

Climate Challenges in the Massively Parallel Era

Henry Tufo

Computer Science Section Head
Computational and Information Systems Laboratory
National Center for Atmospheric Research

Associate Professor and BoCCE Director
Department of Computer Science
University of Colorado at Boulder



NCAR

W⁵

- ❑ The National Center for Atmospheric Research (NCAR) is a Federally Funded Research and Development Center primarily funded by the National Science Foundation and located in Boulder, Colorado.

- ❑ NCAR's mission is to plan, organize, and conduct atmospheric and related research programs in collaboration with the universities and other institutions, to provide state-of-the-art research tools and facilities to the atmospheric sciences community, to support and enhance university atmospheric science education, and to facilitate the transfer of technology to both the public and private sectors.

- ❑ The Computer Science Section (CSS) is responsible for:
 - ❑ Applications, Algorithms, and Implementations.
 - ❑ Experimental Systems / Technology Tracking.
 - ❑ Grid Computing / Cyberinfrastructure.

- ❑ The focus of today's talk is on developing climate models that are capable of exploiting the (massively parallel) petascale systems coming on line in the near future.

Fundamental Tenets

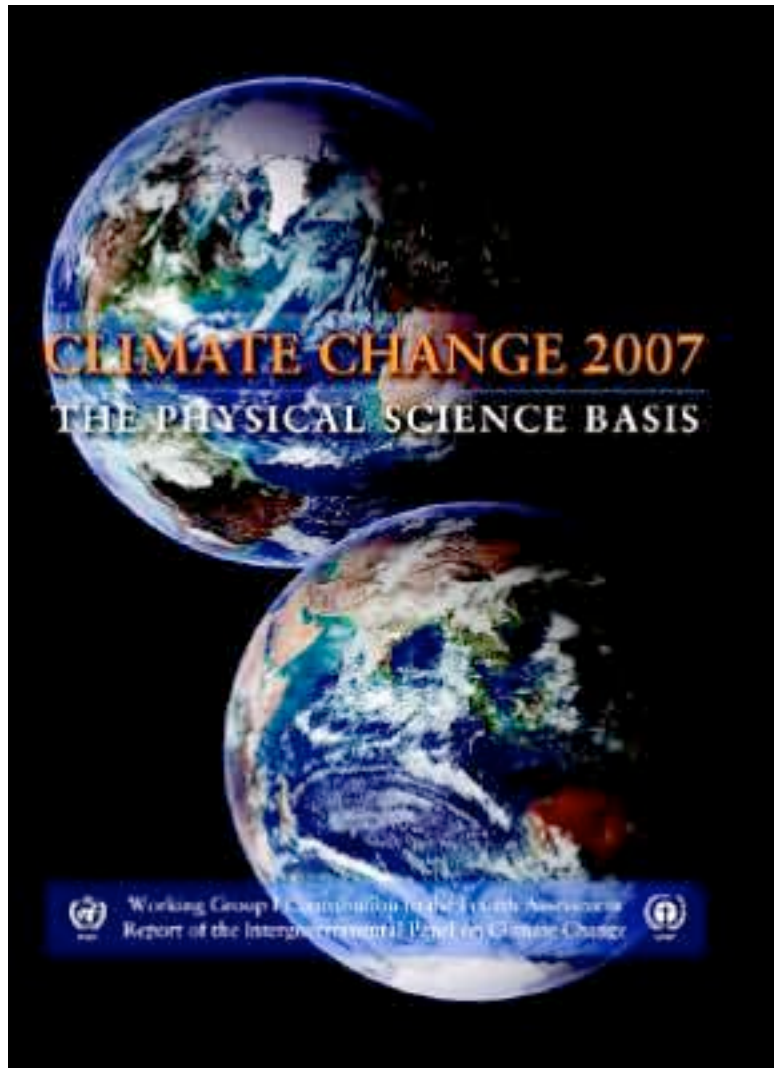
- ❑ We have little to no influence on the computing technology available to the community for performing science (i.e., we're the grass, not the wind).
- ❑ Given the fact that thread speeds have stalled, parallelism is the only viable option for achieving the level of simulation fidelity and integration rates required to investigate climate change.
- ❑ The last ten years told the story inter-node parallelism and we now have a good understanding of how to scale applications to $O(100K)$ nodes and build such systems.
- ❑ The next ten years holds the promise of significantly increased intra-node parallelism.

Outline

- ❑ Motivation – Climate Change
- ❑ Motivation - Parallelism
- ❑ Community Climate System Model (CCSM)
- ❑ From $O(100)$ to $O(100K)$ Nodes
 - ❑ Working with the clay
 - ❑ Building anew
- ❑ Next 1000x?
- ❑ Take Home Messages / Acknowledgements

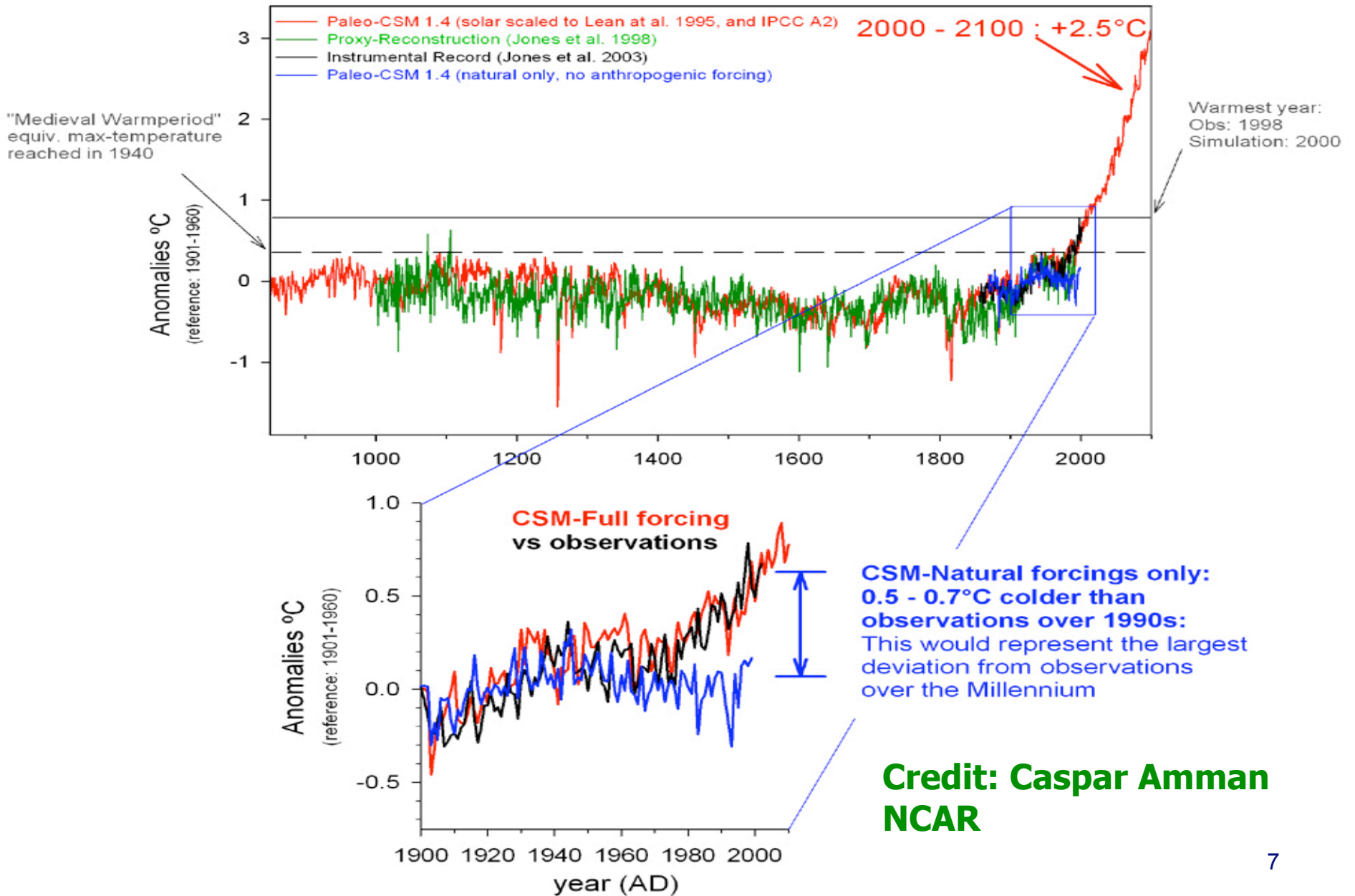
Motivation - Climate Change

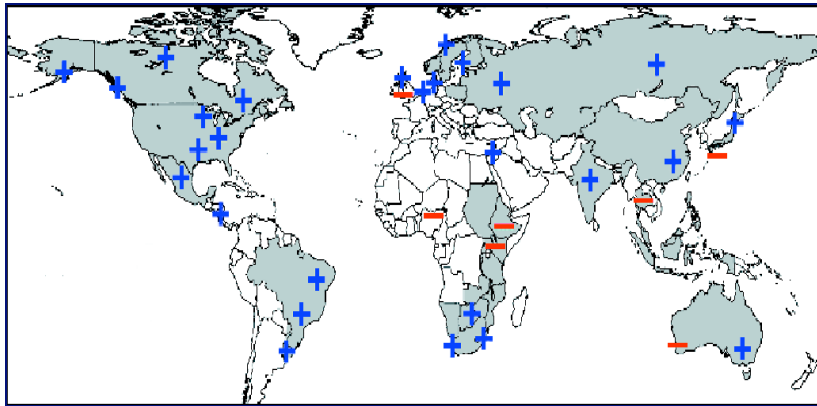
IPCC AR4 - 2007



- o IPCC AR4: "Warming of the climate system is un-equivocal" ...
- o ...and it is "very likely" caused by human activities.
- o Most of the observed changes over the past 50 years are now simulated by climate models adding confidence to future projections.
- o Model Resolutions: O(100 km)

Climate Change





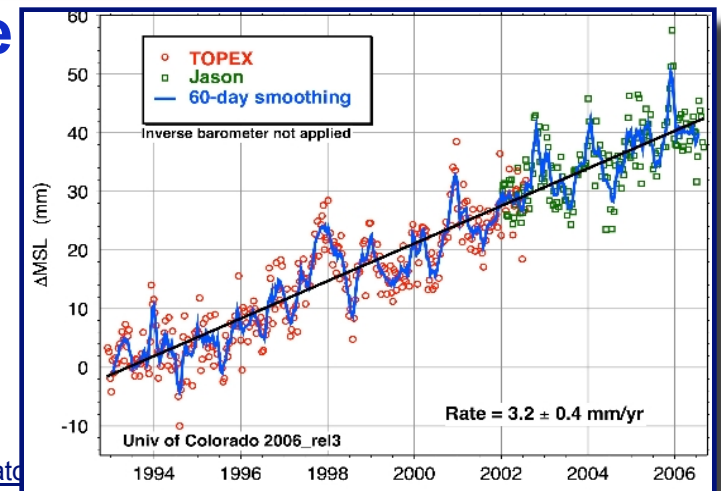
Change observed across scales

Effect on Extreme Precipitation



Global-wide Glacier Collapse

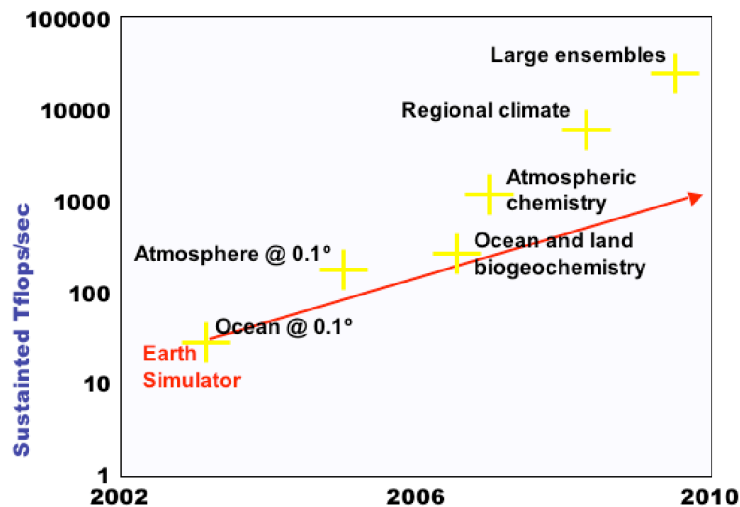
Global Sea Level Rise



Motivation - Parallelism

Projected Computational Requirements

<i>Issue</i>	<i>Motivation</i>	<i>Compute Factor</i>
Spatial resolution	Provide regional details	10^3 - 10^5
Model completeness	Add “new” science	10^2
New parameterizations	Upgrade to “better” science	10^2
Run length	Long-term implications	10^2
Ensembles, scenarios	Range of model variability	10
Total Compute Factor		10^{10} - 10^{12}



A Science Based Case for Large-Scale Simulation
(SCaLeS), SIAM News, 36(7), 2003 - David Keyes

Establishing a PetaScale Collaboratory for the Geosciences
UCAR/JOSS, May 2005

Computational and Information Systems Laboratory
National Center for Atmospheric Research

(Slide courtesy of John Drake, ORNL)

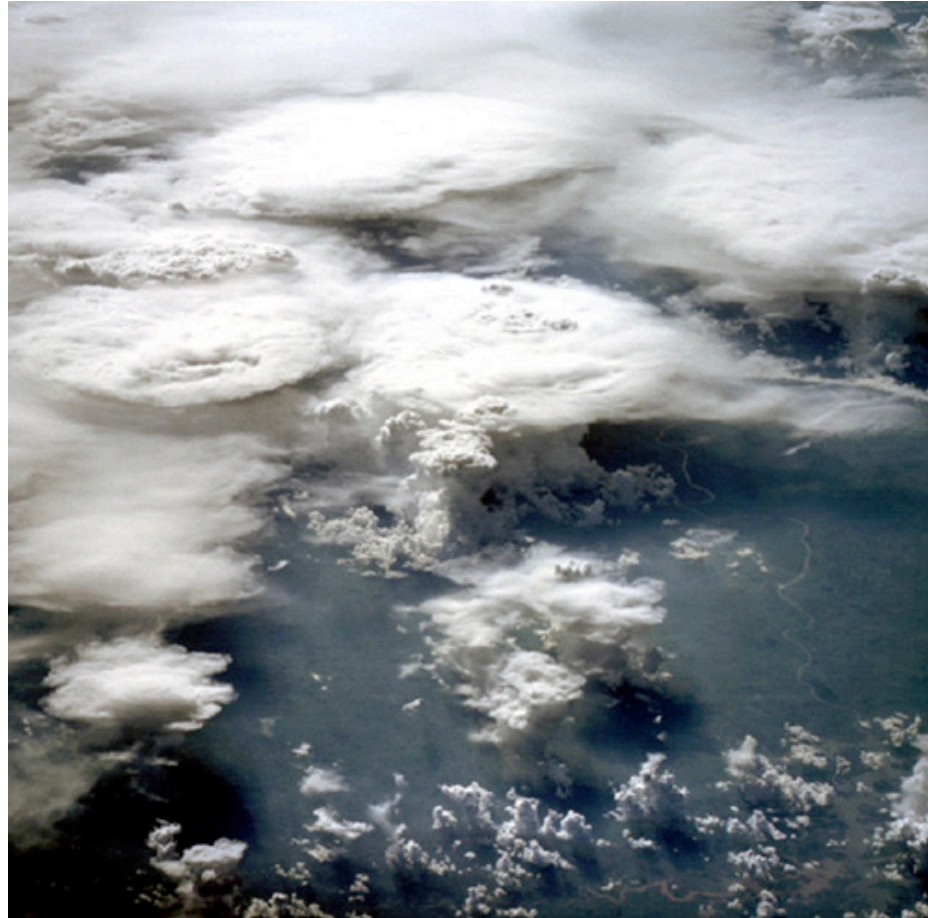
Goal:

- City/cloud-scale simulations

Strategies:

- *New dynamical formulations*
- *First-principle calculations -- no look-ups!*
- *Commensurate grids for all components*

Convective Scales in the Atmosphere are Tiny: basically $O(1 \text{ km})$



NASA Satellite Imagery

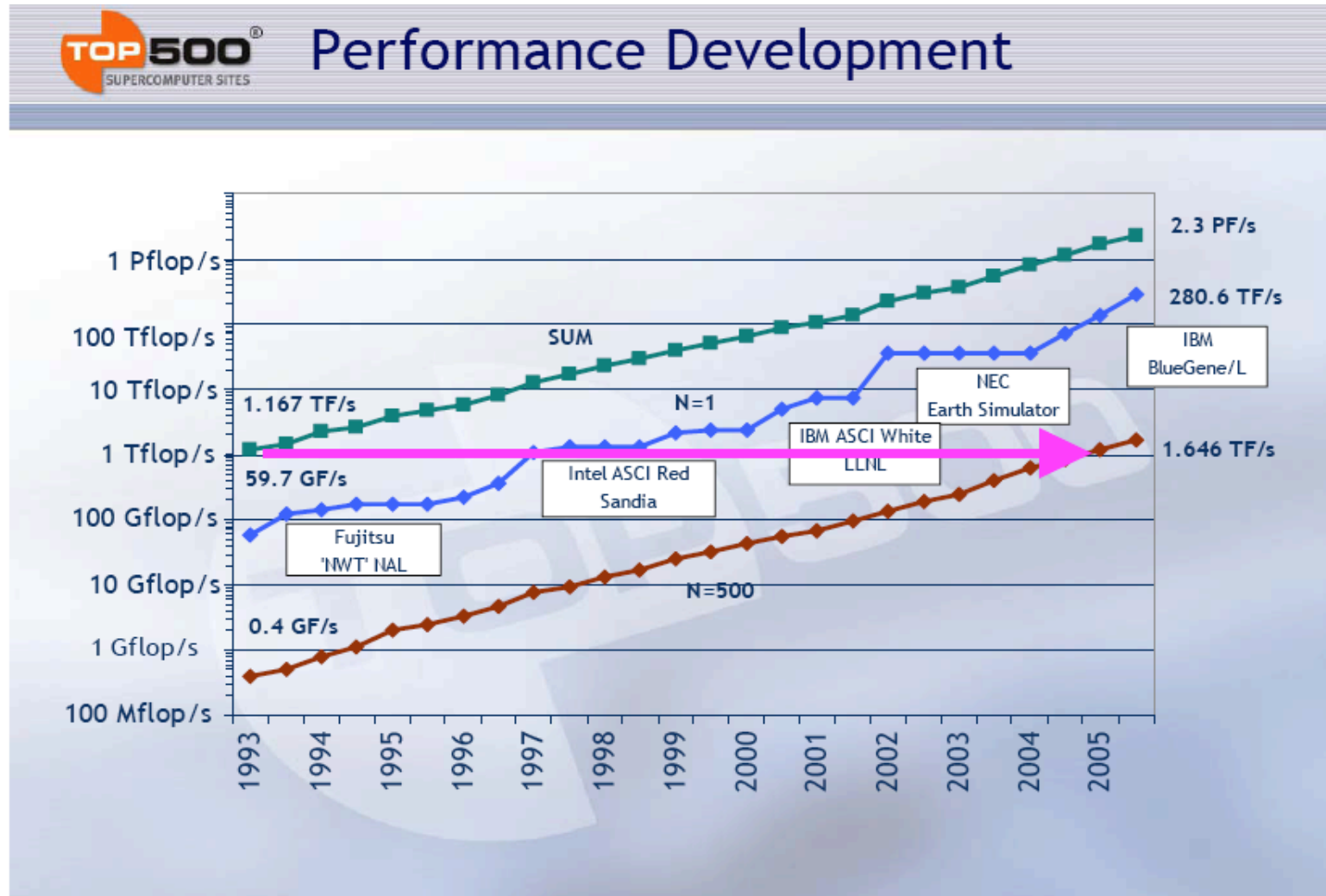
Computational and Information Systems Laboratory
National Center for Atmospheric Research

Estimating the number of FLOP's in a century experiment

- ❑ Area of the earth = $5.112 \times 10^8 \text{ km}^2$
- ❑ Number of levels ~ 100
- ❑ Number of tracers ~ 100
- ❑ Seconds in a century = $3.15 \times 10^9 \text{ s}$
- ❑ Flops/site per timestep ~ 5000
- ❑ FLOPS/century = $8 \times 10^{23}/(\text{dx}^2 \cdot \text{dt})$
- ❑ Roughly a **mole of FLOPS per century** for a 1 km atmospheric model.

**Thus a century run of a 1 km model takes ~10 wall
clock days at 10^{18} FLOPS sustained!**

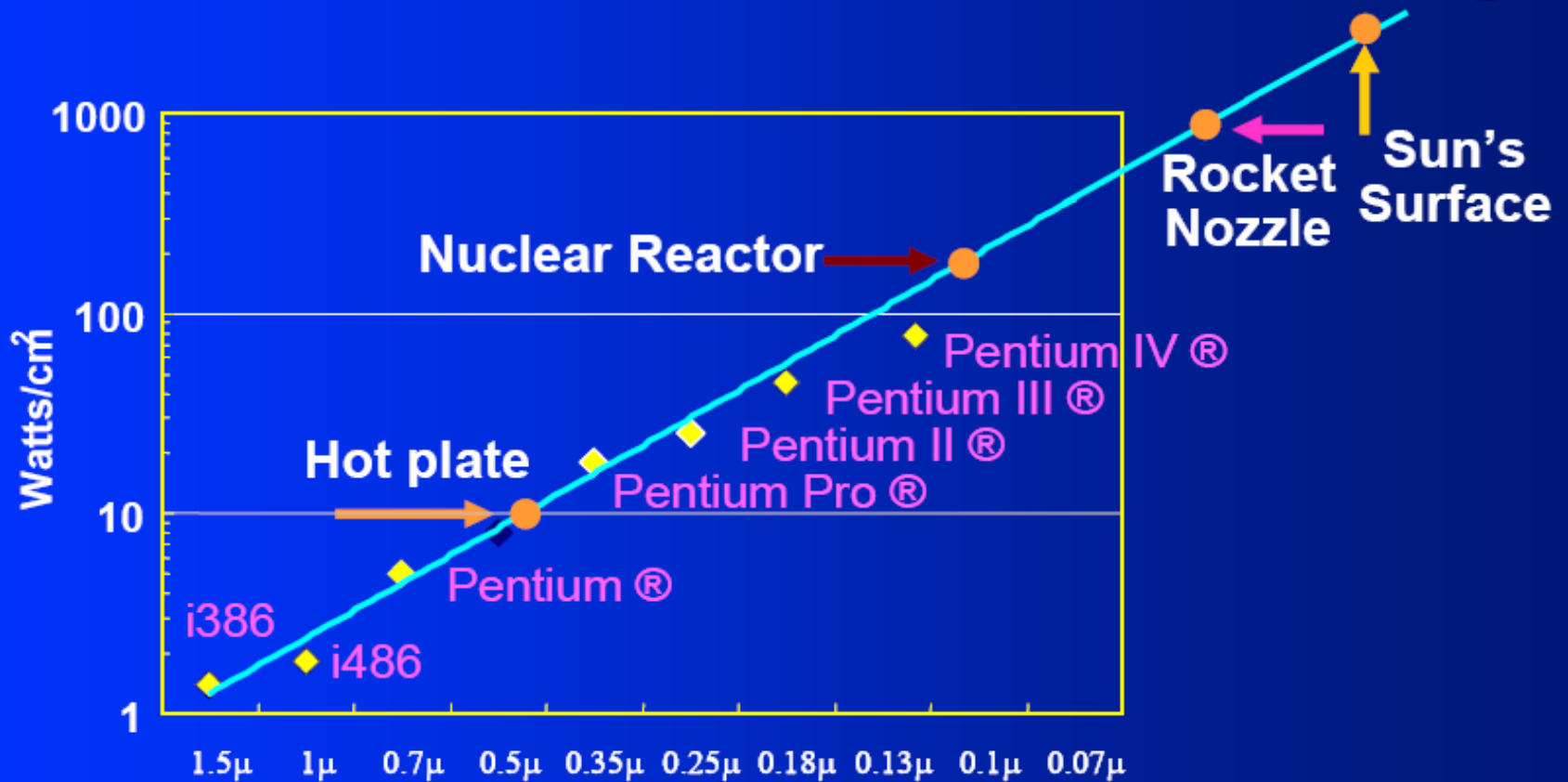
Performance Improvements are not coming fast enough!



...suggests 10^7 improvement will take 30 years

Computational and Information Systems Laboratory
National Center for Atmospheric Research

Relentless rise of power density



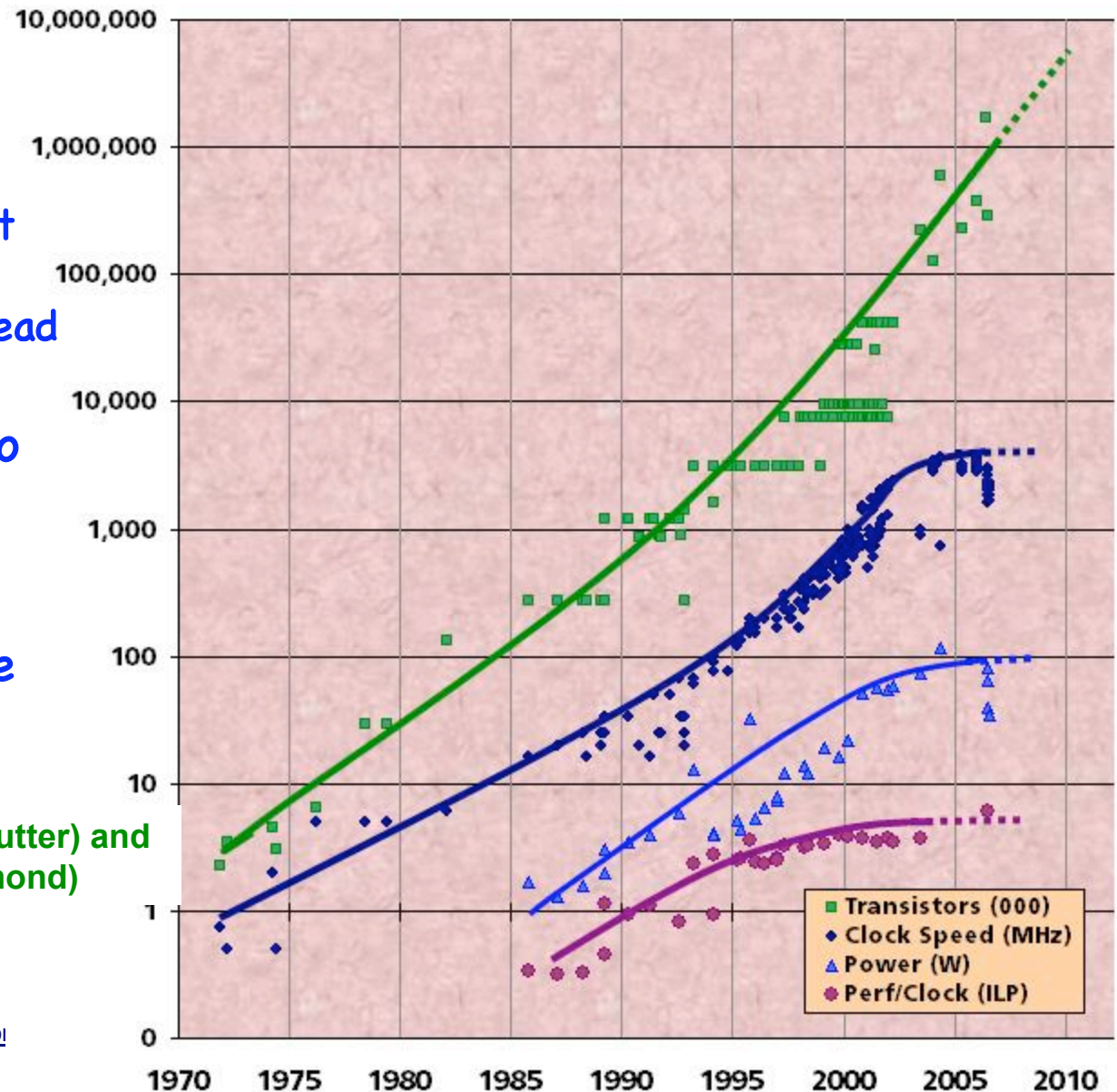
- 80% increase in power density/generation
- Voltage scales by ~0.8
- 225% increase in current consumption/unit area !

Chip Level Trends: Stagnant Clock Speed

- ❑ Chip density is continuing increase
~2x every 2 years
 - ❑ Clock speed is not
 - ❑ Number of cores are doubling instead
- ❑ There is little or no additional hidden parallelism (ILP)
- ❑ Parallelism must be exploited by software

Source: Intel, Microsoft (Sutter) and Stanford (Olukotun, Hammond)

[Col](#)



Key Point: Using the existing approaches on a bigger parallel machine can't be the answer

- ❑ **Moore's Law** is not fast enough.
 - ❑ This suggests a giant machine is required
- ❑ However, **Amdahl's Law** is formidable opponent.
 - ❑ **How long does it take you synchronize 1 billion threads?**
- ❑ **Dynamical timestep goes like N^{-1}**
 - ❑ Merciless effect of Courant limit
 - ❑ The cost of dynamics relative to physics increases as N
 - ❑ e.g. if dynamics takes 20% at 25 km it will take 86% of the time at 1 km
- ❑ **Traditional parallelization of horizontal leaves $O(N^2)$ per thread cost (vertical x horizontal)**
 - ❑ Must inevitably slow down as long as we have stalled thread speeds

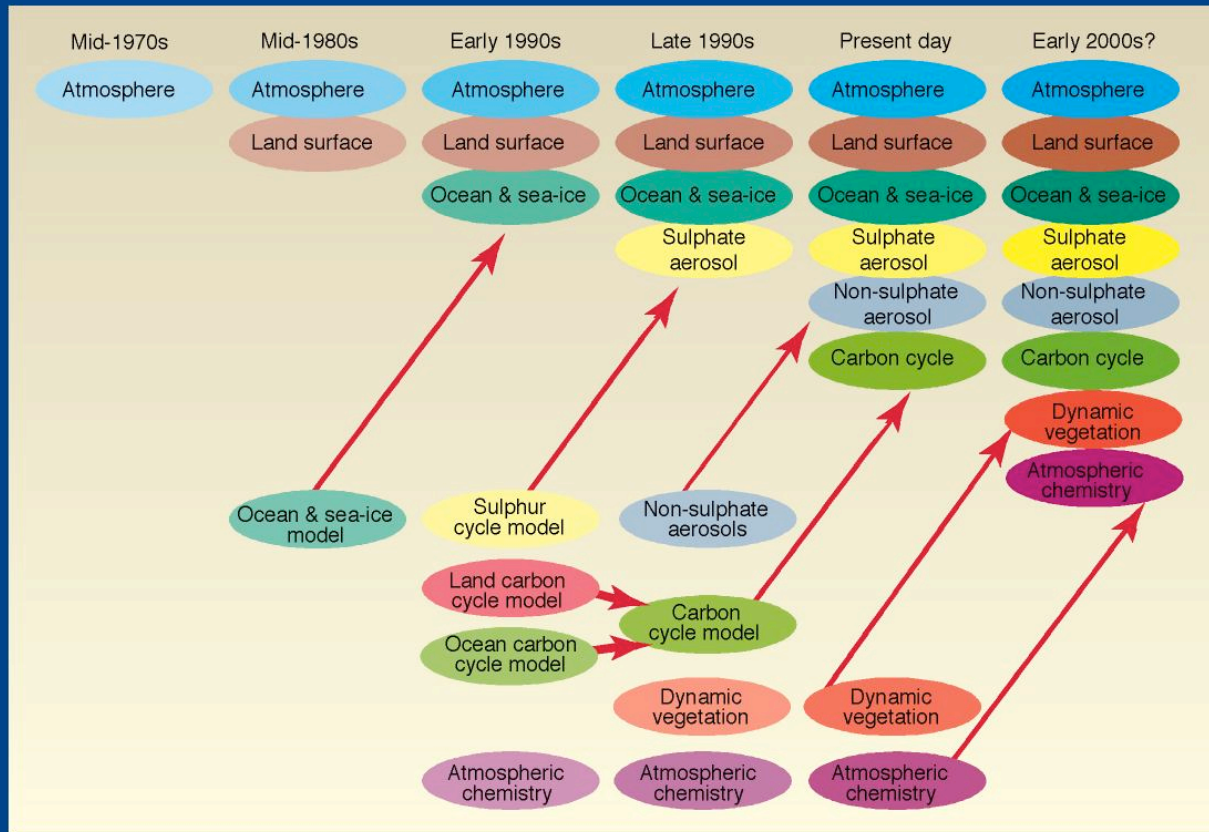
Options for Application Acceleration

- ❑ **Scalability**
 - ❑ Eliminate communication bottlenecks
 - ❑ New parallelism
 - ❑ Dynamic load balancing algorithms
- ❑ **Algorithmic Acceleration**
 - ❑ **Bigger Time Steps**
 - ❑ Semi-Lagrangian Transport
 - ❑ Implicit or semi-implicit time integration – solvers
 - ❑ **Fewer Points**
 - ❑ Adaptive Mesh Refinement methods
- ❑ **Hardware Acceleration**
 - ❑ **More Threads**
 - ❑ CMP, GP-GPU's
 - ❑ **Faster threads**
 - ❑ device innovations (e.g. high-K)
 - ❑ **Smarter threads**
 - ❑ Architecture – old tricks, new tricks... magic tricks
 - ❑ Vector units

Community Climate System Model

Evolution of Climate Models

The development of climate models, past, present and future



WG1 - TS BOX 3
FIGURE 1

CCSM4: a 1st Generation Earth System Model

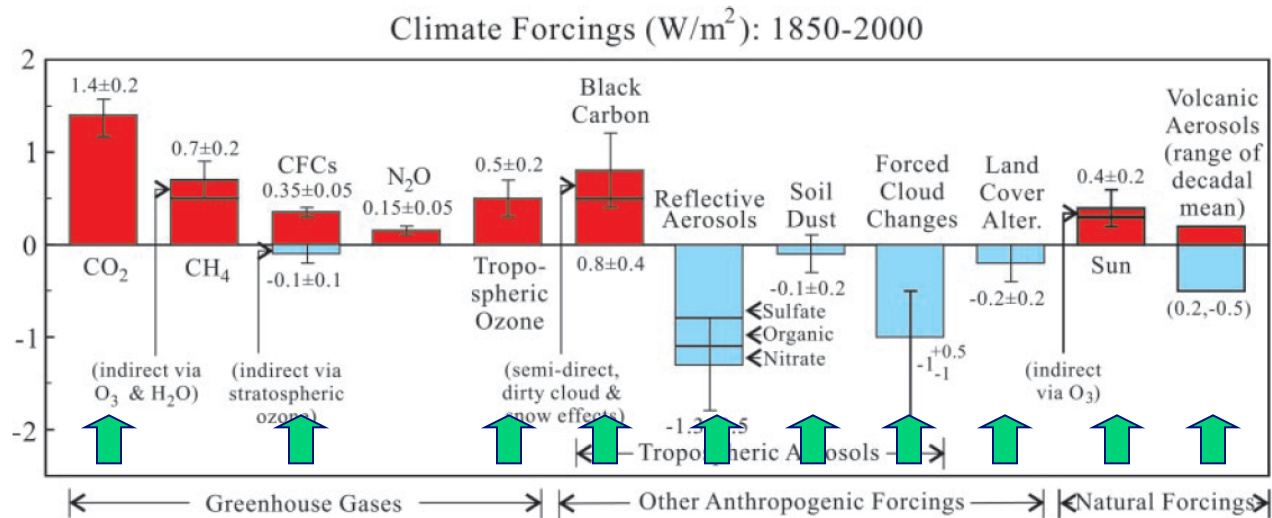
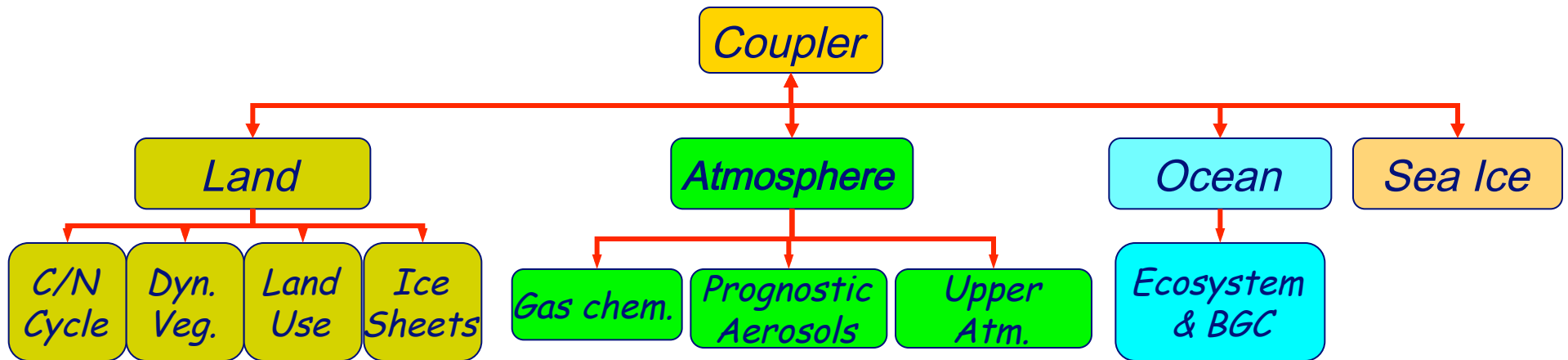
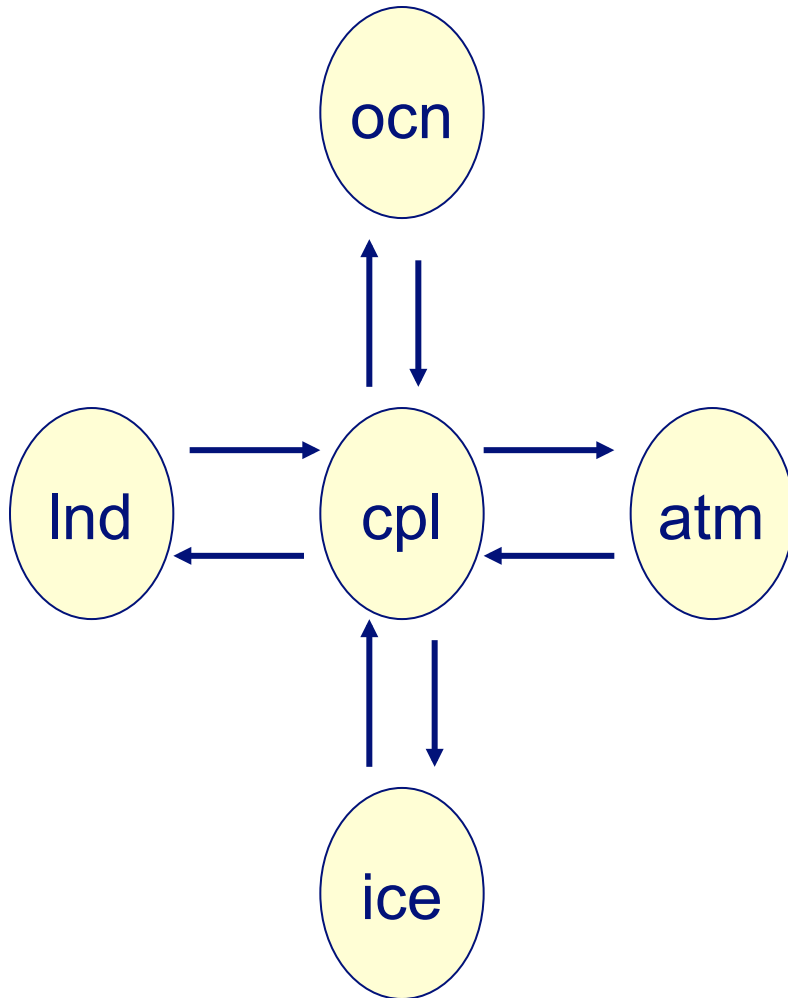


Fig. 1. Estimated climate forcings; error bars are partly subjective 1σ uncertainties.

CCSM Design and Details



- ❑ Hub and spoke design
- ❑ Multiple executables (60K-200K lines each)
 - ❑ Ocean (ocn): POP
 - ❑ Atmosphere (atm): CAM/HOMME
 - ❑ Sea Ice (ice): CICE
 - ❑ Land (Ind): CLM
 - ❑ Coupler (cpl): MCT
 - ❑ (Single executable in beta.)
- ❑ Need 5 simulated years/day implies that we must run at “low” resolution.
- ❑ Typical configuration run on **O(200) processors**. Key question is whether we can scale up the individual components without adding work.
- ❑ Target Petascale Configuration:
 - ❑ CAM - 30 km, L26
 - ❑ POP, Sea Ice, and Land - 0.1°

From $O(100)$ to $O(100K)$ Nodes

Working With the Clay

CPL7

CCSM4 (CPL7) Design 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) Design Goals (con't)

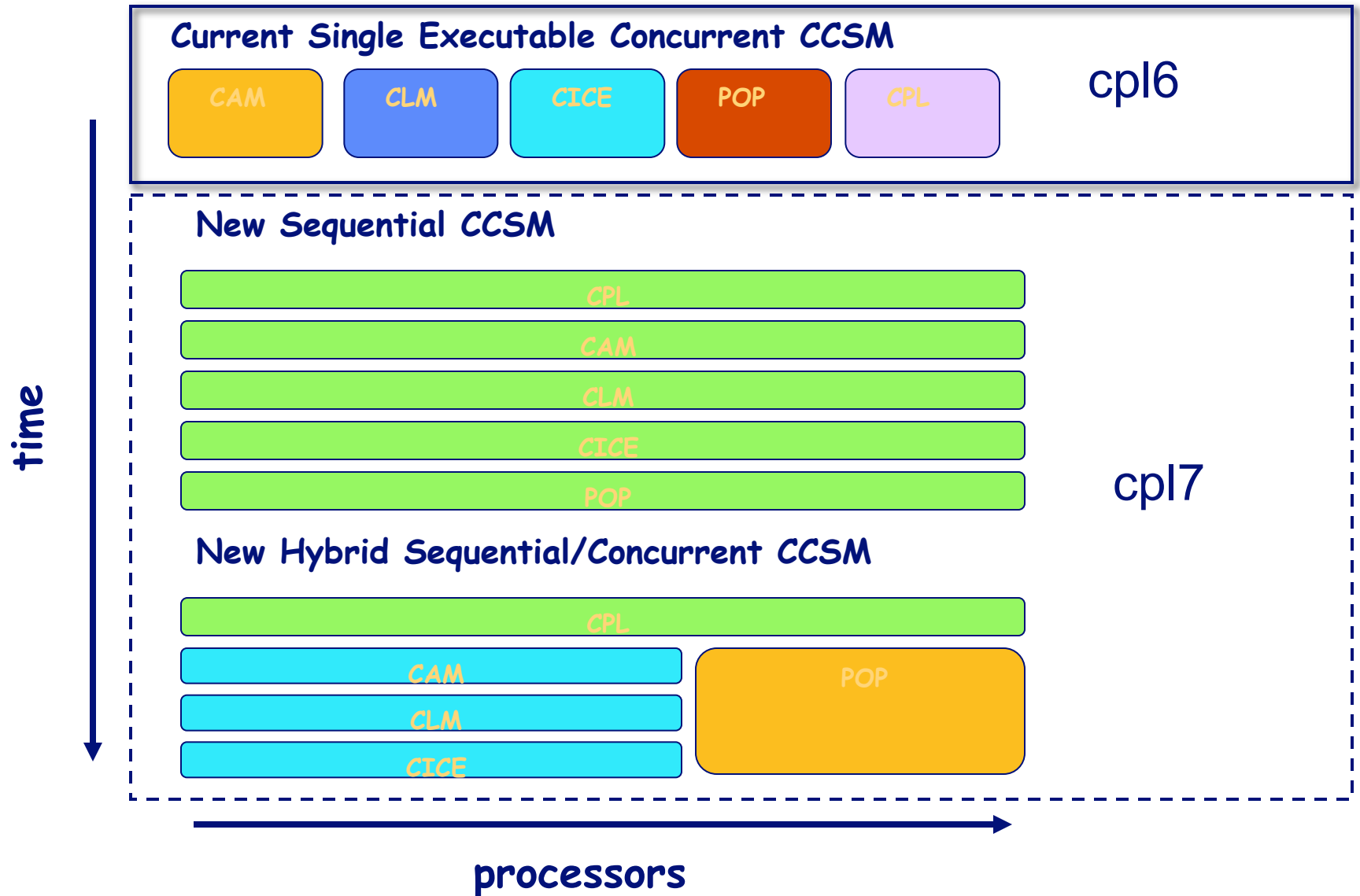
- ❑ Target new flexibility in run time configuration of system
 - ❑ Mixed sequential/concurrent component sequencing
 - ❑ Make it easier to add new components vs. CCSM3

- ❑ 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, J. Dennis

CPL6 -> CPL7 Design



POP

POP (Parallel Ocean Program)

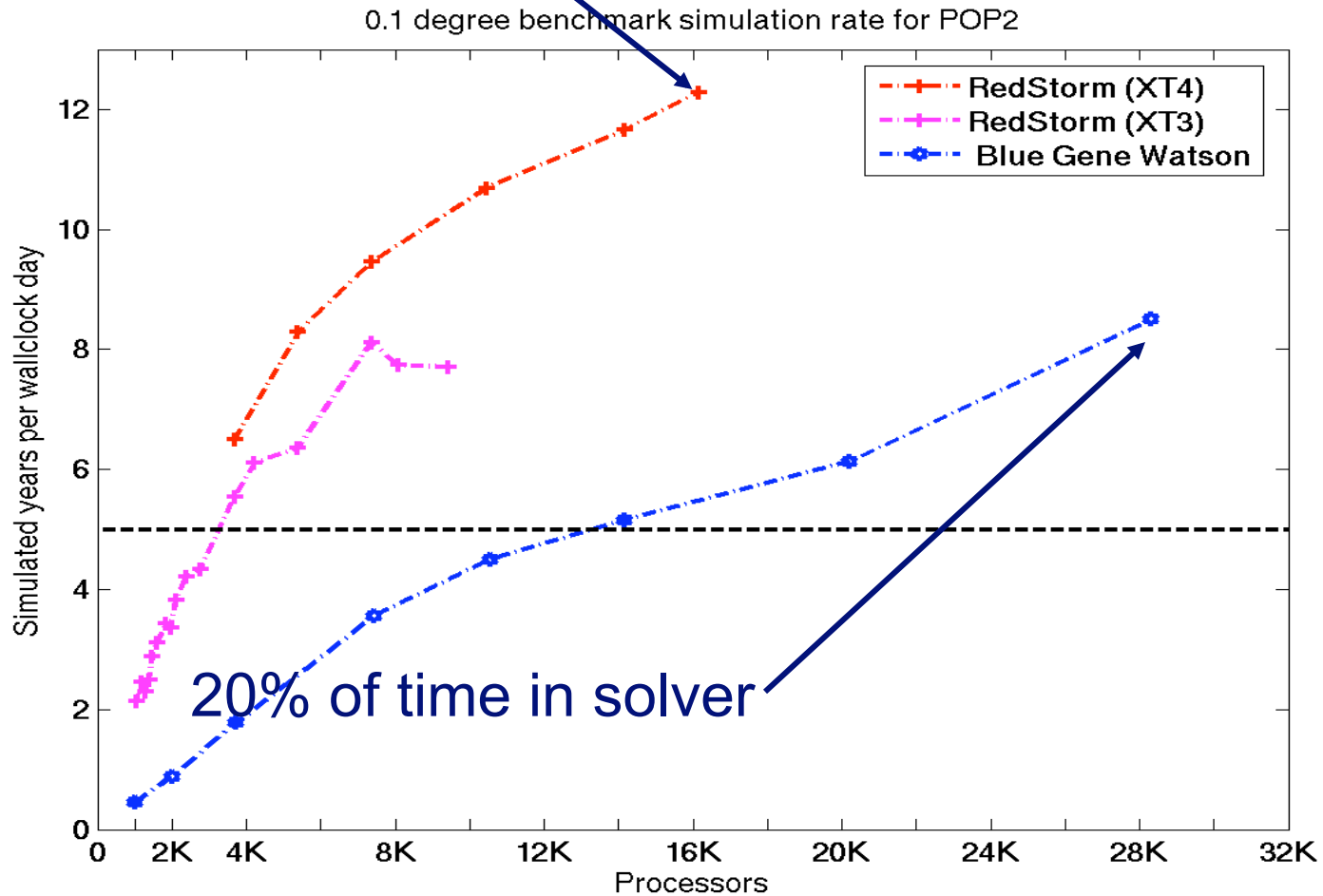
- ❑ Developed at LANL
- ❑ Two components:
 - ❑ **Baroclinic**: Finite difference
 - ❑ **Barotropic**: Solve surface pressure (2D) with PCG (diagonal preconditioning)
- ❑ Number of improvements to POP base code
 - ❑ Improve partitioning/load-balancing using space-filling curves.
 - ❑ Aggregated 3-D boundary exchange.
 - ❑ Dipole grid:
 - ❑ Redesign of the barotropic solver using 1D data structures
 - ❑ Tripole grid:
 - ❑ Reworking of tripole boundary exchange [Jones]

Status of POP

- ❑ Simulation rate of offline LANL-POP2 benchmark
 - ❑ Blue Gene +50%
 - ❑ Cray XT4 +33%
 - ❑ Does not include MPI_reduce fixes [P. Worley]
- ❑ Reunite LLNL & CCSM POP2 code base
 - ❑ LLNL [scalable + 0.1 degree support]
 - ❑ CCSM [ocean model used in CCSM 3.5]
- ❑ Won BGW cycle allocation
 - ❑ 110 Rack Days/ 5.4M CPU hours
 - ❑ Completed 9.5 year of spin up [7600 processors]
- ❑ Prototype version of PIO [binary]
 - ❑ restart: CCSM-POP & LANL-POP + private mods
 - ❑ tavg: LANL-POP + private mods

POP2 0.1° benchmark

71% of time in solver

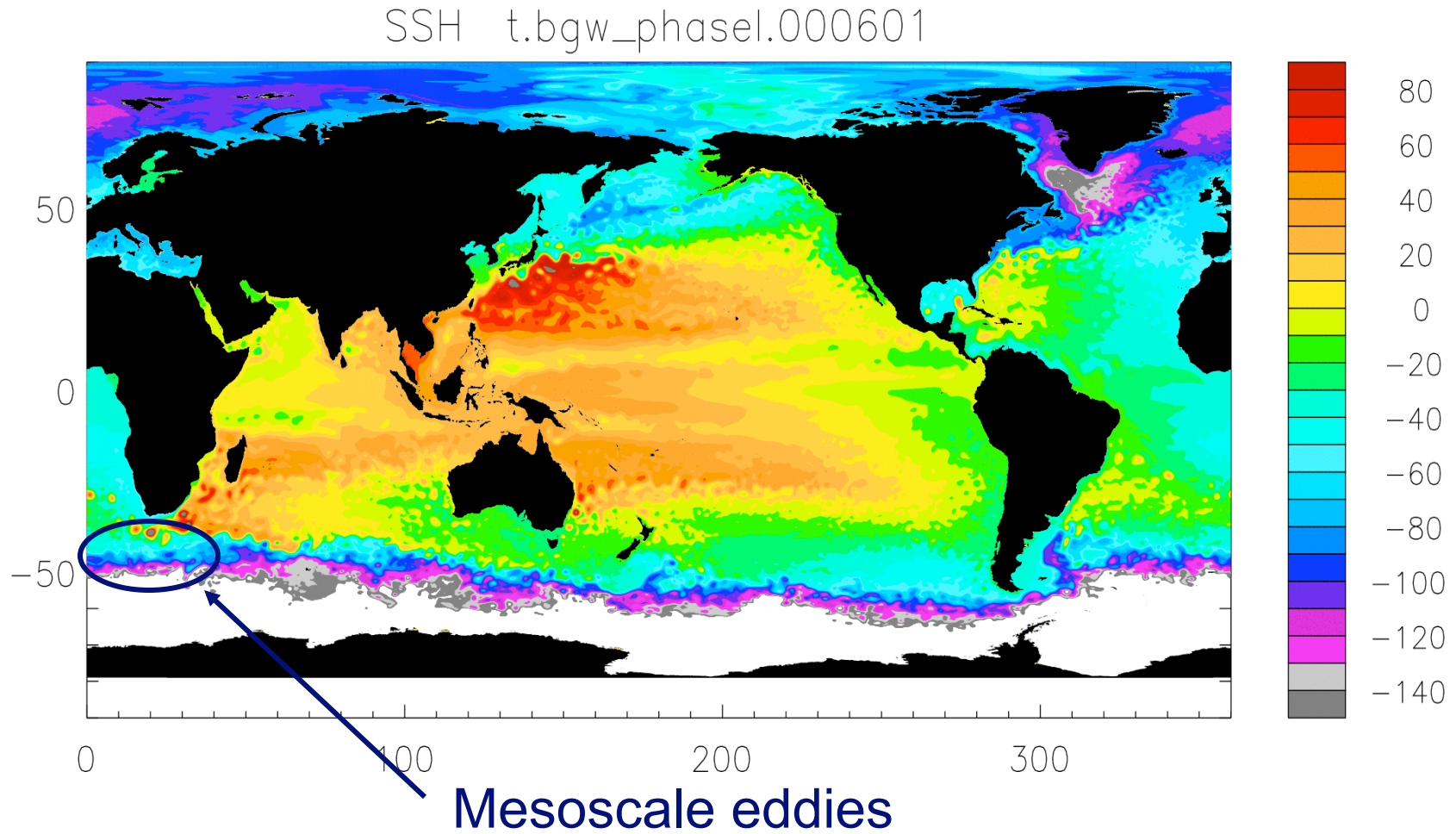


20% of time in solver

RedStorm data courtesy of M. Taylor

Computational and Information Systems Laboratory
National Center for Atmospheric Research

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

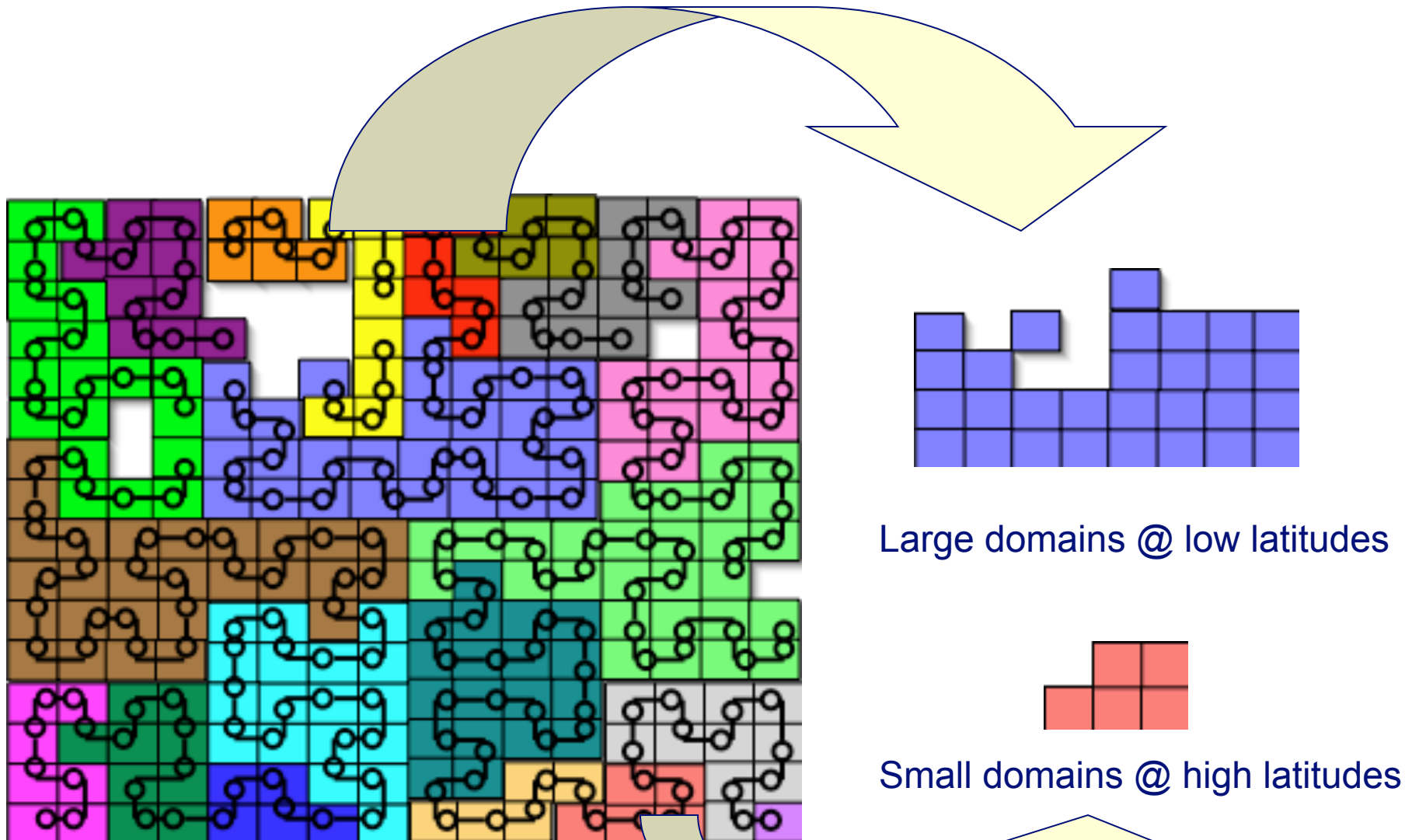


CICE

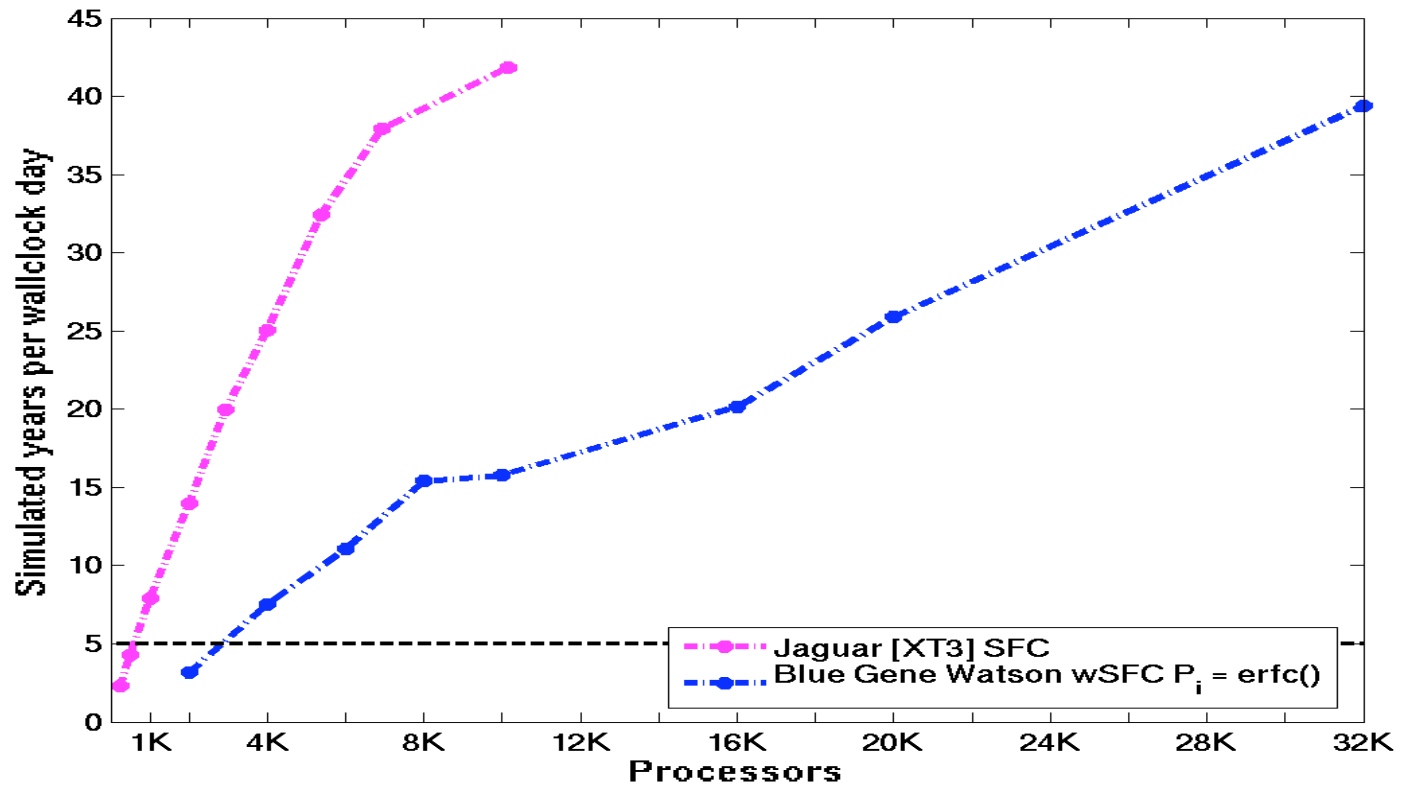
CICE4

- ❑ 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 time step/ no Forcing
 - ❑ 10K Cray XT3 processors
 - ❑ 40K Blue Gene/L processors

1° CICE4 on 20 processors

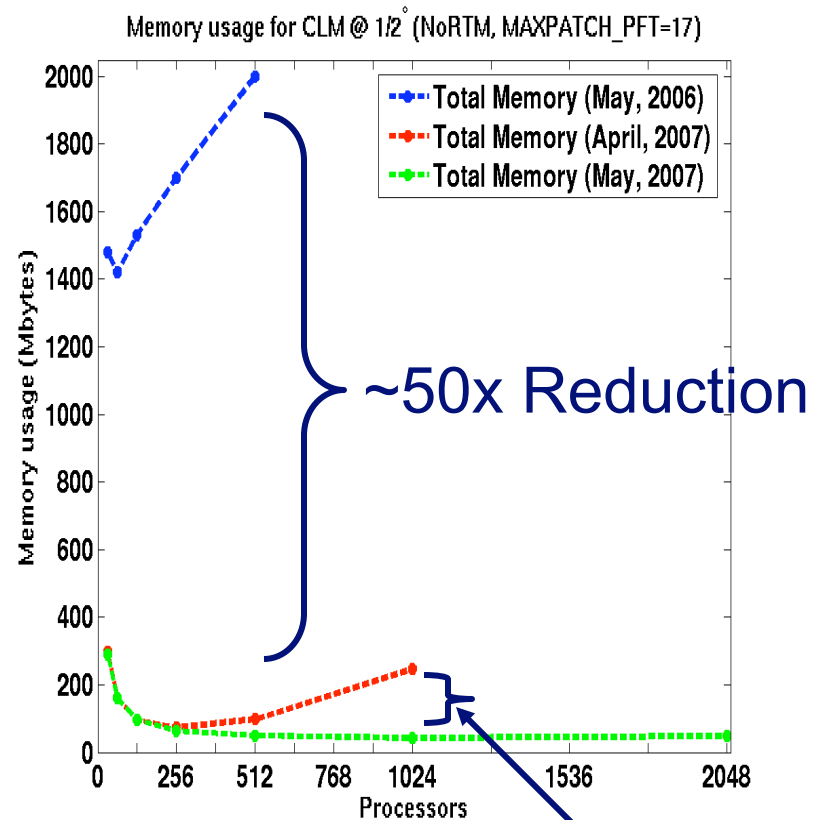


CICE4 @ 0.1°



CLM/MCT

Status of CLM/MCT



- Work of T. Craig
- Elimination of global memory
 - Reworking of decomposition algorithms
- Resources for 1/6° global run
 - May 2006: 512 processors, ~18 Gbytes per proc
 - May 2007: 512 processors, ~495 Mbytes per proc
- Future Work
 - Investigation scalability at 1/6° & 1/10°
 - Addition of PIO

Patch to MCT

From $O(100)$ to $O(100K)$ Nodes

Building Anew

HOMME Framework

- ❑ HOMME = High-Order Methods Modeling Environment
- ❑ Framework for developing scalable and efficient Atmospheric General Circulation Models (AGCMs) to support climate science.
- ❑ Serves as a prototype for the Community Atmospheric Model (CAM) component of the Community Climate System Model (CCSM).
- ❑ Designed for high-order methods (e.g., spectral element and discontinuous Galerkin methods) on the cubed-sphere.
- ❑ Configurable for shallow water and (dry/moist) primitive equations (hydrostatic).
- ❑ Support for:
 - ❑ Explicit and semi-implicit time stepping.
 - ❑ Several vertical discretization schemes (e.g., Lin vertical Lagrangian method).
 - ❑ Geometrically non-conforming elements and dynamically adaptive meshes (AMR).
- ❑ Proven to efficiently scale to 10,000's of processors (2001 Gordon Bell Honorable Mention).

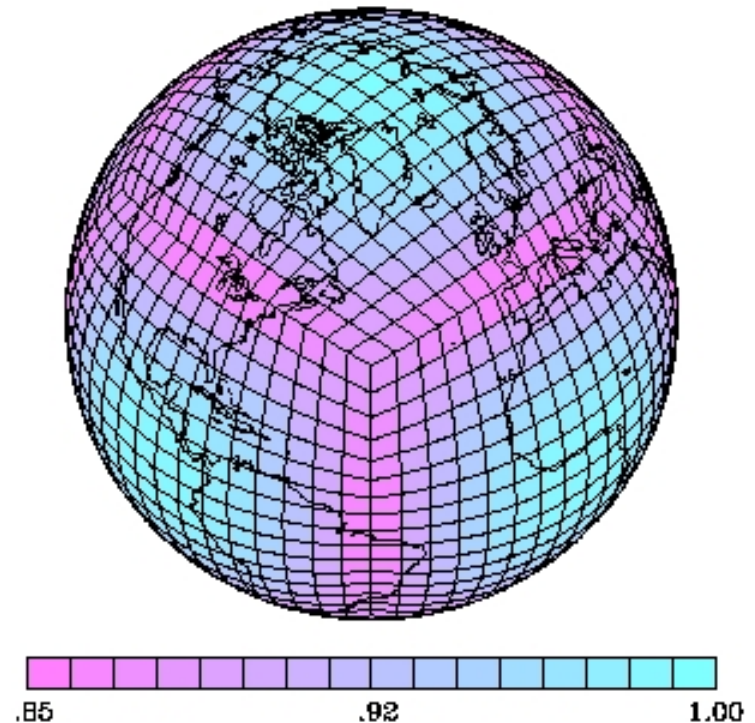
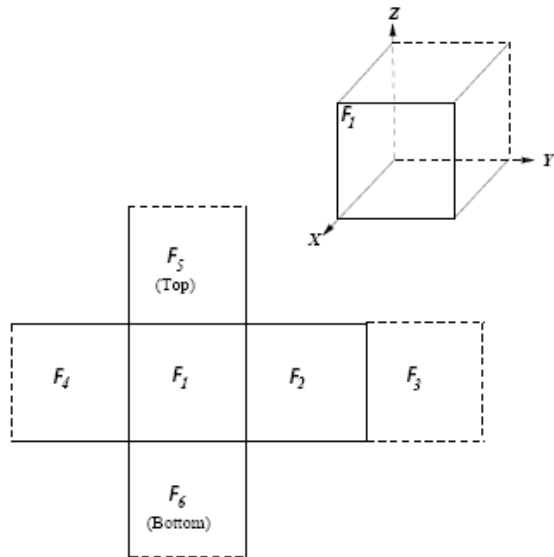
Advantages of High-Order Methods

- ❑ Algorithmic Advantages of High Order Methods:
 - ❑ h-p element-based method on quadrilaterals ($N_e \times N$)
 - ❑ Exponential convergence in polynomial degree (N)

- ❑ Computational Advantages of High Order Methods:
 - ❑ Naturally cache-blocked $N \times N$ computations
 - ❑ Nearest-neighbor communication between elements (explicit)
 - ❑ Well suited to parallel μ processor systems

Geometry - Cube-Sphere

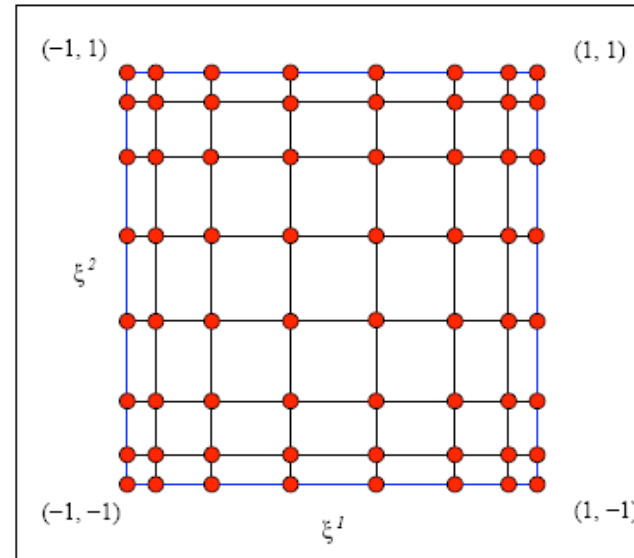
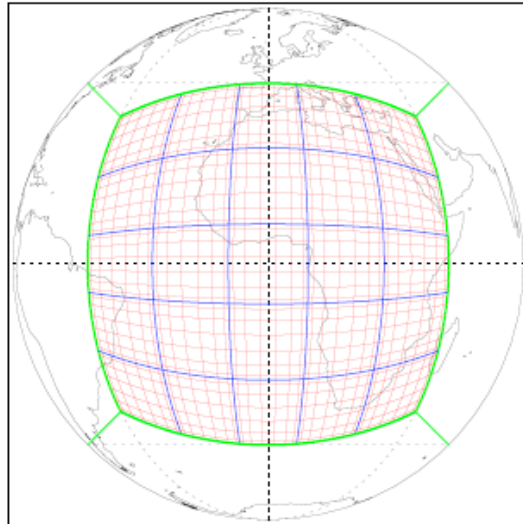
- ❑ Sphere is decomposed into 6 identical regions using a central projection (Sadourny, 1972) with equiangular grid (Rancic et al., 1996).
- ❑ Avoids pole problems, quasi-uniform.
- ❑ Non-orthogonal curvilinear coordinate system with identical metric terms



Ne=16 Degree of non-uniformity

Computational Mesh

Cubed-Sphere ($N_e = 5$) with 8×8 GLL points



- ❑ Elements:
 - ❑ A quadrilateral “patch” of $N \times N$ gridpoints
 - ❑ Gauss-Lobatto Grid
 - ❑ Typ. $N=8$

- ❑ Cube
 - ❑ $N_e =$ Elements on an edge
 - ❑ $6 \times N_e \times N_e$ elements total

Domain Decomposition

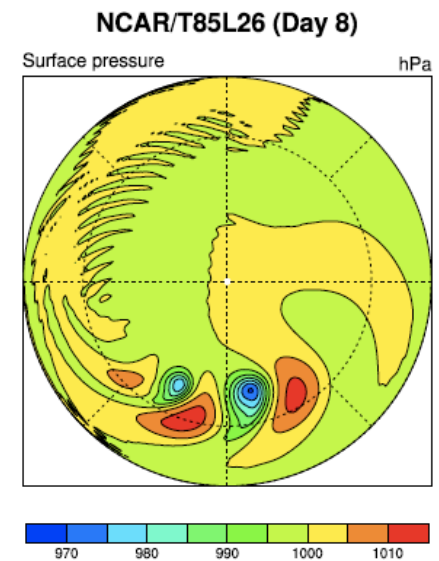
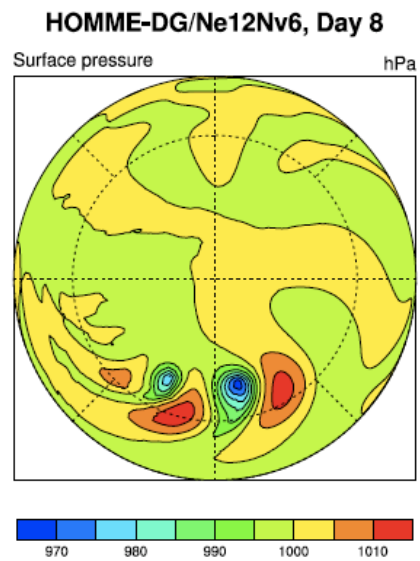
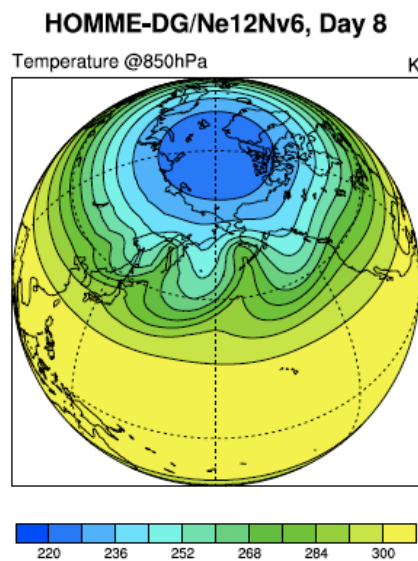
- ❑ Mapping the elements to processors is achieved using Hilbert space-filling curves (Sagan, 1994; Dennis et al., 2006).
- ❑ Generates the best partitioning mappings when $N_e = 2^n 3^m 5^k$ where n , m , and k are positive integers.
- ❑ (Have also examined Metis and Chaco but we've found SFC to be superior at large processors counts.)

Key Points

- ❑ Only C^0 continuity or flux conservation is enforced across element interfaces.
- ❑ Locally the mesh is structured with solution, data, and geometry expressed as sums of Nth-order tensor-product Lagrange polynomials based on the Gauss or Gauss-Lobatto quadrature points.
- ❑ Globally the mesh is an unstructured array of deformed quadrilaterals (layered in 3D).
- ❑ Exponential convergence (large N ideal for transitional flows because of minimal numerical dispersion and dissipation).
- ❑ Geometrically nonconforming formulation provides additional meshing flexibility and adaptivity.

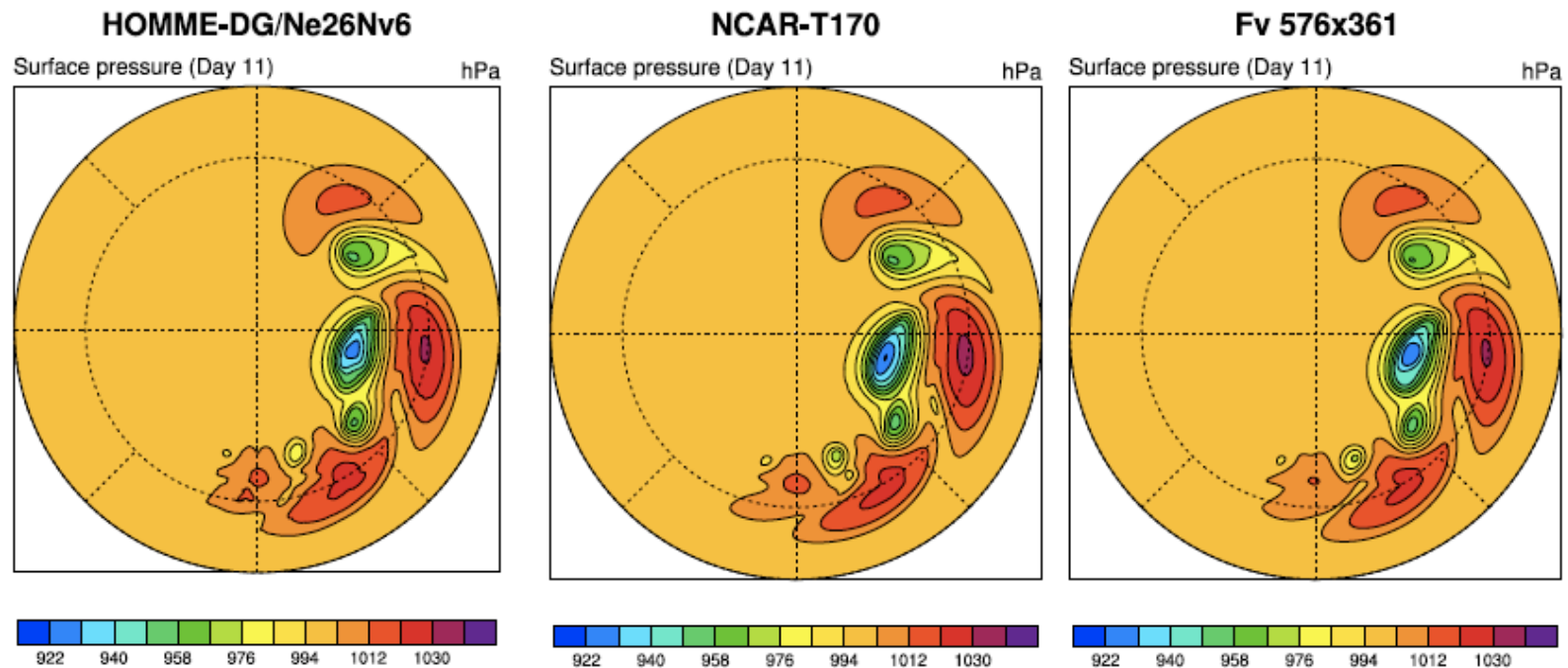
DG@HOMME vs. NCAR Global Spectral Model (T85)

Temperature field at 850 hPa and corresponding surface pressure for the baroclinic test after 8 days (left panels). Reference surface pressure with the NCAR global spectral model (T85). Both use 26 levels and an approximate horizontal resolution of 1.4° .



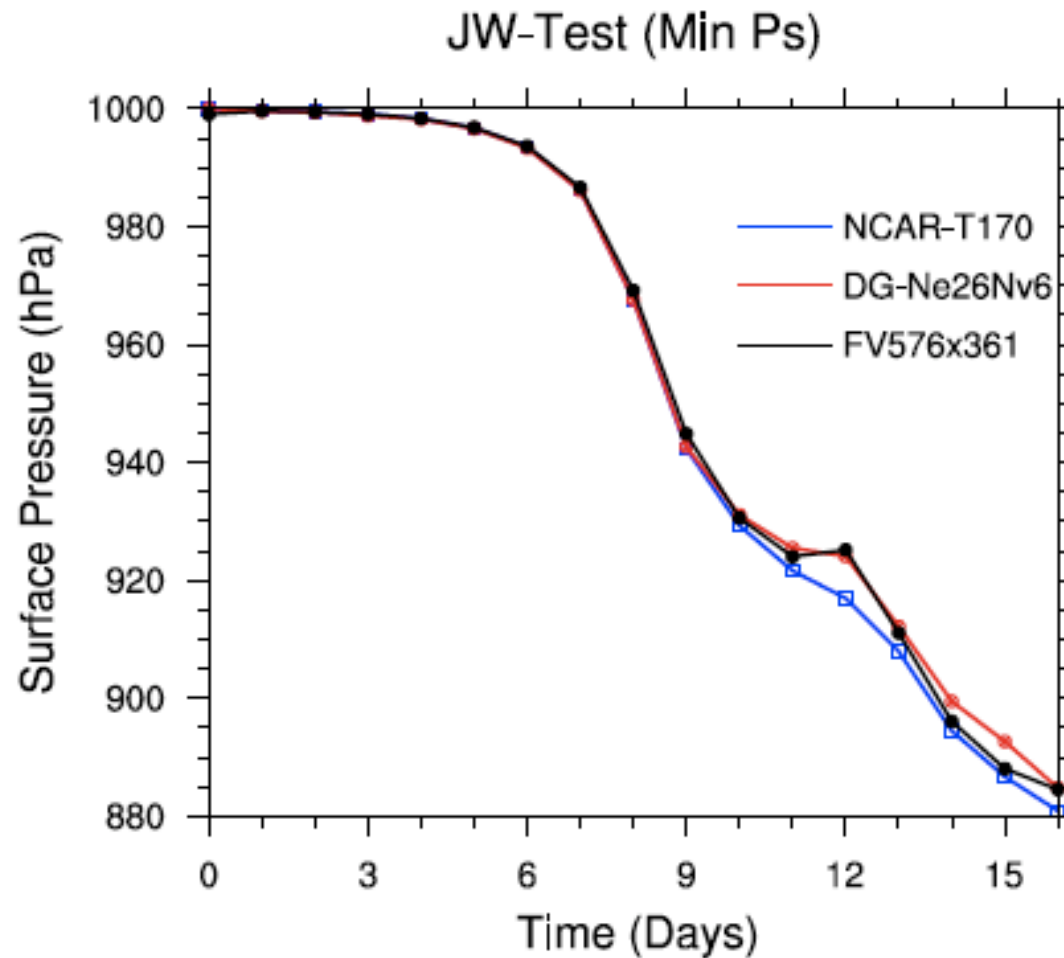
DG@HOMME vs. NCAR Climate Models (T170)

Surface pressure evolution after 11 days of model integration. DG baroclinic solution ($N_e = 26$, $N_v = 6$). Reference solutions with the NCAR global spectral model (T170) and the finite-volume (FV) dynamical cores. All use 26 vertical levels, the horizontal resolution for the DG and T170 solutions is approximately 0.7° and $0.5^\circ \times 0.625^\circ$ for the FV solution..

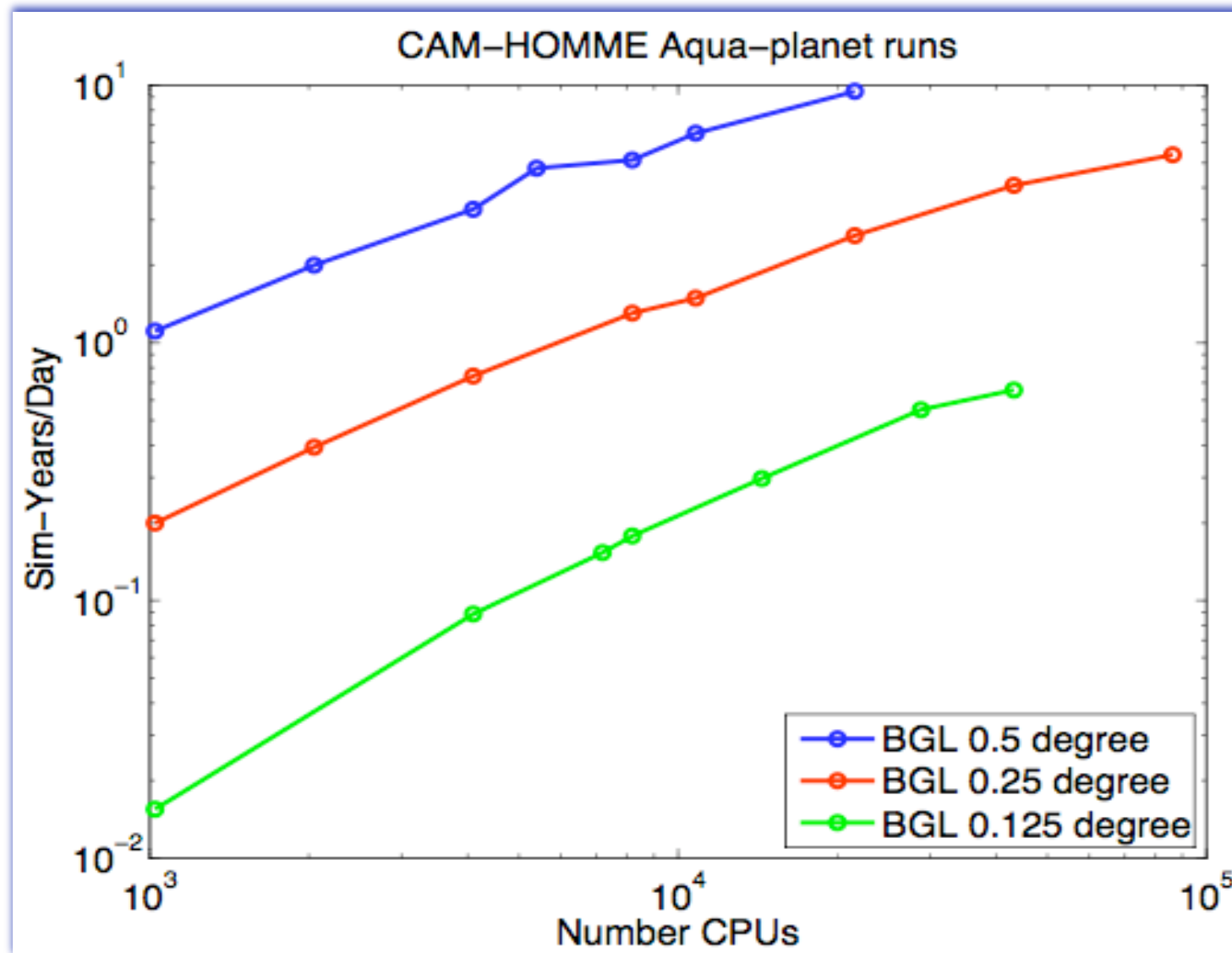


DG@HOMME vs. NCAR Climate Models (T170)

16-day time trace of minimum surface pressure for all three models.

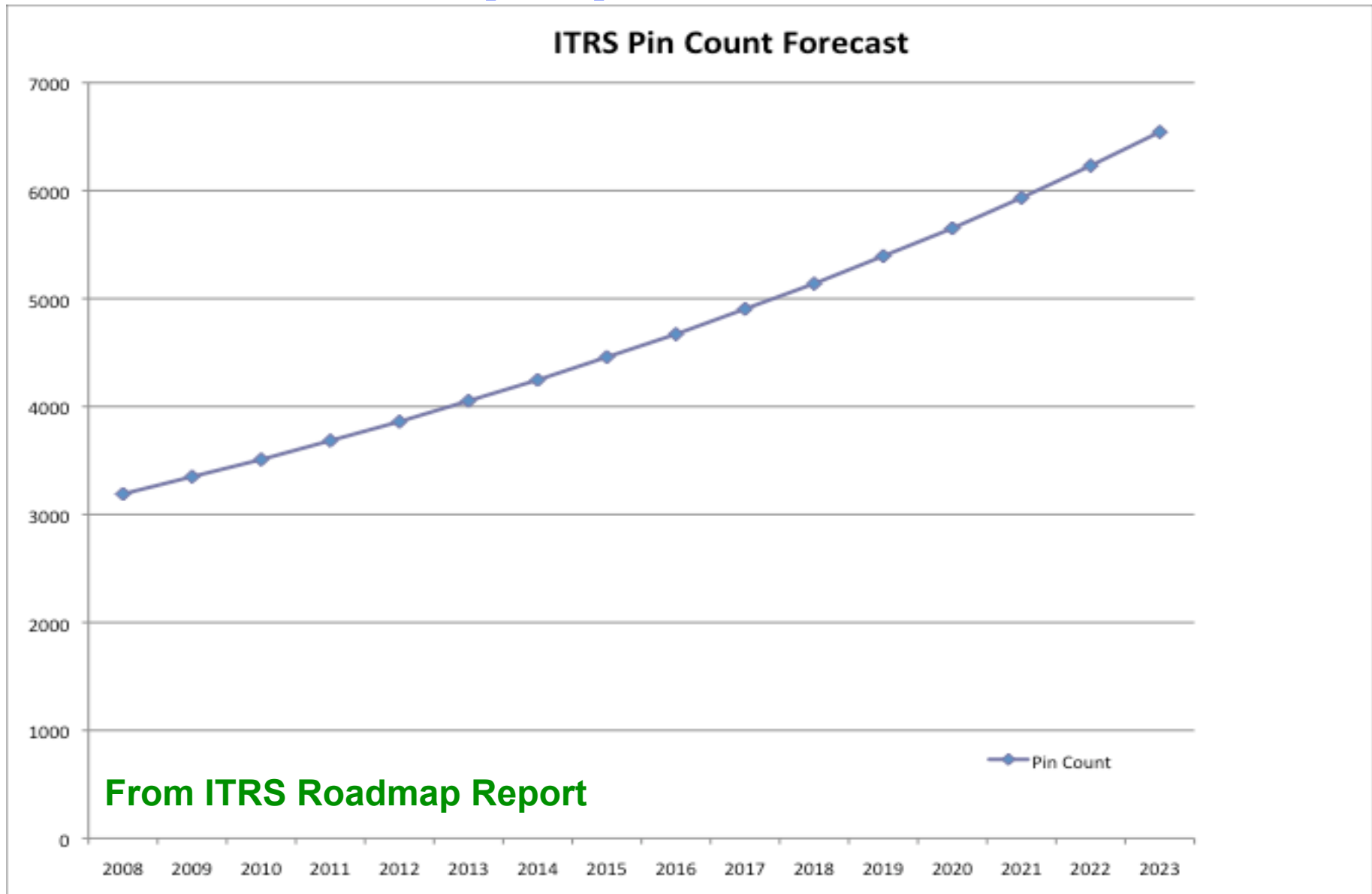


Scalability of Cube-Sphere Dycore

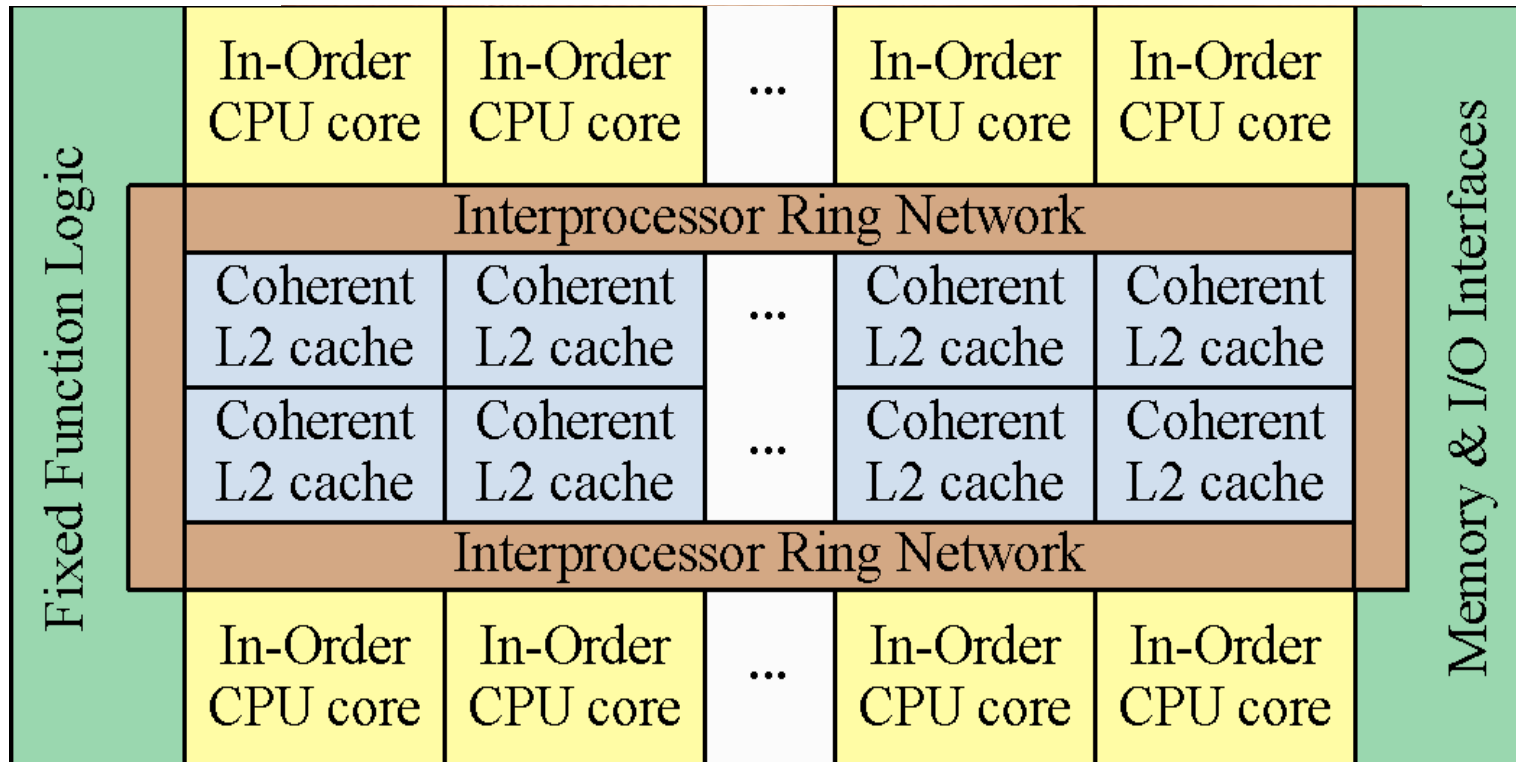


The Next 1000x?

Pin count (BW) is not projected to keep up with core count.



Accelerator Research



- ❑ **Graphics Cards – Nvidia 9800/Cuda**
 - ❑ **Measured 109x on WRF microphysics on 9800GX2**
- ❑ **IBM Cell Processor – 8 cores**
- ❑ **Intel Larrabee**

Computational Intensity (CI)

- ❑ Compute Intensity:

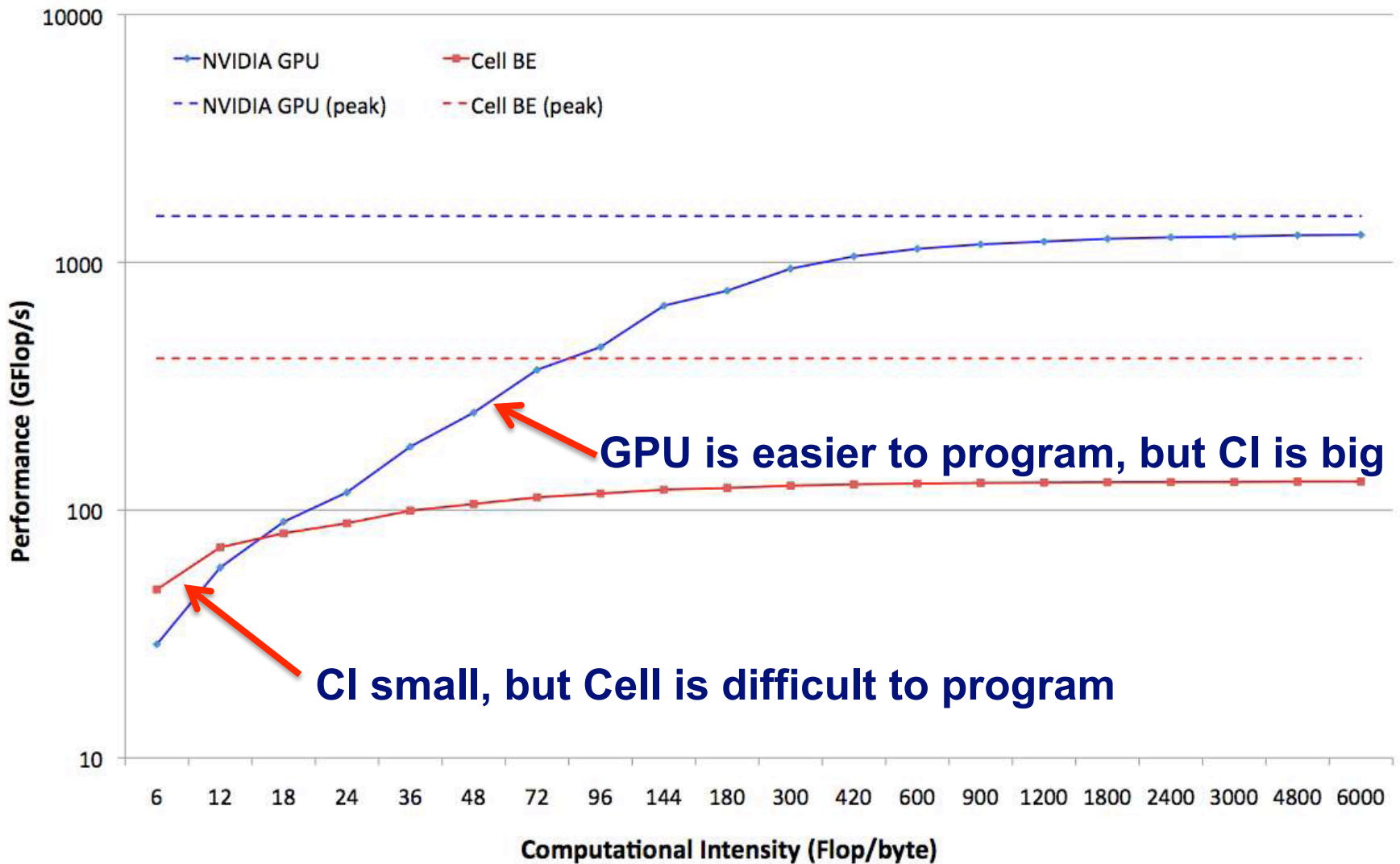
$$CI = \text{Total Operations} / (\text{Input} + \text{Output data})$$

- ❑ $\text{GFLOPS} = CI * \text{Bandwidth}$

- ❑ Bandwidth expensive, flops cheap

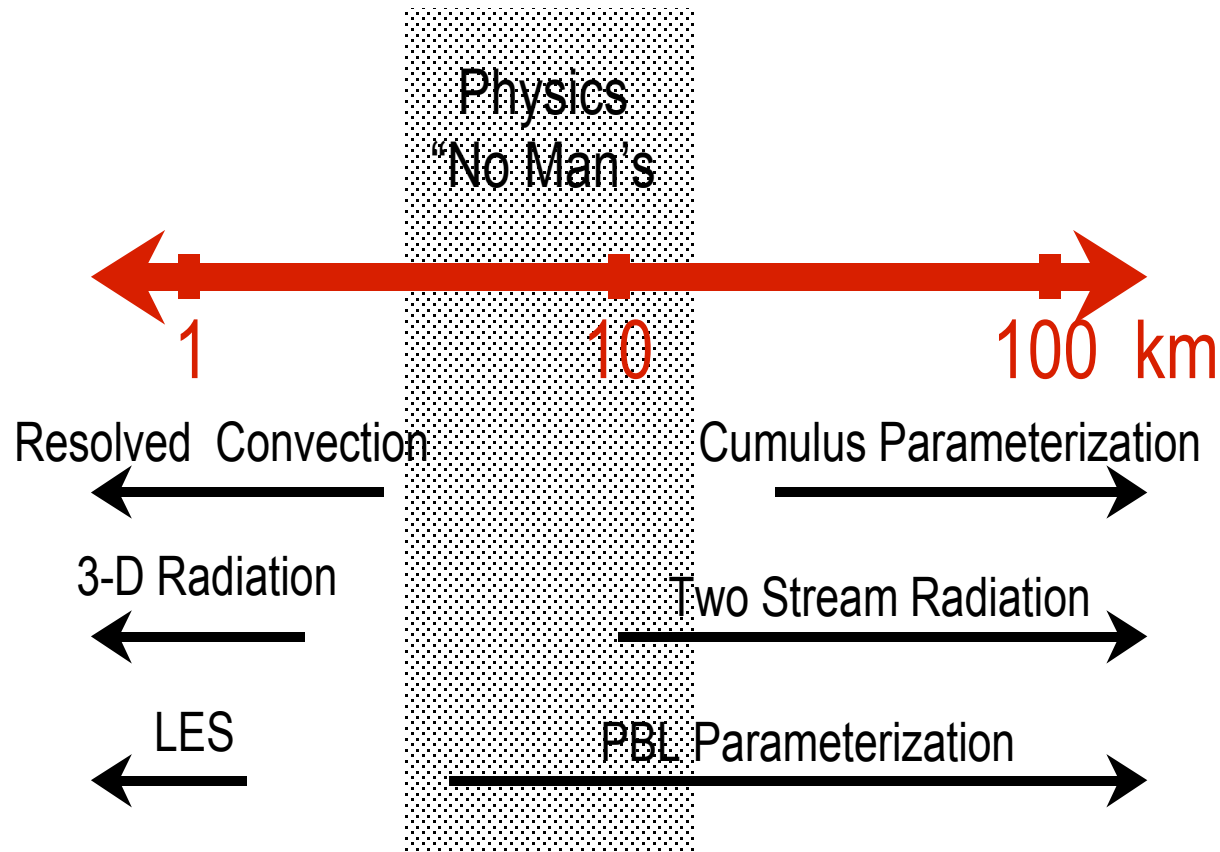
- ❑ The higher the CI, the better we're able to exploit this state of affairs

Where's the tipping point for Cell and GPU's?



Huge “Modeling Gap” Between Current Climate Resolutions and the Cloud Resolving Scale

Challenges in High Resolution Numerical Weather Prediction



Courtesy Joe Klemp, NCAR

[Computational and Information Systems Laboratory](#)
National Center for Atmospheric Research

Super-parameterization:

An intermediate step in the quest for a cloud-system-resolving AGCM

- ❑ The computational cost of the following simulations are approximately the same:
 - ❑ A **millennium**-long simulation using a traditional climate model.
 - ❑ A **few years**-long simulation using a traditional climate model with CRCP
 - ❑ A **day**-long simulation of a cloud-system-resolving AGCM O(few km)

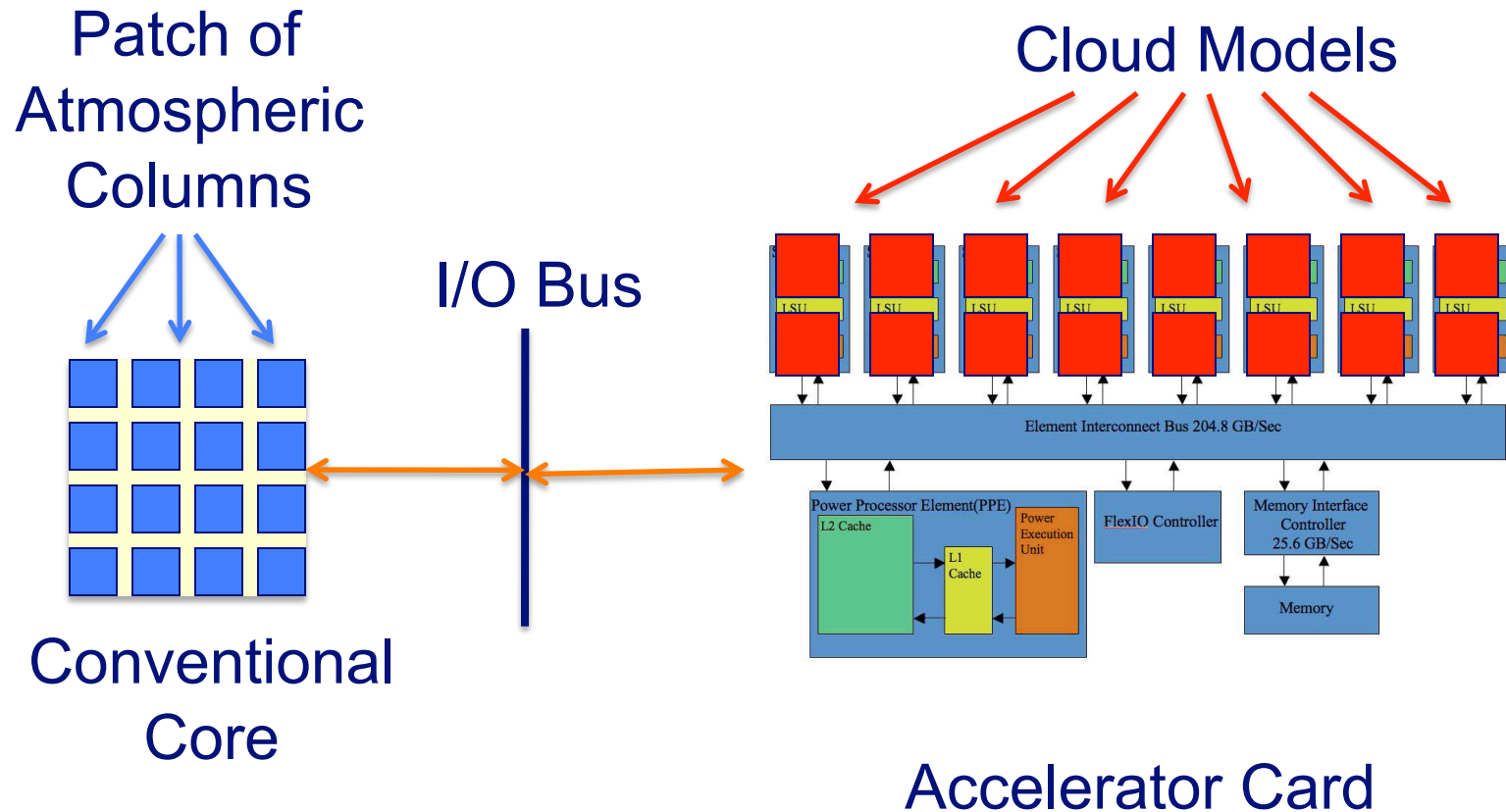
Super-parameterization:

What is “super” about it?

- ❑ The basic idea is to represent cloud dynamics of the sub grid-scales of the AGCM by embedding a 2D or 3D **cloud resolving model in each column** of the large-scale model.
- ❑ Involves thousands of cloud resolving models interacting in a way consistent with large scale dynamics.
- ❑ **Embarrassingly parallel but extremely expensive.** (~150x more expensive than traditional physics)

Running

“Super Parameterization” on an Accelerator Cluster



Conclusions:

Investments Required in

- ❑ People and infrastructure for...
 - ❑ Algorithm Research
 - ❑ Software Engineering
 - ❑ Computational Science Research
- ❑ Architectural innovations including...
 - ❑ Global reduction networks
 - ❑ Robust I/O subsystems
 - ❑ Tighter integration of accelerators and μ procs
- ❑ New modeling schemes, such as...
 - ❑ pairing sub-grid-scale models with accelerators
 - ❑ Parallelize the vertical

What We Need From You (Really!)

- ❑ Keep providing scalable node architectures.
- ❑ Provide programming models, software environment, hardware integration so that we can take advantage of accelerator technologies.
- ❑ Power efficiency – maximize useful flops (better yet, the amount of science accomplished) per unit of energy.
- ❑ Work with us to fundamentally restructure our codes and lower barriers to experimenting with new computational paradigms.

Key Point: Working on Climate Scalability Requires Big Interdisciplinary Teams – e.g.

Contributors:

D. Bader (ORNL)	S. Mishra (NCAR)
D. Bailey (NCAR)	S. Peacock (NCAR)
C. Bitz (U Washington)	K. Lindsay (NCAR)
F. Bryan (NCAR)	W. Lipscomb (LANL)
T. Craig (NCAR)	R. Loy (ANL)
A. St. Cyr (NCAR)	J. Michalakes (NCAR)
J. Dennis (NCAR)	A. Mirin (LLNL)
J. Edwards (IBM)	M. Maltrud (LANL)
B. Fox-Kemper (MIT,CU)	J. McClean (LLNL)
E. Hunke (LANL)	R. Nair (NCAR)
B. Kadlec (CU)	M. Norman (NCSU)
D. Ivanova (LLNL)	T. Qian (NCAR)
E. Jedlicka (ANL)	C. Stan (COLA)
E. Jessup (CU)	M. Taylor (SNL)
R. Jacob (ANL)	H. Tufo (NCAR)
P. Jones (LANL)	M. Vertenstein (NCAR)
J. Kinter (COLA)	P. Worley (ORNL)
	M. Zhang (SUNYSB)

Funding:

- DOE-BER CCPP Program Grant
 - DE-FC03-97ER62402
 - DE-PS02-07ER07-06
 - DE-FC02-07ER64340
 - B&R KP1206000
- DOE-ASCR
 - B&R KJ0101030
- NSF Cooperative Grant NSF01
- NSF PetaApps Award

Computer Time:

- Blue Gene/L time:
 - NSF MRI Grant
 - NCAR
 - University of Colorado
 - IBM (SUR) program
 - BGW Consortium Days
 - IBM research (Watson)
 - LLNL
 - Stony Brook & BNL
- CRAY XT time:
 - NICS/ORNL
 - NERSC
 - Sandia

A winter scene featuring snow-covered evergreen trees in the foreground and middle ground. A wooden fence is visible in the lower center. In the background, a mountain peak is visible under a clear blue sky with a bright sun or moon. The text "Thanks! Any Questions?" is overlaid in the center in a bold, blue font.

Thanks!
Any Questions?