IS-ENES2 Final General Assembly
16-18th January 2017
Paris, France

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Giovanni Aloisio – CMCC

and

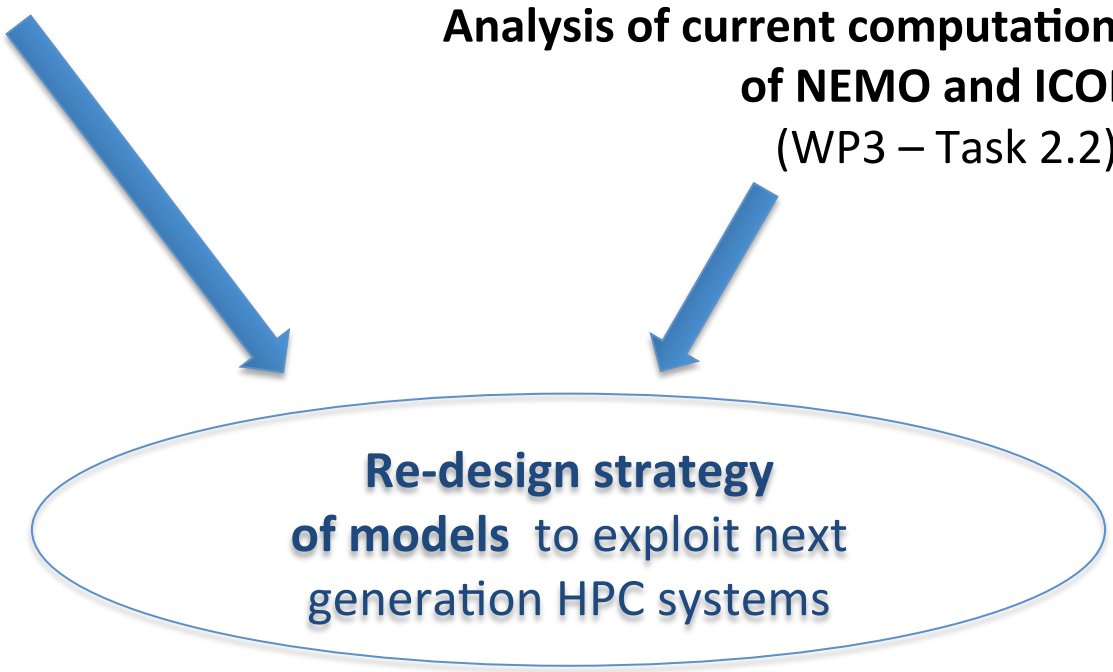
from interaction with vendors?

Joachim Biercamp - DKRZ

WP3/NA2 institutions

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Technology tracking and investigation on European and global initiatives focusing on the models improvement at exascale (WP3 – Task 3)



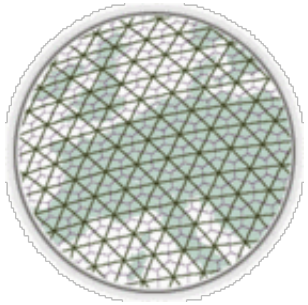
Technology tracking: what have we learnt for climate models (NEMO-ICON)?

NEMO (Nucleus for European Modeling of the Ocean)

- Finite-difference on tri-polar ORCA grid
- Central to a number of European projects
- Fortran90, MPI only
- Highly portable
- ~20 years of development

ICON (Icosahedral non-hydrostatic general circulation model)

- Non-hydrostatic global model with a local zoom function
- Icosahedral grid
- Fortran90/95 and C
- Parallelized with MPI and OpenMP



Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Emerging HPC systems are based on ***many-core*** and ***hybrid processors***, with **limited memory per core!**

What do we need?

- new algorithms to improve the **memory scaling**
- new algorithms to limit the **data movement overhead**
- new **hybrid programming models** (e.g. MPI+X) to better exploit hybrid architectures
- improving the **code vectorization** level to fully exploit the available vector units
- new approaches to increase the **computation per byte of accessed data**

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

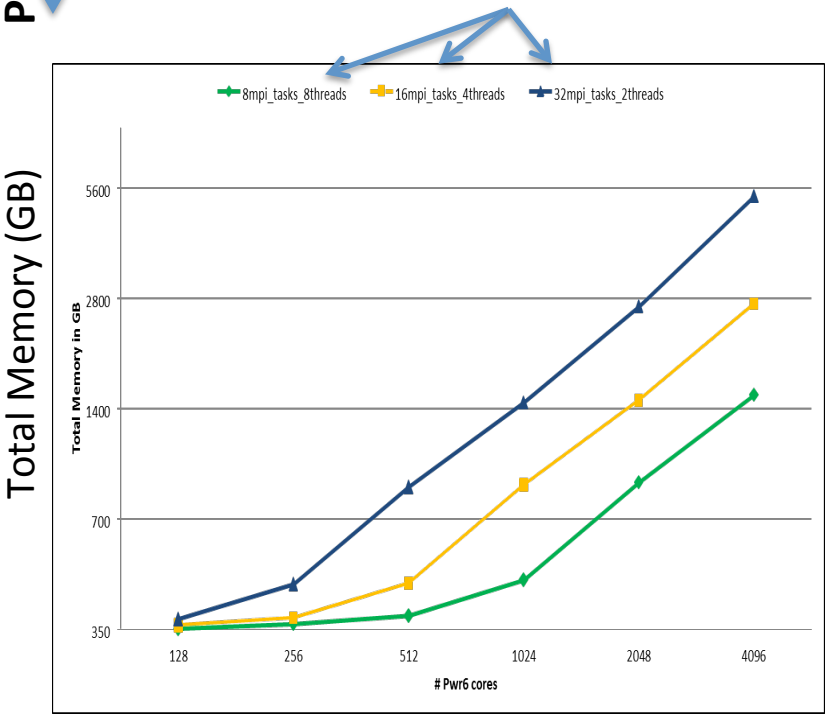
Memory Scaling issue

- The analysis of ICON lead to recommendations for improving memory scaling
- *The allocated memory increases with the number of MPI tasks, due to the replication of some global data structure on all the parallel processes*

Promlem

Solution

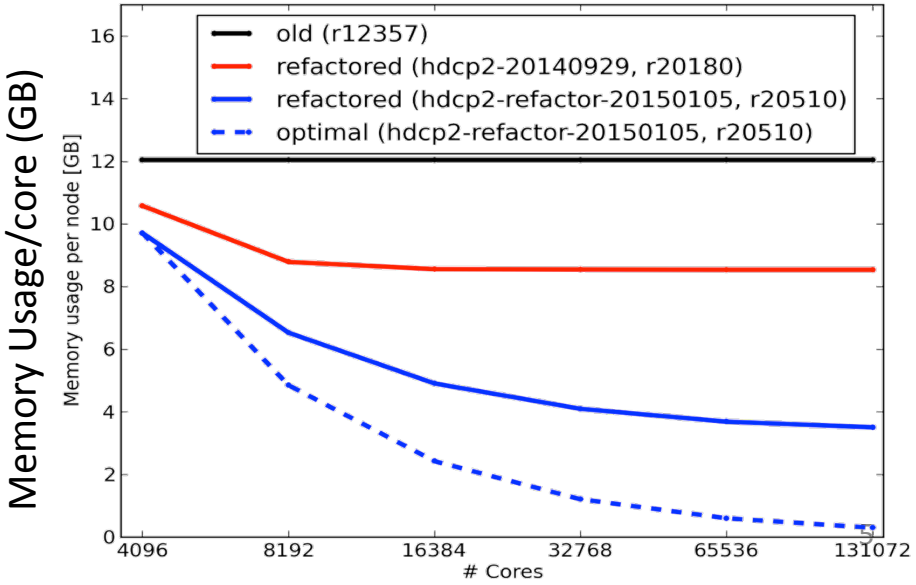
processes (# MPI tasks * # threads) per node



Total Memory (GB)

#of IBM P6 cores

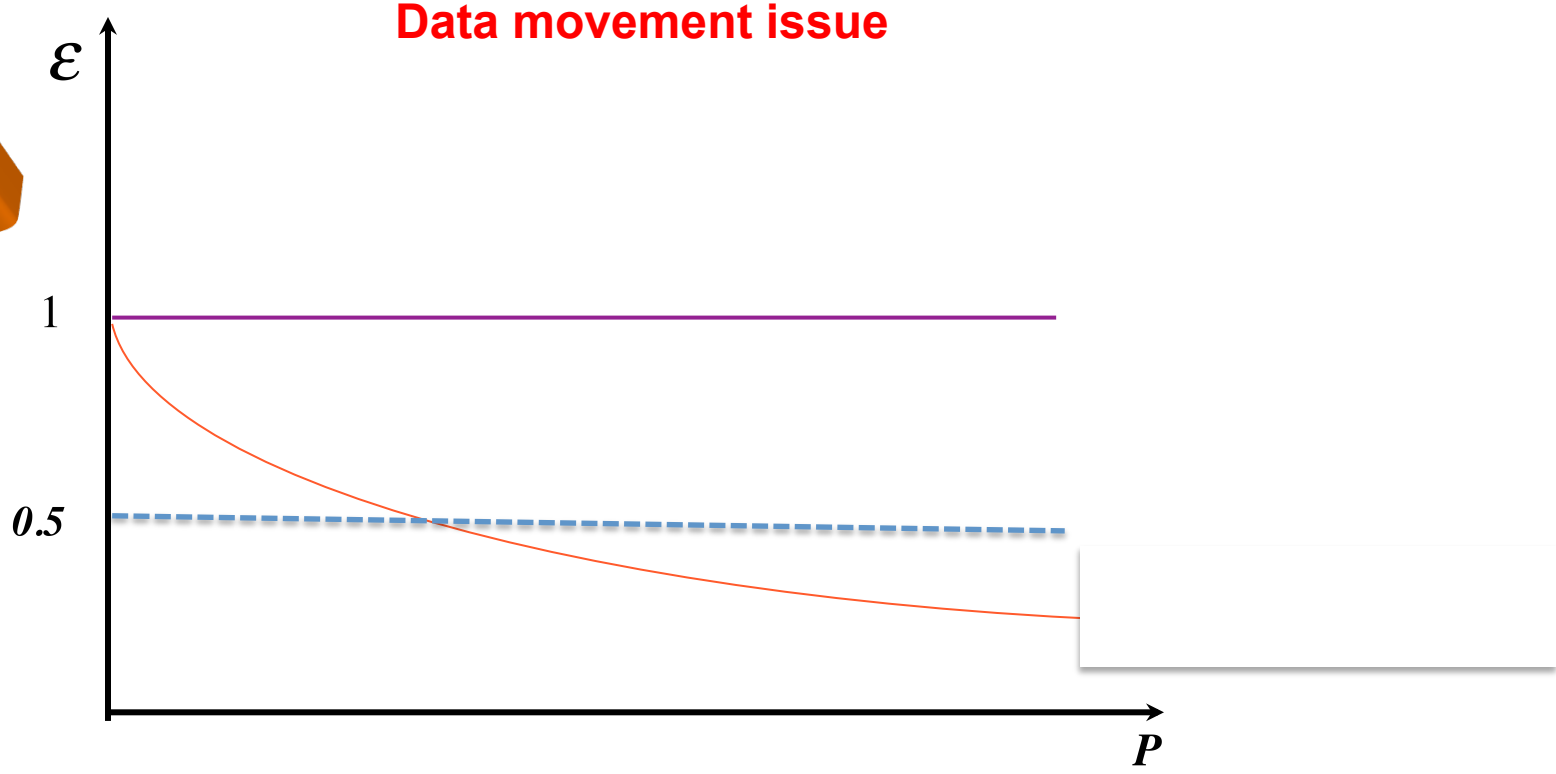
*New parallel algorithms are needed to decompose the input data over all processes of the model and hence reduce memory consumption per process (**memory scaling**)*



Memory Usage/core (GB)

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Data movement issue



D = problem dimension

n = grain size = D/P

$$\varepsilon = \frac{S}{P} = \frac{1}{1 + f_c} \qquad f_c = \frac{T_{comm}}{T_{comp}} \approx \frac{1}{n}$$

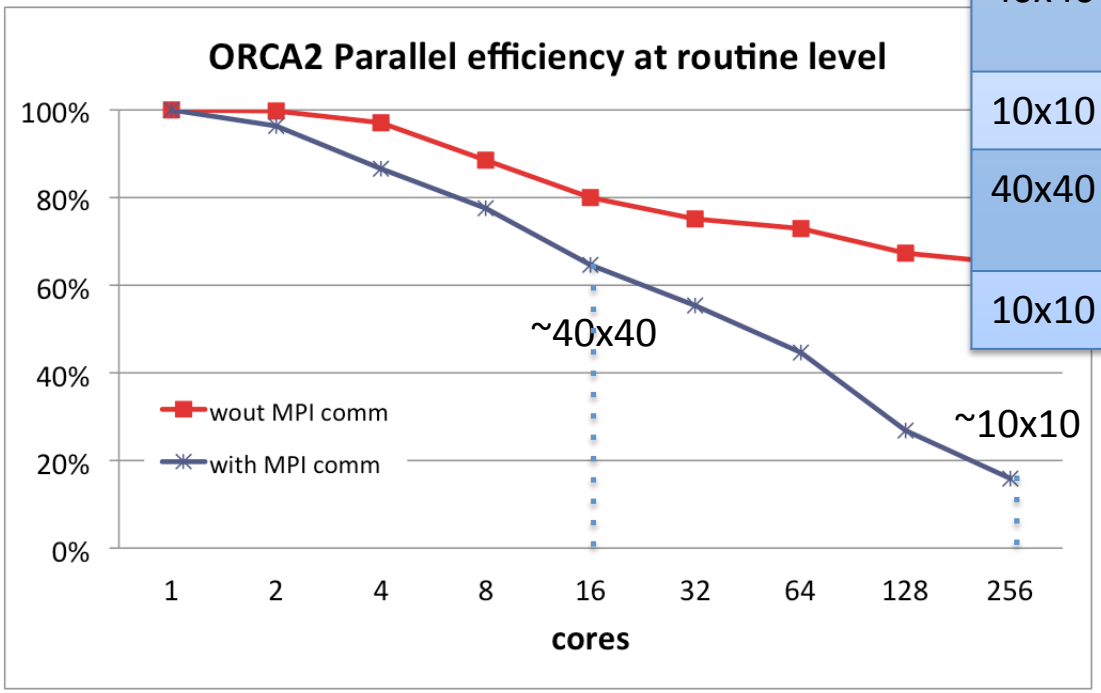
$$0 \le f_c < 1$$

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Data movement issue

NEMO currently only scales to local domain sizes of ~40x40.
The target for the future should be 10x10.

Domain size	Configuration	Scalability limit (cores)
40x40	ORCA2 (200 Km)	16
10x10	ORCA2	256
40x40	ORCA025 (25 Km)	1024
10x10	ORCA025	16384
40x40	ORCA12 (10 Km)	8000
10x10	ORCA12	131000



NEMO scalability is strongly affected by the communications overhead

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Solution: Limit the data movement overhead

- *Ghosting technique to reduce the communication frequency*
  - Increasing the halo size allows MPI processes to exchange data less frequently
- *Overlap computation and communication*
  - Use of non-blocking communications allows to overlap the halo exchange with the inner domain computation
- *MPI3 collective neighbors communications instead of point-to-point*
  - A single collective communication involving the neighboring processes can replace the 4 point-to-point communications activated at each exchange
- *Message aggregation to maximize the utilization of the full bandwidth*
  - Two or more consecutive exchanges can be aggregated to reduce the message start-up overhead
- *Use of asynchronous I/O servers (such as XIOS) to decouple the I/O tasks from the computing processes*

New hybrid programming models to better exploit hybrid architectures

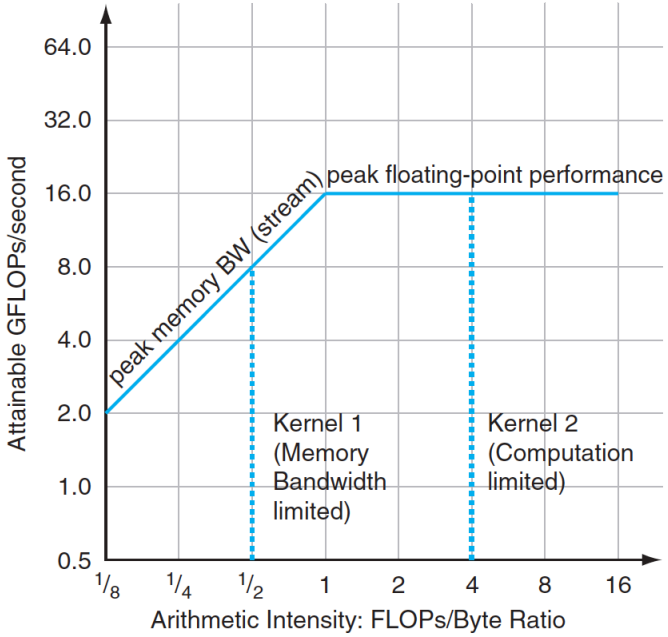
- *Introduction of a second level of parallelism in NEMO based on the OpenMP shared memory paradigm*
  - Increase of the parallelization level
  - Reduction of the memory requirements
  - Exploitation of MIC architectures
    - Execution of the hybrid code in native mode
    - High-resolution configurations can be tested on KNL, able to access 384 GB DDR4 memory
    - Code vectorization can be increased by using OpenMP4.0 explicit SIMD directives
- *Decoupling of the radiation component from the dynamical core in ICON*
  - Re-design of the current parallelization concept, from being based purely on domain decomposition towards a mixed concept, which in addition enables functional/task parallelism
  - Exploitation of hybrid architectures by placing the separated functional components on different devices (e.g. CPUs and GPUs)

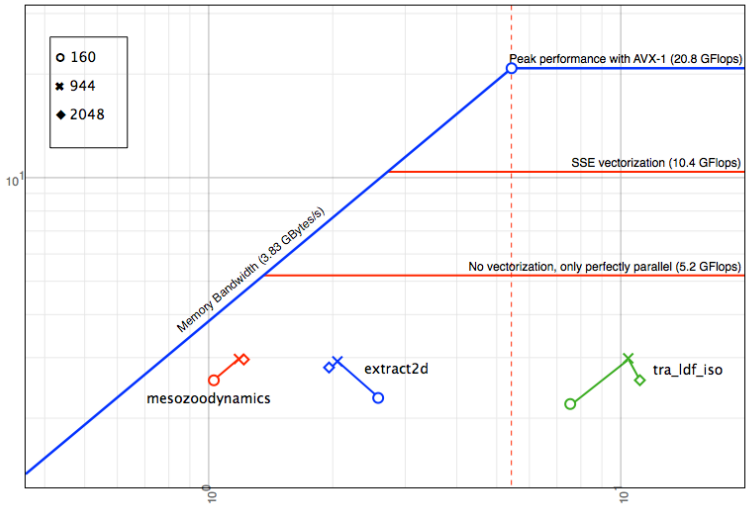
Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Increasing the computation per byte of accessed data

Analysis by using the roofline model

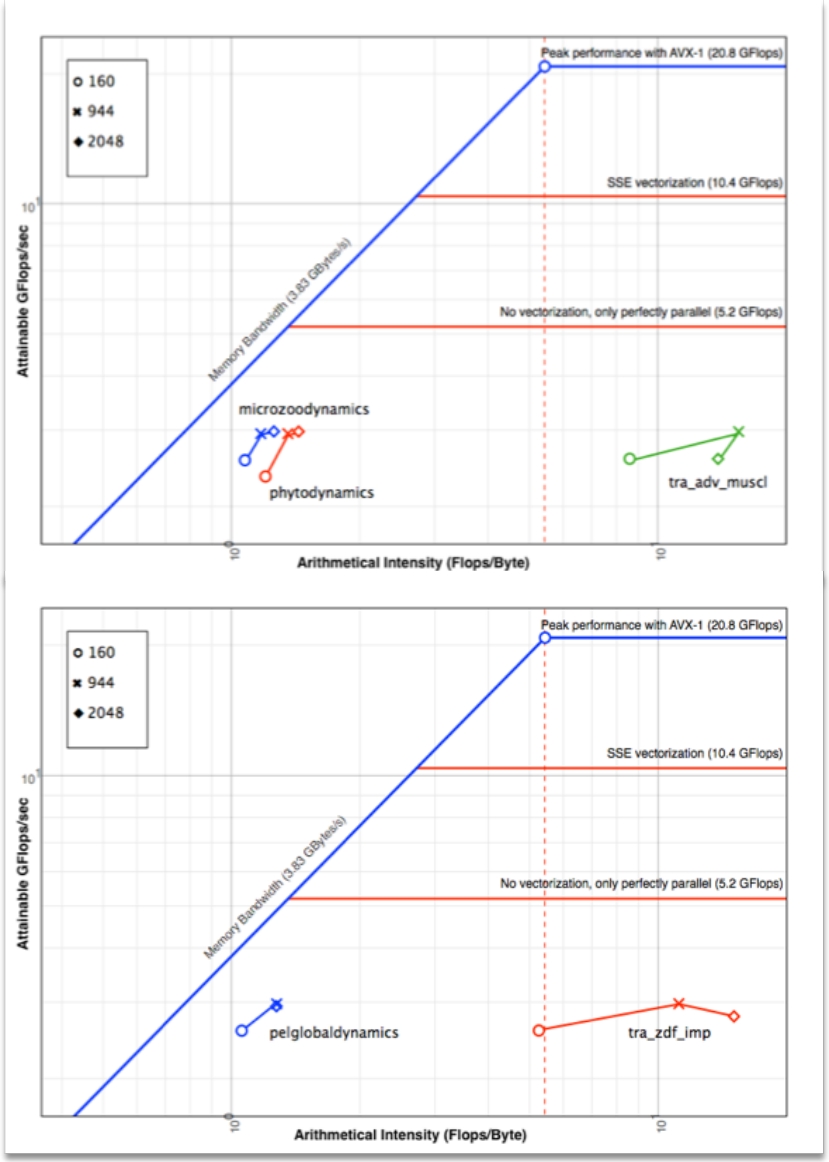
- Measuring of the hardware performance counters (at routine level) by executing the sequential version of the code (to avoid the communication overhead)
- Definition of new coding approaches to improve the computation per byte of accessed data (e.g. cache-blocking, SIMD instructions, …)
- Test on benchmark configurations

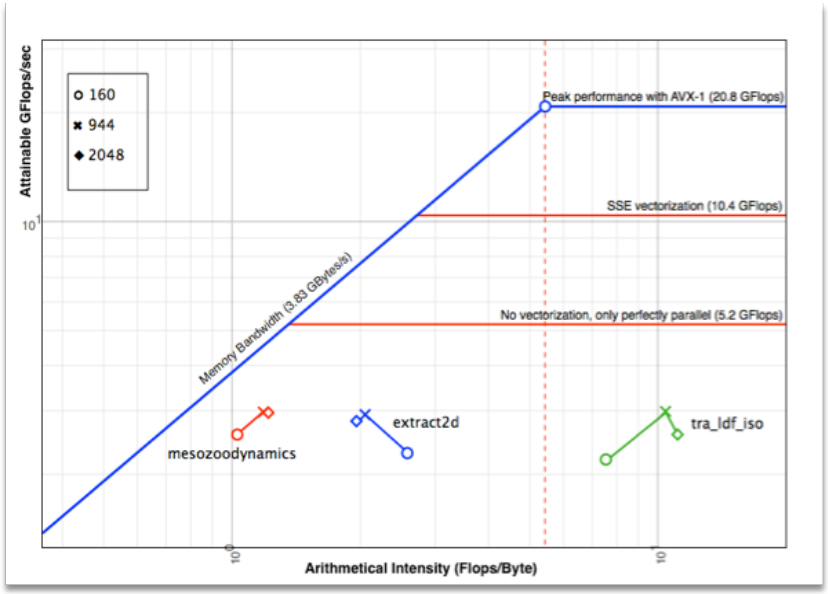




NEMO Roofline on Athena

The roofline model





At the left of the *ridge point* (memory limit)

- To improve pre-fetching with compiler
- To improve memory affinity (binding)

At the right of the *ridge point* (computation limit)

- To improve vectorization (SIMD)

Disruptive approaches to increase models scalability

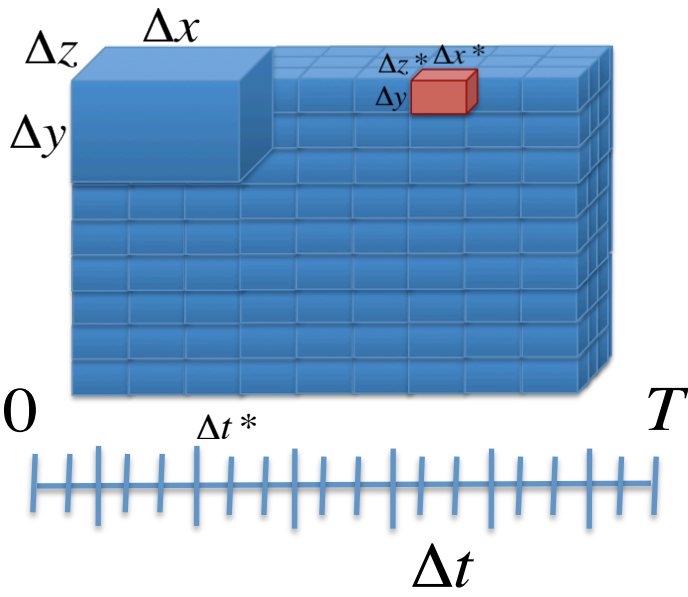
Parallelization in Time

Increasing the resolution ***r*** of a factor ***d*** means that the computational workload increases by a d^4 factor

Spatial parallelization allows to reduce the workload on three dimensions only

Parallel-in-time (PINT) methods aim at filling this gap

A deep re-design of the solvers at numerical level is needed



CFL-Condition

$$\frac{\Delta t}{\min(\Delta x, \Delta y, \Delta z)} < C$$

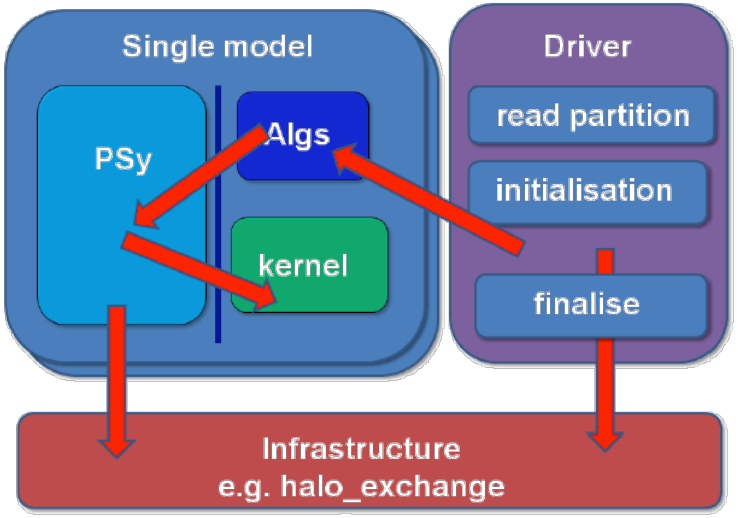
Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Disruptive approaches to increase performance portability

Separation of concerns

Separation of the computational aspects from the numerics

Decoupling methodology experienced by UK team during the Gung-Ho and gOcean projects.



- Ease development of new algorithms
- Ease maintenance
- Automated procedure for code generation (generation of hardware specific code)

- Need to rewrite the code (long-term activity)
- Need to maintain legacy code
- To few results to evaluate performance benefits

Conclusions

Future machines provide the computational power needed to increase both the resolution and the complexity of climate models

What have we learnt?

- Need to introduce techniques to reduce the data movement overhead and to increase memory scaling
- Need to adopt hybrid parallel approaches to increase the parallelism (also considering the *time* direction)
- Need to analyze and improve the computation per byte of accesses data
- Need to decouple the computational aspects from the numerics

Technology tracking: what have we learnt for climate models (NEMO-ICON)?

Giovanni Aloisio – CMCC

and

from interaction with vendors?

Joachim Biercamp - DKRZ

WP3/NA2 institutions

cmcc
MAX-PLANCK-GESELLSCHAFT
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
DKRZ
DEUTSCHES KLIMARECHENZENTRUM
Science & Technology Facilities Council
cnrs
Institut Pierre Simon Laplace
Met Office
MANCHESTER 1824
The University of Manchester
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Technology tracking: what have we learnt from interaction with vendors?

- Discussion with vendors are necessary and useful for both sides
  - These discussions happen on several levels between (IS)-ENES partners also outside IS-ENES
  - However, projects like IS-ENES and "official" bodies like the HPC-TF can provide a forum to have coordinated and thus more efficient and visible information exchange
- Vendors interested in having climate (and weather) applications as showcases
  - For the HPC branches of the vendors climate and weather are key markets
  - However HPC has to follow other markets (it's not big enough in itself)
  - HPC companies become integrators of technologies driven by other markets
- The free lunch **is** over
  - Computers are becoming broader not faster
  - Investment costs for new systems will probably no longer decrease (per e.g. FLOP)
- "Co-design" (potentially triggering innovation) is possible depending how it is defined at what is targeted

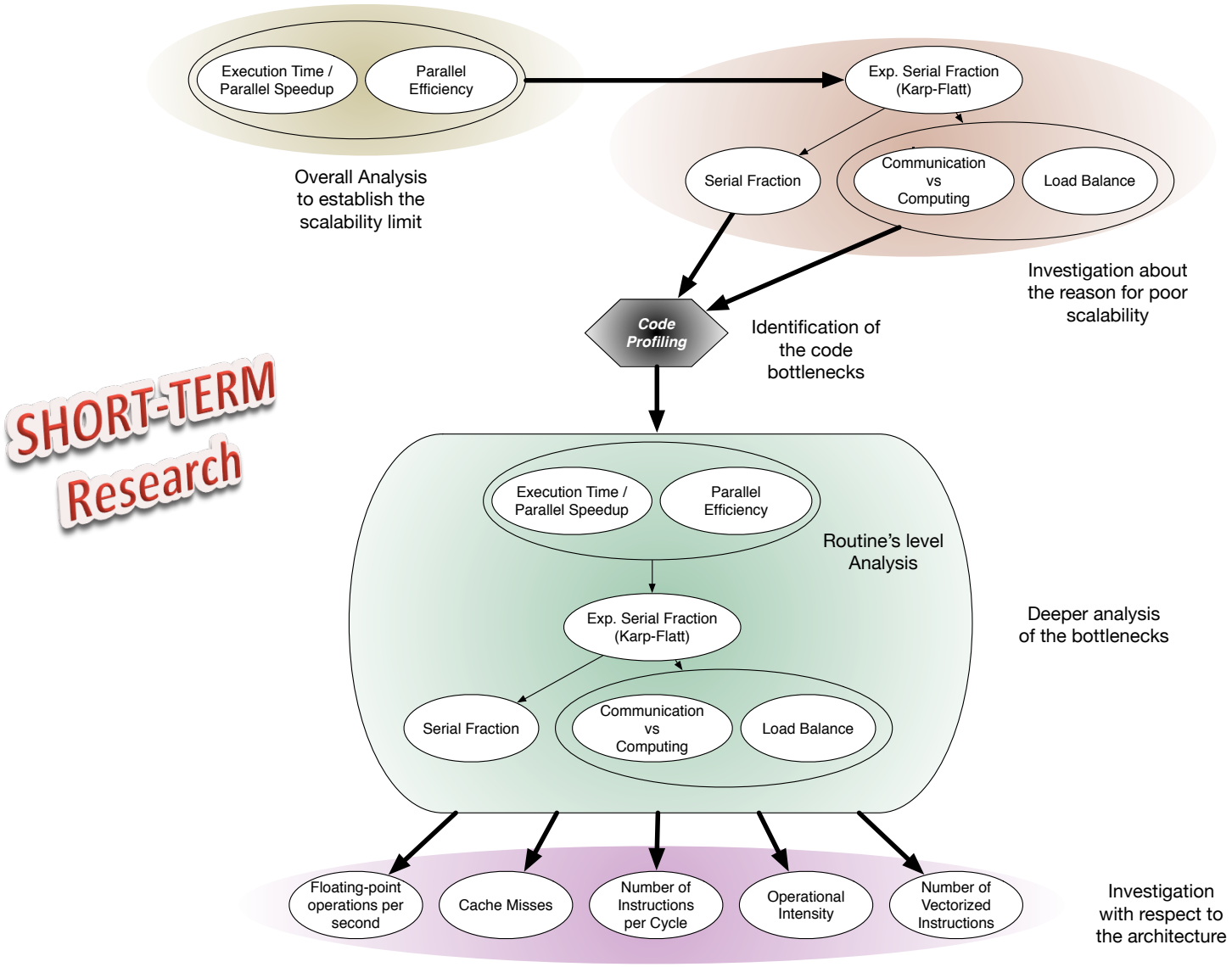
Technology tracking: what have we learnt from interaction with vendors?

Co-design …

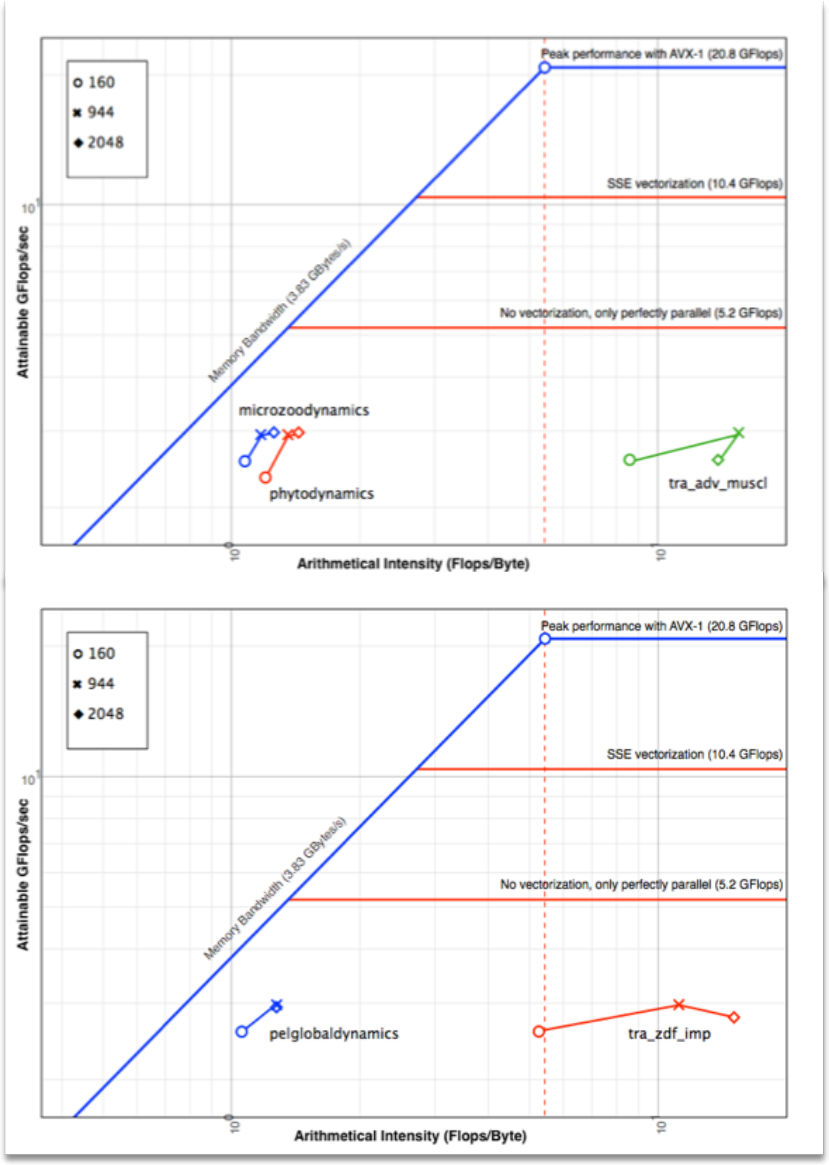
- Chances to influence HW-development are very limited
  - our market share is to small
  - there are technical limitations
- We might influence system configuration and software stack
  - EsD discussion. User requirements influence design decisions
  - choice of functional units (CPU, GPU, PGA, Mem Controller, Mem hierarchies, NICs ….) on a chip ?
  - provide good benchmarks
- We might influence the supported Examples
  - Fortran statements asked for by HPC-TF and provided by some vendors
  - possible Grid Tools support by NVIDIA
  - ACME being one application driving Coral configuration; and again EsD
- We need to influence decisions of domain scientists limited by available computing environment
  - high res vs ensembles vs data mining vs
  - e.g. computability of EPECC type simulations -> ESiWACE

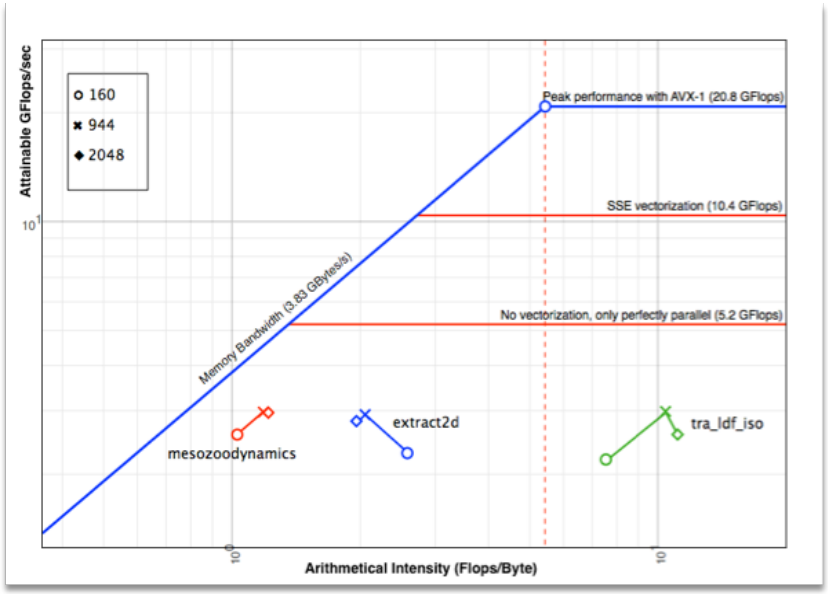
Backup Slides

The Methodology used for the model optimization



The roofline model





At the left of the *ridge point* (**memory limit**)

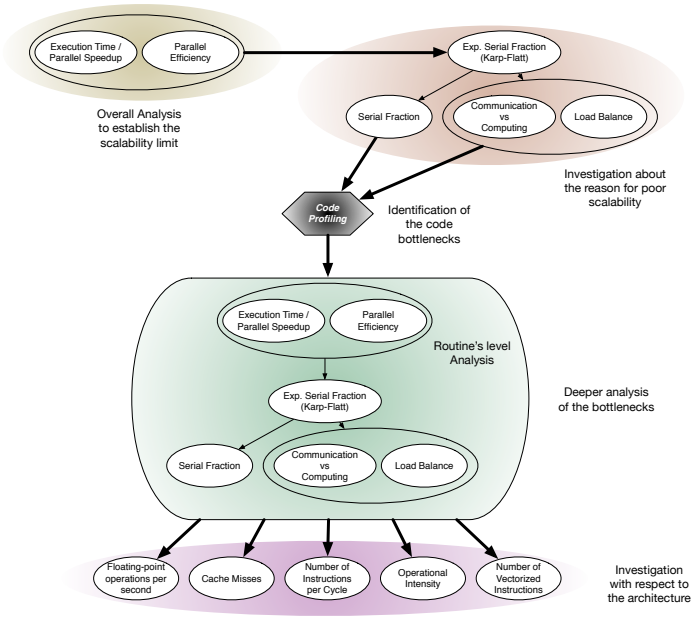
- To improve pre-fetching with compiler
- To improve memory affinity (binding)

At the right of the *ridge point* (**computation limit**)

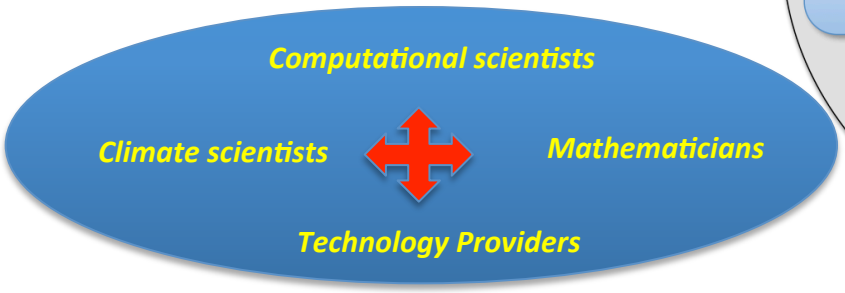
- To improve vectorization (SIMD)

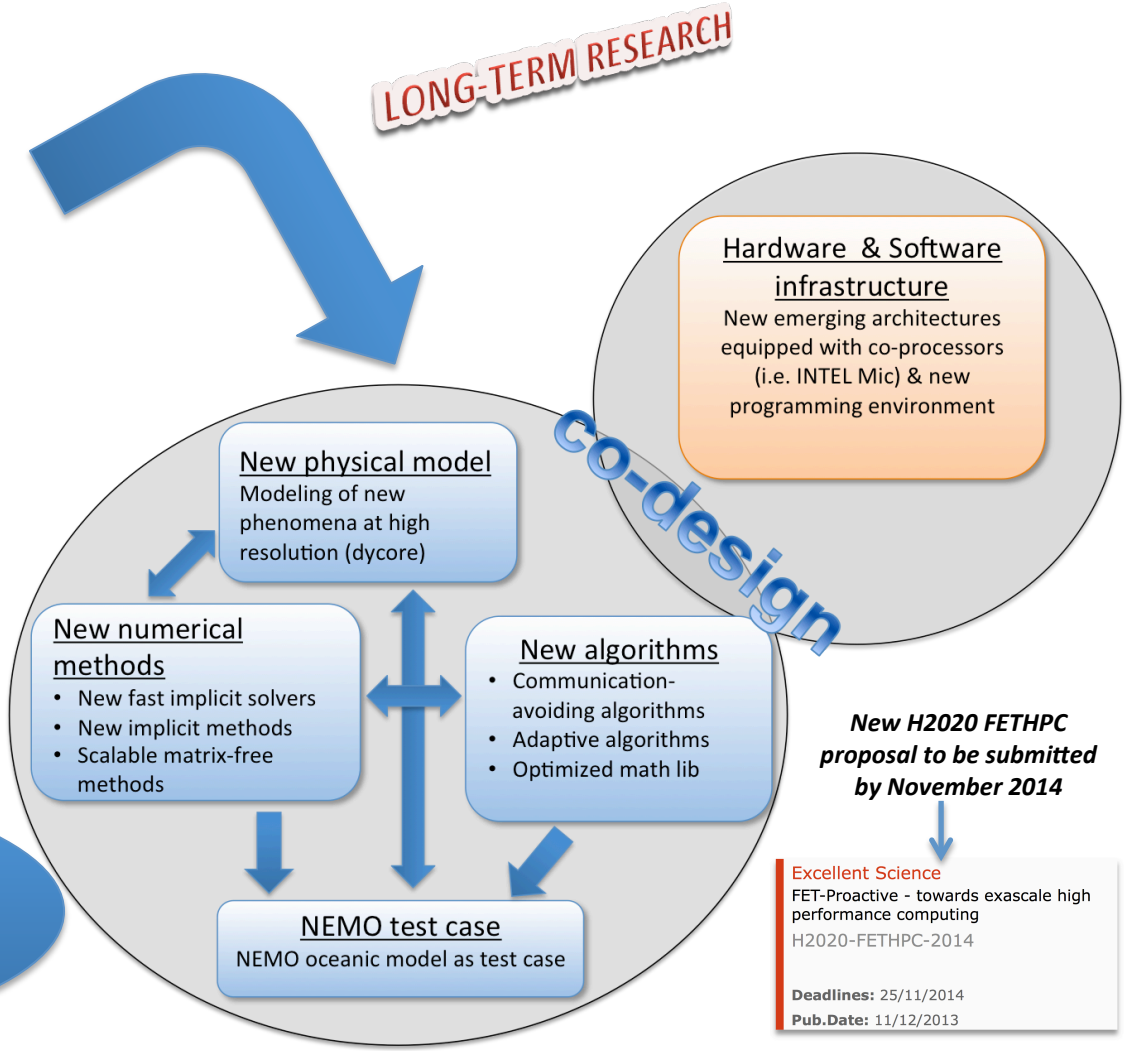
CHANCE

(Co-design of High performance Algorithms & Numerics for oCean models at Exascale)



Joint effort on *OME* *by C²MT groups*



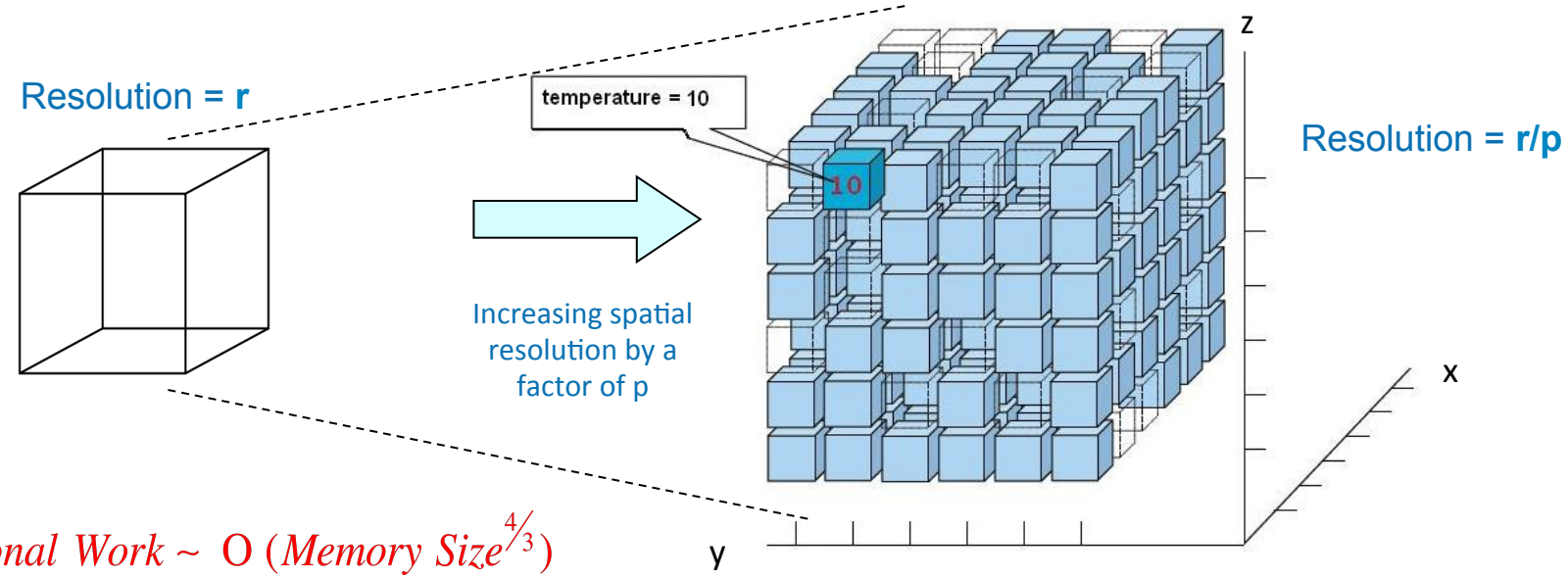


New H2020 FETHPC proposal to be submitted by November 2014

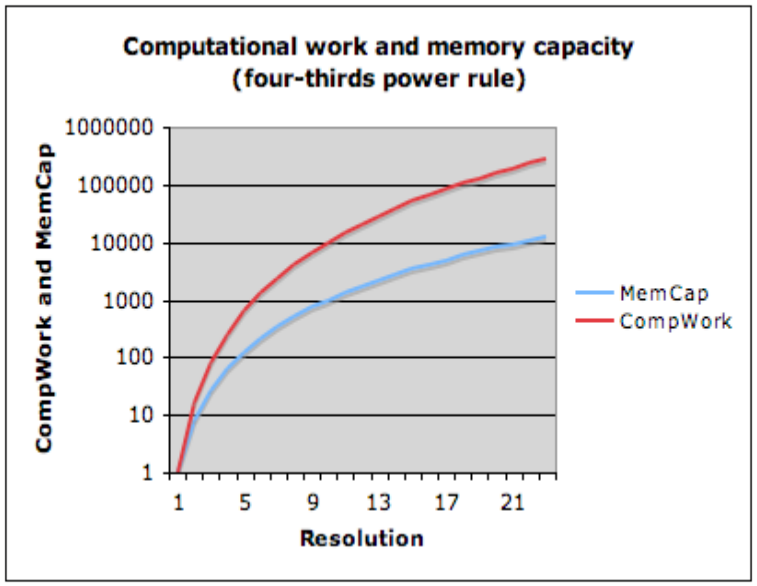
Excellent Science
FET-Proactive - towards exascale high performance computing
H2020-FETHPC-2014

Deadlines: 25/11/2014
Pub.Date: 11/12/2013

Memory requirements for high resolution climate models

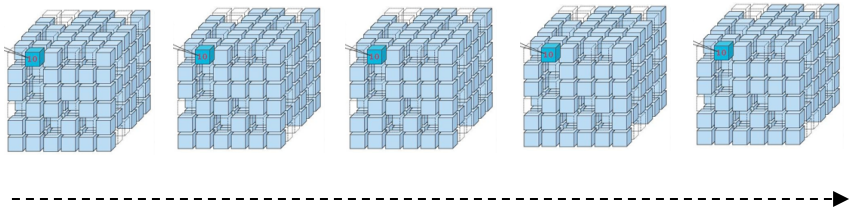


$$Computational\ Work \sim O(Memory\ Size^{4/3})$$



Four-thirds power rule

Considering p=5, it will require 125 times the memory capacity and 5 times iterations to reach the same accuracy



p=5 Timesteps

Climate simulations: Resolution increasing and memory requirements

Let's suppose a p-fold increase in the resolution: $r \rightarrow \frac{r}{p}$

⬇

a **p**-fold increase in the computing iterations and a **p^3** increase in data :

$$Computational\ Work = \alpha * p^4 \rightarrow p = \frac{1}{\alpha} * Computational\ Work^{1/4}$$

$$Memory\ Size = \beta * p^3 \rightarrow p = \frac{1}{\beta} * Memory\ Size^{1/3}$$

$$Computational\ Work^{1/4} = \frac{\alpha}{\beta} * Memory\ Size^{1/3}$$

$$Computational\ Work = \frac{\alpha}{\beta} * Memory\ Size^{4/3}$$

⬇

The computational work requirements scale as the four-thirds power of the memory size