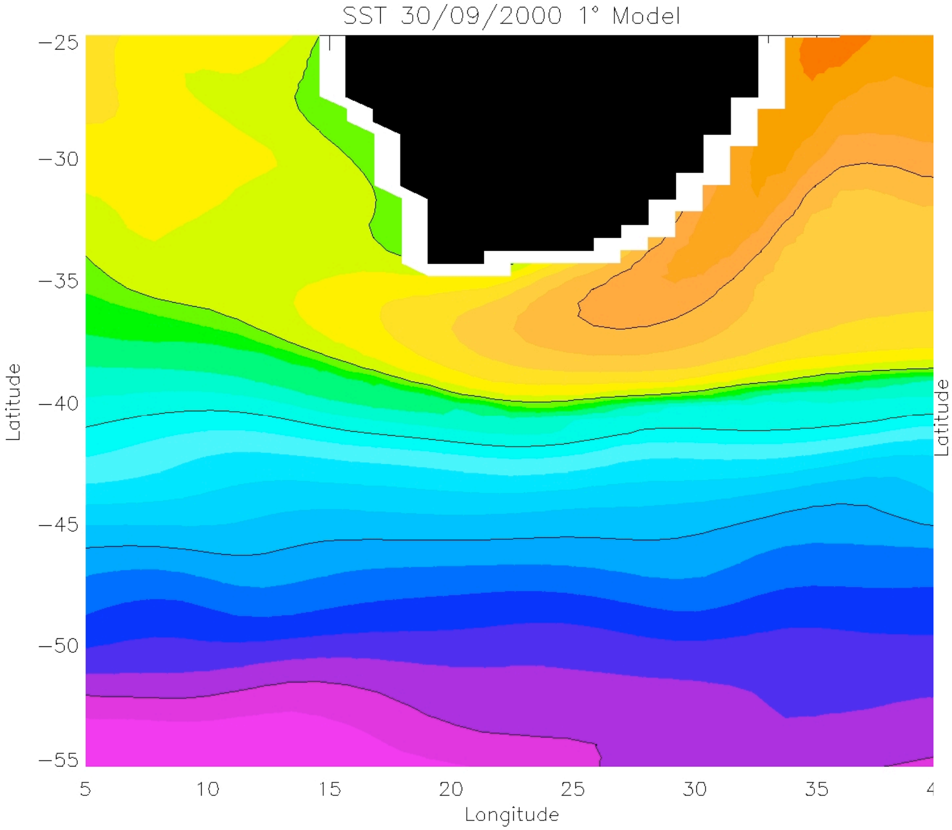
Optimizing High-Resolution Climate Variability Experiments on Cray XT4 and Cray XT5 Systems at NICS and NERSC

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

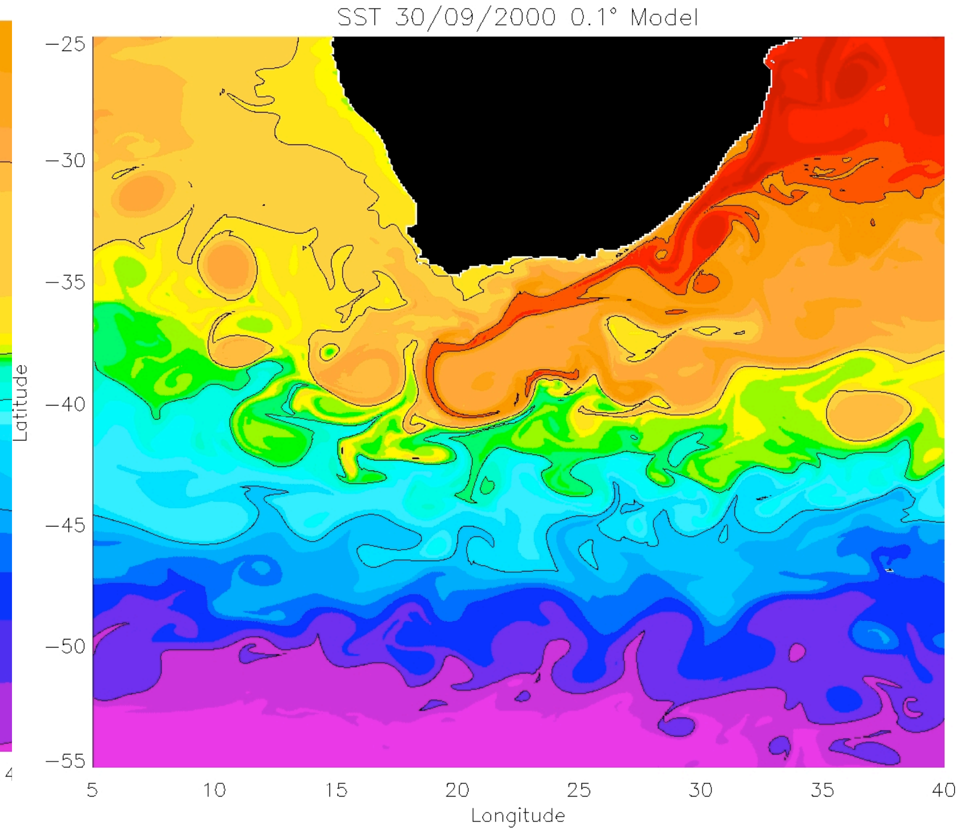
Outline

- Science Motivation
- Computing Systems Used
- CCSM Coupled System Optimization
- Scaling and Efficiency Results

Why High Resolution? Resolving Ocean Mesoscale Eddies



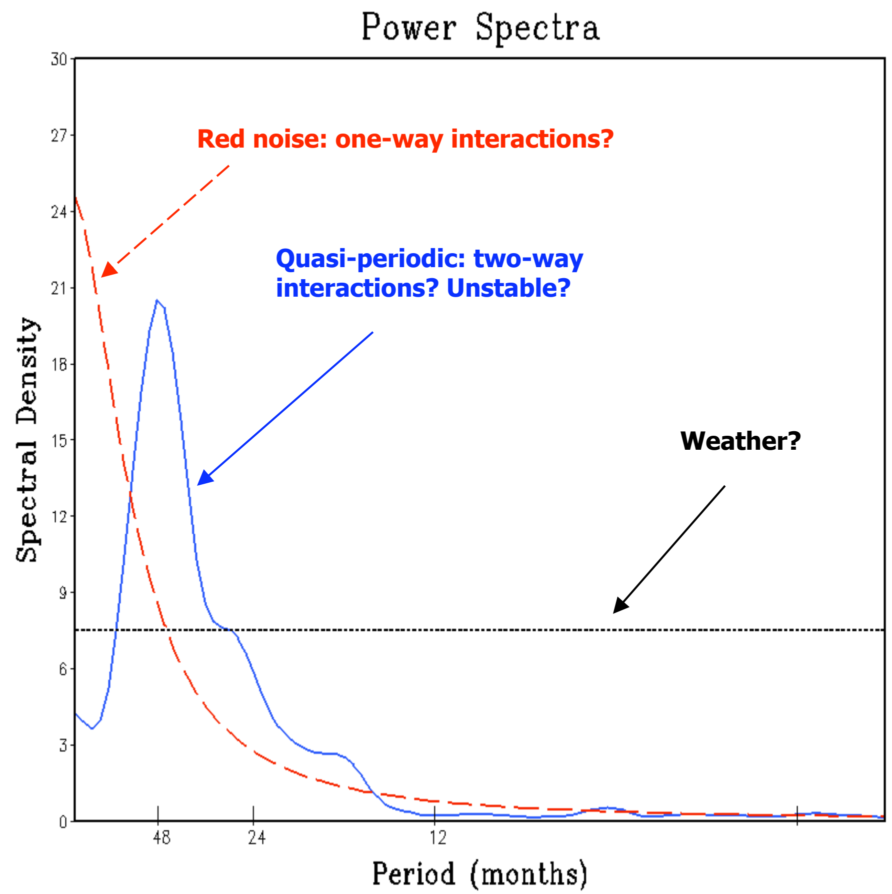
Ocean component of CCSM (Collins et al, 2006)



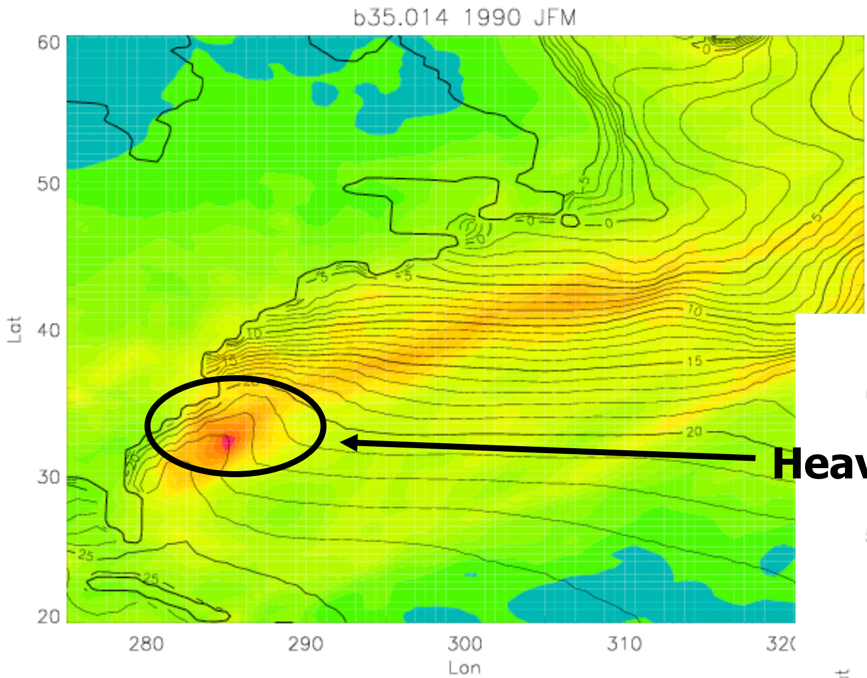
Eddy-resolving POP (Maltrud & McClean,2005)

Understanding Weather-Climate Interactions

- One-way air-sea interactions (stochastic atmosphere, aka weather noise, forces ocean)
  - Ocean as thermodynamic “red filter” -- Hasselmann (1976)
  - Ocean-dynamics: preferred low frequency time scale(s)
- One-way air-sea interactions (stochastic ocean forces atmosphere)
  - Tropical instability waves
  - Kuroshio current extension
- Two-way air-sea interactions
  - (Stable) coupled feedbacks + weather noise (MJO, WWB)
  - (Stable) coupled feedbacks + weather noise + dynamics
  - Unstable coupled feedbacks + weather noise + dynamics



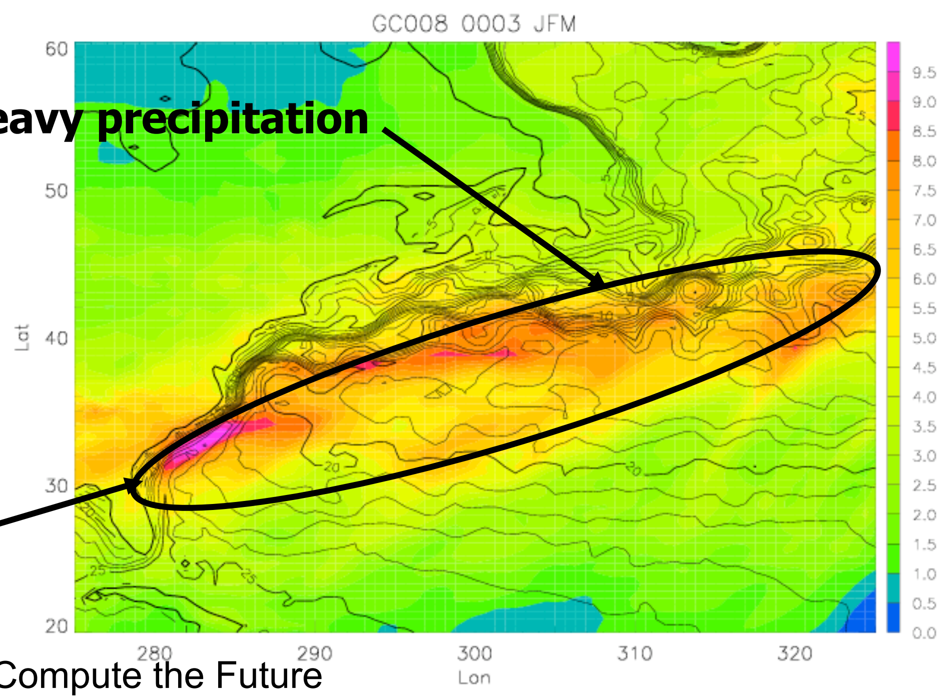
Ocean-Atmosphere Interactions: North Atlantic Winter Storm Track



0.5° atm + 1° ocn

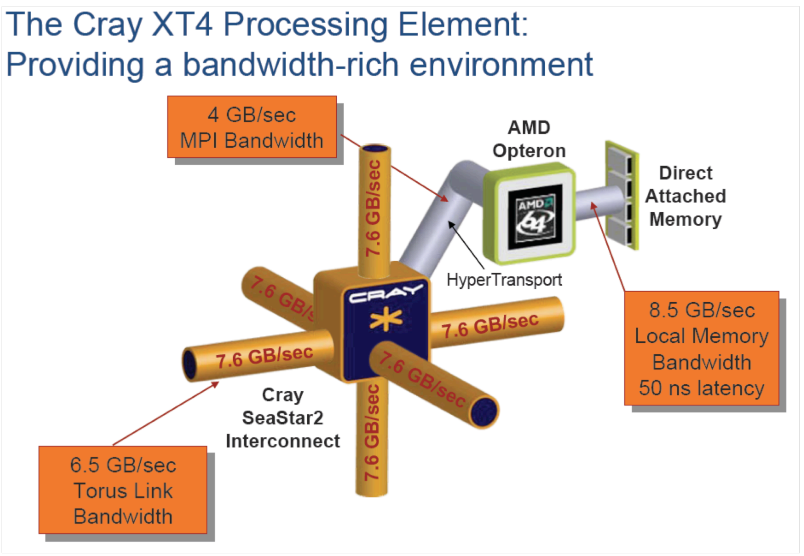
Heavy precipitation

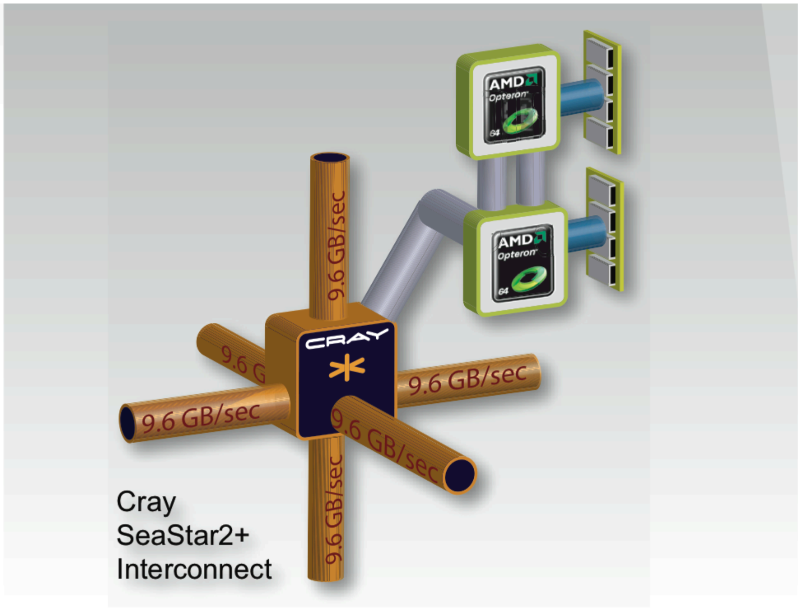
0.5° atm + 0.1° ocn



Strong SST gradient

Cray XT4 & XT5 Architectures





Courtesy of Cray, Inc.

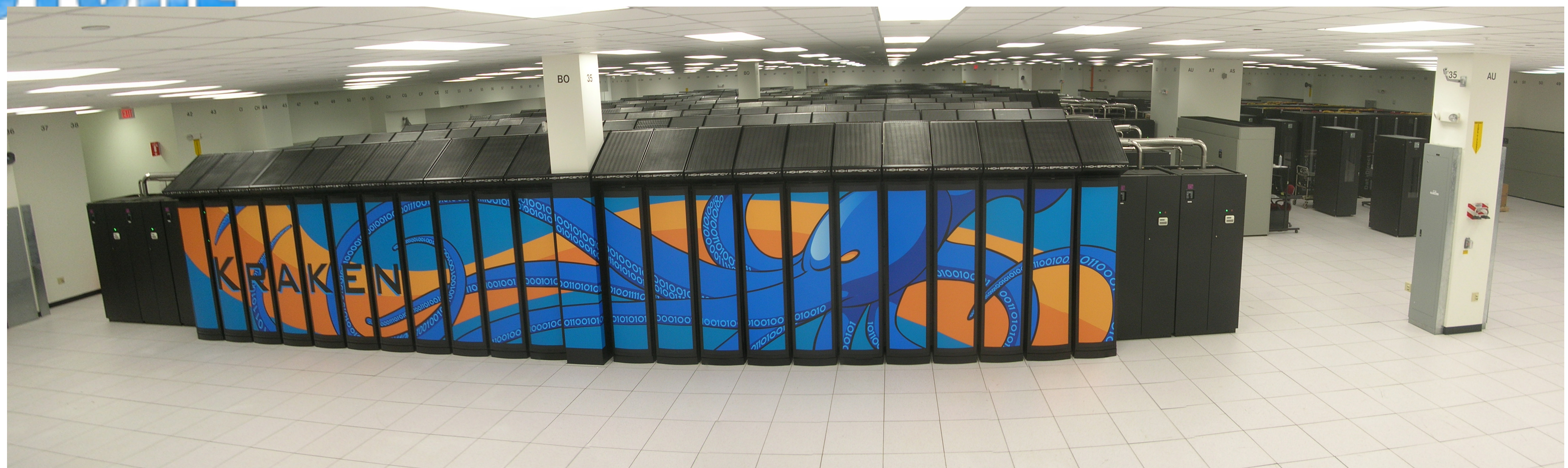
Franklin Cray XT4 at NERSC

Courtesy NERSC

Franklin Cray XT4 at NERSC

- Node:
  - One socket/node
  - AMD Opteron Quad Core 2.3 GHz
  - 8 GB/node (2 GB/core)
- Network:
  - Cray SeaStar2 Router
  - 3D Torus dimensions:(17x24x24)
- Aggregate:
  - Core count: 38,640 (9660 nodes)
  - 356 TFLOPS peak
  - Main Memory: 78 TB

Kraken XT-5 at NICS



Courtesy of Pat Kovatch, NICS

Kraken Cray XT4 at NICS

- Node:
  - One socket/node
  - AMD Opteron Quad Core 2.3 GHz
  - 4 GB/node (1 GB/core)
- Network:
  - Cray SeaStar2 Router
  - 3D Torus dimensions: (12x16x24)
- Aggregate:
  - Core count: 18,048 (4,512 nodes)
  - 166 TFLOPS peak
  - Main Memory: 18 TB

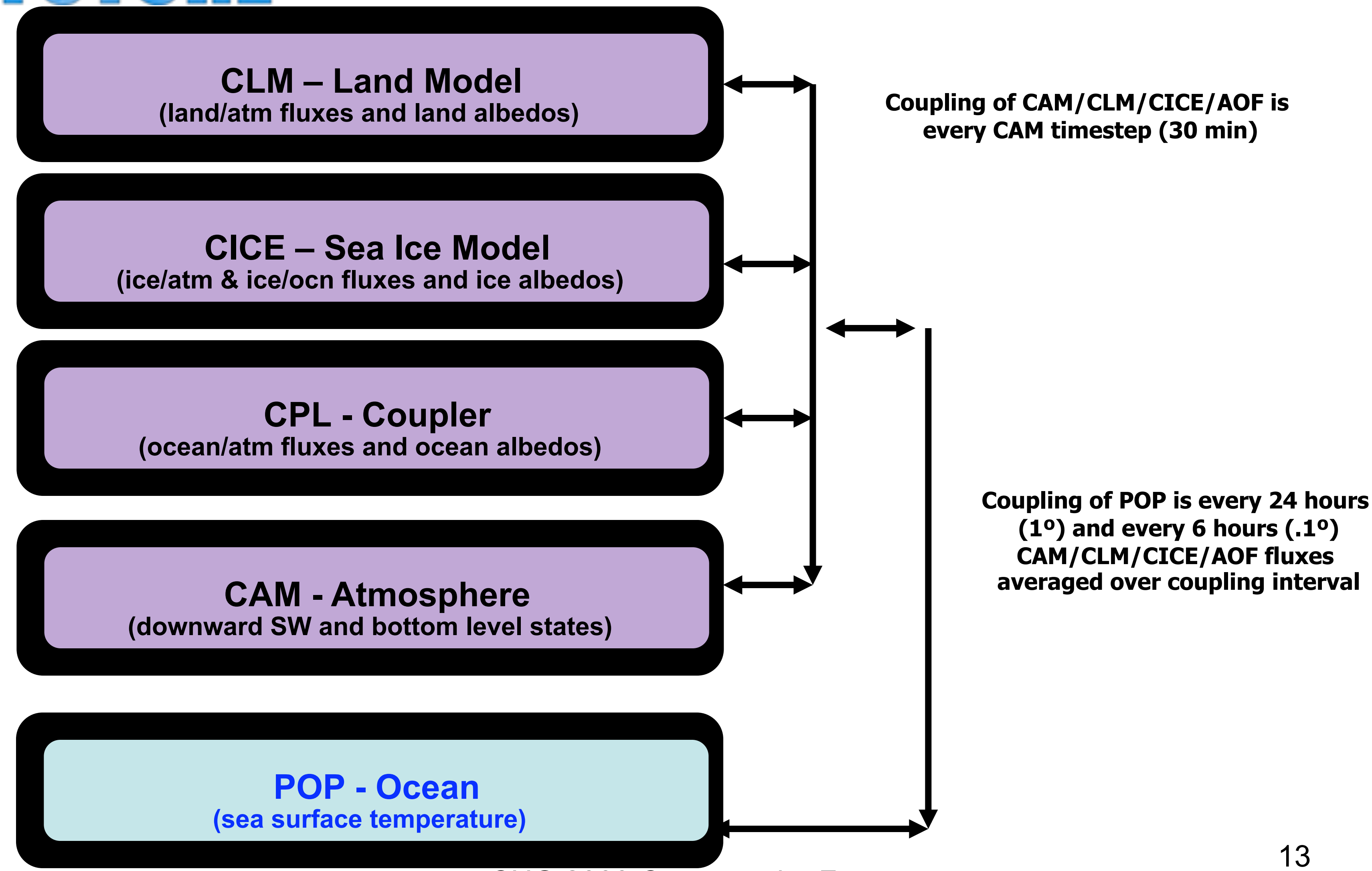
Kraken Cray XT5 at NICS

- Node:
  - Two sockets/node
  - AMD Opteron Quad Core 2.3 GHz
  - Memory:
    - 3,840 nodes with 8 GB (1 GB/core)
    - 4,416 nodes with 16 GB (2 GB/core)
- Network:
  - 3D Torus dimensions: (22x16x24)
- Aggregate:
  - Core count: 66,048 (8,256 nodes)
  - 608 TFLOPS peak
  - Main Memory: 100 TB

Community Climate System Model (CCSM)

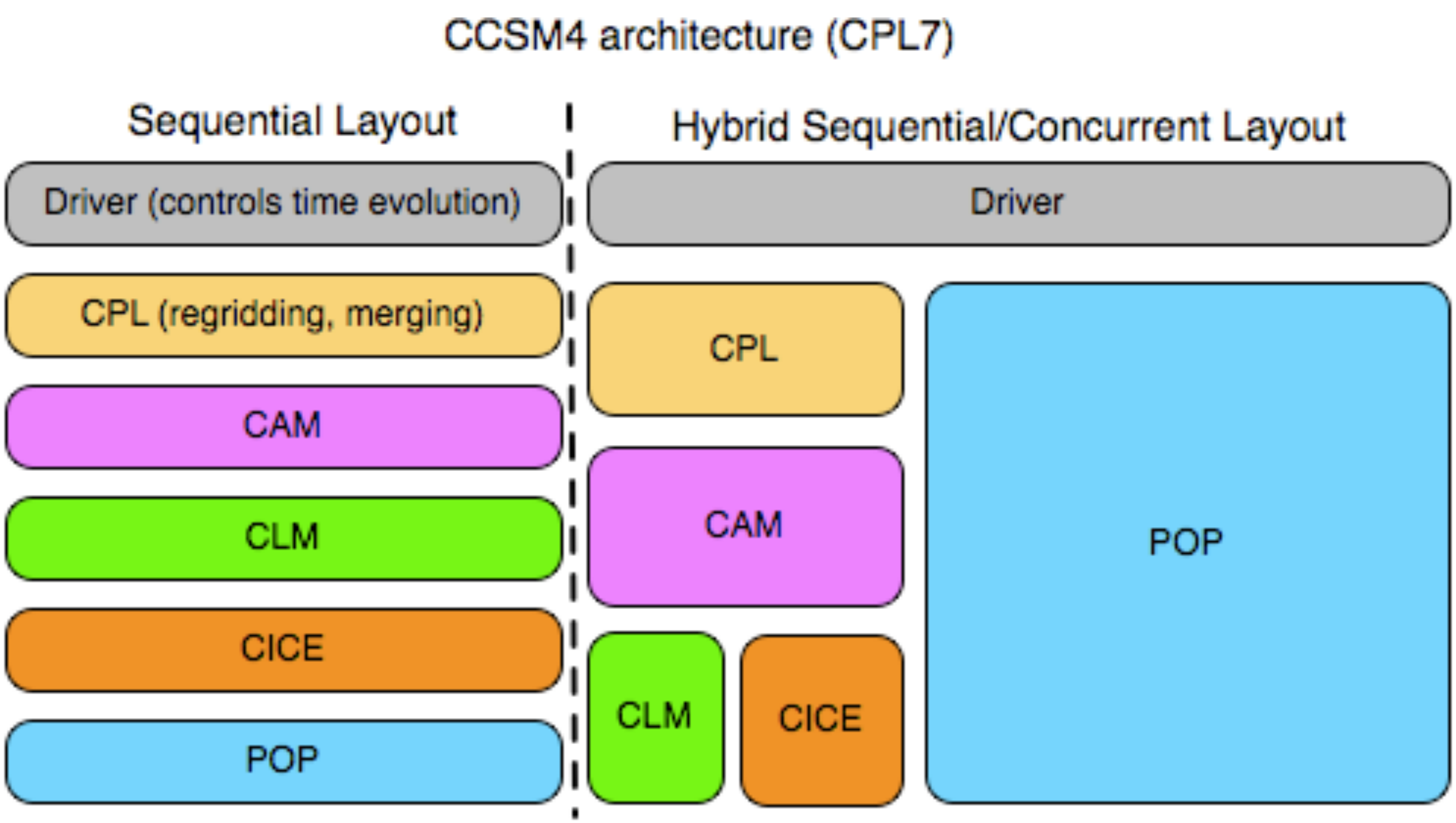
- Multiple component models on different grids
- Flux and state between components [CPL]
- Large code base: >1M lines
  - Developed over 20+ years
  - 200-300K lines are critically important --> no comp kernels, need good compilers
- Demanding on networks:
  - need good message latency + bandwidth

CCSM Coupling and Execution Flow

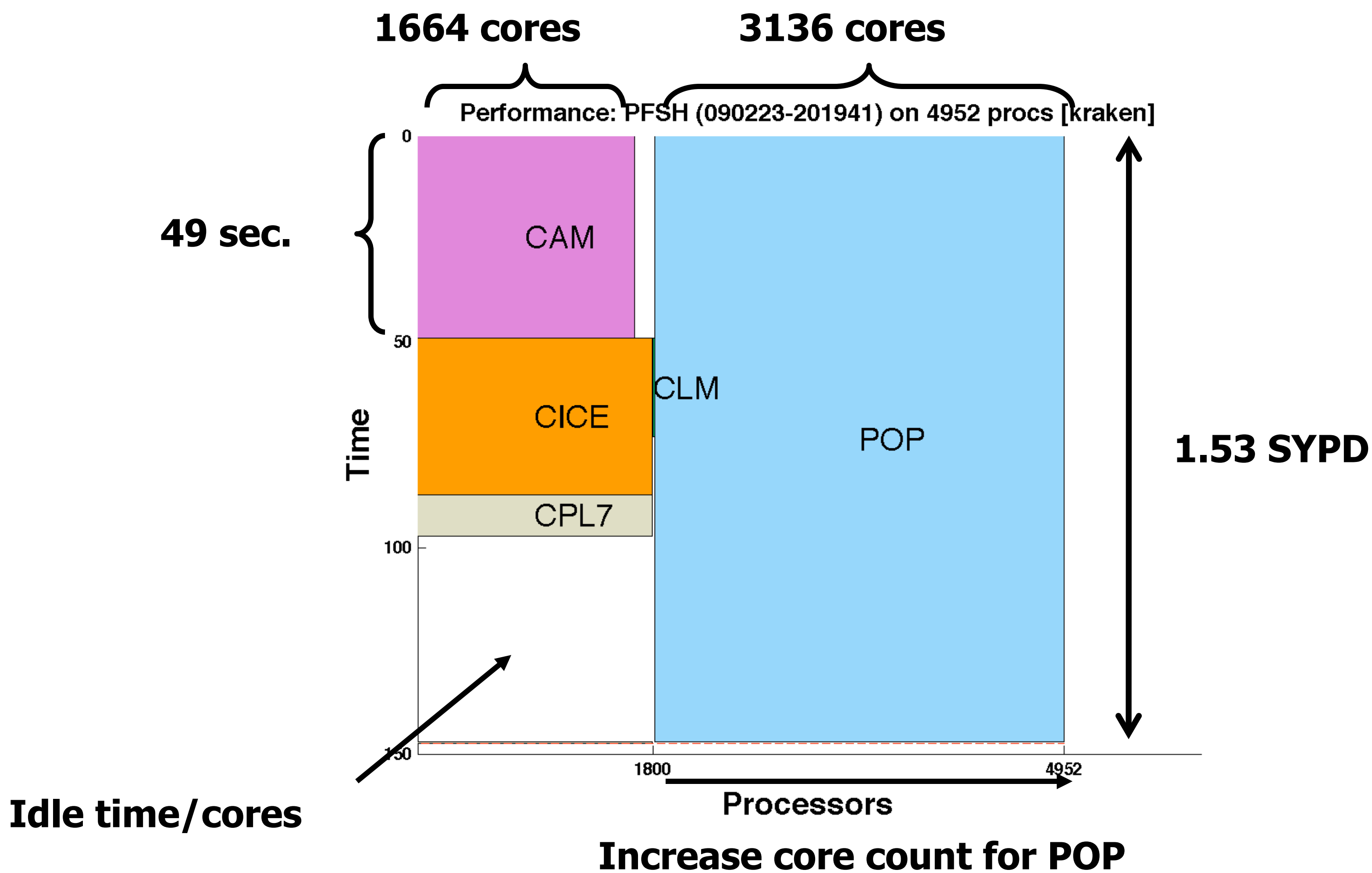


Courtesy Vertenstein, CGD/NCAR

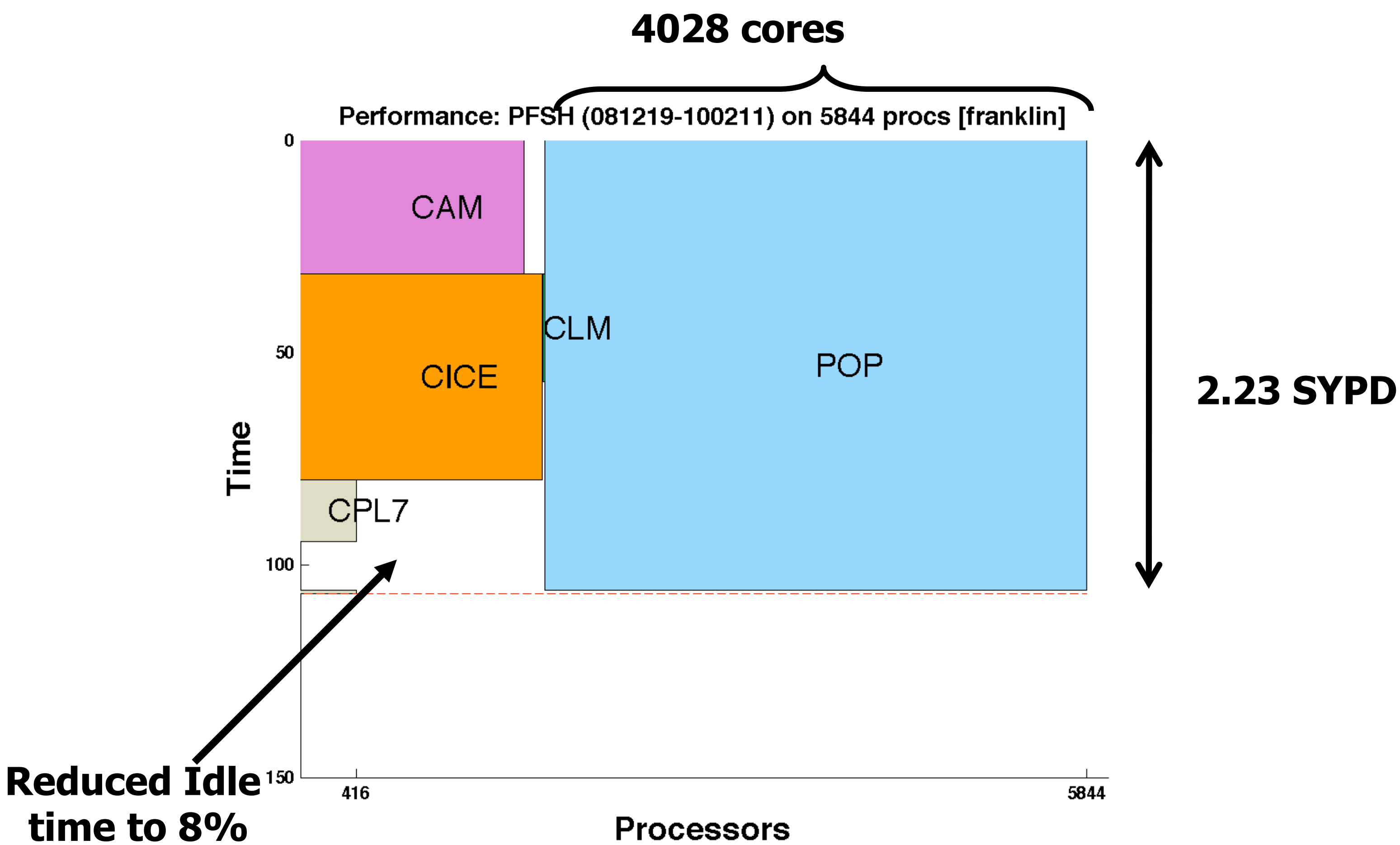
CCSM CPL7 architecture



CCSM4_alpha on 4952 Cores



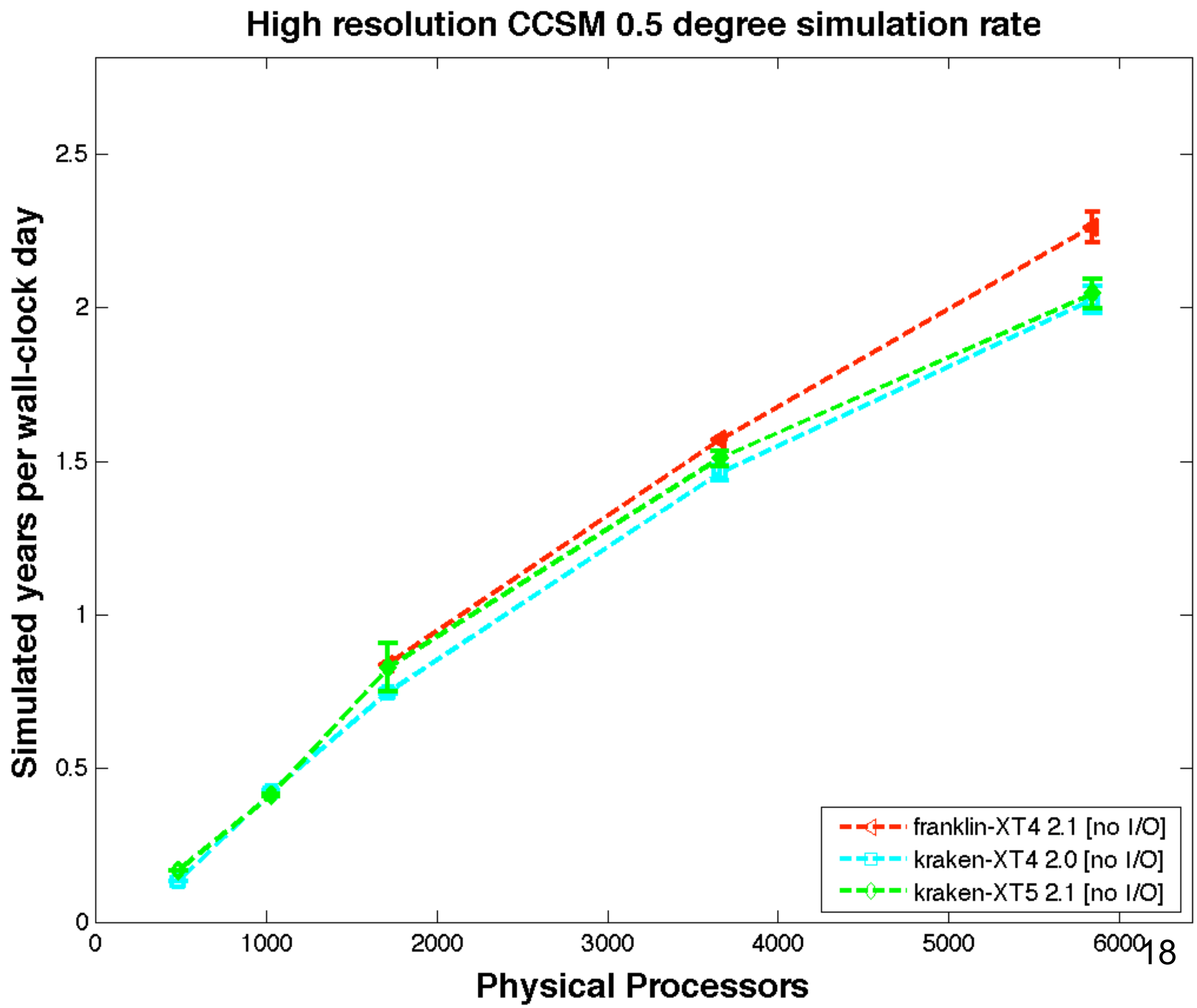
CCSM4_alpha on 5844 Cores



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 Cray XT Scalability (no I/O)



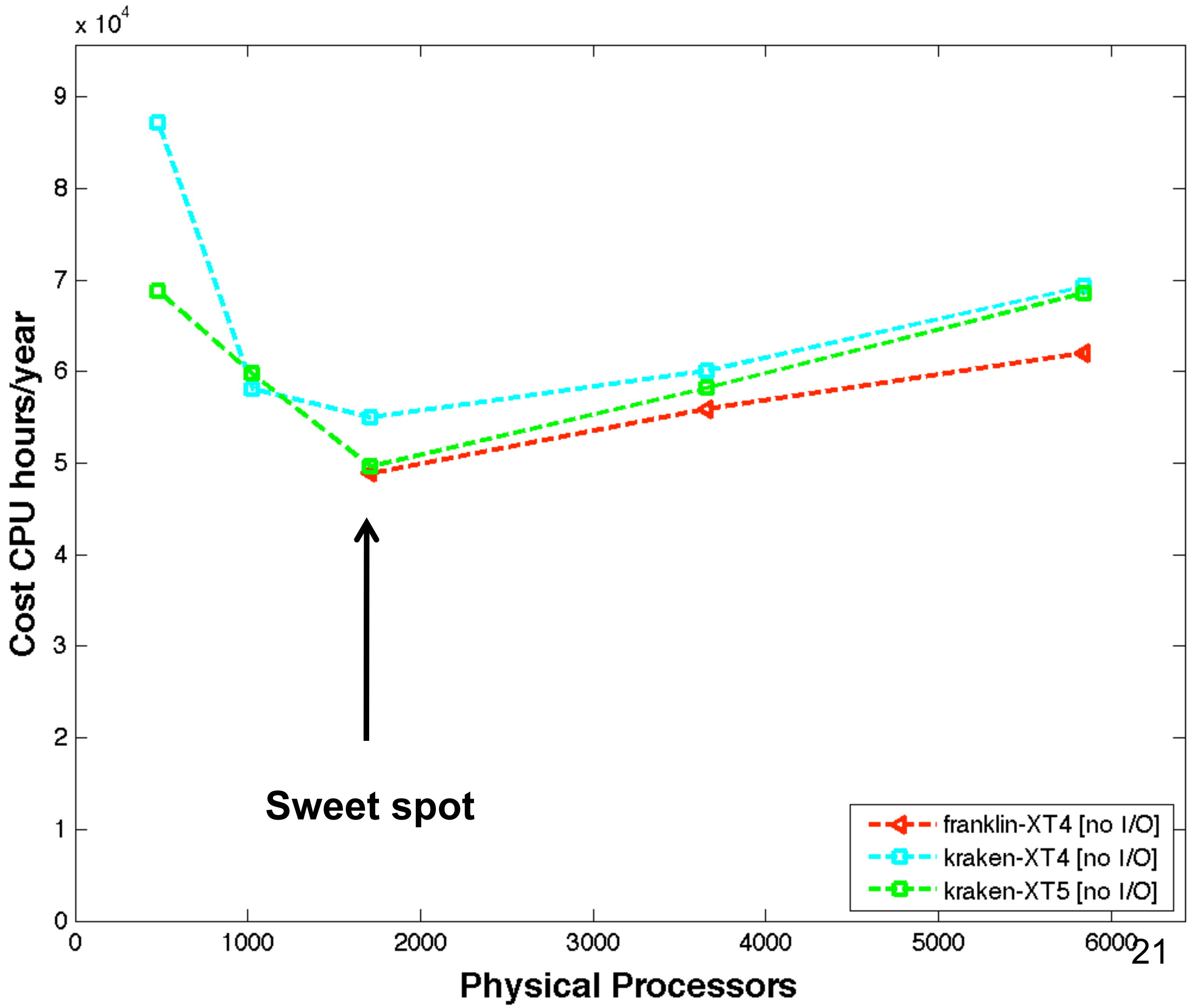
Why the XT4/XT5 Scaling Differences ?

- XT4 Differences
  - Franklin scales better than Kraken
  - Nearly identical systems
  - different OS's (CNL 2.0.62 vs CLE 2.1.56HD)
  - POP highly sensitive to OS jitter (Ferriera and Brightwell)
  - Different levels of kernel level noise between CNL 2.0 and CLE 2.1?

Why the XT4/XT5 Scaling Differences ?

- XT4 – XT5 Differences
  - CCSM scales better on Franklin XT4 than Kraken XT5
  - Apparently identical OS's.
  - Dual socket bandwidth issues?
  - Standalone POP benchmarks seem to rule out node bandwidth issues on XT5.
  - Hardware latency issues?

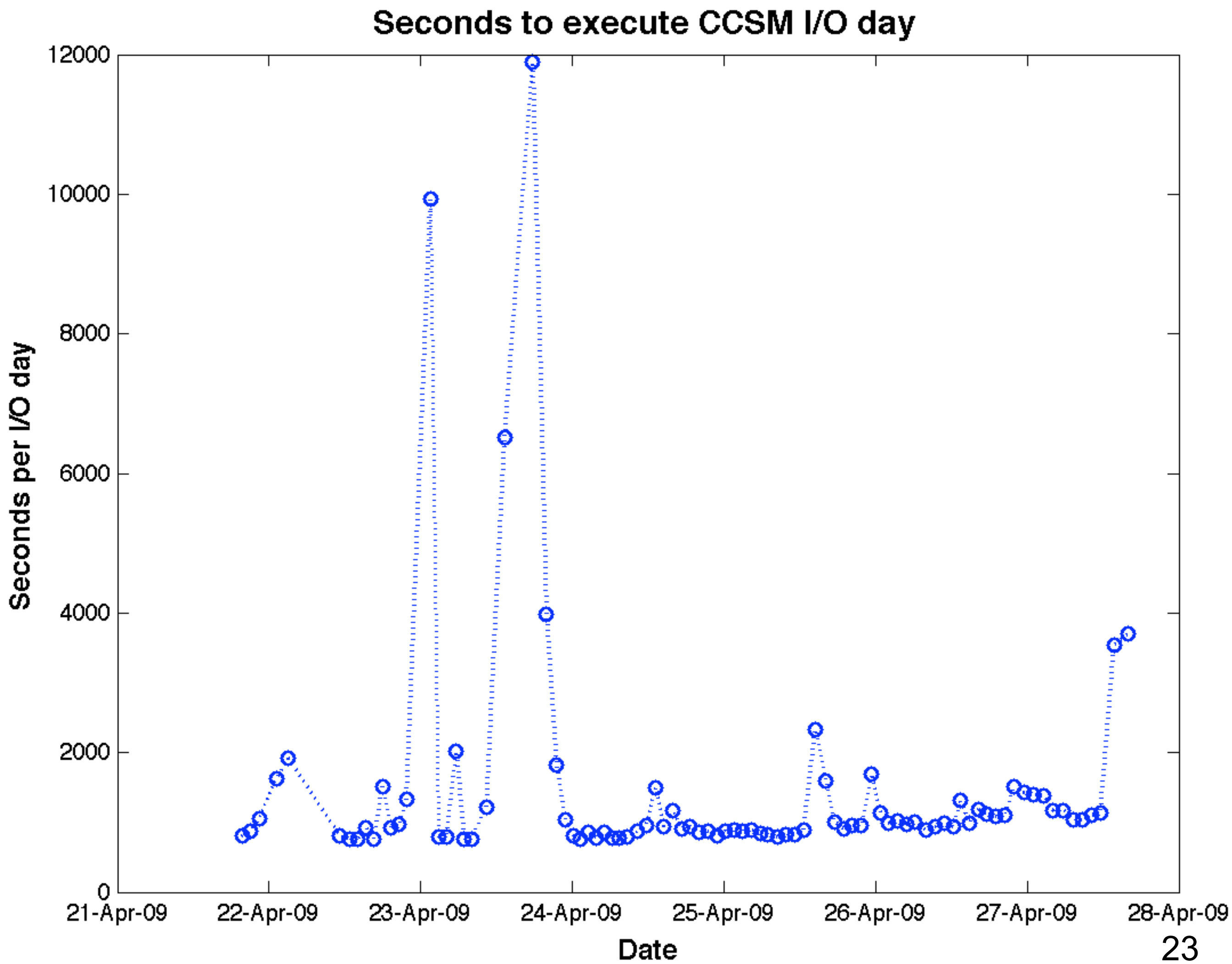
CCSM4_alpha Computational Costs (No I/O)



What About IO?

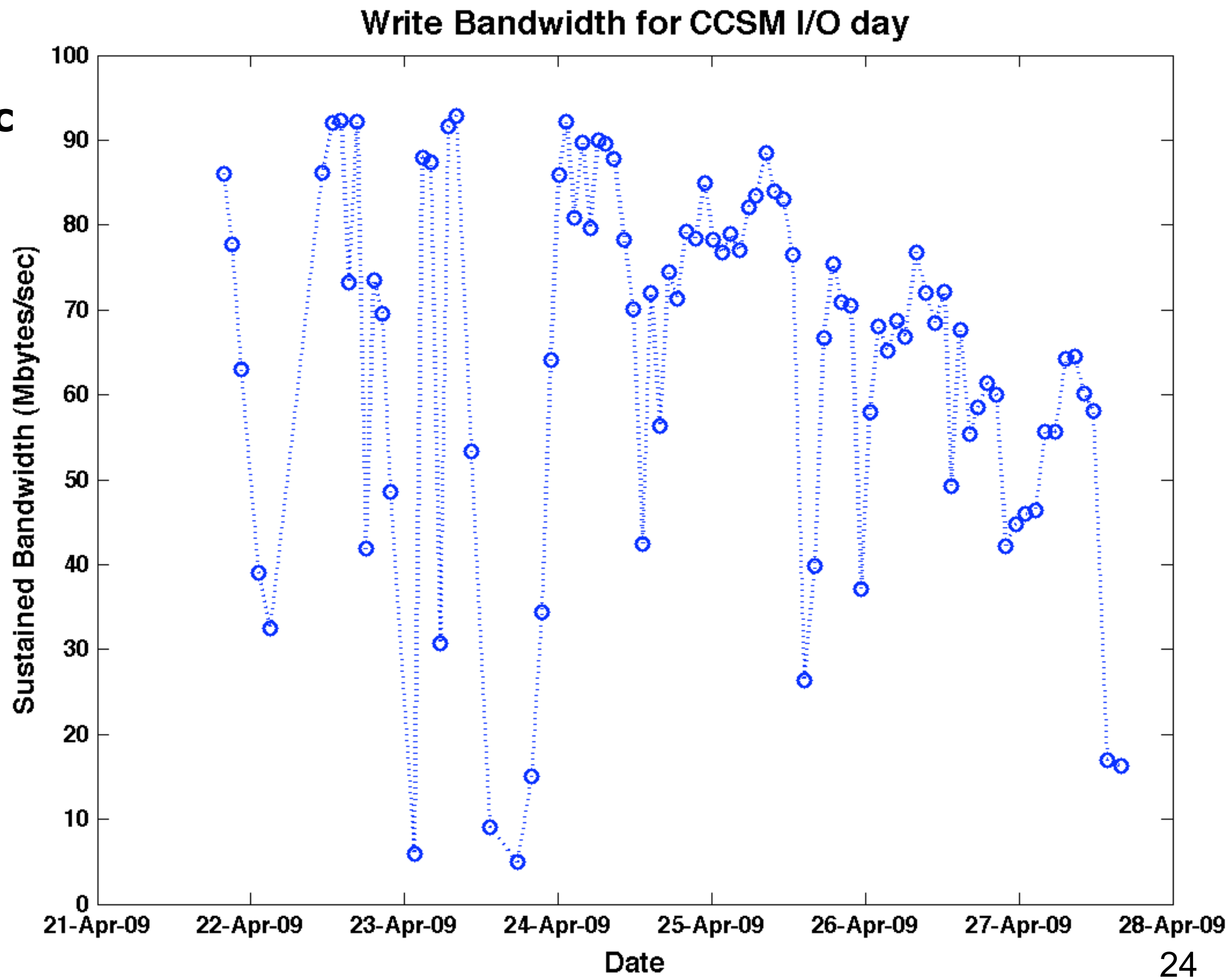
- CCSM I/O is currently serialized from each component.
- Total monthly output data = 57.9 GB
- File size ranges from 95 MB to 24 GB.
- “I/O” times aggregate per component MPI-based gather operations and write costs.
- Write sizes range from 864 KB to 1.4 GB.

Variability of CCSM File Write Times on Kraken



CCSM Sustained Output Bandwidth on Kraken

High = 92 MB/sec



Low = 5 MB/sec

Simulation Costs with Serial I/O Included

Cost to simulate 7.25 years	CPU hours	% of cost
Computational Cost	605K	76.6%
Serial Output Overhead [@92 MB/sec]	89K	11.2%
Output Variability Overhead	96K	12.2%
Total Output Overhead	185K	23.4%
Actual Total Cost	790K	100%

Plans to Understanding and Address I/O Issues

- Investigate possible issues with component gathers
- Profile the writes to identify any possible write latency issues
- Understand the sources of any Lustre file system variability
- Replace serial parallel I/O in CCSM with parallel I/O (in progress).

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]
    - OCE-0825754
    - CNS-0421498
    - CNS-0420873
    - CNS-0420985
- Computer Allocations:
  - TeraGrid TRAC @ NICS
  - DOE INCITE @ NERSC
  - LLNL Grand Challenge
- Thanks for Assistance:
  - Cray, NICS, and NERSC