Refactoring CESM for Exascale

Dr. Richard Loft
Director, Technology Development
Computational and Information Systems Laboratory
National Center for Atmospheric Research

IS-ENES Conference
Hamburg, Germany
March 17, 2014

CISL Computational & Information Systems Laboratory

NCAR

Outline

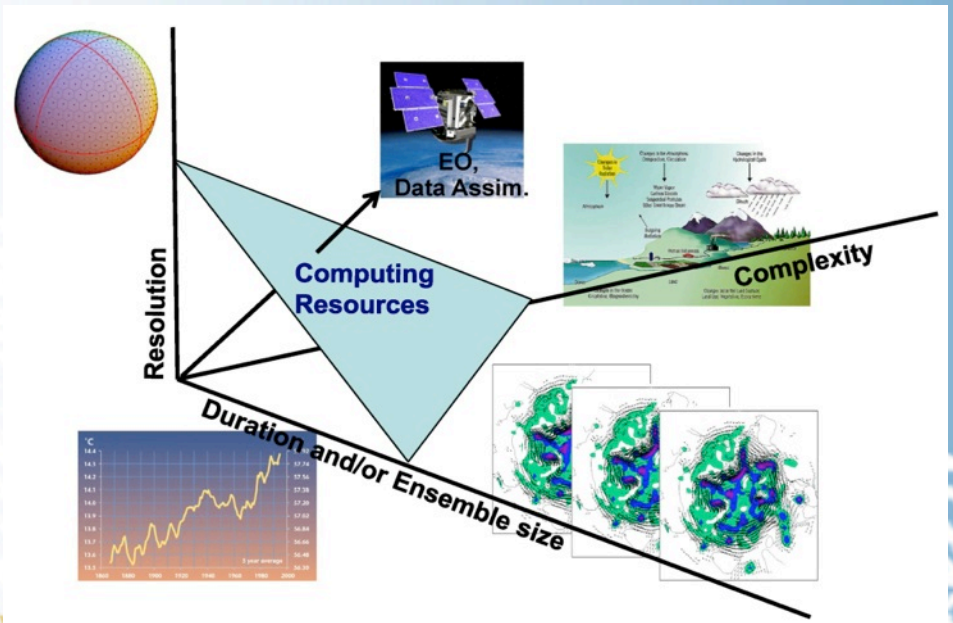
- **What will we run at exascale?**
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- **NCAR Exascale strategy**
  - Computational optimizations
  - Data optimization
- **Summary**

What will we run at exascale?

- **Single ultra high-resolution models**
  - Capture convective-scale processes
  - Non-hydrostatic; 1 km resolution
  - Most vulnerable to a system component failure
- **Climate system data assimilation**
  - Study predictability on seasonal to decadal scales
  - EKF scalable but very data-intensive: will stress memory – I/O subsystem
  - EKF can be made fault resilient

What will we run at exascale?

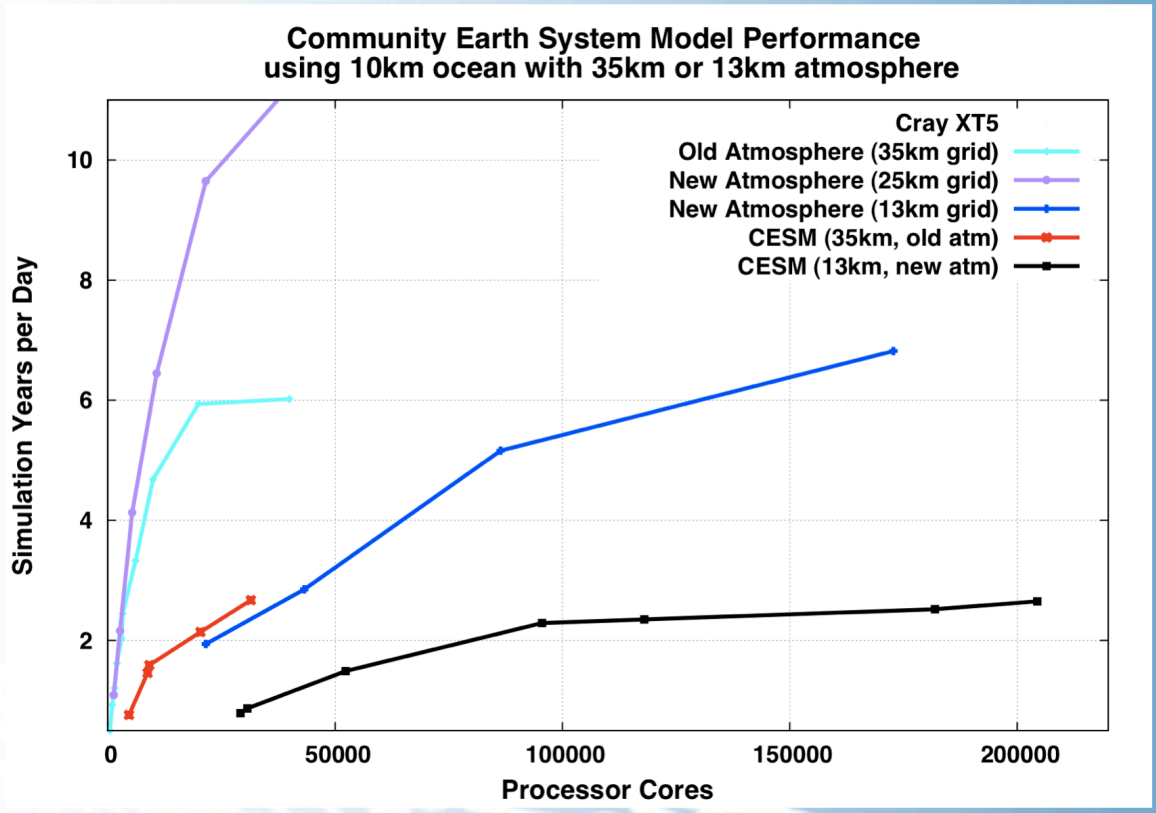
- **Very large ensembles of "low-res" models**
  - Study natural variability and extremes
  - High I/O volumes – I/O byte/flop goes up for low-res
  - Fault tolerant – resubmit ensemble members that fail.



Outline

- **What will we run at exascale?**
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- **NCAR Exascale Strategy**
  - Computational optimizations
  - Data optimization
- **Summary**

CESM Computational Performance
(courtesy of Pat Worley)

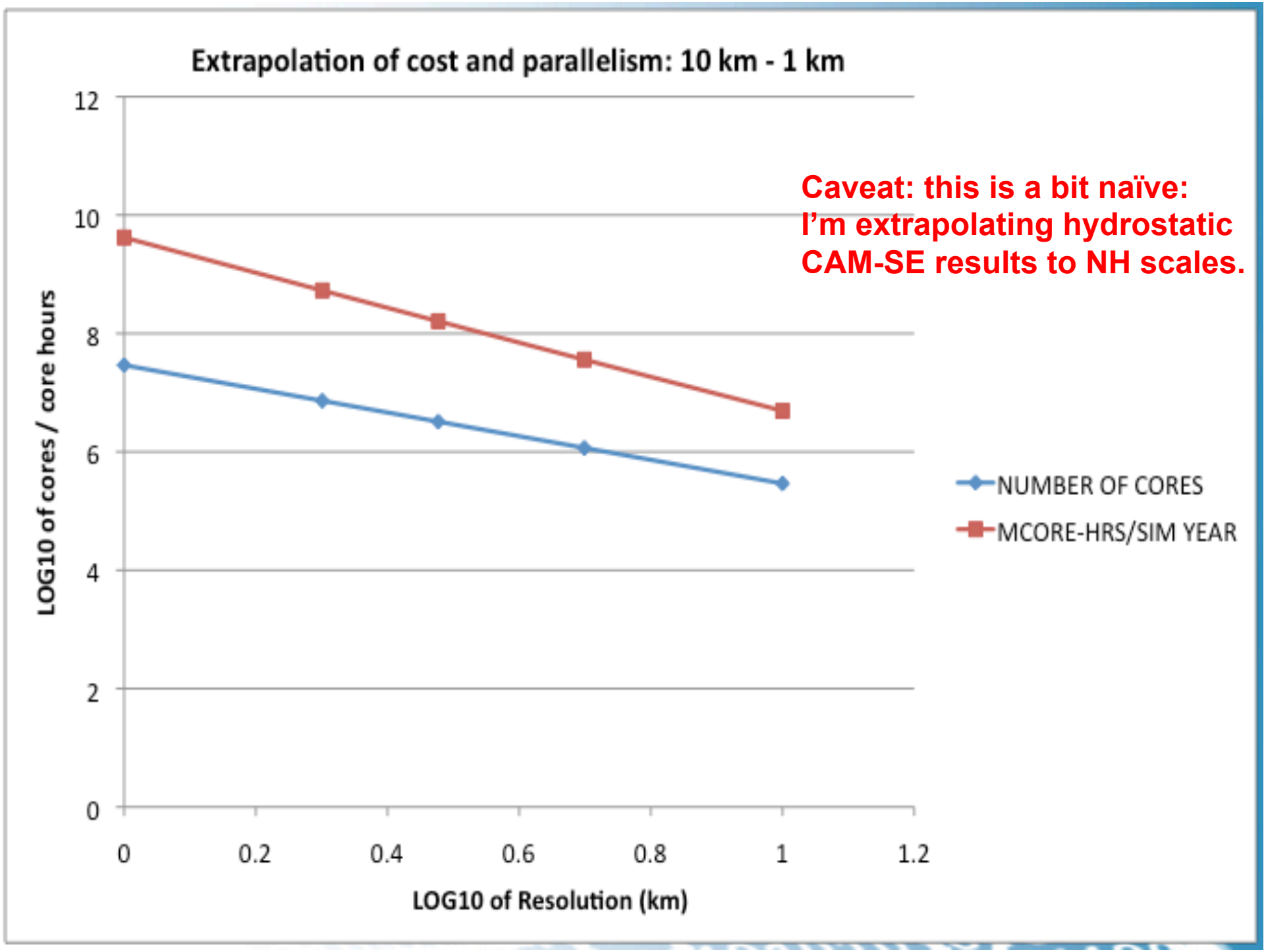


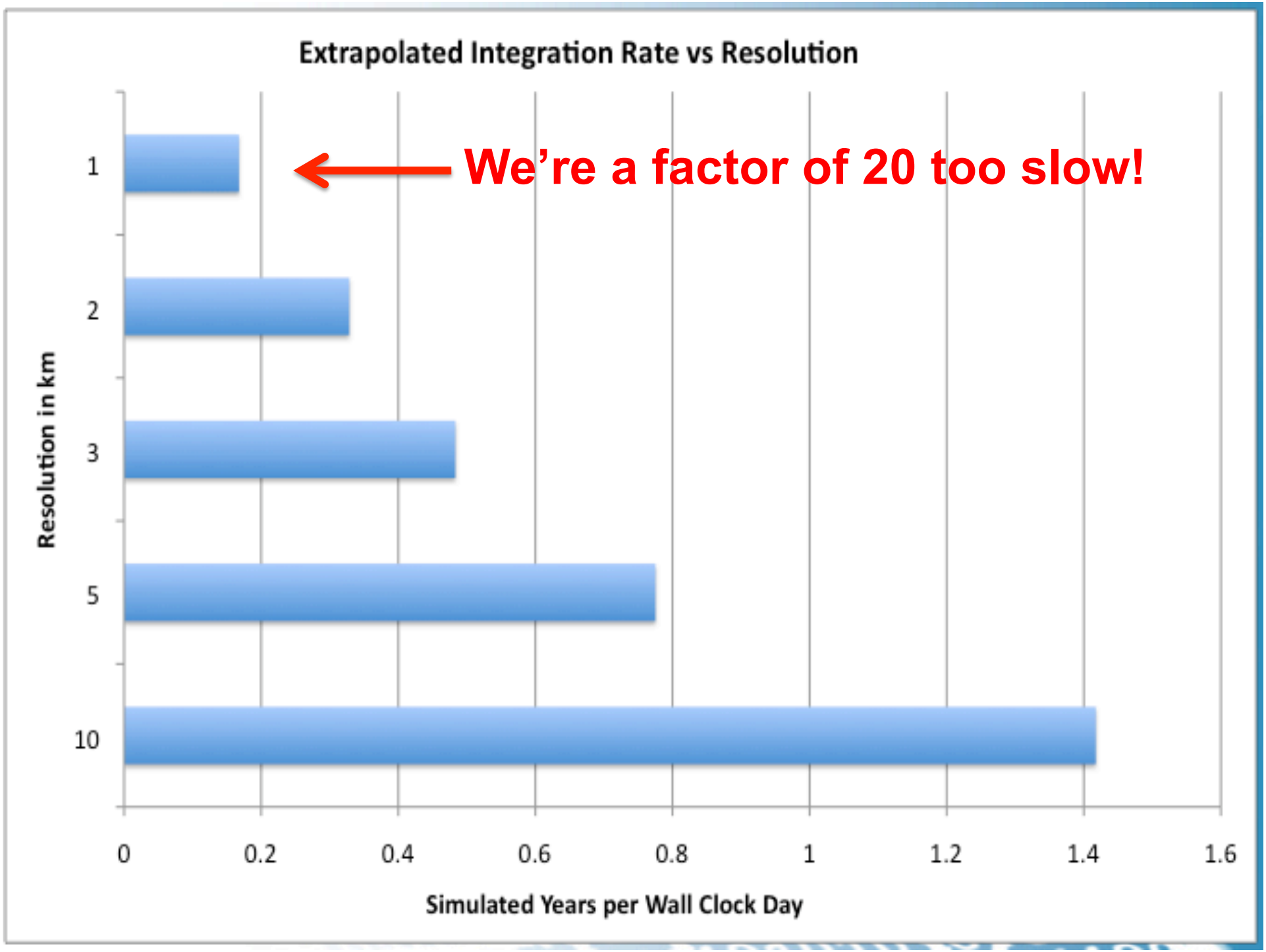
JaguarPE

For 35km CAM / 10km POP/CICE, CESM constrained by CAM, CICE and POP
For 13km CAM / 10KM POP/CICE, CESM constrained by CAM
Spectral element-based atmospheric dynamics permits scalable CESM performance at high resolution.

CISL Computational & Information Systems Laboratory

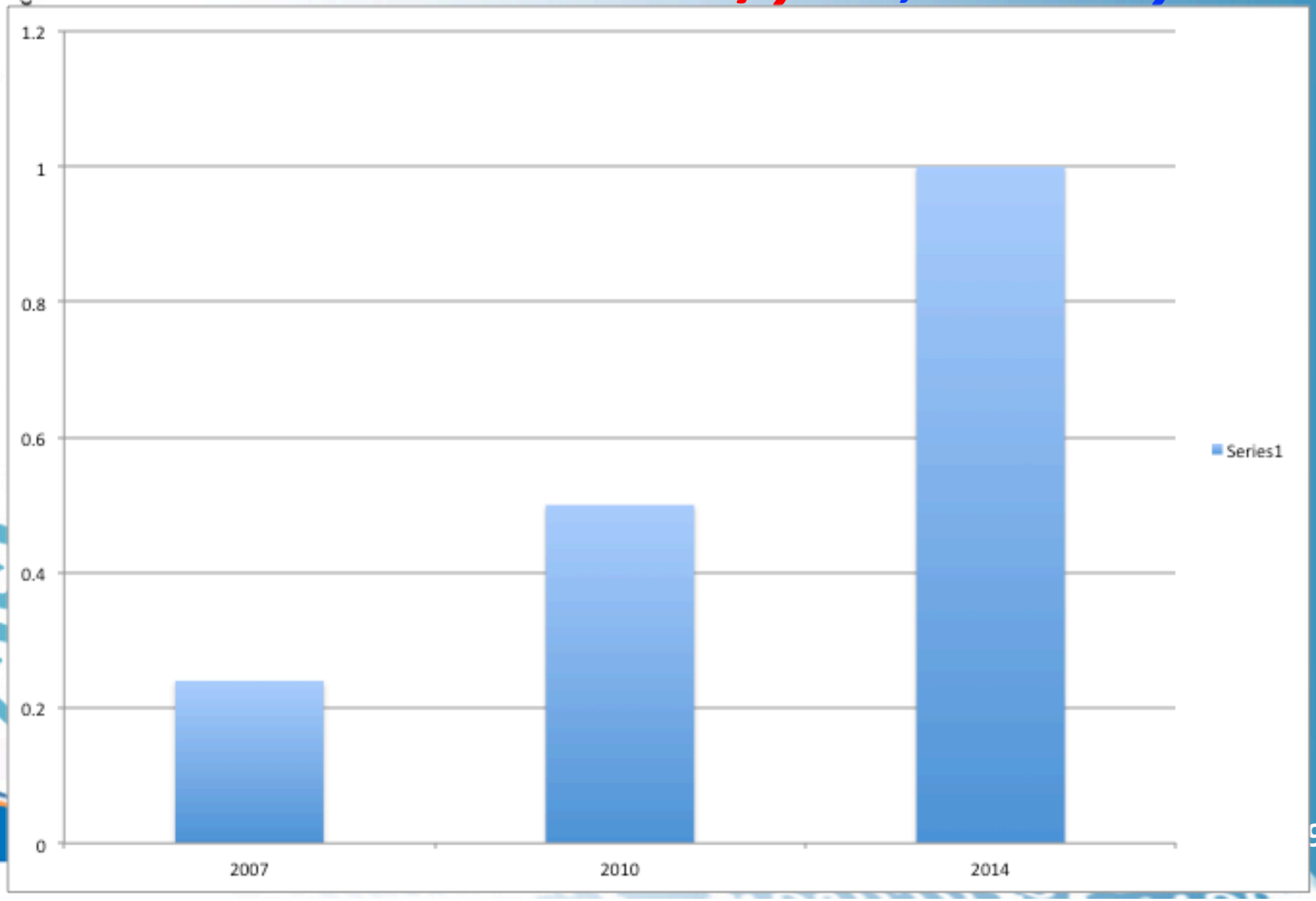
NCAR


Extrapolation of cost and parallelism: 10 km - 1 km
LOG10 of cores / core hours
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
1.2
LOG10 of Resolution (km)
NUMBER OF CORES
MCORE-HRS/SIM YEAR
Caveat: this is a bit naïve:
I'm extrapolating hydrostatic
CAM-SE results to NH scales.

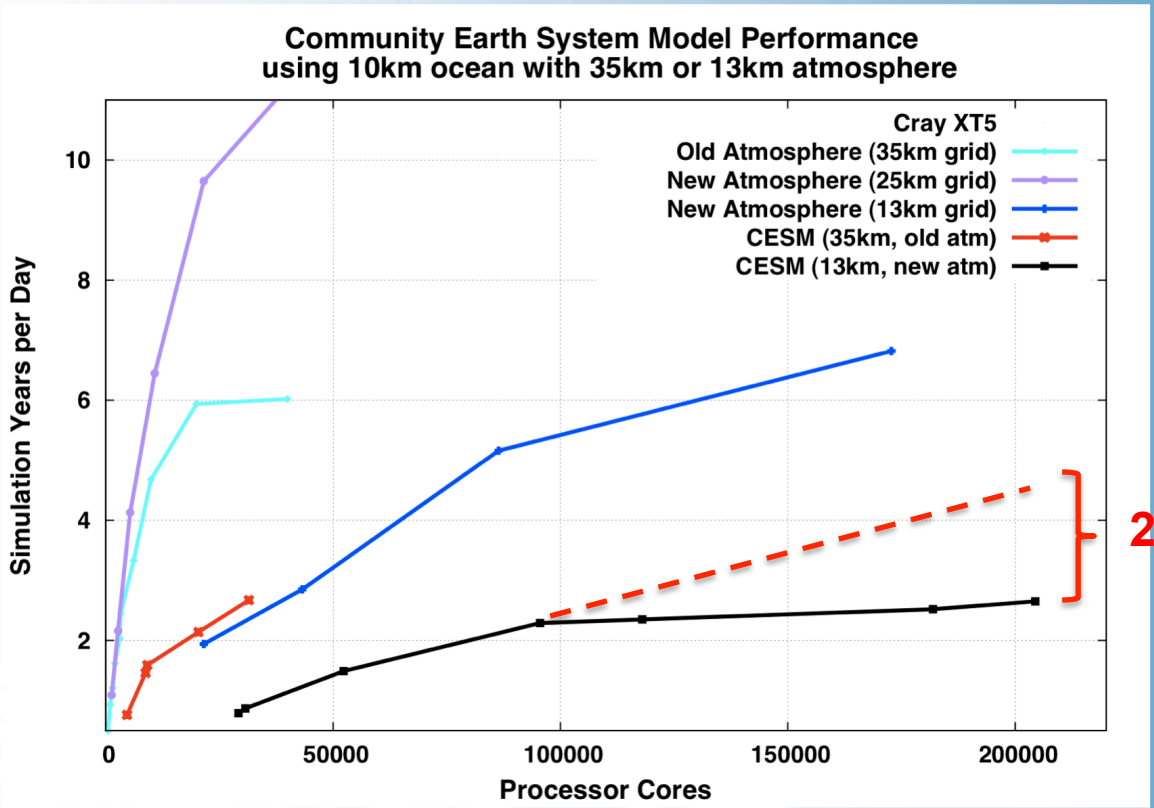

Extrapolated Integration Rate vs Resolution
We're a factor of 20 too slow!
Resolution in km
1
2
3
5
10
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Simulated Years per Wall Clock Day

Rate of improvement of memory bandwidth

IBM POWER series: 22.6%/year; ~3.4x by 2020



CESM Computational Performance
(courtesy of Pat Worley)



Spectral element-based atmospheric dynamics permits scalable CESM performance at high resolution.

Further improvements will require optimization of ocean and sea ice models

CISL Computational & Information Systems Laboratory

NCAR

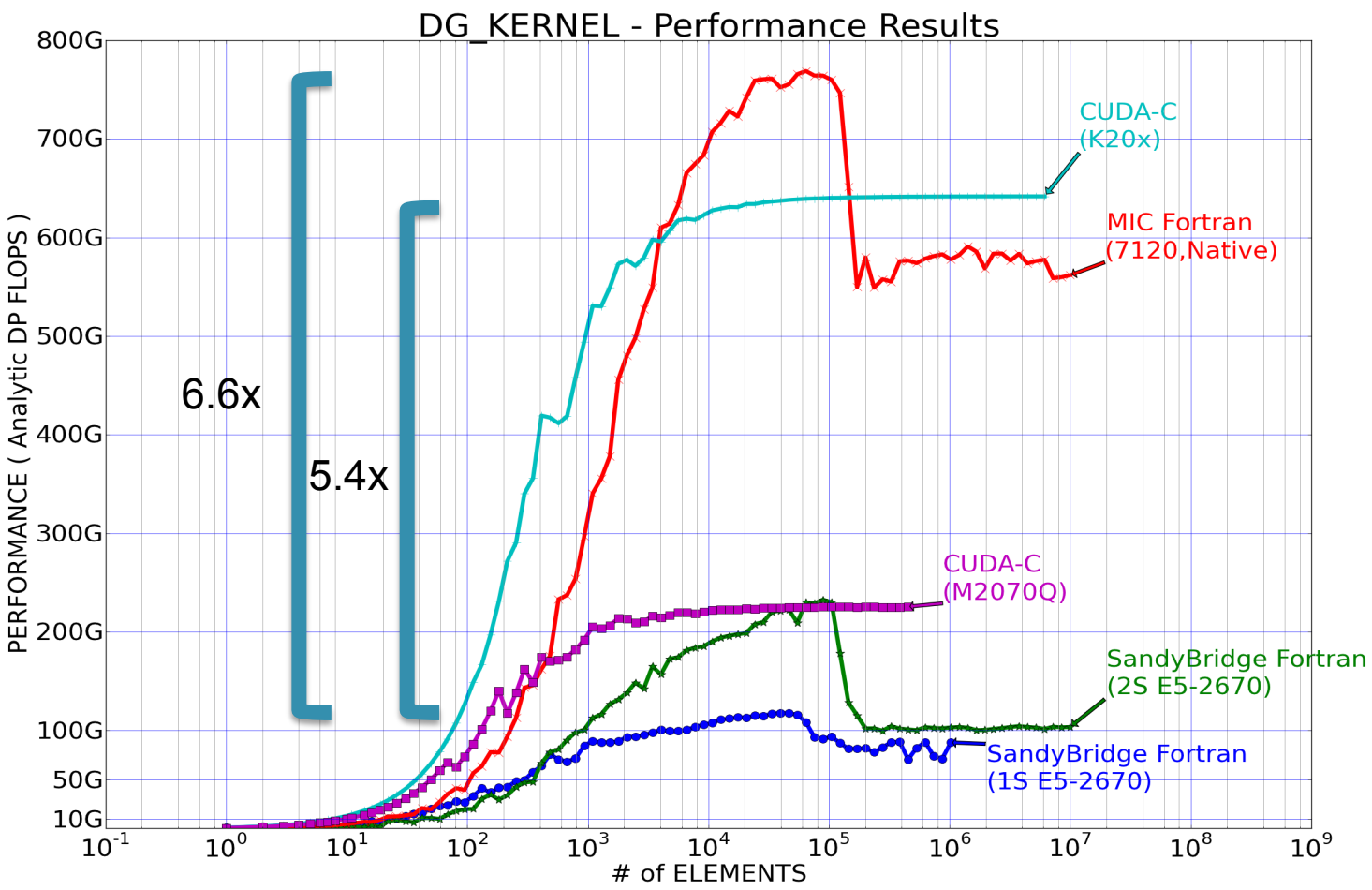
Can accelerators fill in the missing performance 3-4 x?

Discontinuous Galerkin Gradient Kernel

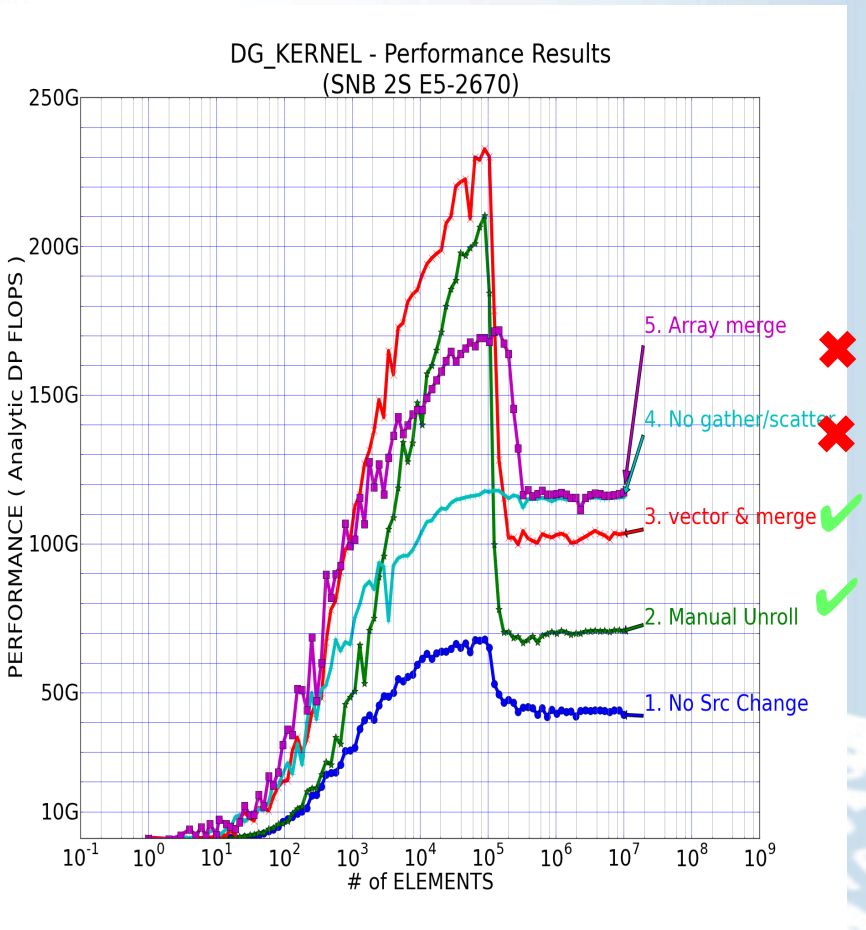
- **Analogous to vector calculus calculation in the atmospheric dynamics of CAM-SE component of CESM**
- **DG and CG both highly scalable, element-based method***
- **Procedure for N elements:**
  - Each 4x4 element contains a scalar quantity
  - Do two 4x4 double precision matrix multiplies of a derivative matrix with the element and its transpose to get x and y components of gradient
  - Use gradients to update flux vectors
  - Repeat

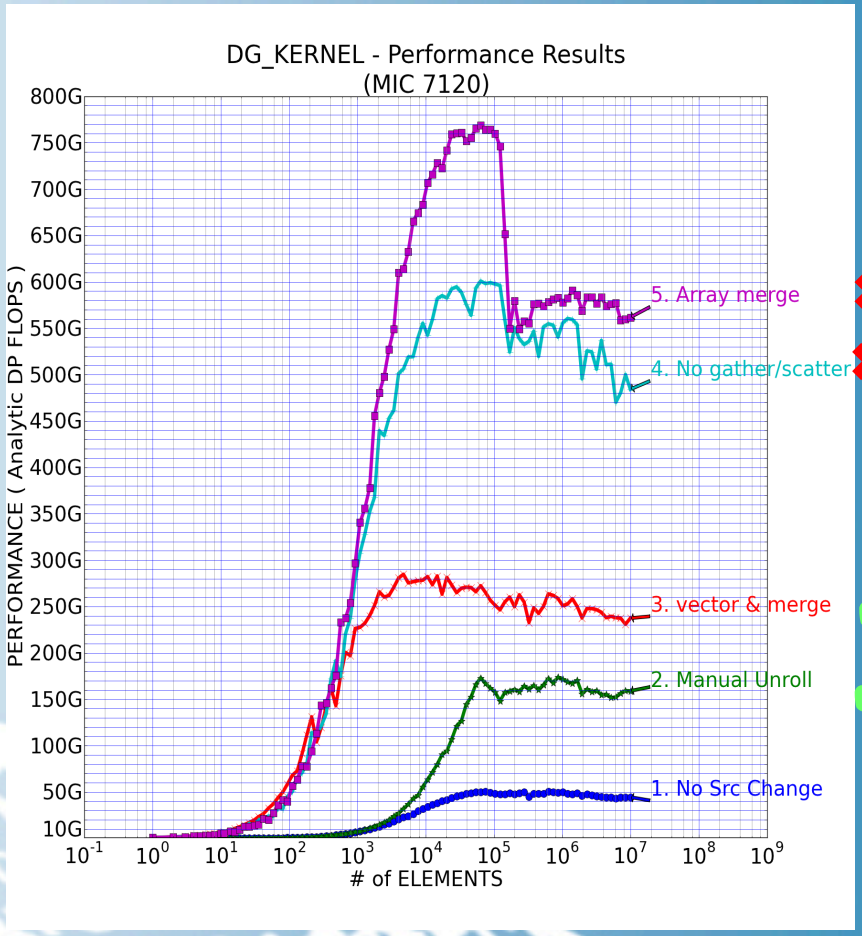
* R. D. Nair, Stephen J. Thomas, and Richard D. Loft: A discontinuous Galerkin global shallow water model, Monthly Weather Review, Vol. 133, pp 876-888

The best performance results from CPU, GPU, and MIC



Optimization portability between Xeon and Xeon Phi





Generally, performance tuning on a micro-architecture also helps to improve performance on another micro-architecture. However, it is not always true.

NCAR

How do you get beyond kernels?

- **Lots of code (1.5M lines in CESM)**
  - Single source multiple architectures
  - portability & performance resilience
- **Code is rapidly changing**
  - Optimization cycle of 1 month – OK
  - Optimization cycle of 6 months – start over
- **Verification/validation requirements**
  - Most restrictive in climate but some sort of software verification/science validation is always required.

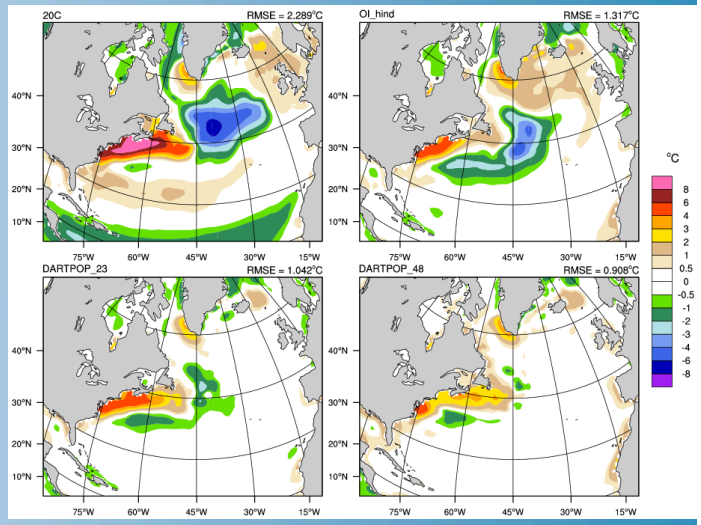
Outline

- **What will we run at exascale?**
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- **NCAR Exascale Strategy**
  - Computational optimizations
  - Data optimization
- **Summary**

What about Data Assimilation Using Ensemble Kalman Filter Methods?

Decadal prediction problem

- Fully coupled CESM
- 10 km resolution
- 50 Ensembles members
- 10 Different Start dates
- 10 Years in length
- State initialized using the Data Assimilation Research Testbed (DART)



Ensemble Ocean Reanalysis IPCC Decadal Forecast ICs

50 members x 200,000 cores = 10 million cores

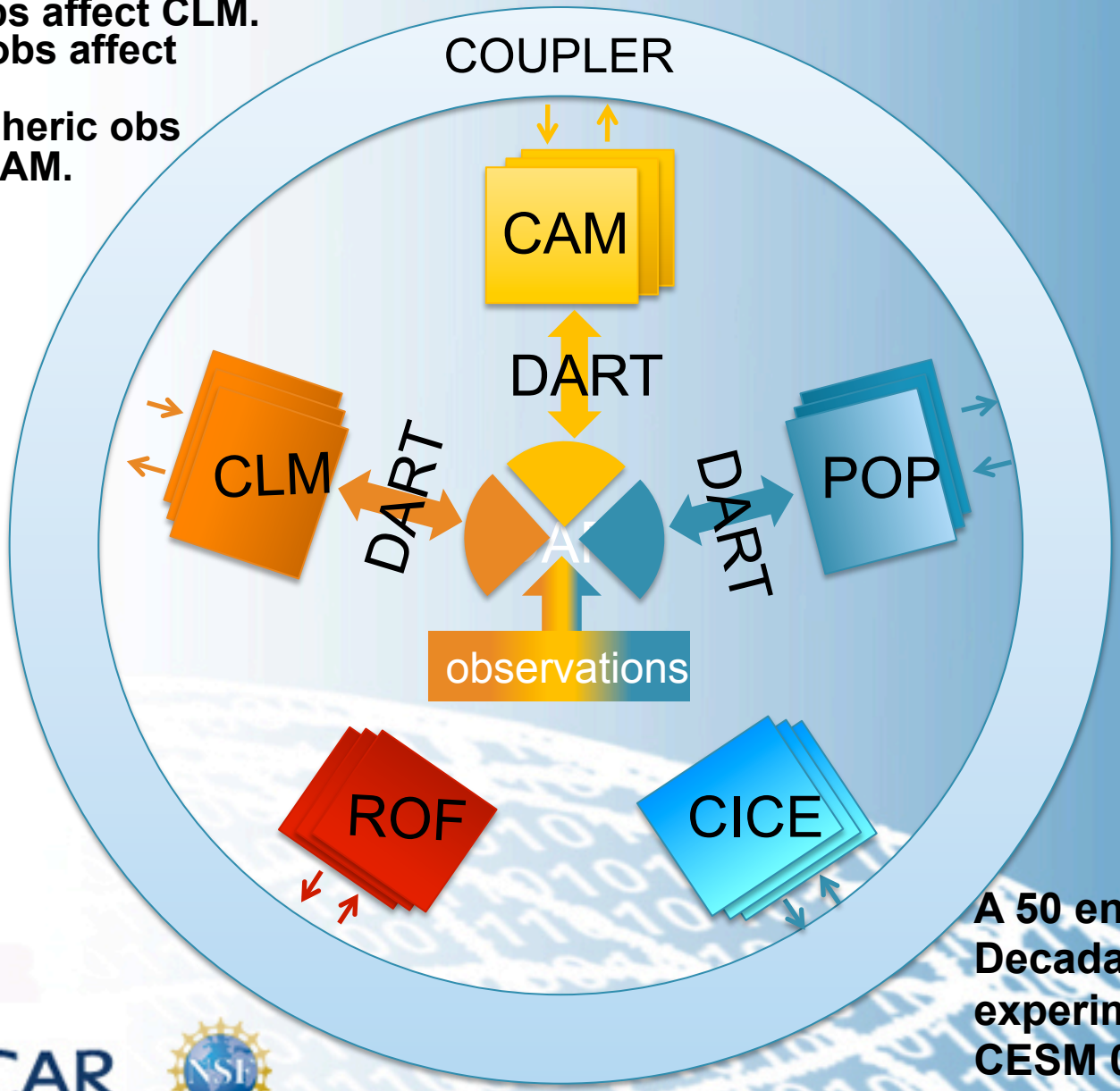
Computational & Information Systems Laboratory

CISL

NCAR

Climate Data Assimilation with DART…

- Land obs affect CLM.
- Ocean obs affect POP.
- Atmospheric obs affect CAM.

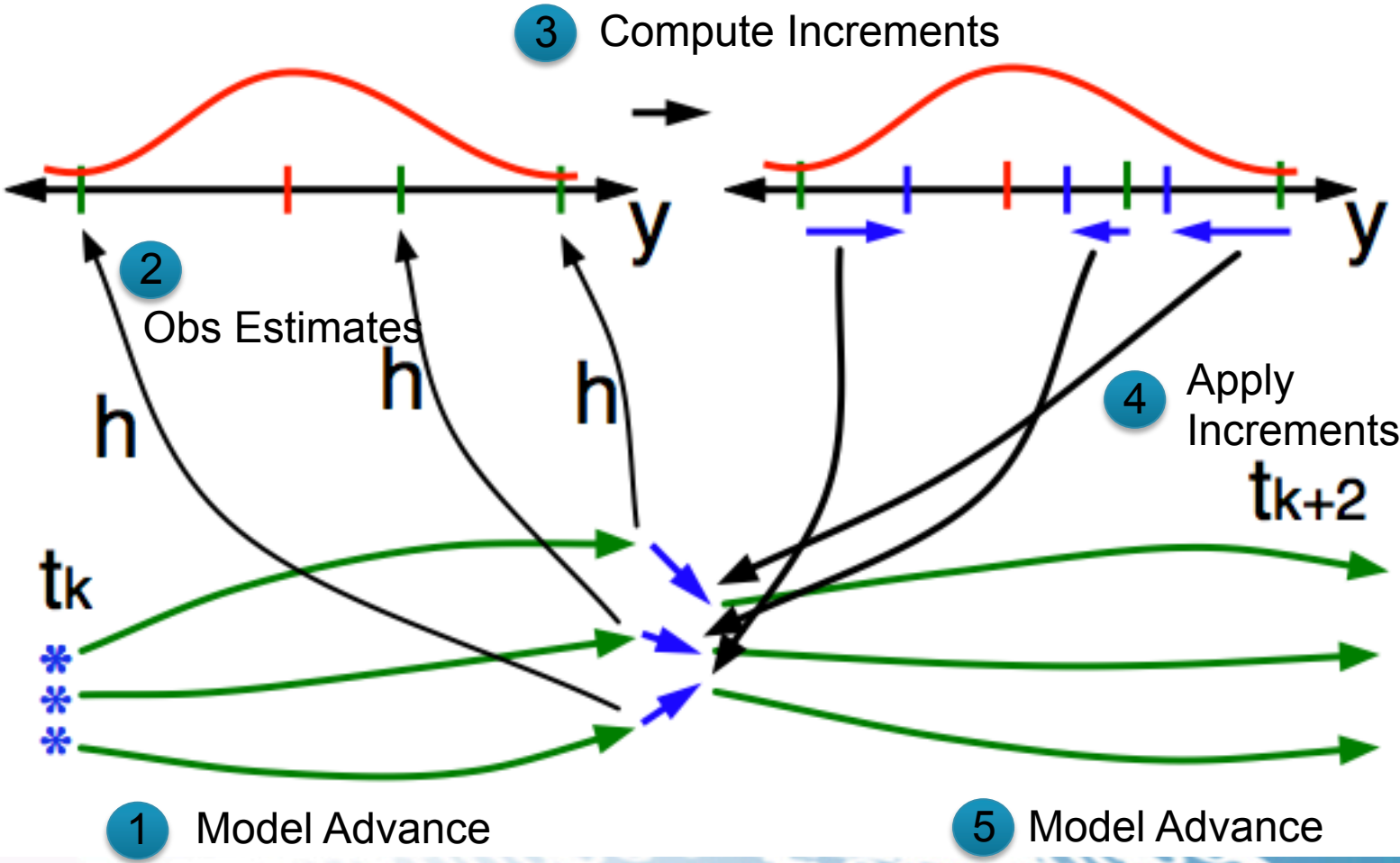


A 50 ensemble member Decadal prediction experiment with CESM 0.1° needs a scalable DART

CISL Computational & Information Systems Laboratory

NCAR

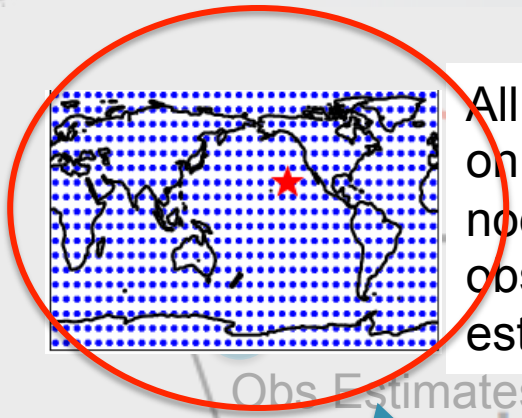
DART's EKF Data Assimilation



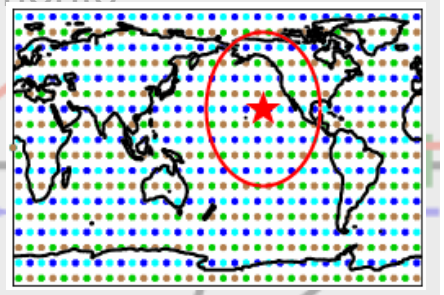
EKF Scalability Challenges

- **Data Distribution**
  - Observations non-uniformly distributed on globe
  - Load imbalances
- **Data Movement**
  - At least 3 distinct data decompositions in the current software implementation
  - Message passing costs
- **Data Volume**
  - Running 50-100 copies of a large model means 50-100x the data input and output
  - Each 10 km instance = O(10) TB/simulated year
  - 10 simulated years x 50 member ensembles = O(5) PB/decadal forecast

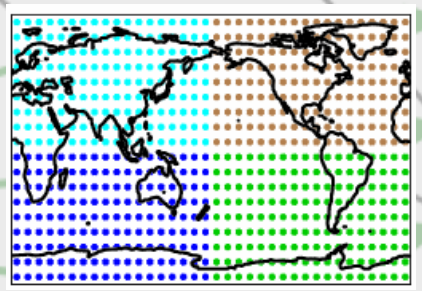
Current DART Data Decompositions



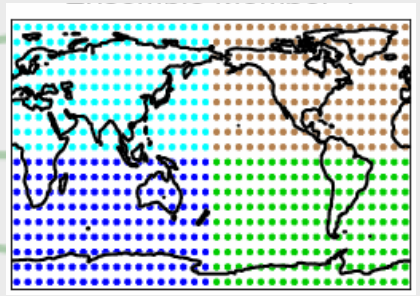
All data on same node for observation estimates



Interleave ensemble member data for load balancing



Spatially local data decomp

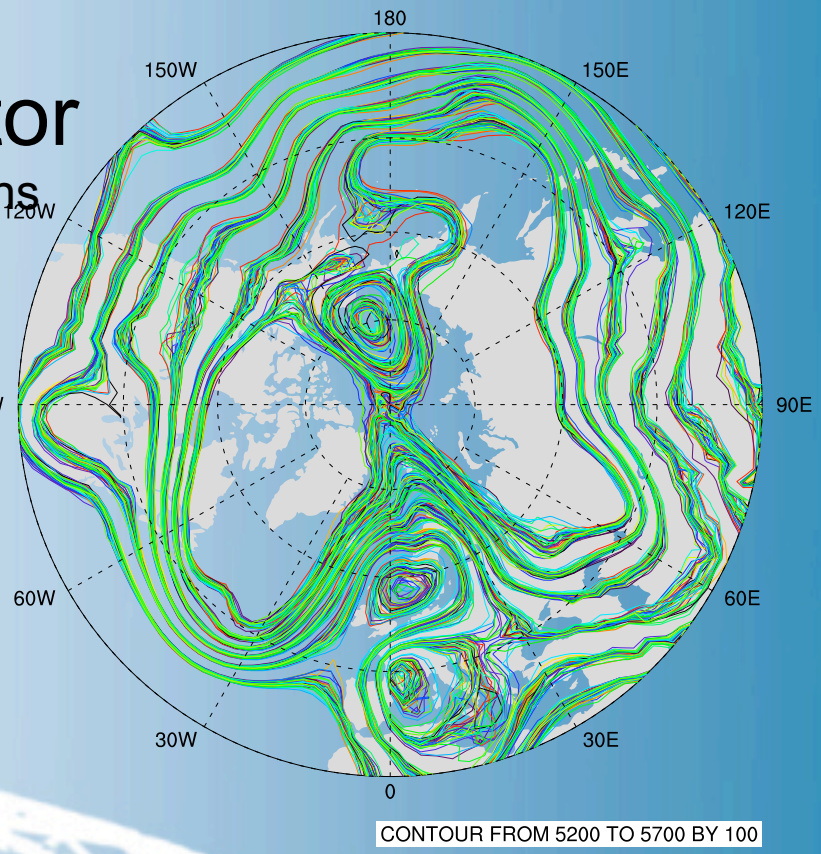


DART scalabilty needs work

CAM FV forward operator

Specific humidity only : 23 090 observations

processors	512	4096
DART (original)	1.01s	0.96s



WRF forward operator

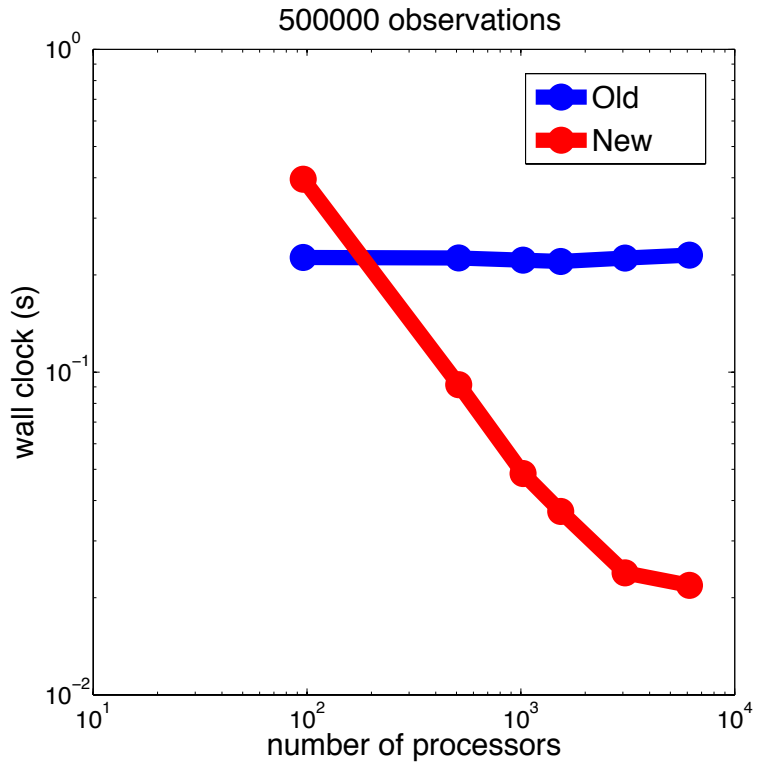
54, 400 observations

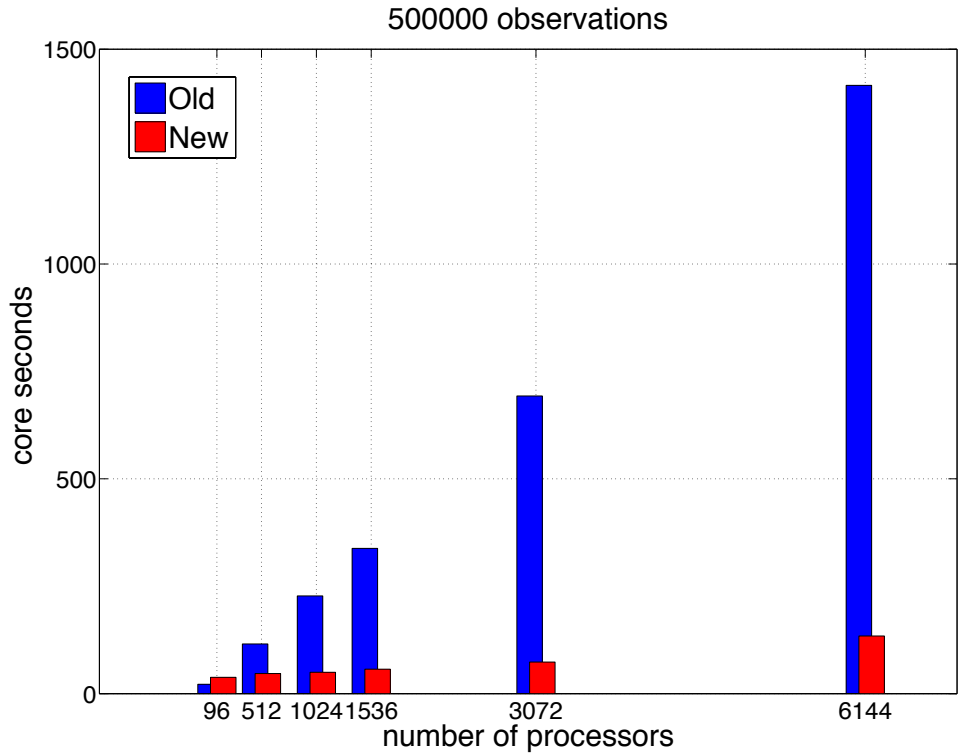
processors	1024	4096
DART (original)	0.6s	0.6s

CISL Computational & Information Systems Laboratory

NCAR

DART: Parallelization of fwd operator: Lorenz_96 test case



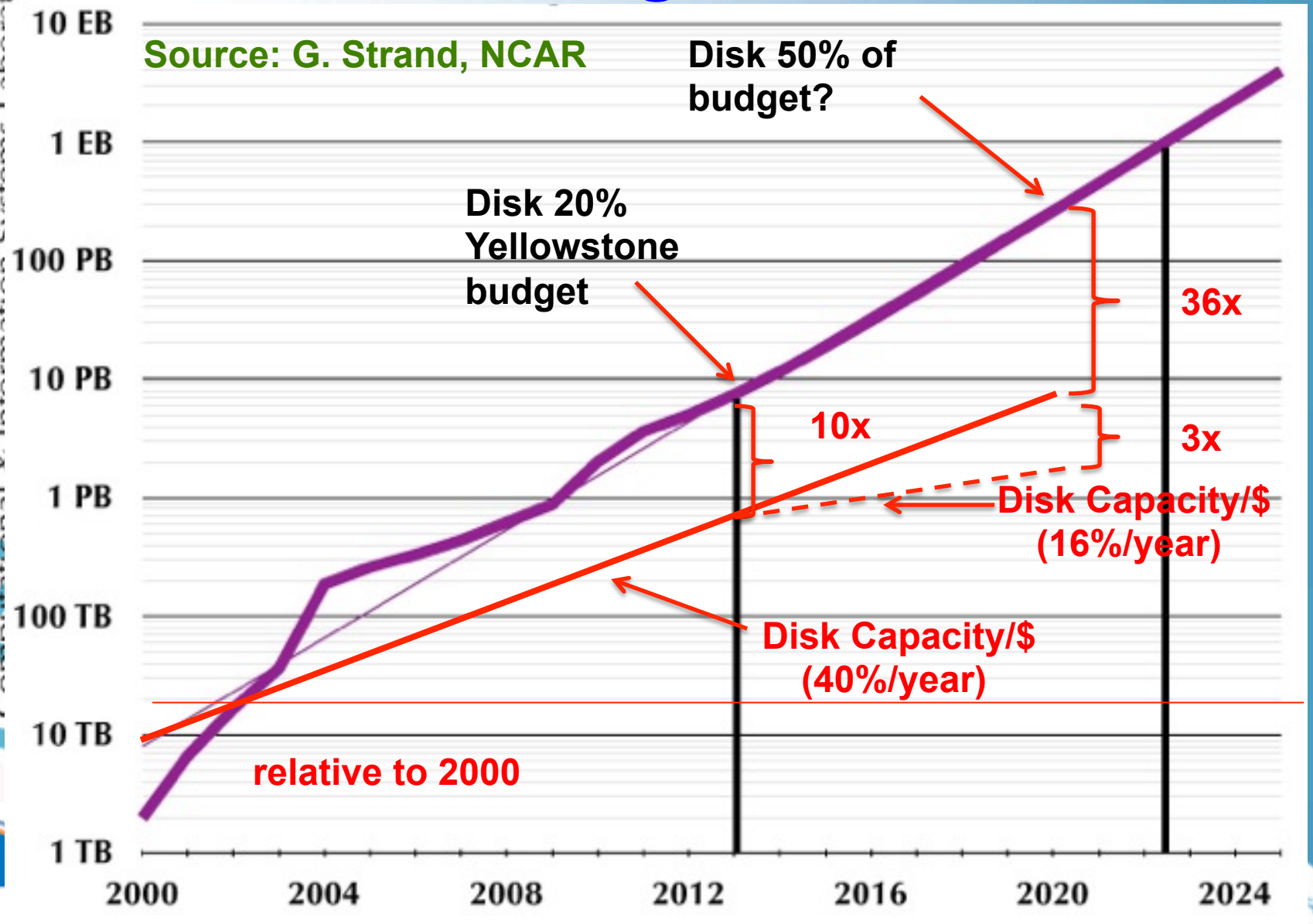


Time **Cost**

Outline

- What will we run at exascale?
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- NCAR Ezascale Strategy
  - Computational optimizations
  - Data optimization
- Summary

Historical and Projected Volume of CESM Modeling Data: shortfalls



Outline

- What will we run at exascale?
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- **NCAR Exascale Strategy**
  - NCAR computational optimizations
  - Data optimization
- Summary

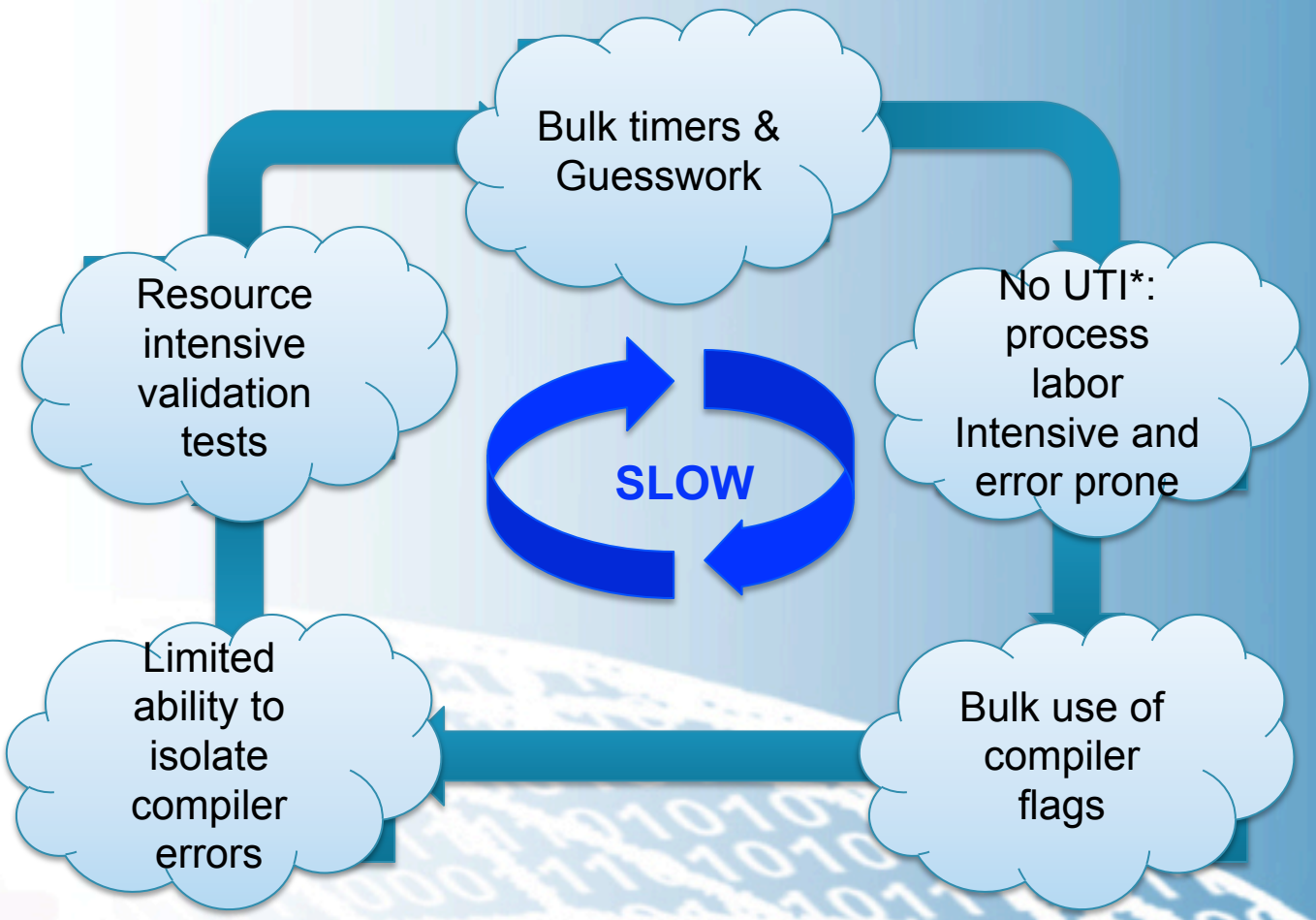
Strategy Overview

- **Investments**
  - 4 new staff in last 18 months. More needed!
- **Partners**
  - DoE (long-standing)
  - **G8 ECS project** (3 years)
  - **Intel Parallel Computing Center** (new)
- **Workshops/Meetings**
  - NCAR workshop on "Programming weather, climate, and earth-system models on heterogeneous multi-core platforms"
- **Platforms**
  - Locally: 16-node Caldera (NV), Pronghorn (Xeon Phi)
  - U.S.: Titan, Stampede, etc…

Outline

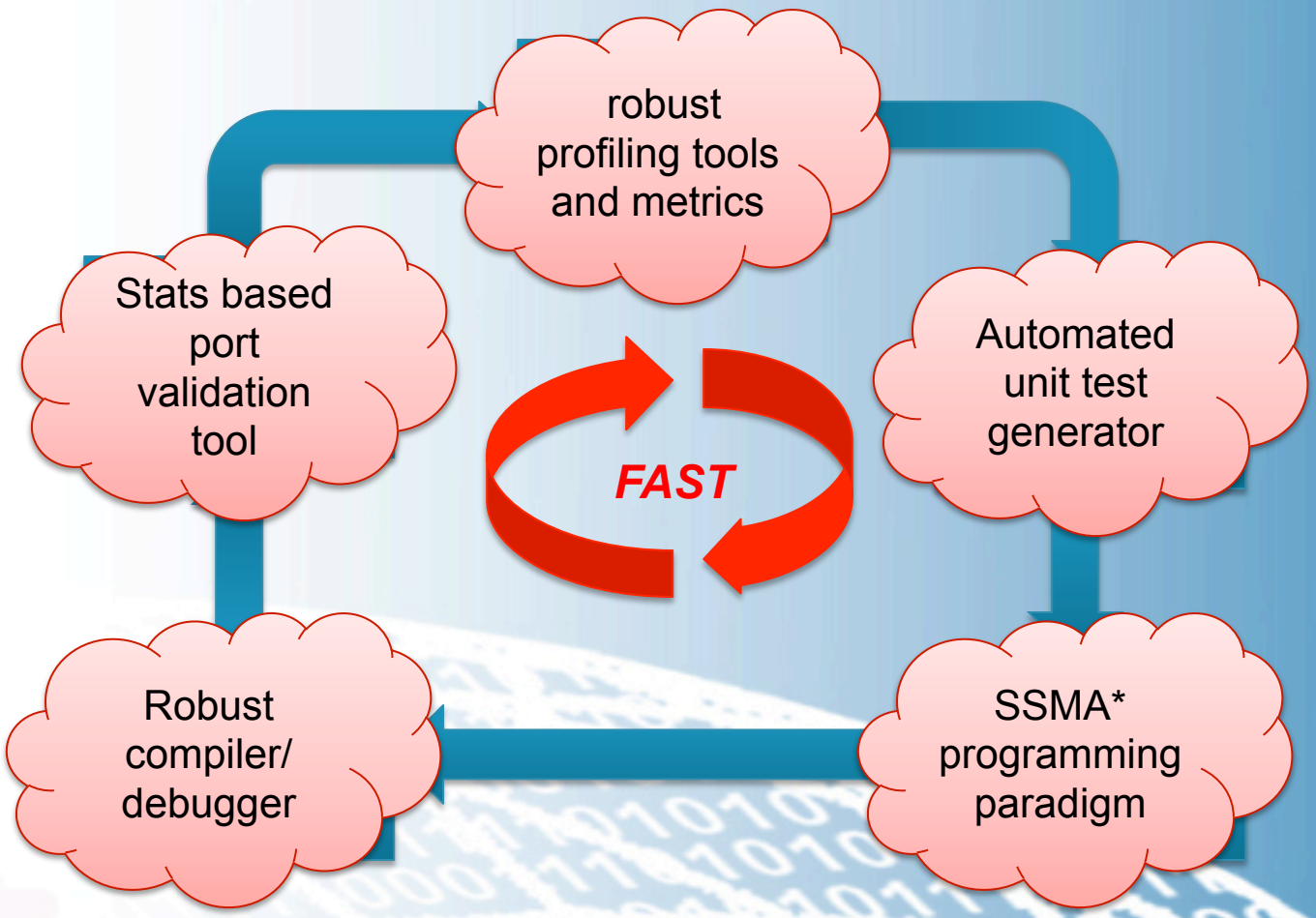
- What will we run at exascale?
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- NCAR Exascale Strategy
  - Computational optimizations
  - Data optimization
- Summary

Life without a Performance Enhancement Methodology



* unit test infrastructure

Performance Enhancement Methodology: a virtuous cycle for code improvement



Computational & Information Systems Laboratory

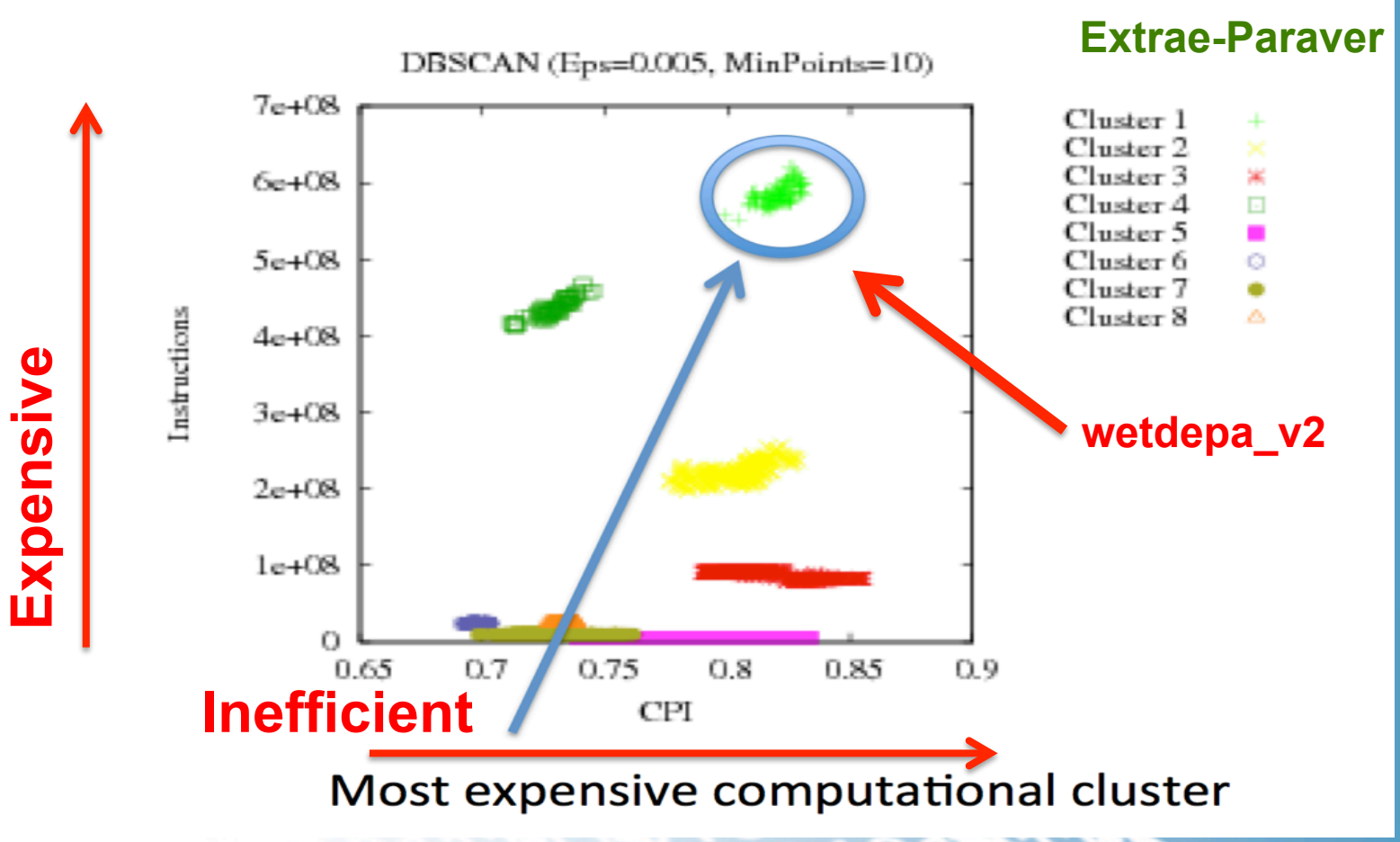
CISL

NCAR

*single source multiple architecture

BSC tools helping to find high priority sections that are expensive *and* inefficient.

Extrae-Paraver



- Result of an Extrae trace of CESM on Yellowstone.
- Vertical ~ time; horizontal ~ 1/flops

Now use what we've learned from dg_kernel to speed up wetdepa_v2… to get fast and right!

Single core

	Intel Phi (Intel 13.1.1)			Intel Sandybridge (Intel 13.1.2)		
	-O2	-O3	-O3 -fast	-O2	-O3	-O3 -fast
orig	42.85	41.24	3.74	3.43	3.32	0.97
mod	6.50	6.61	4.58	1.09	1.12	1.04

- wetdepa is small only ~600 lines
- Restructured branched loops + promoted scalars to vectors.
- -O3 fast for original code gave incorrect results
- 2.5% to 0.7% of code execution time = $222K savings

Significant gains possible from code refactoring!

Outline

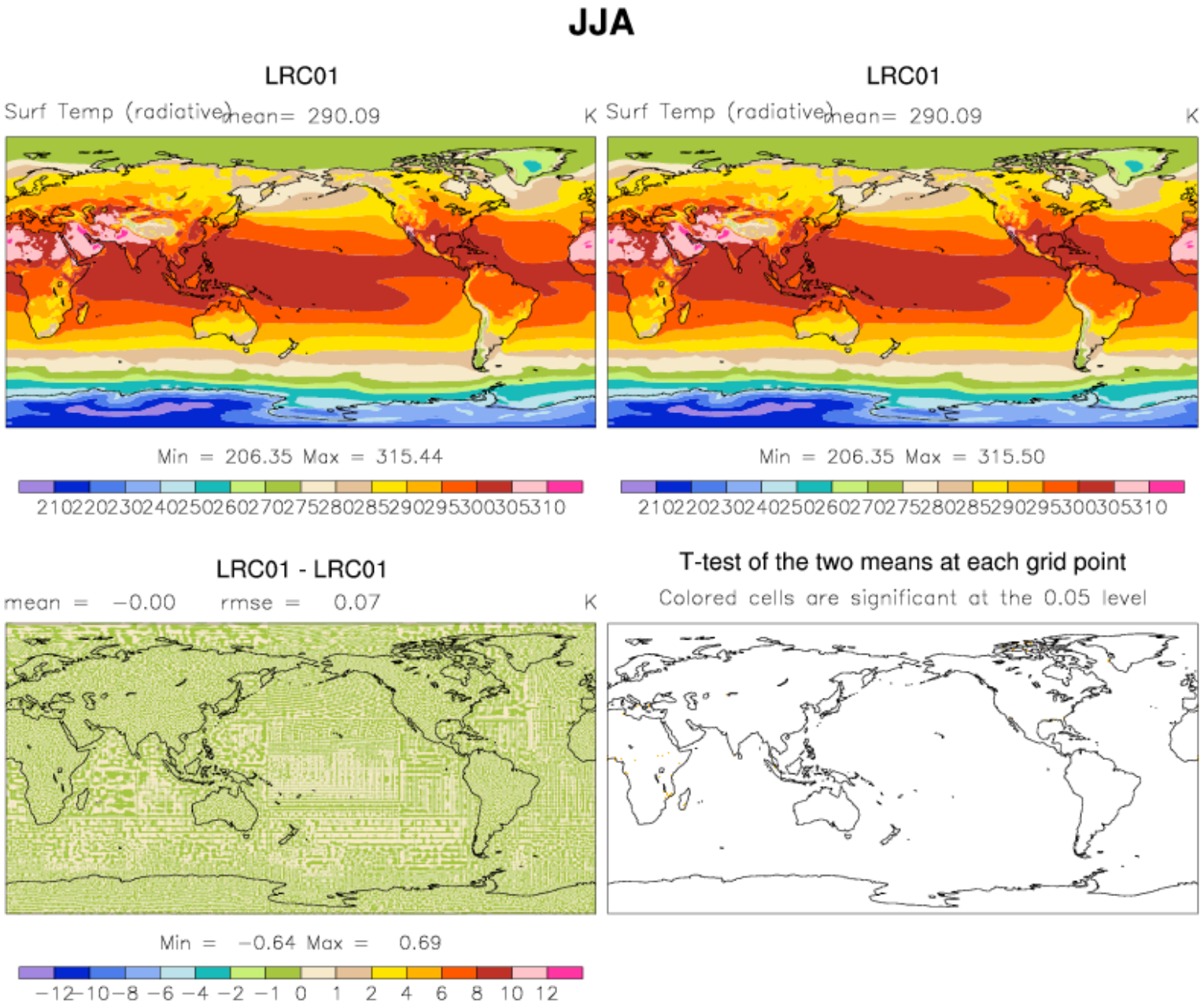
- What will we run at exascale?
  - Ultra-high resolution models
  - Data assimilation
  - Very large ensembles of low-res models
- NCAR Exascale Strategy
  - Computational optimizations
  - Data optimization
- Summary

What can lossy data compression get us?

Lossy data compression evaluation metrics

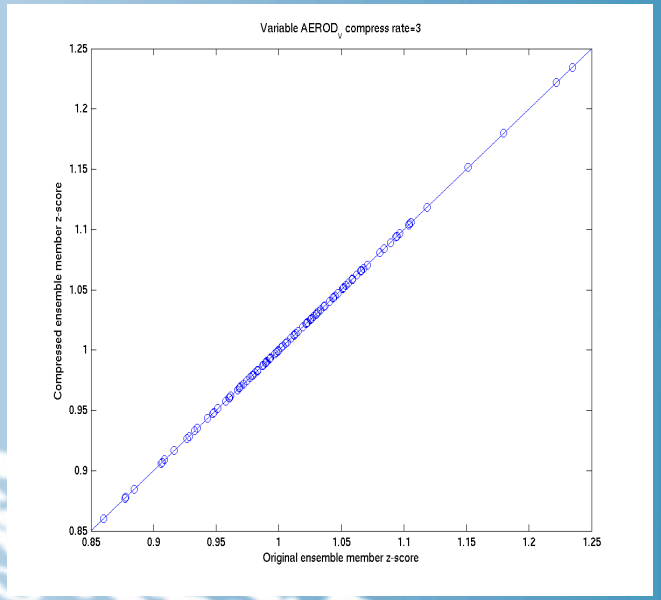
- **Pearson Correlation**: < 0.99999
- **RSMZ-ensemble test**
  - Choose single ensemble member
  - Compress/decompress member
  - Does decompressed members z-score still belong to ensemble?
- **RSMZ-bias test**
  - Compress/decompress all members
  - Calculate z-score versus uncompressed ensemble
  - Compare z-score of compressed versus original
  - Does compression/decompression introduce bias?
- **Ermax ensemble test**
  - Compress/decompress all members
  - Calculate Ermax for uncompressed ensemble
  - Calculate Ermax for original ensemble

T_S (Surface Temperature)



Lossy Compression – 5x

- **Using fpzip**
- **For a variable (u) choose highest compression rate such that all pass**
  - Pearson Correlation ✔
  - RMSZ-ensemble test ✔
  - Ermax-ensemble test ✔
  - RMSZ-bias test ✔
- **170 variables**
  - 7 variables: fpzip-32 [68%]
  - 50 variables: fpzip-24 [39%]
  - 113 variables: fpzip-16 [15%]
- **Overall 18% of original file size**



Outline

- What will we run at exascale?
- Ultra-high resolution models
- Data assimilation
- Very large ensembles of low-res models
- Strategy
- NCAR computational optimizations strategy
- NCAR data optimization strategy
- **Summary**

Conclusions for 2020

- **20x integration rate shortfall can be addressed through improvements in scalability, memory bandwidth and node performance optimizations (with accelerators?)**
- **3.6x climate data storage shortfall can be addressed with lossy compression without sacrificing climate information**
- **We are ramping up significant new staff/tech investments to accelerate our rate of progress in both areas.**

Thanks!