The Abdus Salam International Centre for Theoretical Physics

“New paradigms” for the CP climate simulation: GPU, C++

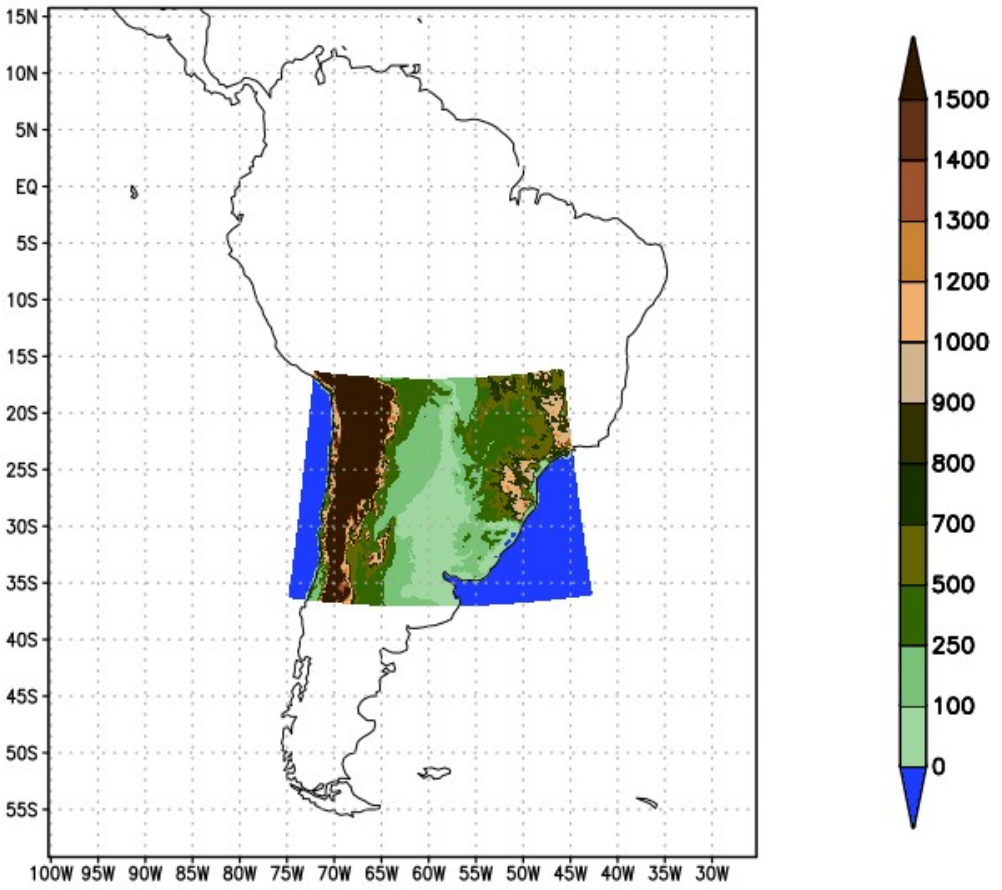
Graziano Giuliani
ggiulian@ictp.it
Francesca Raffaele
ICTP ESP

The Abdus Salam International Centre for Theoretical Physics

CP Simulation

DOMAIN dimension:

iy = 570
jx = 730
kz = 41
ds = 4.0



Run realization credit:
Francesca Raffaele
<fraffael@ictp.it>

The Abdus Salam International Centre for Theoretical Physics

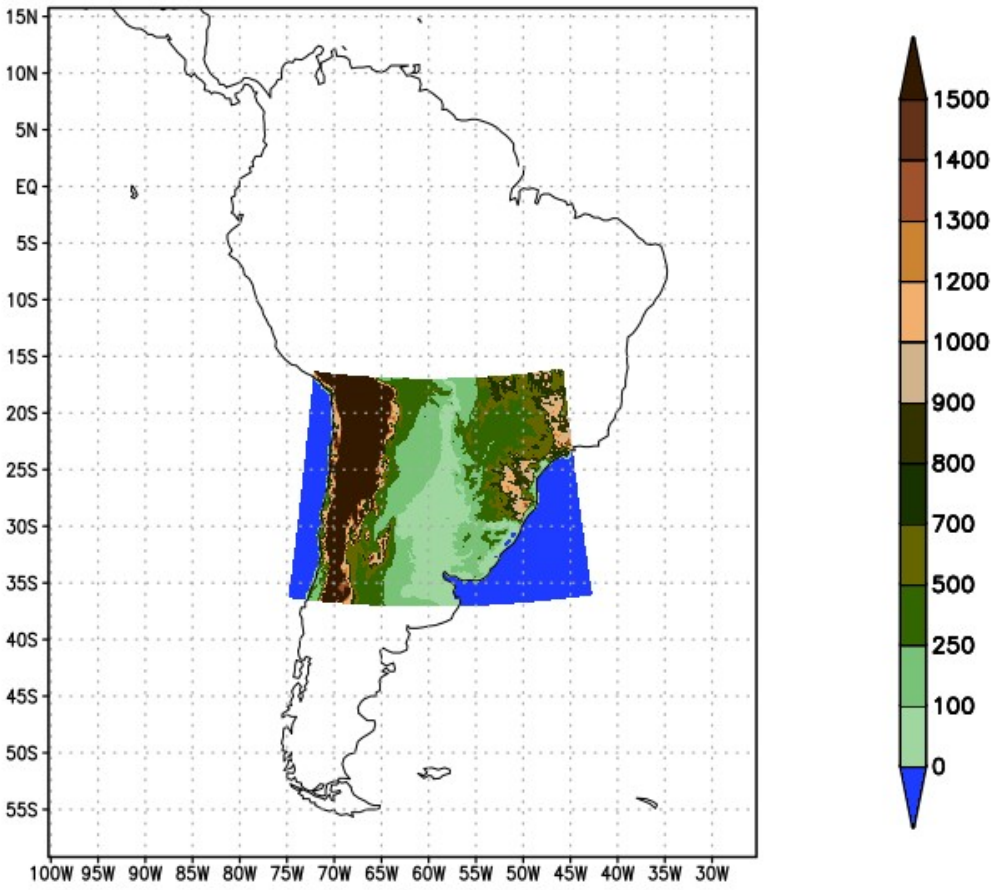
CP Simulation

DOMAIN dimension:

iy = 570
jx = 730
kz = 41
ds = 4.0

CORDEX dimension:

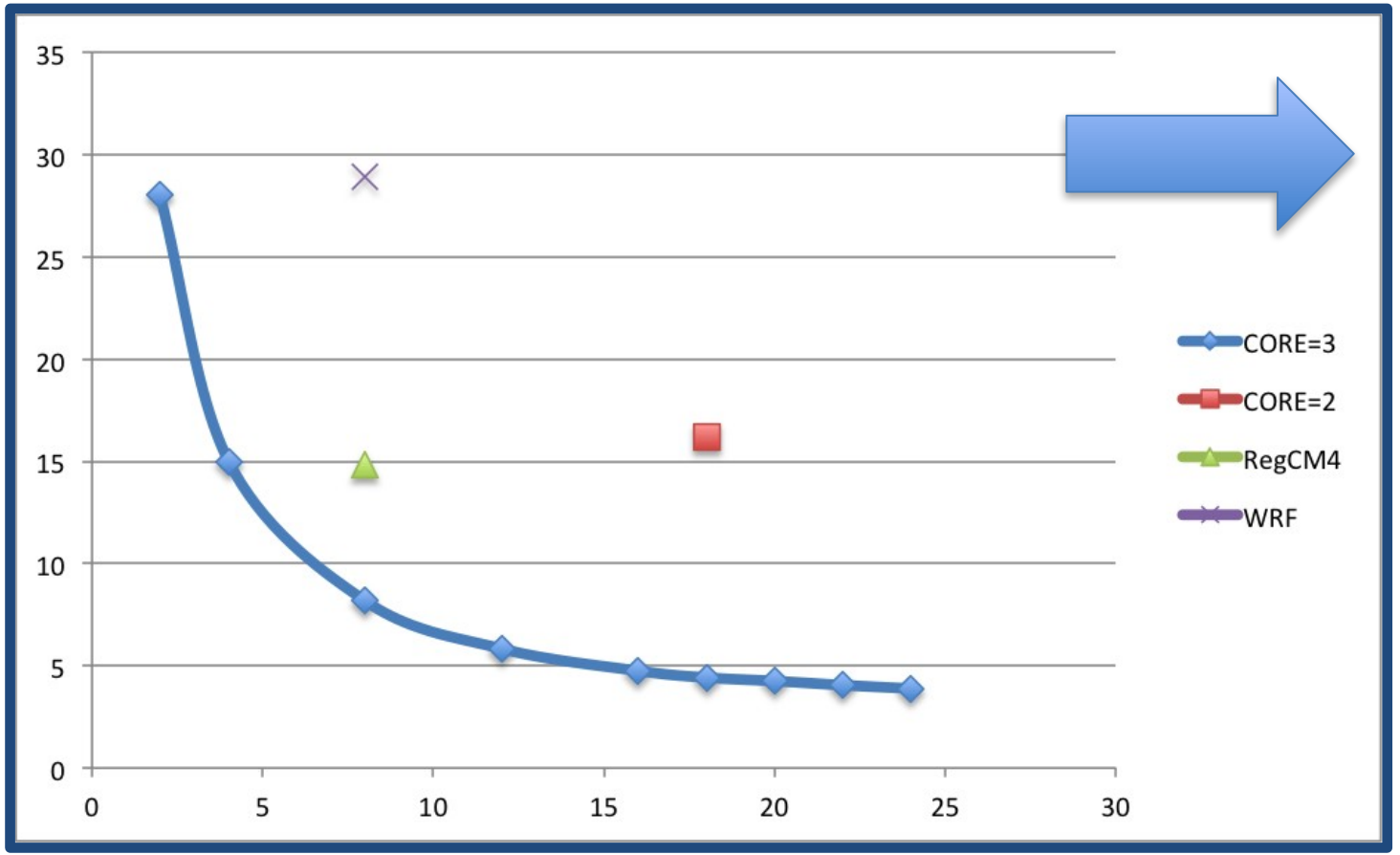
iy = 500
jx = 480
kz = 23
ds = 25.0



Run realization credit:
Francesca Raffaele
<fraffael@ictp.it>

The Abdus Salam International Centre for Theoretical Physics

Performance



1 month ~ 4 hours
1 year ~ 2 days
150 years ~ 1 year!

Run performance credit:
Francesca Raffaele
<fraffael@ictp.it>

The Abdus Salam International Centre for Theoretical Physics

HPC trends

NCAR CDC 6600

Control Data Corporation
In use: December 30, 1965 - May 20, 1977
Production use
Peak teraflops: 0.00
Processors: 1.00
Clock speed: 0.01GHz
Memory: 0.00TB
Predecessor: CDC 3600
Successor: CDC 7600

https://www2.cisl.ucar.edu/ncar-supercomputing-history/cdc6600

The Abdus Salam International Centre for Theoretical Physics

ECMWF

Cray XC40 cluster

- Aries™ interconnect.
- Intel Xeon processors 2.1 GHz
- 8.49 Teraflops
- 129,960 cores over 3610 nodes

Linux Cluster
General purpose processor
MPI interconnect

https://www.ecmwf.int/en/computing/our-facilities/supercomputer

The Abdus Salam International Centre for Theoretical Physics

JMA

Hitachi

Cray XC50

Linux Cluster
General purpose processor
MPI interconnect

https://www.jma.go.jp/jma/en/News/JMA_Super_Computer_upgrade2018.html

The Abdus Salam International Centre for Theoretical Physics

NCAR

NCAR CHEYENNE

SGI ICE XA Cluster
In use from January 2017
Peak petaflops: 5.34
Processors: 145,152 over 4,032 nodes
Clock speed: 2.3 GHz (Intel Xeon)
Memory: 313TB
Mellanox EDR Infiniband

Linux Cluster
General purpose processor
MPI interconnect

https://www2.cisl.ucar.edu/computing-data/computing#cheyenne

The Abdus Salam International Centre for Theoretical Physics

DKRZ Levante

- BullSequana XH2000
- Two partitions (CPU and GPU)
- 2832 nodes with 362496 CPU AMD EPYC
- 240 NVIDIA A100 GPUs
- 130 Petabyte filesystem

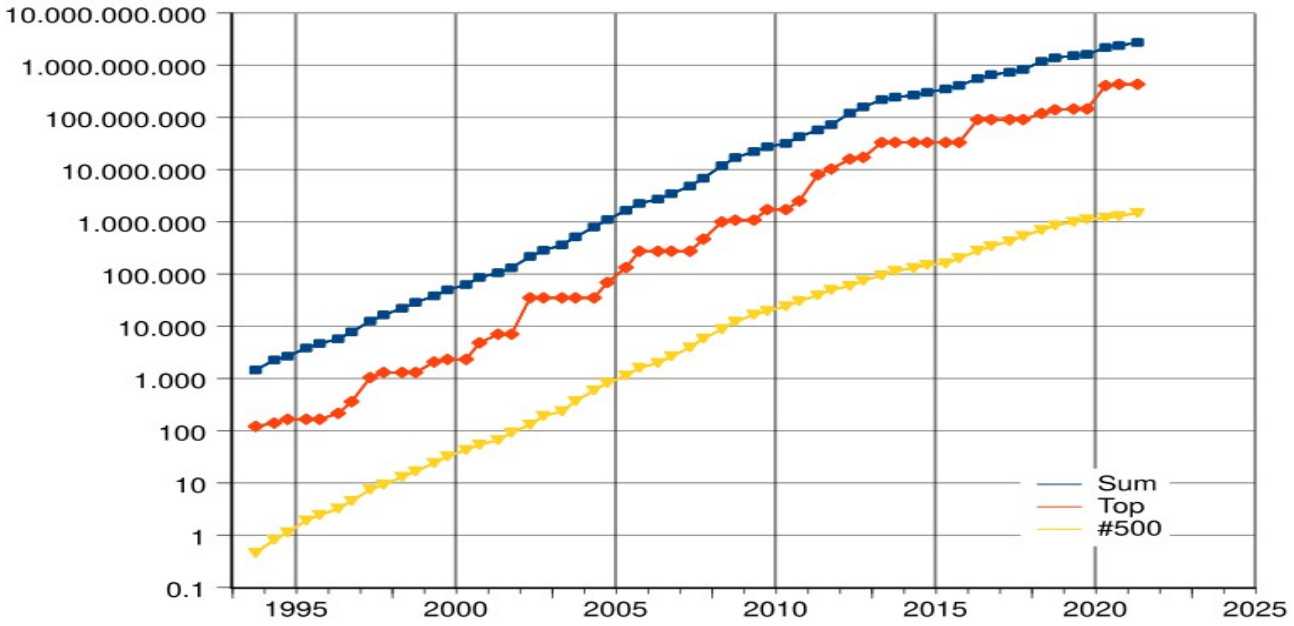
Linux Cluster
General purpose processor
MPI interconnect

https://www.dkrz.de/en/systems/hpc

The Abdus Salam International Centre for Theoretical Physics

Super Computers

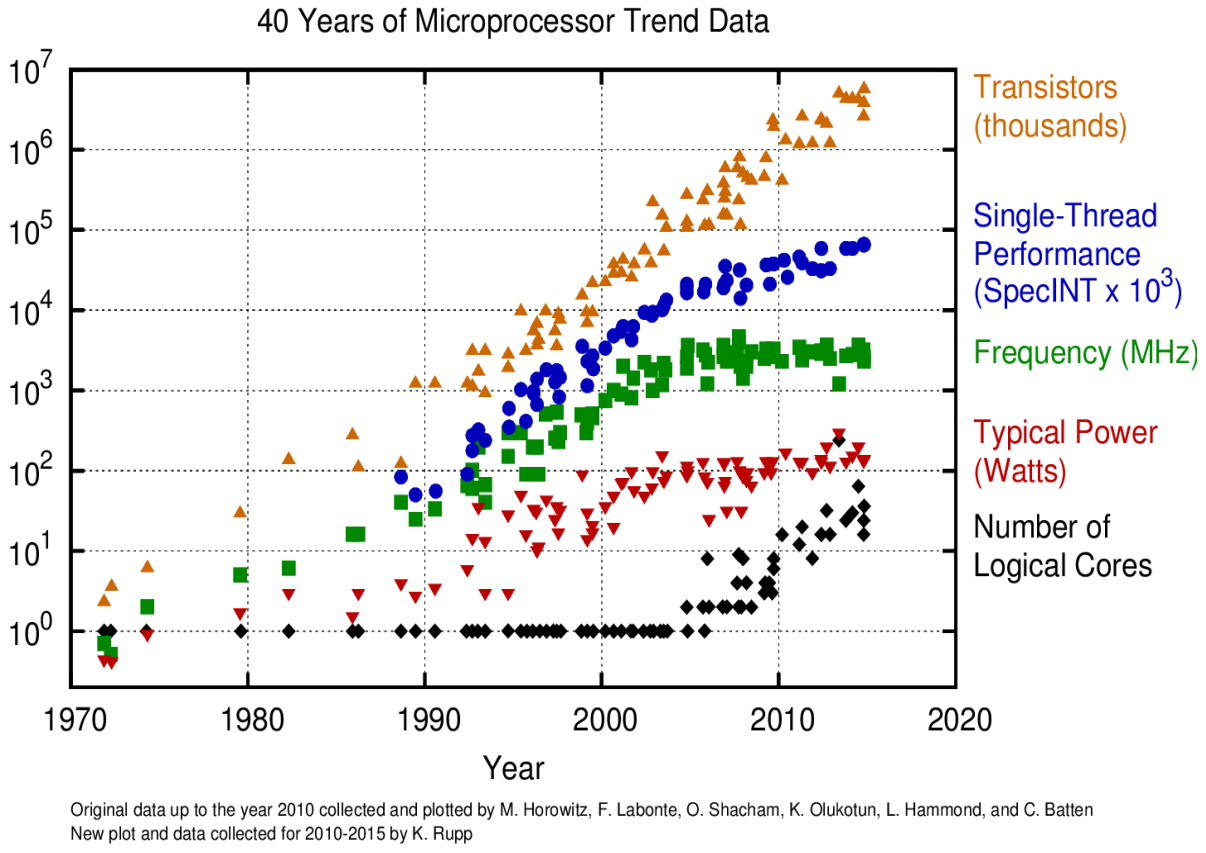
TOP500



The Abdus Salam International Centre for Theoretical Physics

CPU trend

- CPU market has moved to laptop/mobile
  - Low power
  - mix high and low performance cores on one chip
- HPC exa-scale systems will be power hungry hogs and energy consumption MUST be minimized (convergence on different architecture? ARM? RISC-V?)



The Abdus Salam International Centre for Theoretical Physics

GPU

- *Hardware acceleration is the use of computer hardware designed to perform specific functions more efficiently when compared to software running on a general-purpose CPU. [Wikipedia]*
- Out of gaming and into GPGPU
  - 2001: programmable shaders and floating point support on graphics processors
  - 2003: GPU-based solution of general linear algebra problems: reformulating computational problems in terms of graphics primitives

The Abdus Salam International Centre for Theoretical Physics

Cost

- An HPC GPU has a price of 10 - 25k
- An HPC CPU has a price in the range 1 – 10k

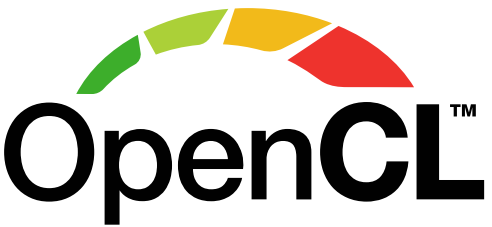
- Cost effective:

Cost is in the mean an order of magnitude greater: code running on a GPU must run at least 10x faster than the same code on CPU. To be included in vendor success story a 100x is required.

The Abdus Salam International Centre for Theoretical Physics

Low level tools

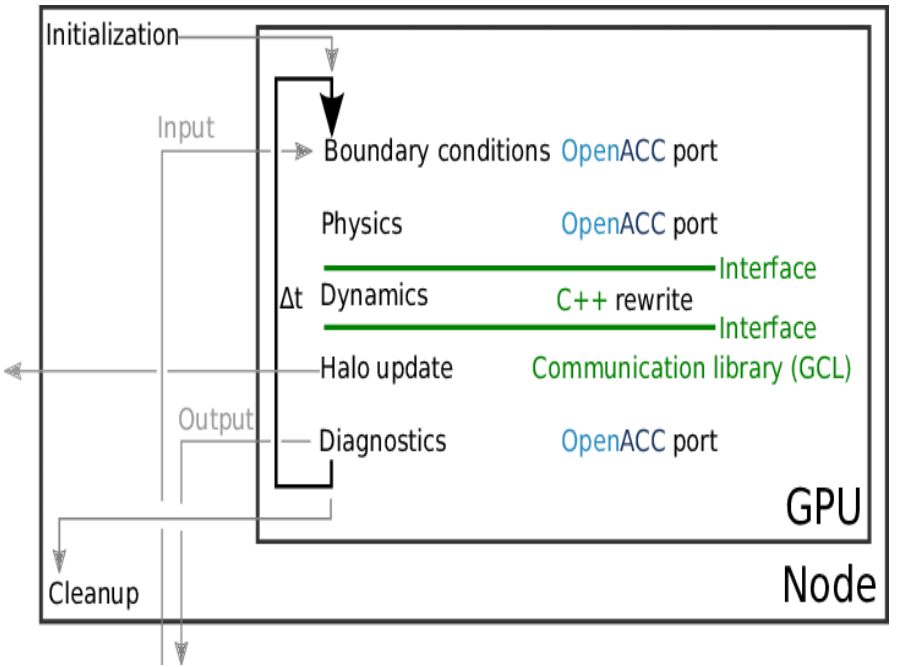
- Vendor usually provide SDK for the GPU programming allowing low level access.
  - To get the level of performances expected the computation must be on the GPU
  - Memory transfer between GPU memory and system memory are a bottleneck and should be reduced
  - Only latest cards allow GPU-GPU direct transfer and MPI aware GPU only recently is an option
- SDK are hardware vendor provided and push for a lock into the hardware solution
- Vendor agnostic solutions use only an intersection subset of the available resources
- Complete control over the hardware for optimization is usually provided through a C/C++ library directly interacting with the low-level hardware driver. Fortran interface again is usually late coming, has reduced capabilities and is vendor locked.
- Missing direct generic code optimization for accelerator as the one given for CPU targets



The Abdus Salam International Centre for Theoretical Physics

Accelerate weather?

- MeteoSwiss - Cray CS-Storm supercomputer at CSCS with 96 GPUs (old K80)
- First national meteorological service using GPUs for operational NWP in 2016
- 4 years of development with strong investment from NVIDIA



Leutwyler, D., Fuhrer, O., Lapillonne, X., Lüthi, D., and Schär, C.: Towards European-scale convection-resolving climate simulations with GPUs: a study with COSMO 4.19, Geosci. Model Dev., 9, 3393–3412, https://doi.org/10.5194/gmd-9-3393-2016, 2016.
Fuhrer, O., Chadha, T., Hoefler, T., Kwasniewski, G., Lapillonne, X., Leutwyler, D., Lüthi, D., Osuna, C., Schär, C., Schulthess, T. C., and Vogt, H.: Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0, Geosci. Model Dev., 11, 1665–1681, https://doi.org/10.5194/gmd-11-1665-2018, 2018.

The Abdus Salam International Centre for Theoretical Physics

C++?

- Specialized compilers composed from domain- and target-specific dialects implemented on top of a shared infrastructure.
  - Tobias Gysi, Carlos Osuna, Oliver Fuhrer, Mauro Bianco, and Thomas C. Schulthess. 2015. STELLA: a domain-specific tool for structured grid methods in weather and climate models. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '15). Association for Computing Machinery, New York, NY, USA, Article 41, 1–12. https://doi.org/10.1145/2807591.2807627
  - Osuna, C., Wicky, T., Thuering, F., Hoefler, T., & Fuhrer, O. (2020). Dawn: a High-level Domain-Specific Language Compiler Toolchain for Weather and Climate Applications. Supercomputing Frontiers and Innovations, 7(2), 79–97. https://doi.org/10.14529/jsfi200205

```
%0 = stencil.apply seq(dim = 2, range = 0 to 60, dir = -1)
  (%arg0 = %rhs : !stencil.temp<?x?x?xf64>,        sequential loop
   %arg1 = %sup : !stencil.temp<?x?x?xf64>) -> !stencil.temp<?x?x?xf64> {
  %1 = stencil.index 2 [0, 0, 0] : index
  %cst = constant 59 : index                       get current position
  %2 = cmpi "eq", %1, %cst : index
  %3 = scf.if %2 -> (f64) {                        boundary condition
    %4 = stencil.access %arg0 [0, 0, 0] : (!stencil.temp<?x?x?xf64>) -> f64
    scf.yield %4 : f64
  } else {                                         main stencil
    %5 = stencil.access %arg0 [0, 0, 0] : (!stencil.temp<?x?x?xf64>) -> f64
    %6 = stencil.access %arg1 [0, 0, 0] : (!stencil.temp<?x?x?xf64>) -> f64
    %7 = stencil.depend 0 [0, 0, 1] : f64
    %8 = mulf %6, %7 : f64                         access output 0 at offset [0, 0, 1]
    %9 = addf %8, %5 : f64                         (loop carried dependency)
    scf.yield %9 : f64
  }
  stencil.return %3 : f64
}
```

The Abdus Salam International Centre for Theoretical Physics

Accelerate climate?

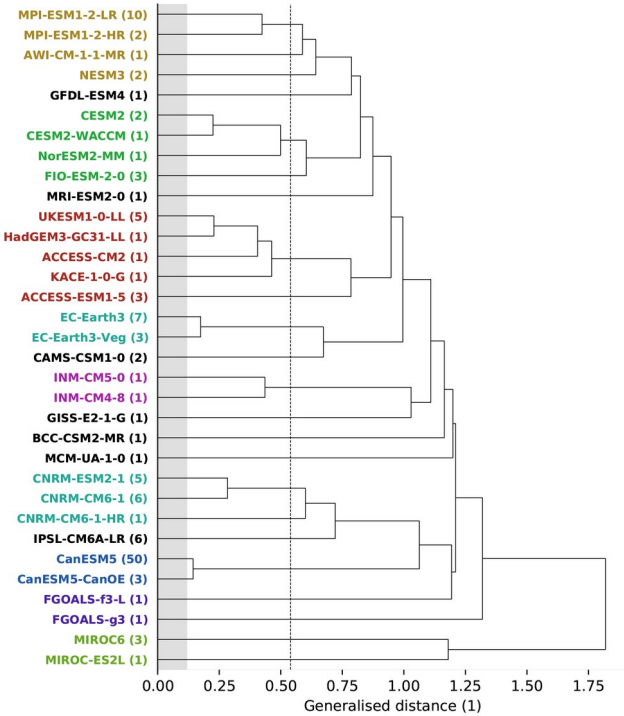
- NCAR Casper cluster : specialized data analysis and visualization resources; large-memory, multi-GPU nodes; and high-throughput computing nodes
- ECMWF : Virtual GPU features 2x5 NVIDIA Tesla V100 cards targeting Machine Learning workloads.
- DKRZ Levante : CPU and GPU partition
  - Atos Bull Sequana XH2000 370,000+ cores
    - CPU : 2,832 compute nodes AMD 7763 64 cores → 14 Petaflops
    - GPU : 60 nodes with 4 A100 GPUs each → 2.8 Petaflops

Giorgetta, M. A., Sawyer, W., Lapillonne, X., Adamidis, P., Alexeev, D., Clément, V., Dietlicher, R., Engels, J. F., Esch, M., Franke, H., Frauen, C., Hannah, W. M., Hillman, B. R., Kornblueh, L., Marti, P., Norman, M. R., Pincus, R., Rast, S., Reinert, D., Schnur, R., Schulzweida, U., and Stevens, B.: The ICON-A model for direct QBO simulations on GPUs (version icon-cscs:baf28a514), EGU sphere [preprint], https://doi.org/10.5194/egusphere-2022-152, 2022.
Jae Youp Kim, Ji-Sun Kang, Minsu Joh, GPU acceleration of MPAS microphysics WSM6 using OpenACC directives: Performance and verification, Computers & Geosciences, Volume 146, 2021, 104627, ISSN 0098-3004, https://doi.org/10.1016/j.cageo.2020.104627.

The Abdus Salam International Centre for Theoretical Physics

One model to rule them all

- We treat model as independent entities, amenable to statistical treatments such as averaging or taking standard deviations to identify uncertainty, even though the models represent only limited explorations in a potentially huge conceptual space, but...
  - Shared history of climate models[1]: inappropriate to treat different climate models as randomly sampled independent draws from the hypothetical model space
- The actual future may well be outside the set of modeled futures
- Nevertheless, using different models gives us an idea about uncertainty: the more the models, the more the information, as long as their independence is not hindered[2]



1) Masson, D., and Knutti, R. (2011), Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864.
2) Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth Syst. Dynam., 11, 995–1012 , https://doi.org/10.5194/esd-11-995-2020, 2020.

The Abdus Salam International Centre for Theoretical Physics

Climate modeling

We, therefore, the Representatives of the united Modelling Groups of the World, in AGU Congress, Assembled, appealing to the Supreme Judge of the model ensemble for the rectitude of our intentions, do, in the Name, and by Authority of the good People of these modeling Centers, solemnly publish and declare, That these disparate Models are, and of Right ought to be Free and Independent Codes, that they are Absolved from all Allegiance to NCAR, GFDL and Arakawa, and that all algorithmic connection between them and the Met Office of Great Britain, is and ought to be totally dissolved; and that as Free and Independent Models, they have full Power to run Simulations, conclude Papers, contract Intercomparison Projects, establish Shared Protocols, and to do all other Acts and Things which Independent Models may of right do. — And for the support of this Declaration, with a firm reliance on the protection of Divine PCMDI, we mutually pledge to each other our Working Lives, our Git Repositories, and our sacred H-Index.

https://www.realclimate.org/index.php/archives/2018/07/model-independence-day/

The Abdus Salam International Centre for Theoretical Physics

Inclusive by design

- As long as they produce good and independent results, different models should not be phased out because they do not perform
- Using just a single model would just not do.

https://docs.esmvaltool.org/en/latest/recipes/recipe_climwip.html

The Abdus Salam International Centre for Theoretical Physics

Today dilemma

- HPC Tier-0 centers are moving to GPUs
  - All rising or stable #top500 systems are GPU based
  - AI is the new Buzzword and HPC supports AI workloads on GPU
  - Example: Italy CINECA: EU and Italian funds for innovation and research, 112 members for 4295 users
    - 2012 IBM Blue Gene/Q ← ICTP RegCM running
    - 2016 Lenovo Marconi Xeon Cluster + Phi accelerators ← ICTP RegCM running
    - 2020 IBM Marconi-100 Power9 + Nvidia Tesla ← ICTP RegCM not qualified
    - 2022 Leonardo Xeon + 14000 Nvidia Ampere ← ICTP RegCM not qualified

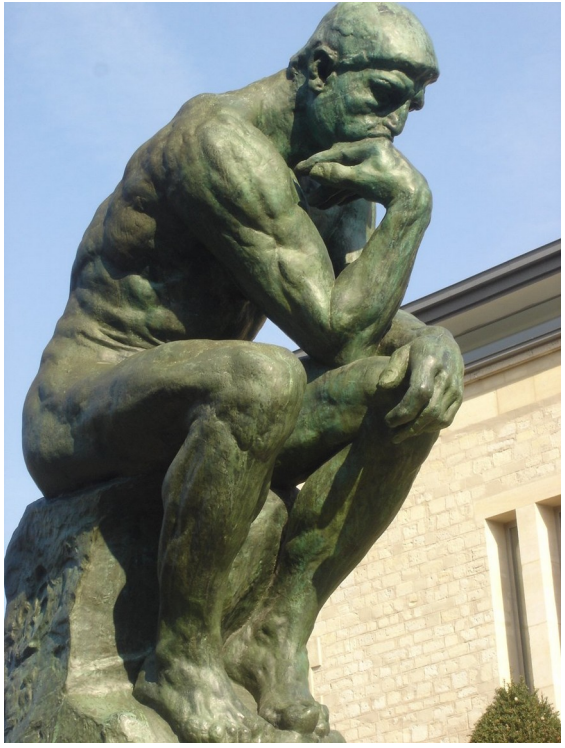
CINECA

The Abdus Salam International Centre for Theoretical Physics

GPU code

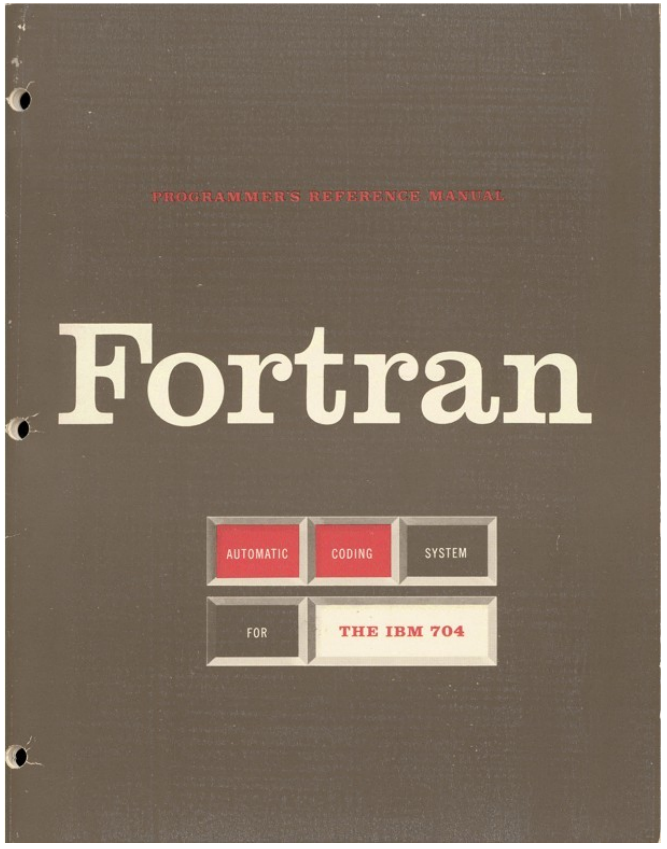
Without measurable GPU performances, ICTP RegCM will not be allowed access to new pre-exascale CINECA computing resources.

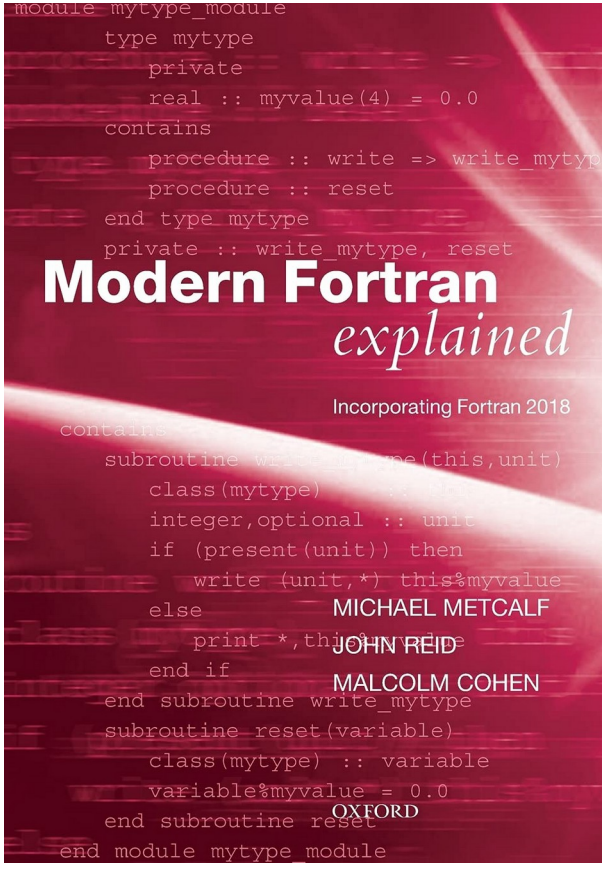
- What is the effort of the “porting to GPU”?
- Do we have enough human and financial resources?



The Abdus Salam International Centre for Theoretical Physics

Climate Programming





The Abdus Salam International Centre for Theoretical Physics

Why stick with Fortran?

- Longevity – old code still works 50+ years on
- First class arrays
- Performance
- Numerical libraries
- MPI compatibility
- Parallelization STANDARD (co-arrays) never picked up: missing implementations supporting the standard out of the multi-core environment and into the multi node environment. Good idea but GPU vendors lack of interest dooming the implementation to remain viable only for multi-threaded scenarioes.

Babi Hijau - Photo taken by own, Pubblic domain, https://commons.wikimedia.org/w/index.php?curid=2801324

The Abdus Salam International Centre for Theoretical Physics

C++

- Born 1979 as C extended to OO features
- System, performance, design
- ISO standard
- Interoperability
- Libraries

The Abdus Salam International Centre for Theoretical Physics

Fortran/C++ interoperability

- Fortran 90 modules share functionality with classes and namespaces in C++
  - Example: function add_abs in module mymod becomes:

    __mymod_MOD_add_abs
- C++ namespaces are encoded in symbols
  - Example: int func::add_abs(int,int) becomes:

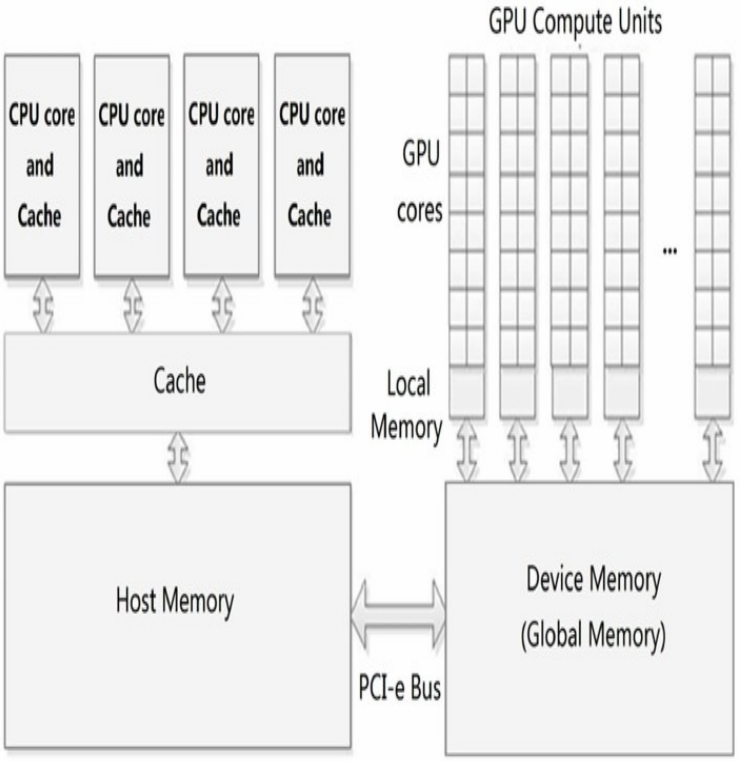
    _ZN4funcL7add_absEii
- C++ classes are encoded the same way. No overloading is possible.
- Figuring out which symbol to encode into the object as undefined is the job of the compiler. Different compilers will differently mangle the symbols in produced objects.

NOT EASILY PORTABLE

The Abdus Salam International Centre for Theoretical Physics

C++ and GPU

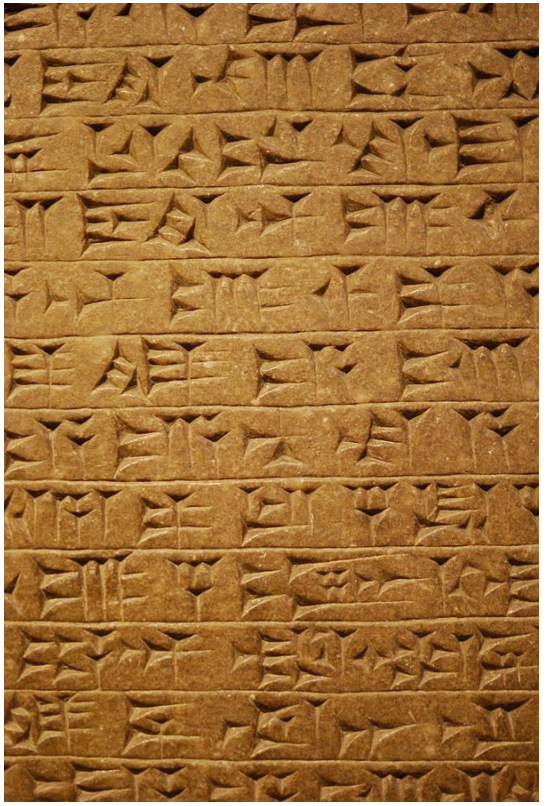
- Direct access to the hardware through vendor provided libraries
- Must create computation Kernels: part of the code that will be executed on the GPU using memory on the accelerator
- Reduction and optimization of the data transfer through PCI bus CPU memory ↔ GPU memory
- Domain Specific Languages



The Abdus Salam International Centre for Theoretical Physics

Porting Fortran → C++

- Time consuming : 4-5 years time-frame
- Programmers Team : C++ developers are not cheap and are not easily enticed in academic environment
- GPU vendor help is linked to the big contract on new hardware acquisition or the viability of the technology for a community as a buzzword to attract users
- Model code may become “unreadable” for scientists: requires generational and curriculum changes



The Abdus Salam International Centre for Theoretical Physics

ICTP RegCM and ESIWACE2

- ESIWACE: push the global high-resolution models towards production on European pre-exascale systems, which are aimed to be set up in 2021
- Six months of consulting and help from GPU experts on how to help us on GPUs
- Project expected to complete by the end of fall and the code contributed to RegCM github official repository.

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

The Abdus Salam International Centre for Theoretical Physics

Phase 1: Code evaluation

- Code MUST be OpenSource and on a public repository.
- Model must coded in a compiled language such as Fortran, C or C++
- The produced modification during the project MUST be included in the public model repository



The Abdus Salam International Centre for Theoretical Physics

Phase 2: Methodology

- Porting to C++: not feasible in the allotted time-frame and provided resources
- Duplicate the code with a GPU separate code-base: not accepted because of maintainability concerns
- Add OpenACC pragma (comments) directives to instrument the compiler (if supported by it) to transform part of the code in GPU executable kernels: feasible and accepted by the ICTP

The Abdus Salam International Centre for Theoretical Physics

Phase 3: Implementation

- Identify a limited part of the code to act as a blueprint for RegCM developer to complete the work after project completion: taken the dynamical core module
- Instrument all loops as possible kernels to offload on GPUs
- Asses results are the same on CPU and GPU: expected to have worse performances on GPU compared with CPU
- Merge kernels together to reduce context switching: aim is to increase performances on GPU by reducing MPI tasks: from one task per core to one task per CPU
- Communication optimization: MPI GPU aware: reach the target performance by tuning the GPU-GPU data transfer layer: ONGOING

The Abdus Salam International Centre for Theoretical Physics

Conclusion

- Climate modeling needs to adapt to a changing HPC world where future trends in CPU and accelerators are much less streamlined than they were in the past 20 years
- GPU adoption in multi-purpose national and international HPC center is strongly pushing modeling groups to have model code that can use those resources
- C++ language, even though showing HPC performances equivalent to Fortran and the low level hardware access to directly interface with vendor provided GPU API, is not part of the curriculum of the Atmospheric Physicist. Moving to C++ would require resources not available to most modeling groups and would phase out a big part of the code-base and multi decade expertise with well known and well maintained programming patterns.
- Fortran programming language has no working abstraction for heterogeneous multiprocessing and must use directive based instrumentation which are vendor supported and not part of the programming standard. MPI is going to stay.

The Abdus Salam International Centre for Theoretical Physics

Note for thought: New languages are coming up!

Julia is really promising in terms of performances and we have the first model and tools.

https://clima.caltech.edu
https://juliaclimate.org