

Which computational environment for Climate Change studies?

Stefano Cozzini

CNR-IOM DEMOCRITOS and eXact Lab Trieste

Outline of this talk

- Introduction
 - What do we need to do computational climate science ?
- What is a “Computational Environment” for climate science ?
 - Elements of such environment (HW/SW/BW)
 - The challenges ahead
- Conclusion

Aims of this talk

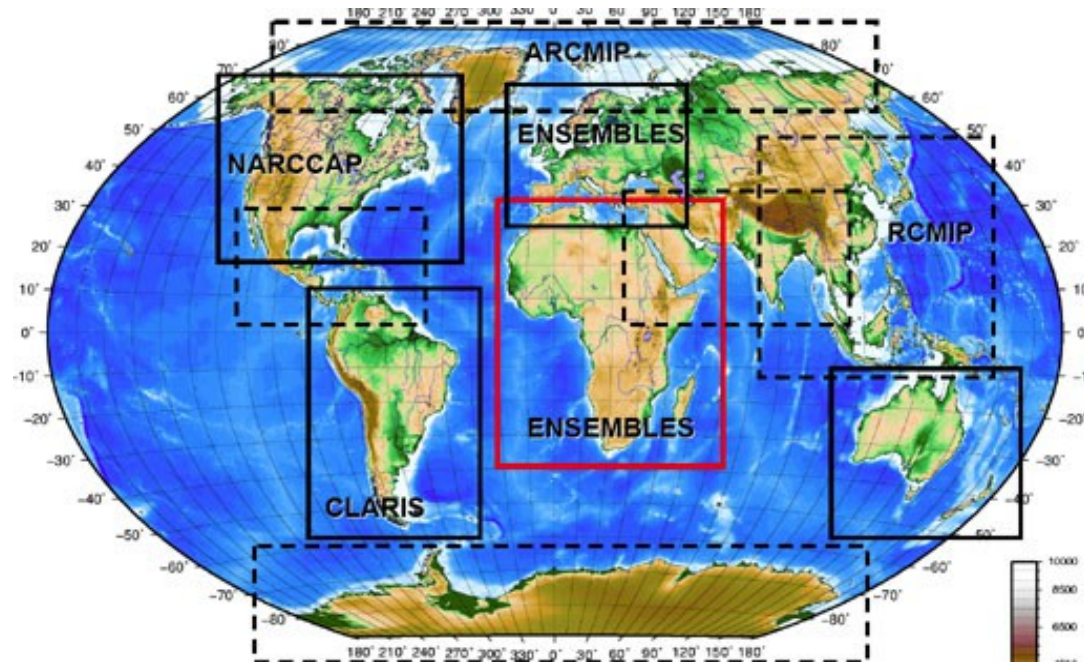
- Give an overview of the complexity of the computational problems involved
- Suggest some modest solutions based on our experiences so far..

A couple of disclaimers

- I am not a climate scientist, but I spent several time with them recently..
- I will use as reference mainly the RegCM4 package for all figures/numbers in term of computational requests

Introduction: What do we need to do computational climate science ?

Let us a start with an example how ESP group is performing CORDEX experiments in Trieste..



A typical CORDEX experiment performed by RegCM4 package...

- Setup several regions over the world
- Run some validation runs on them in order to tune physical parameters
- Run at least a climate scenario for each of them
 - 150 years of simulations (1950-2100)
- Analyze data

How powerful should the computers be ?

- A simulation of one month in seconds on a modern workstation (8 cores)

Name of the domain	Size	CPU time in second
European small	34x64x18	256
Ethiopian	112x128x18	1676
European	160x192x18	3348
African	250x256x18	100030



For a large domain one single 150 years of simulation will take 300 days !!!

Parallel computers are needed..

- On 120 processors of Arctur-1

Name of the domain	Size	CPU time in second
African	250x256x18	1200

For a large domain one single 150 years of simulation will take 25 days !

How large should be the disks ?

- Output data produced in one month of simulation

Name	Grid size	Data size (Gb)
Small European	34×64	0.241
Ethiopian	112×128	1.7
Big European	160×192	3.6
European	128×128	1.81
East Asian	186×224	4.46
Cent		
South		
South America (wet)	202×192	4.38
African	250×256	7.16

More than 13 TERABYTE for the African scenario !

Which kind of computational infrastructure ?

- Massive parallel computers for running the simulations..

for CORDEX experiments: the Arctur-1 system at Gorjansko for three months (~800cores)

- Large SAN (Storage Area Network) for data storage..

– The ICTP SAN of about 200 TB of storage

- Is this enough ?

NOT AT ALL !!

The missing parts..

- Network
 - What about data transfer ?
- Software
 - What about the software ?
- People
 - Who is running the simulations ?
 - Who is maintaining the infrastructure ?

Network issues..

- Moving data from Gorjansko to Trieste..

```
[exact@arctur1 2008]$ scp air.2008.18.nc cozzini@democritos.sissa.it:  
Password:  
air.2008.18.nc          5%   28MB   1.6MB/s   04:53 ETA
```

13,000,000 MB / 1.6 MB ~ 94 days !!!

Lesson learned

It is not just matter of computers and disks !

- Other aspects are actually more important than just the hardware part
- These aspects should be taken carefully into account !

From computational infrastructure to computational environment

- Goal:
provide a *computational environment* to satisfy the all different requirements posed by computational scientists in climate science
- Which kind of requirements ?
All you need to perform a “ *climate computer experiment*”
- The computational environment will comprise
HW SW and BW

Elements of the computational environment

- powerful and modern parallel hardware (HPC)
- pooling of resources geographically distributed (GRIDs)
- Infrastructure as a service (CLOUD)

HARDWARE

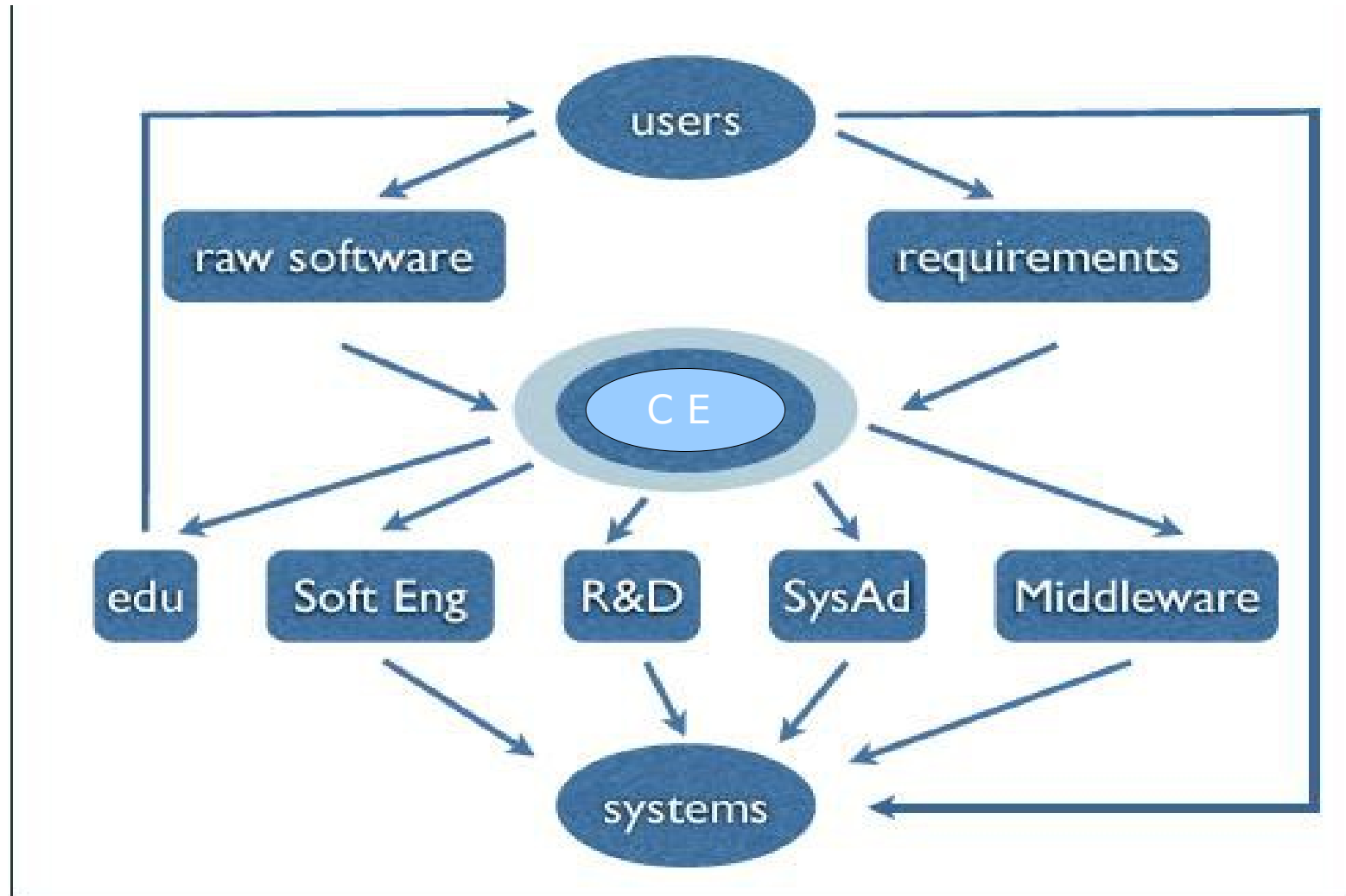
- Scientific software: Models & analysis
- Software for distributed collaboration and data sharing

SOFTWARE

- IT- skilled climate scientists
- Strong partnership between IT people and scientists

BRAINWARE

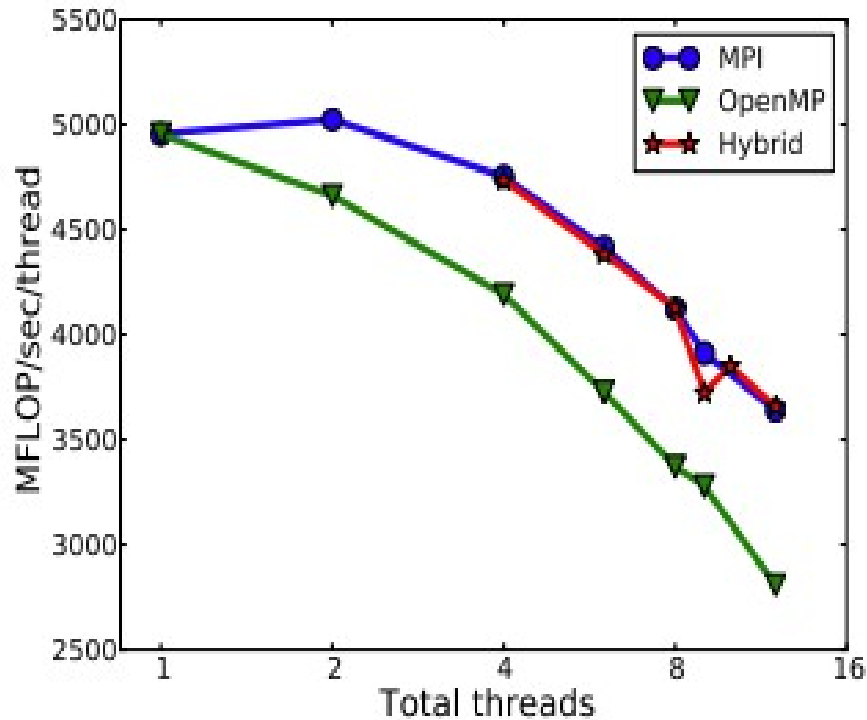
The computational environment



Hardware: challenges/issues ahead

- Exploiting accelerators (GPU/MIC etc.) on new computational nodes:
 - Not an easy task
- Exploiting the pooling of resources for specific climate problems
 - Again not an easy task for complex problems
- Is the cloud approach suitable for climate computational experiments ?

modern parallel computers need highly parallel codes..



- HOMME code behaviour on multicore cpus
- Scalability on multiple threads not so good..

“Evaluating Intel’s Many Integrated Core Architecture for Climate Science” by Theron Voran, Jose Garcia and Henry Tufo, 2011
retrieved from <http://www.tacc.utexas.edu/documents/13601/2ecbc5f2-2519-4650-8b60-298cb035247e>

Pooling of resource for which kind of problems ?

Search this site



climateprediction.net

The world's largest climate forecasting experiment for the 21st century.

Home

News

About

Support


Experiments

Climate Science

weatherathome

Donations

Recent updates

- [Two Weather At Home papers published in Bulletin of the American Meteorological Society](#)
2012-07-13 
- [Server problems resolved](#)
2012-06-18 

Experiment status

Summary	
Model Years	130,429,059
Active Hosts	32,506
Complete Model Simulations	
HadSM3	676,853
HadAM3	17,276
HadAM3P	518,141

Welcome to Climateprediction.net

in [Welcome](#)

What is climateprediction.net?

Climateprediction.net is a distributed computing project to produce predictions of the Earth's climate up to 2100 and to test the accuracy of climate models. To do this, we need people around the world to give us time on their computers - time when they have their computers switched on, but are not using them to their full capacity.

[Read more about the experiments.](#)

What do we ask you to do?

We need you to run a climate model on your computer. The model will run automatically as a background process on your computer whenever you switch your computer on and it should not affect any other tasks for which you use your computer. As the model runs, you can watch the weather patterns on your, unique, version of the world evolve. The results are sent back to us via the internet, and you will be able to see a summary of your results on this web site. Climateprediction.net uses the same underlying software, BOINC, as many other distributed computing projects and, if you like, you can participate in more than one project at a time.

[Read more about BOINC.](#)

Brainware (a.k.a. as People)

- Different classes of people:
 - Users: interested in scientific computational research
 - Developers: interested in scientific packages
 - Planners/maintainers: they define/maintain the computational infrastructure.
- Different level of HPC/IT skills should be taken into account for:
 - Optimal Sharing and exploitation of computational resources
 - Reduce the time to production: time needed to

The challenge: make users with low level of computer skills quite productive..

Elements to consider for software

- Planning and setup advanced solution for scientific computing in climate science
 - Software engineering techniques in climate software
 - Advanced data services

Which kind scientific software for climate simulations ?

- Large scientific climate packages available on the “market”:
 - RegCM4 (ICTP)
 - UM (UK -met office)
 - WRF etc..
- Data processing tools
 - CDO/ NCO/ Netcdf libraries and accompanying tools
- Most of this software is built by the climate scientists themselves, who have little or no training in software engineering.
- Quality of this software varies tremendously..
 - The Climate model tend to be well engineered
 - Some data processing tools are barely even tested.

Climate software: the challenges..

- **Correctness:** How do scientists assess correctness of their code? What does correctness mean to them?
- **Reproducibility:** How do scientists ensure experiments can be reproduced (e.g. for peer review)?
- **Shared Understanding:** How do scientists develop and maintain a shared understanding of the large complex codes they use? E.g. what forms of external representation do they use when talking about their models?
- **Debugging:** How do scientists detect (and/or prevent) errors in the software?

From: Engineering the Software for Understanding Climate Change
available at: <http://www.cs.toronto.edu/~sme/papers/2008/Easterbrook->

08/29/12 Johns-2008.pdf

Our work for RegCM4 development cycle..

- Continuous integration mechanism when developing
 - Each code modification is automatically tested against compilation&run to avoid introducing silly bugs and to guarantee code correctness
- Short runs daily executed and compared against selected tested to guarantee at least verification and correctness
- Scientific validation semi-automated but requires scientists contributions..
- Software is **open source** so everything can be cross-checked by users.

Advanced data services for climate data

- The most important and challenging aspect in the computational environment for climate simulation and analysis..
- New strategies and modern tools are needed in order to coping with the “3 Vs”— **variety, velocity, and volume** — of the big data that climate science generates.

From HPCwire..

August 15, 2012

Climate Science Triggers Torrent of Big Data Challenges

Dawn Levy, OLCF science writer, for HPCwire

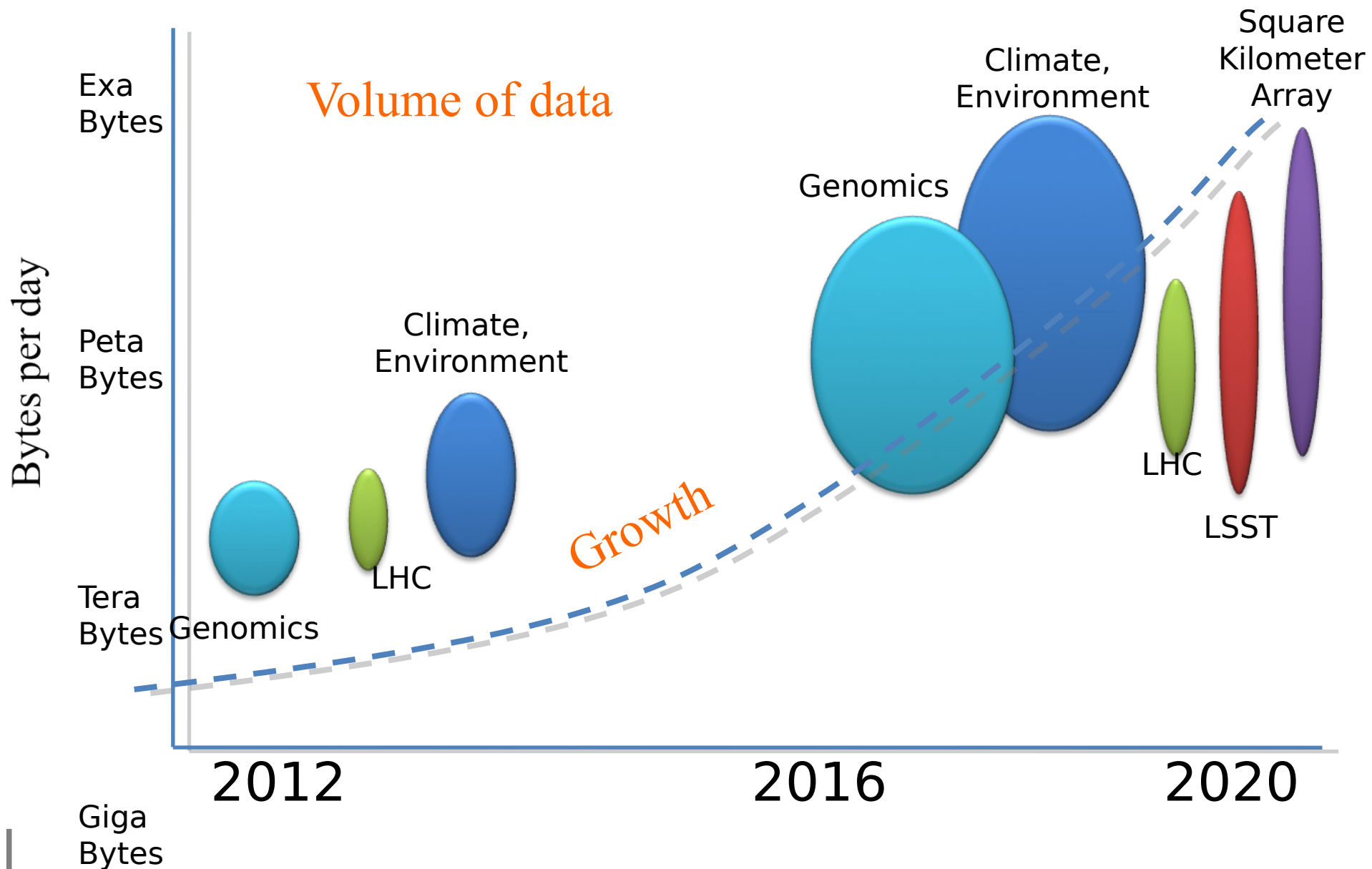
Page: [1](#) | [2](#) | [3](#)

Supercomputers at the Oak Ridge National Laboratory (ORNL) computing complex produce some of the world's largest scientific datasets. Many are from studies using high-resolution models to evaluate climate change consequences and mitigation strategies. The Department of Energy (DOE) Office of Science's [Jaguar](#) (the pride of the Oak Ridge Leadership Computing Facility, or OLCF), the National Science Foundation (NSF)-University of Tennessee's [Kraken](#) (NSF's first petascale supercomputer), and the National Oceanic and Atmospheric Administration's [Gaea](#) (dedicated solely for climate modeling) all run climate simulations at ORNL to meet the science missions of their respective agencies.

The three V's in climate data deluge

- **Velocity:**
 - They are produced at speed higher than the speed you are able to move/analyze and understand them..
- **Variety:**
 - They ranges from simulation datasets from a variety of global, regional, and local modeling simulation packages to remote sensing information
 - datasets come in a variety of data formats and span a variety of metadata standards

Volume : from gigabytes to exabytes



(slide from Tim Killeen, NSF)

Issues on big climate data

- I/O access and bandwidth can't keeping up with computing speed
- Too big to transfer, must move processing to data
- Sensors and models can generate huge datasets easily
- Making huge datasets accessible and useful is difficult
- Other problems: discovery, curation, provenance, organization, integrity..

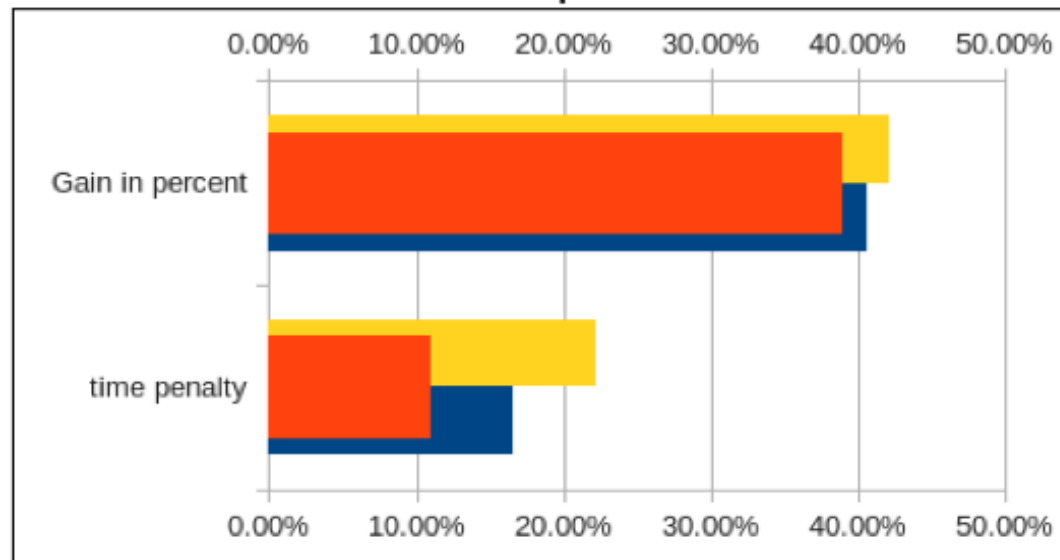
Our efforts in RegCM4 package

- I/O data are all in standard netcdf CF-1.4 compliant format
 - Enable all the netcdf advanced features
- Experiments on setting up an opendap server to store produced data with correct metadata is ongoing:
 - Allows RegCM user community to browse and navigate data easily and download and use only needed subset of datasets.

RegCM4 netcdf I/O features: compression..

- Compression library HDF5 and netcdf4
 - Up to 40% saving in storage space
 - Penalty of 16% in execution speed

Use of HDF5 through NetCDF libraries for
data compression.



RegCM4 netcdf I/O features: remote access through opendap server

- Simply change local file to opendap storage server in the input file:

From:

```
inpglob = '/lustre/dataset/ein15/'
```

To:

```
inpglob = 'http://climadods.ictp.it/dataset/ein15/'
```

- Optimal approach for distributed computing on remote infrastructure
- Burden in I/O data movement hidden within the application itself..

Final considerations

- Supercomputers are not enough to solve computational climate problems
- Software (SW) and Human resource (BW) are more important than raw computing/storage power (HW)
- Only a coordinated effort among these three elements can delivered required performance
- Train/educate “human resources” is a fundamental point

Final slide

- Special thanks to Graziano Giuliani from ESP/ICTP
- Thanks for you attention and patience
- Questions& comments welcome