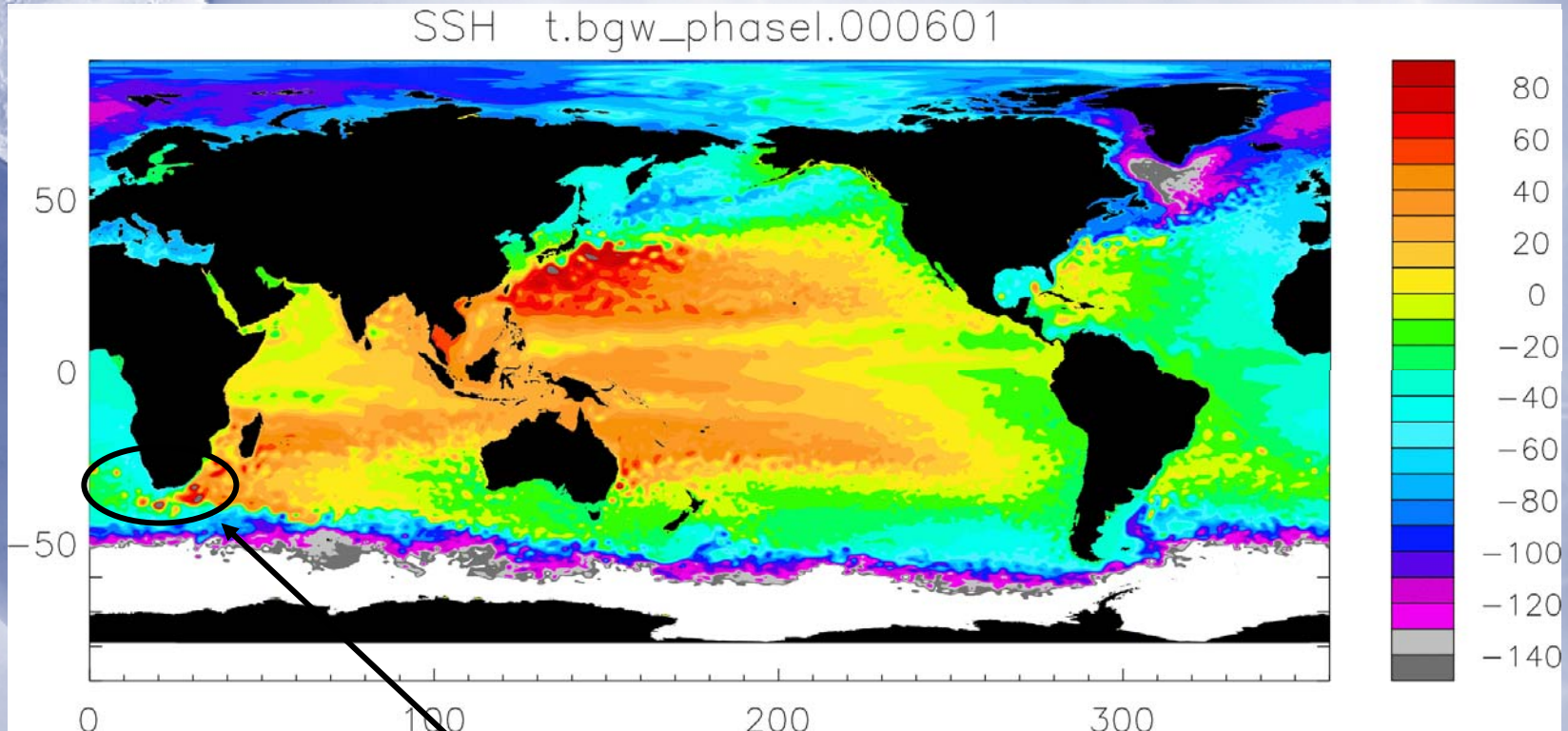
Simulation rate (Blue Gene)





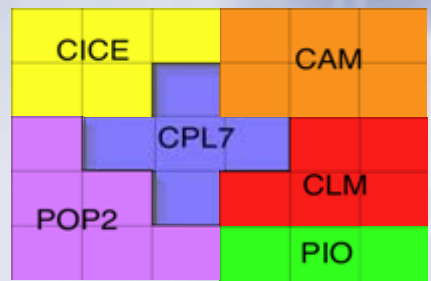
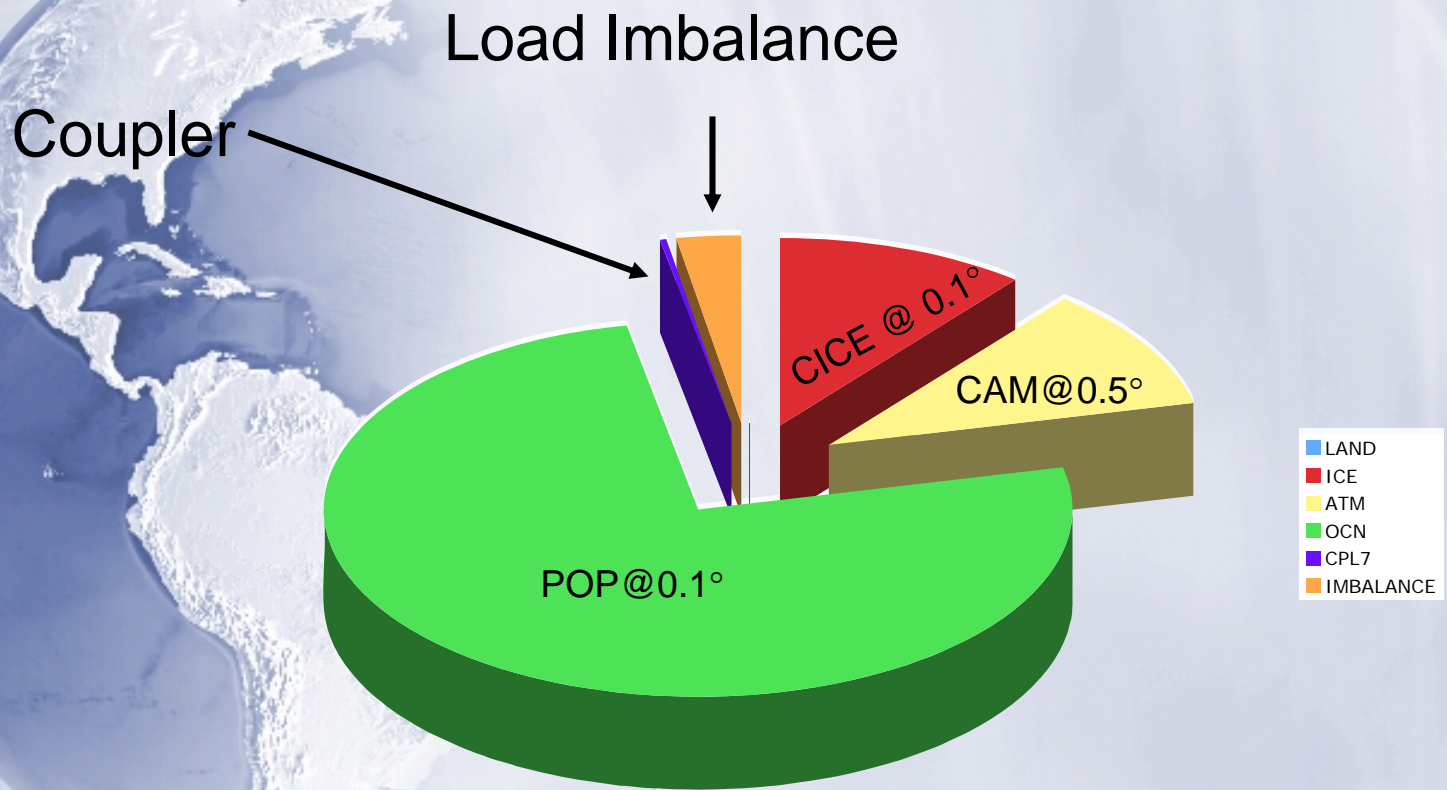
NCAR

Sea-surface height for POP @ 0.1° on Blue Gene Watson



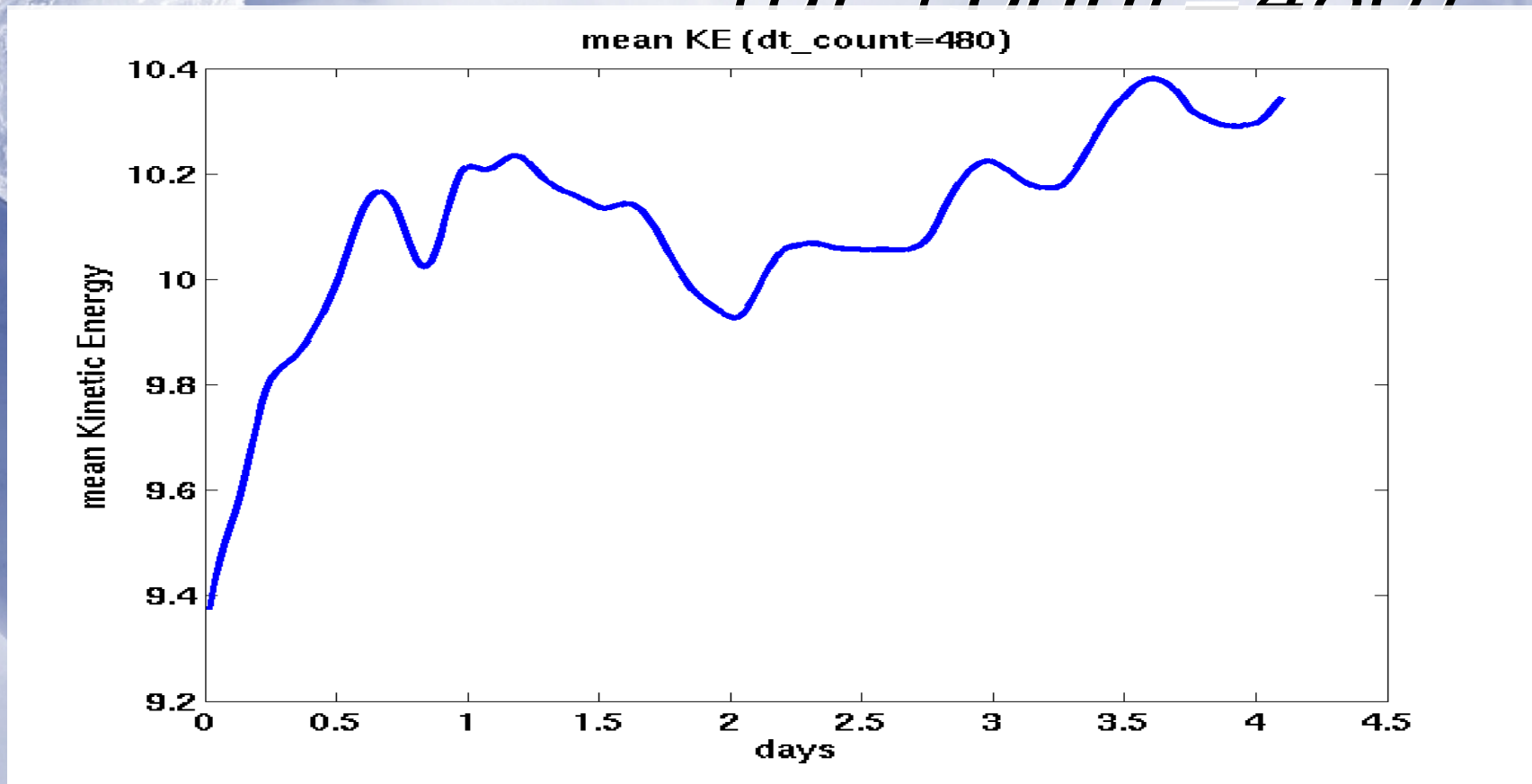
Mesoscale eddies

Ultra-high resolution sequential CCSM on 960 Blue Gene processors (dt_count=200)

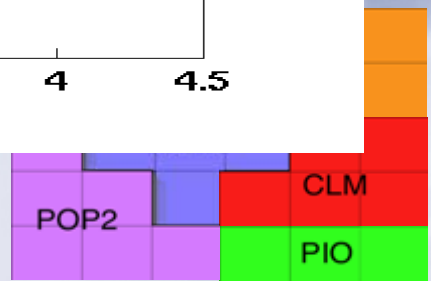
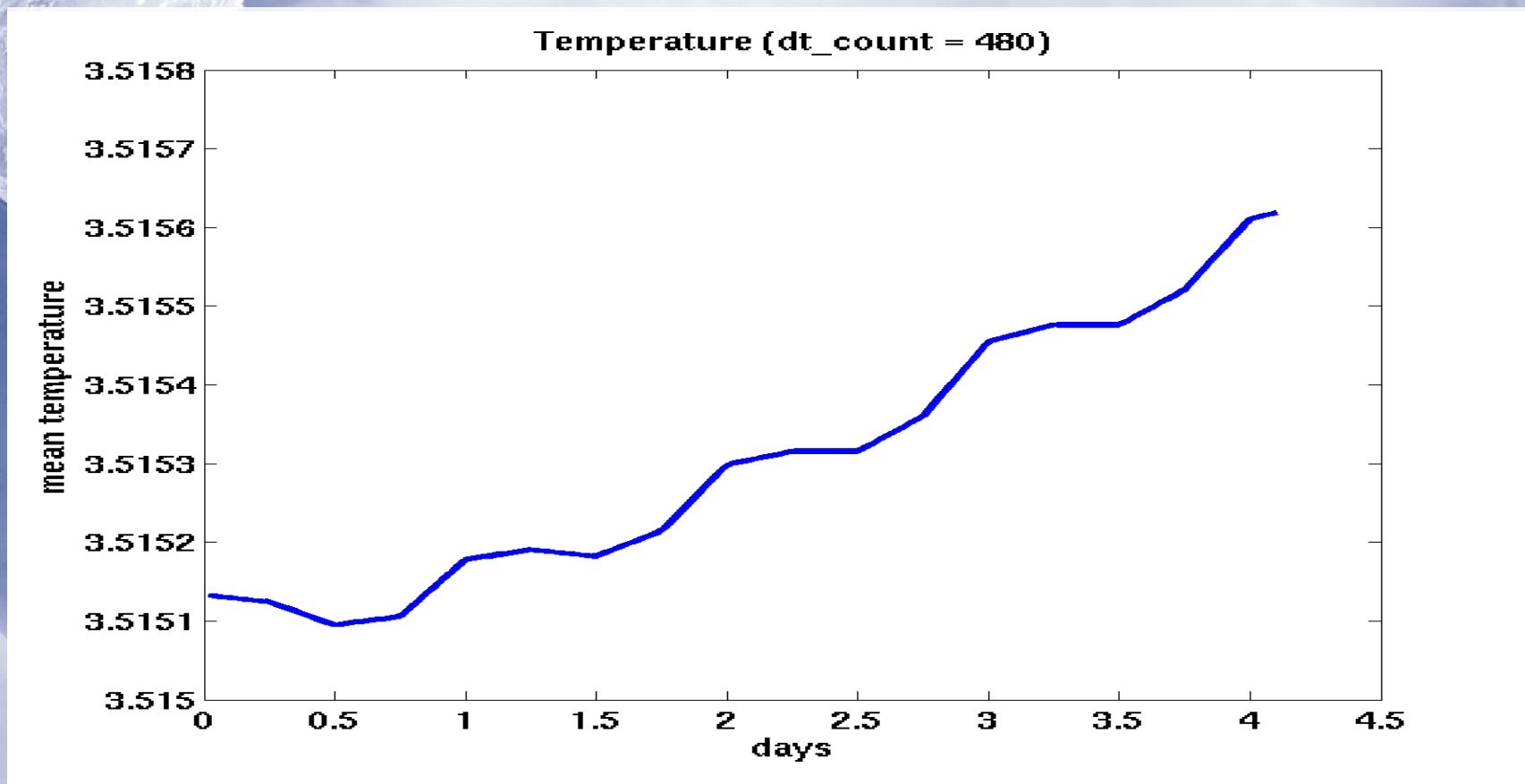


POP mean KE

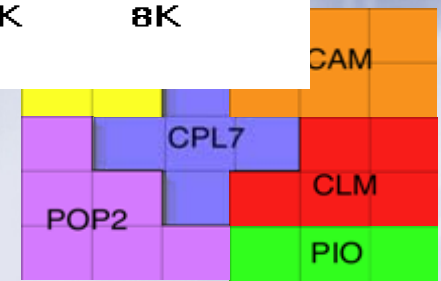
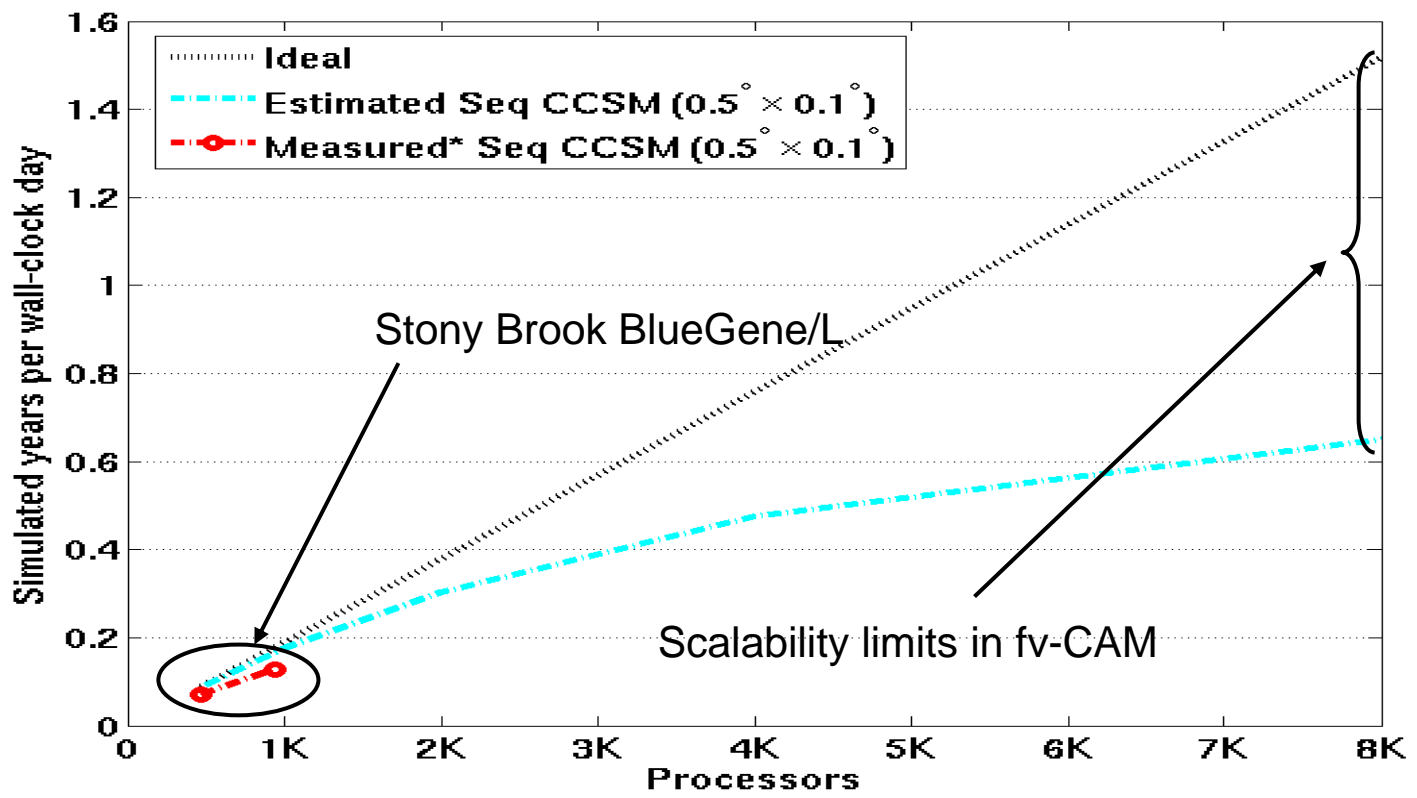
0.47x0.63_tx0.1v2
(dt_count=480)



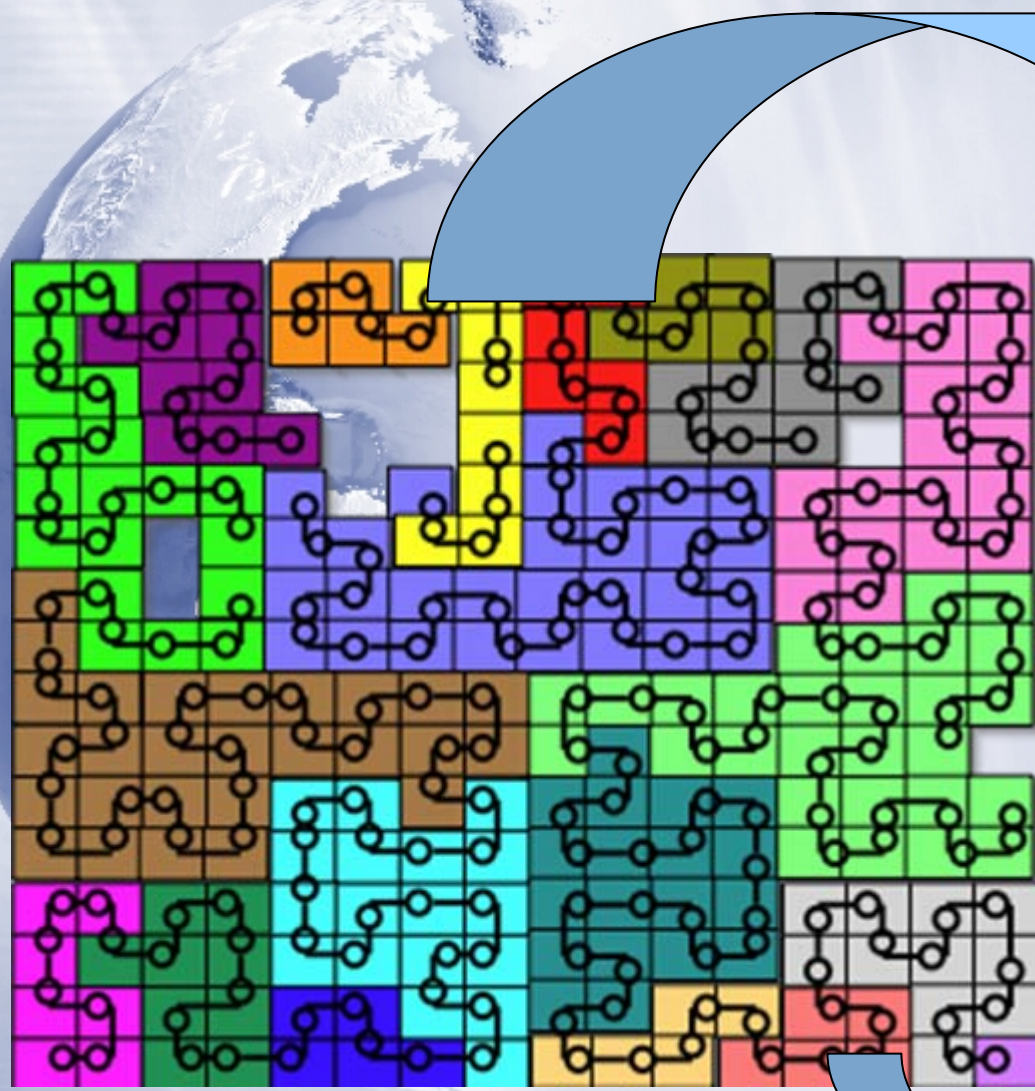
POP mean Temperature *0.47x0.63_tx0.1v2 (dt=480)*



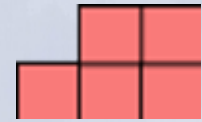
Simulation rate on Blue Gene [dt_count=200]



CICE4 @ 1° on 20 processors

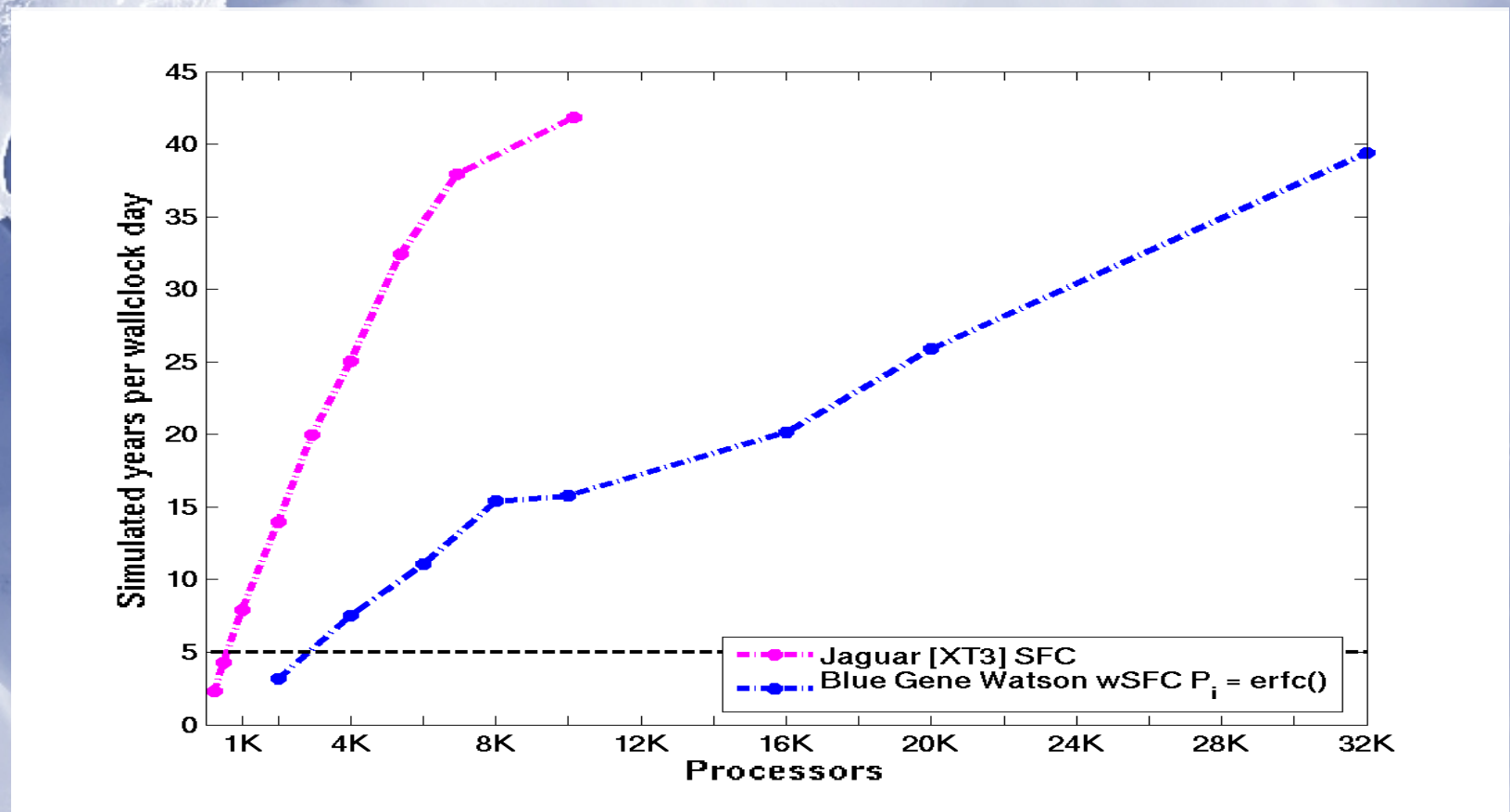


Large domains @ low latitudes



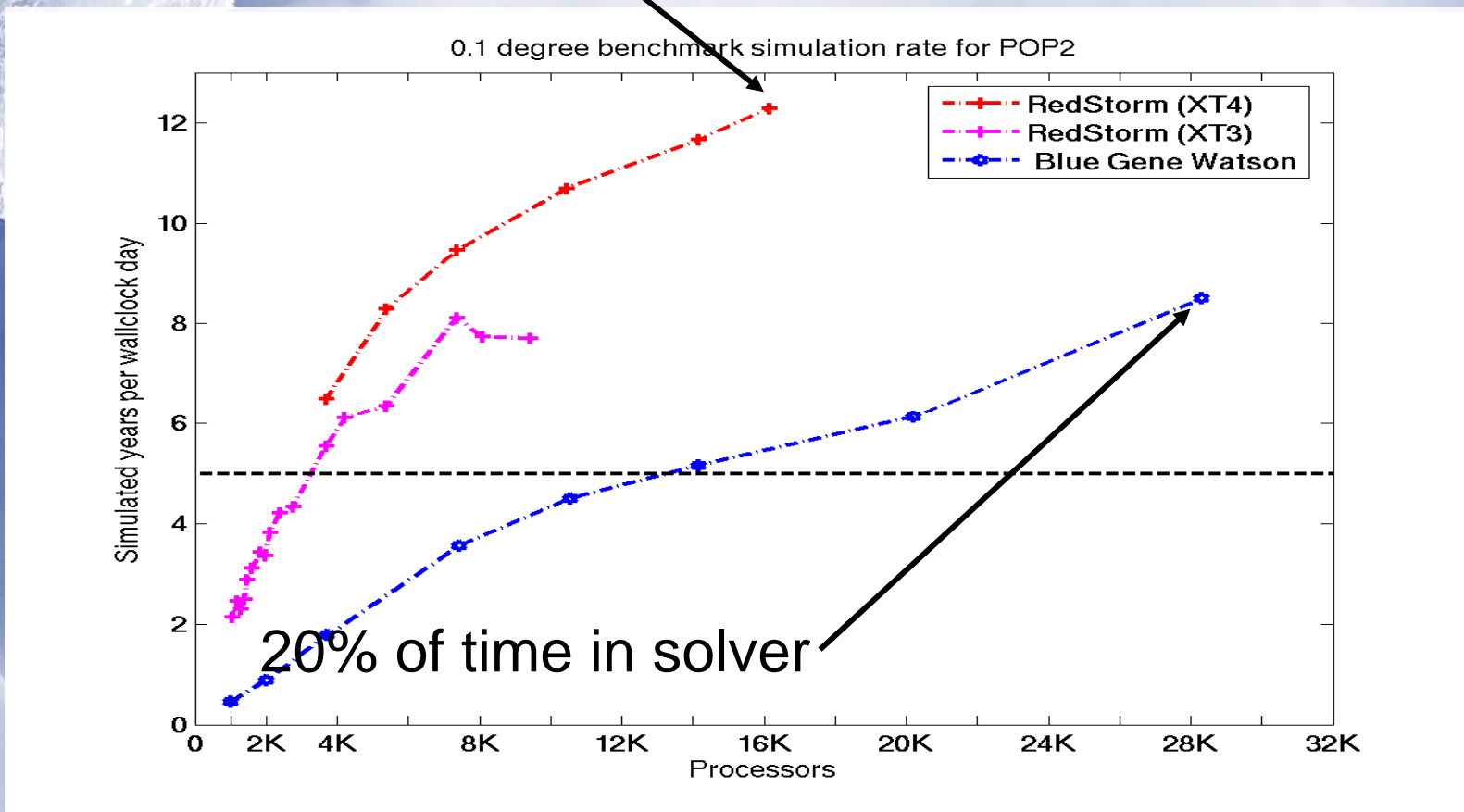
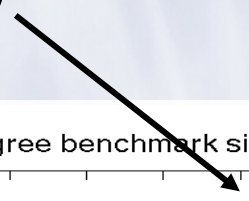
Small domains @ high latitudes

CICE4 @ 0.1°

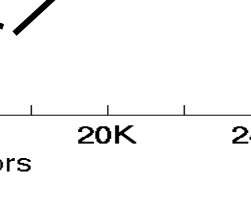


POP2 0.1° benchmark

71% of time in solver



20% of time in solver



Courtesy of M. Taylor