Funded by the European Union

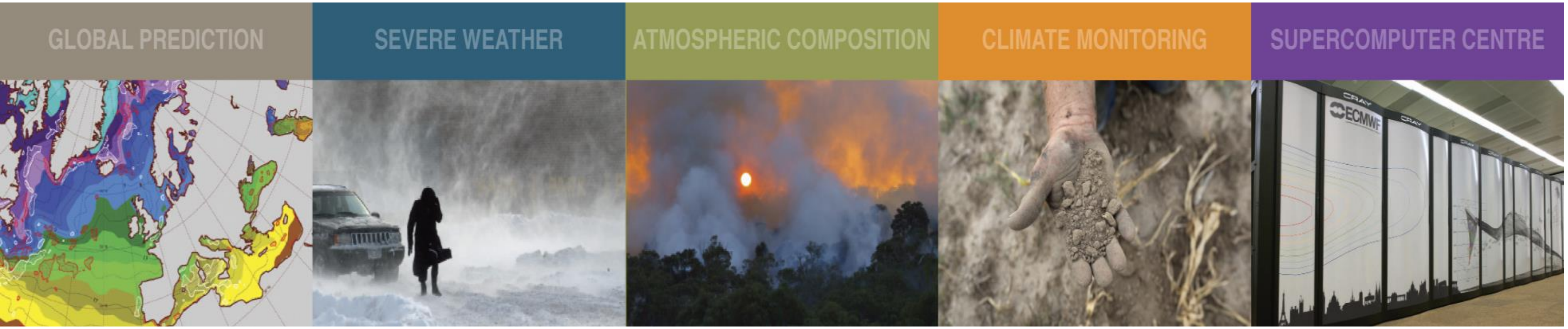
Co-ordinated by ECMWF

ESCAPE

MeteoSwiss DWD DMI METEO FRANCE Toujours un temps d'avance RMI NVIDIA ICHEC Optalysys Bull atos technologies PSNC Loughborough University

Energy-efficient SCalable Algorithms for Weather Prediction at Exascale

Nils P. Wedi, Peter Bauer, A. Mueller, W. Deconinck, ESCAPE project partners
European Centre for Medium-Range Weather Forecasts (ECMWF)

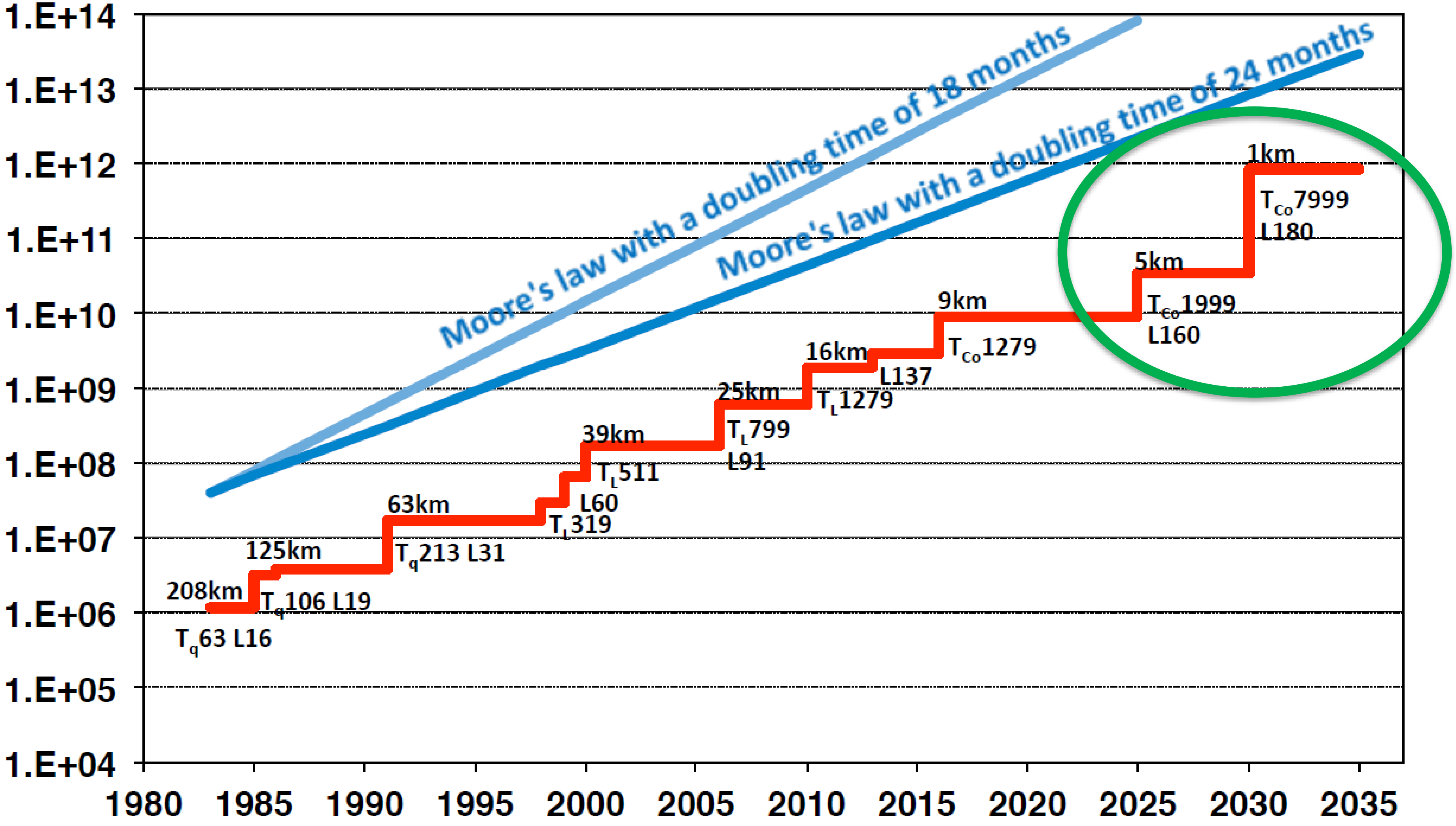


HPC workshop ECMWF, 2018

Outline

- Motivation: Emerging constraints for ensemble-based assimilation and forecasts of Weather & Climate with increasing complexity
- An intermediate goal: globally uniform weather & climate modelling at 1 km horizontal resolution
- *ESCAPE(-2) stands for*
  - Pioneering approaches for refactoring society critical legacy codes
  - Energy-efficient accelerator use in global weather & climate prediction
  - Co-development of novel mathematical algorithms & hardware adaptation
  - Defining and encapsulating the fundamental algorithmic building blocks ("Weather and Climate Dwarfs")
  - Reviewing the need for precision
  - Pioneering algorithm development with hardware adaptation using DSL toolchains
  - A HPCW benchmark and cross-disciplinary Verification, Validation, and Uncertainty Quantification (VVUQ)
  - Application resilience

Computational power drives spatial resolution



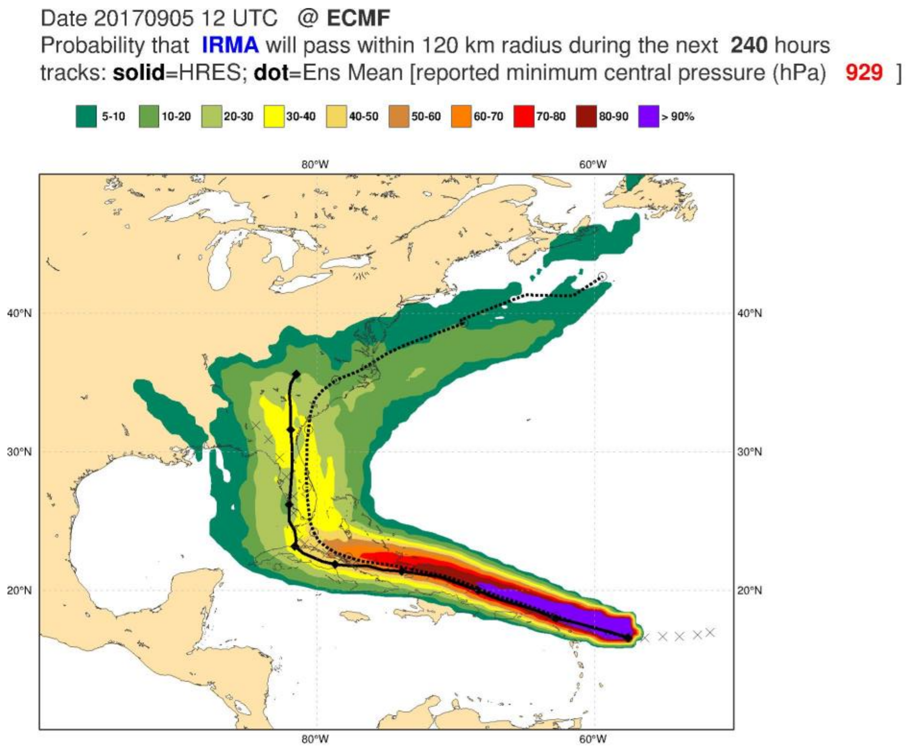
Gap of sustained and peak performance

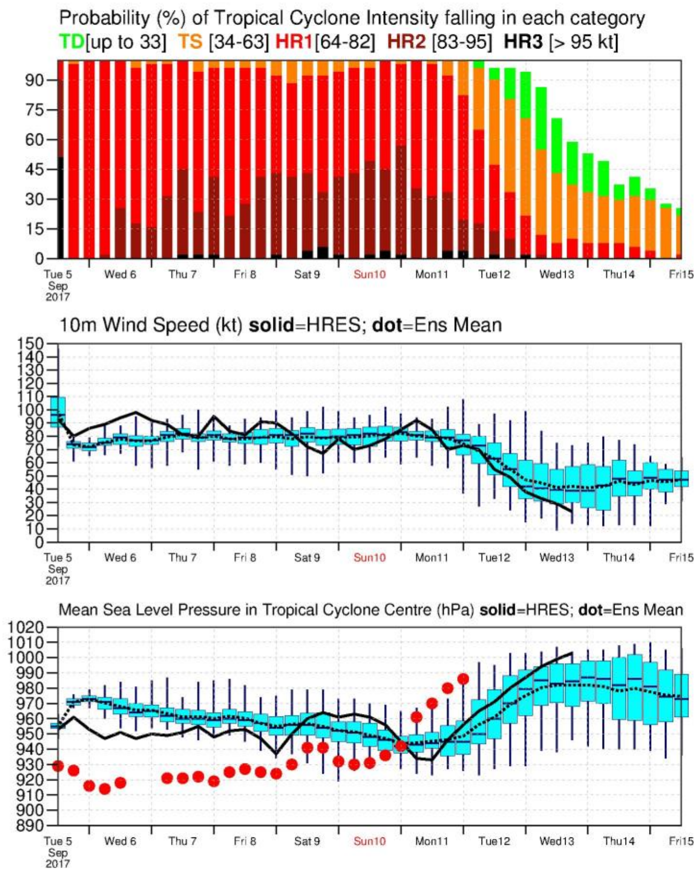
Steepness of gradient from 10km to 1km

(Schulthess et al, 2018)

ECMWF's progress in degrees of freedom
(levels x grid columns x prognostic variables)

Hurricane IRMA 18km vs 5km ensemble

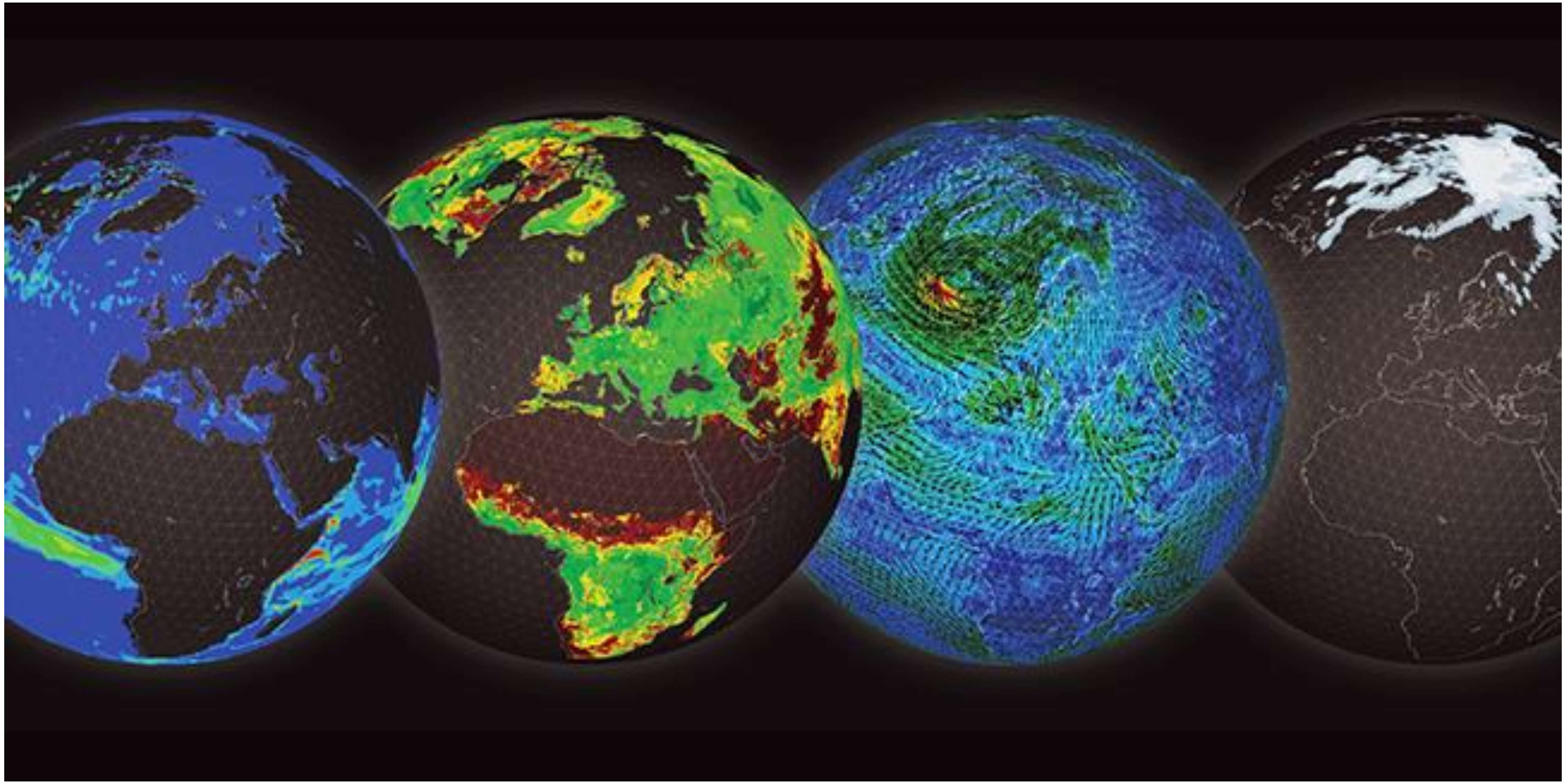




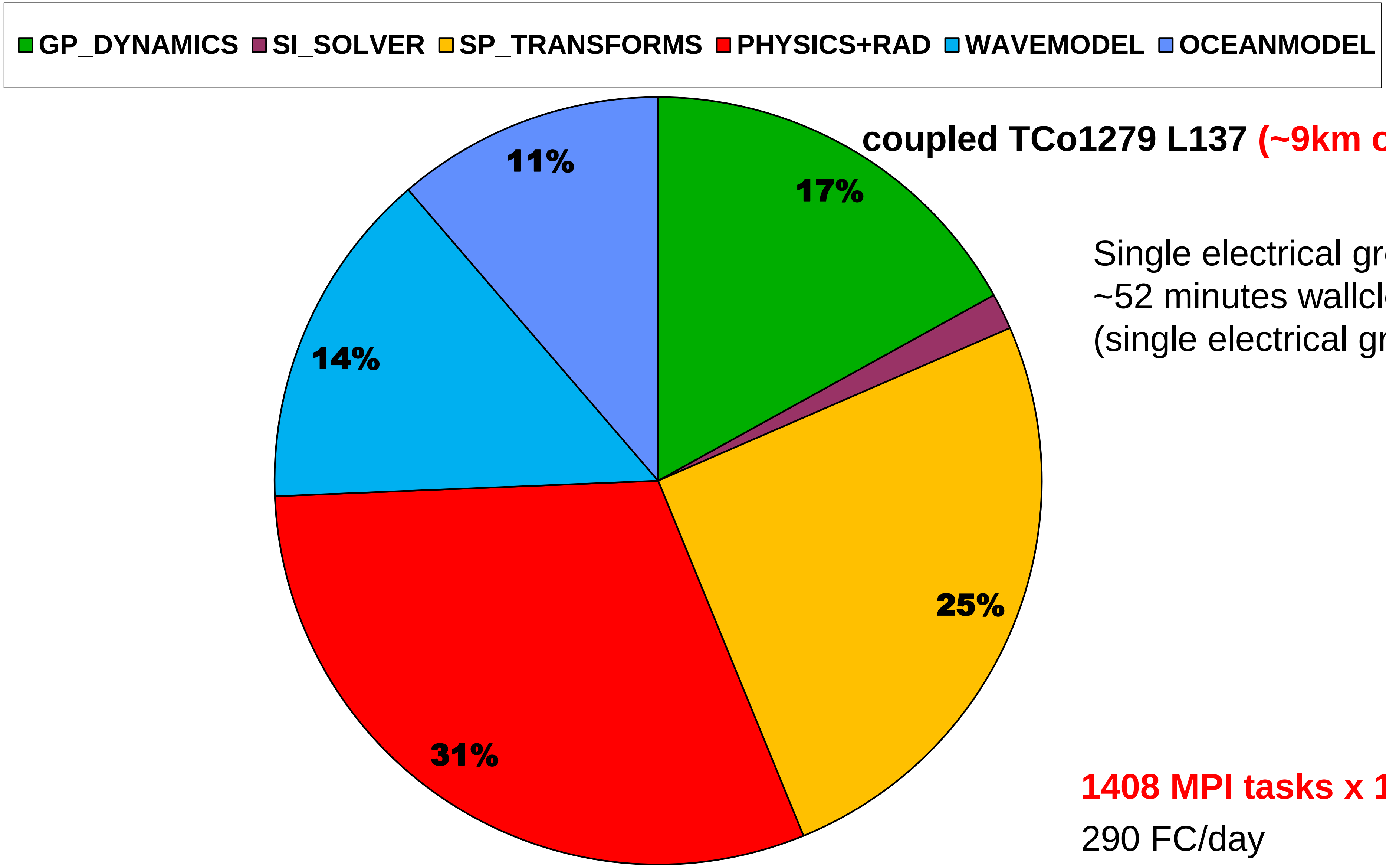
ECMWF Strategy 2025
a 5km ensemble …

S. Lang & L. Magnusson

Ocean – Land – Atmosphere – Sea ice



Where do we spend the time ? Cycle 45r1



coupled TCo1279 L137 (~9km operational) run

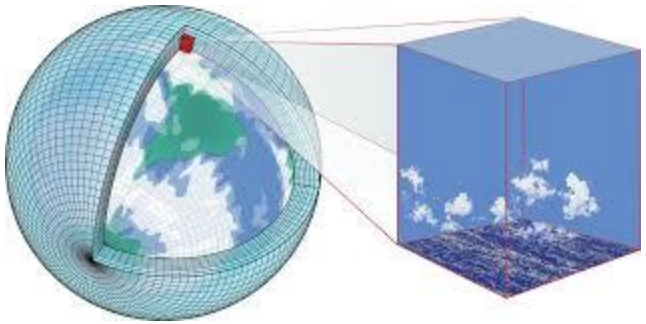
Single electrical group:
~52 minutes wallclock time
(single electrical group==384 nodes)

1408 MPI tasks x 18 threads

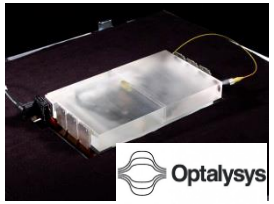
290 FC/day

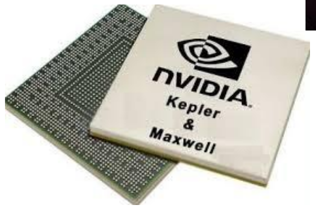
Weather & Climate Dwarfs

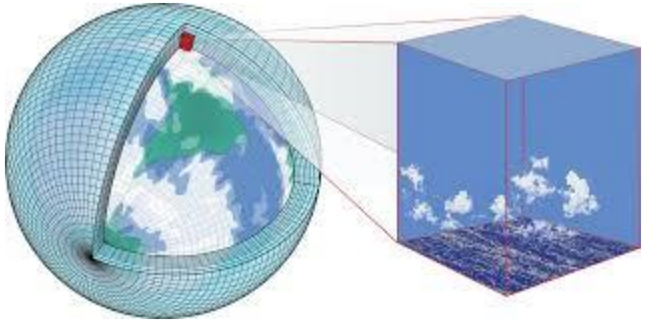
(hpc-escape.eu)



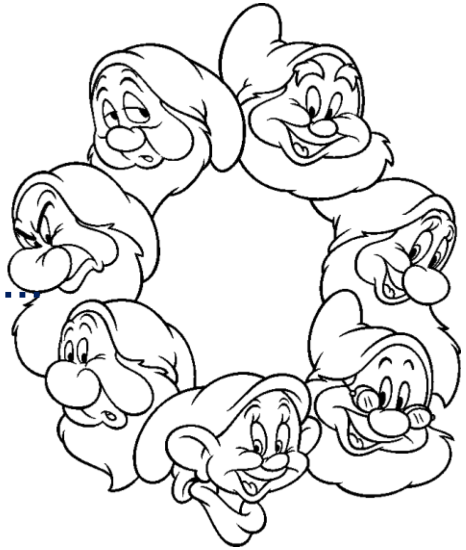
… hardware adaptation …



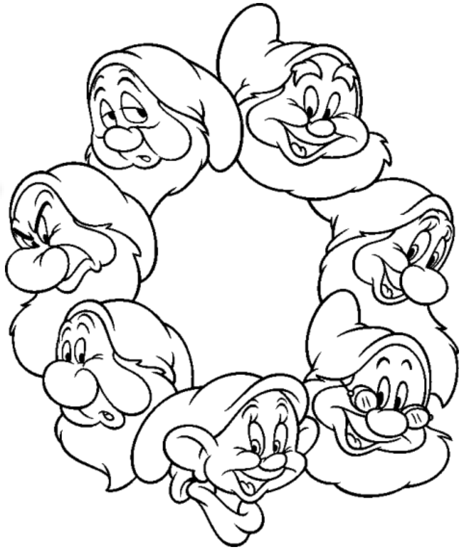




Extract model dwarfs…



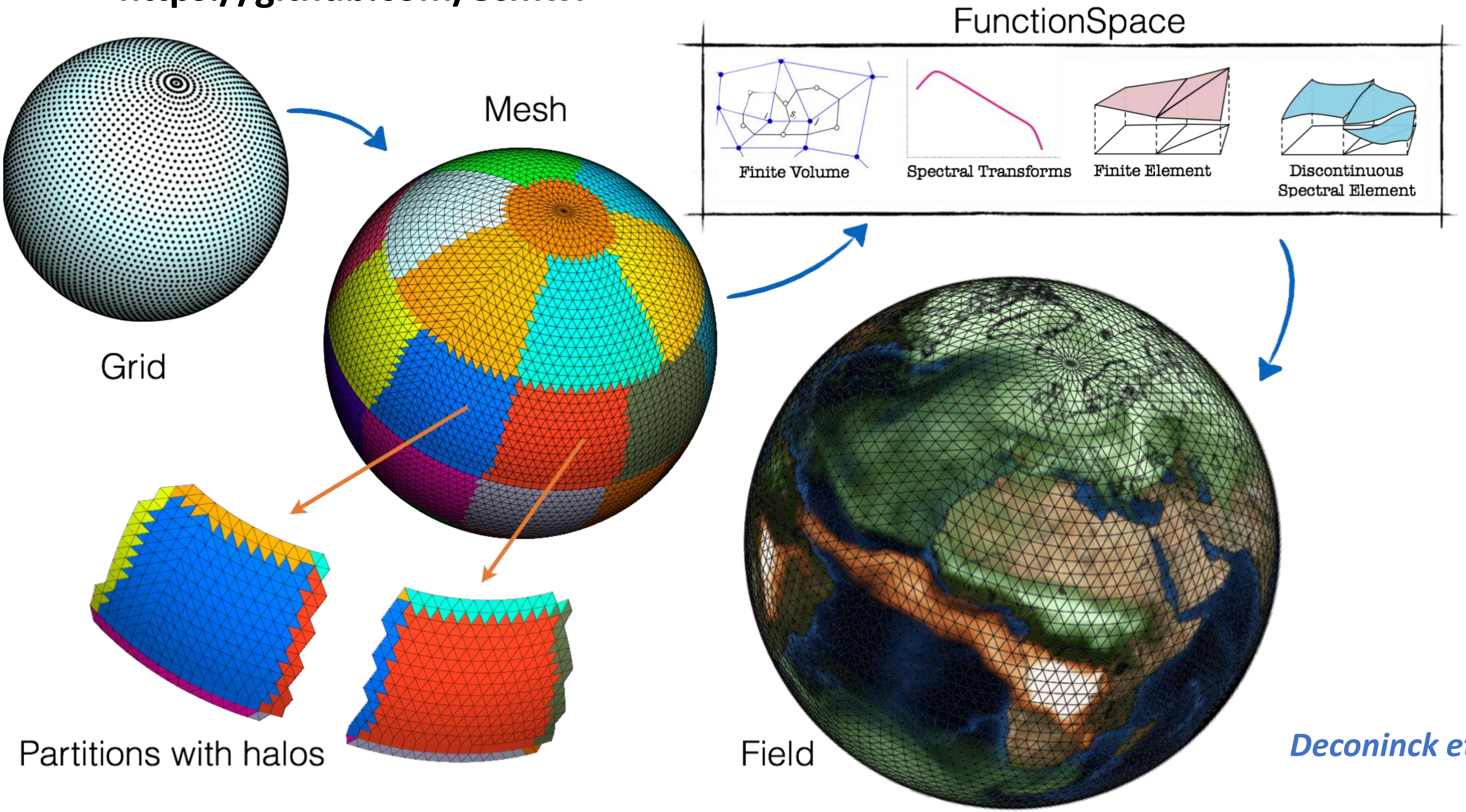
… explore alternative numerical algorithms …



… reassemble model and benchmark

Atlas: a library for NWP and climate modelling

https://github.com/ecmwf



Deconinck et al. 2017

Domain-specific language toolchain

Funded by the European Union

$$F(\Psi_L, \Psi_R, U) \equiv [U]^+\Psi_L + [U]^-\Psi_R \qquad (3a)$$

$$U \equiv \frac{u\delta t}{\delta x}, \quad [U]^+ \equiv 0.5(U+|U|), \quad [U]^- \equiv 0.5($$

Advection (MPDATA)

```
template <uint_t Color> struct upwind_flux {
  using flux = accessor<0, enumtype::inout, icosahe
  using pD =
      in_accessor<1, icosahedral_topology_t::verti
  using vn = in_accessor<2, icosahedral_topology_t

  typedef boost::mpl::vector<flux, pD, vn> arg_lis

  template <typename Evaluation> static void Do(Eva

    constexpr auto neighbors_offsets =
        connectivity<edges, vertices, Color>::offs
    constexpr auto ip0 = neighbors_offsets[0];
    constexpr auto ip1 = neighbors_offsets[1];

    float_type pos = math::max(eval(vn()), (float_
    float_type neg = math::min(eval(vn()), (float_

    eval(flux()) = eval(pos * pD(ip0) + neg * pD(i
  }
};
```

```
ute_upwind_flux(this,pflux,pD,pVn)
e), intent(inout) :: this
t(out) :: pflux(:,:)
t(in)  :: pVn(:,:), pD(:,:)
s, zneg
dges
evels
ge, jlev, ip1, ip2

debug('compute_upwind_flux')

s%dimensions%nb_edges
s%dimensions%nb_levels

DO SCHEDULE(STATIC) PRIVATE(jedge,jlev,ip1,ip2,zpos,zneg)
_edges
ode(1,jedge)
ode(2,jedge)
b_levels
          = max(0._wp,pVn(jlev,jedge))
          = min(0._wp,pVn(jlev,jedge))
jedge)    = pD(jlev,ip1)*zpos+pD(jlev,ip2)*zneg
enddo
!$OMP END PARALLEL DO
end subroutine compute_upwind_flux
```

Complementary skills of CLAW, GridTools (MeteoSwiss) and Atlas (ECMWF)

ECMWF · MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC · Deutscher Wetterdienst Wetter und Klima aus einer Hand · DWD · Optalysys · Bull · PSNC · Loughborough University

Funded by the European Union

Weather & Climate Dwarfs

Dwarf	prototype implemented	docu-mented	based on Atlas	MPI	Open MP	Open ACC	DSL	Optalysys
D - spectral transform - SH	✓	✓	✓	✓	✓	✓		
D - spectral transform - biFFT	✓	✓		✓	✓	✓		✓
D - advection - MPDATA	✓	✓	✓	✓	✓	✓	✓	
D - advection - semi-Lagrangian	✓	✓	✓	✓				
D - elliptic solver - GCR	✓	✓	✓	✓	✓			
P - cloud microphysics - CloudSC	✓	✓		✓				
P - radiation scheme - ACRANEB2	✓	in progress	in progress	✓				
I - LAITRI (3d interpol. algorithm)	✓	✓						
planned next:								
D - advection - discontinousGalerkin	planned	planned	planned	planned				
D - elliptic solver - multigridPrecon	planned	planned	planned	planned				

✓: first version running

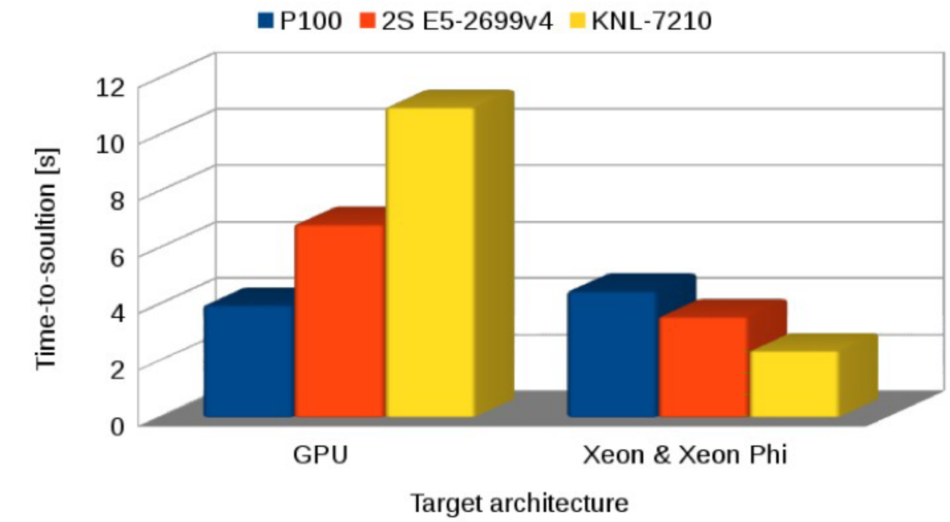
in progress

planned

t

APE

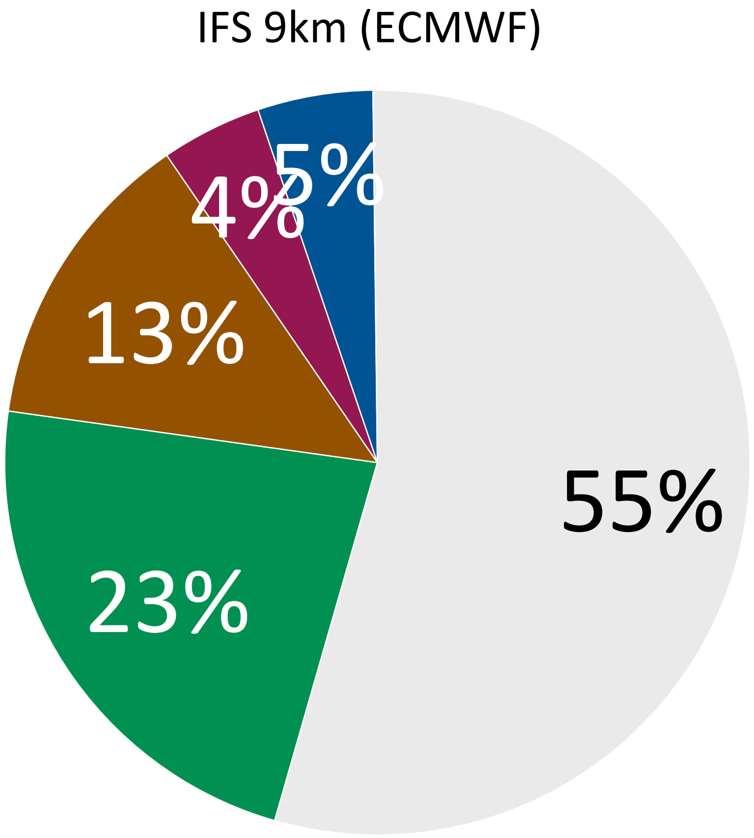
Comparison of software optimized for GPU and Xeon processors

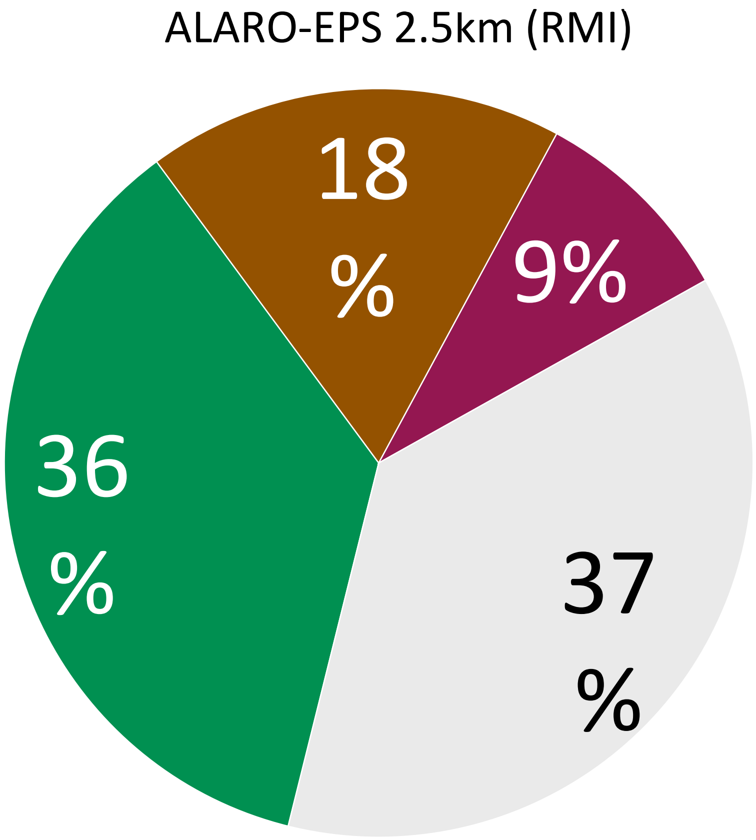


Poulsen & Berg (2017)

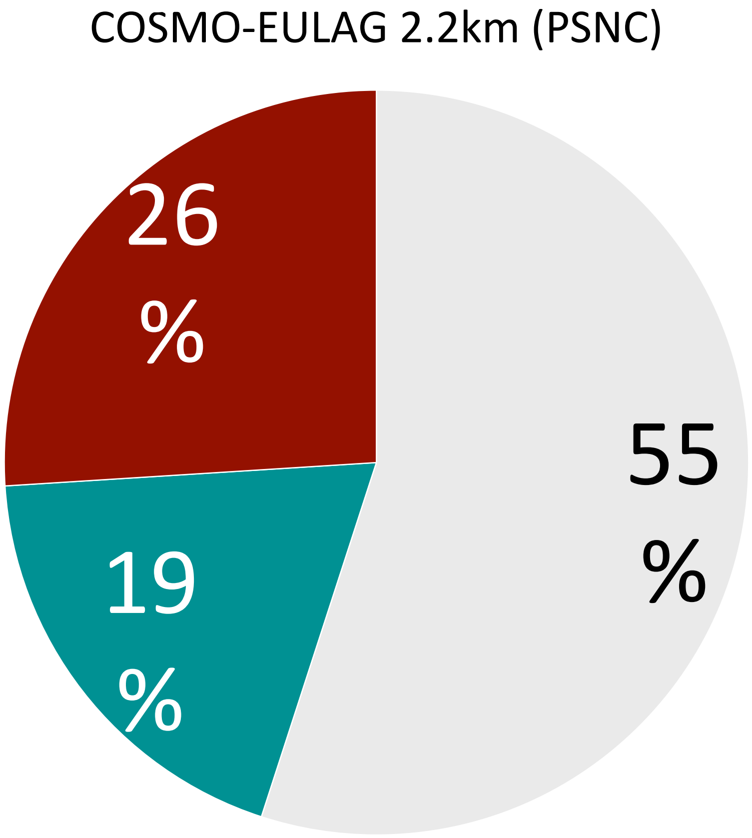
ECMWF MeteoSwiss DMI RMI NVIDIA METEO FRANCE Toujours un temps d'avance ICHEC Deutscher Wetterdienst Wetter und Klima aus einer Hand Optalysys Bull PSNC Loughborough University

dwarfs





spectral transform



GCR solver

ESCAPE

Funded by the European Union

Advection

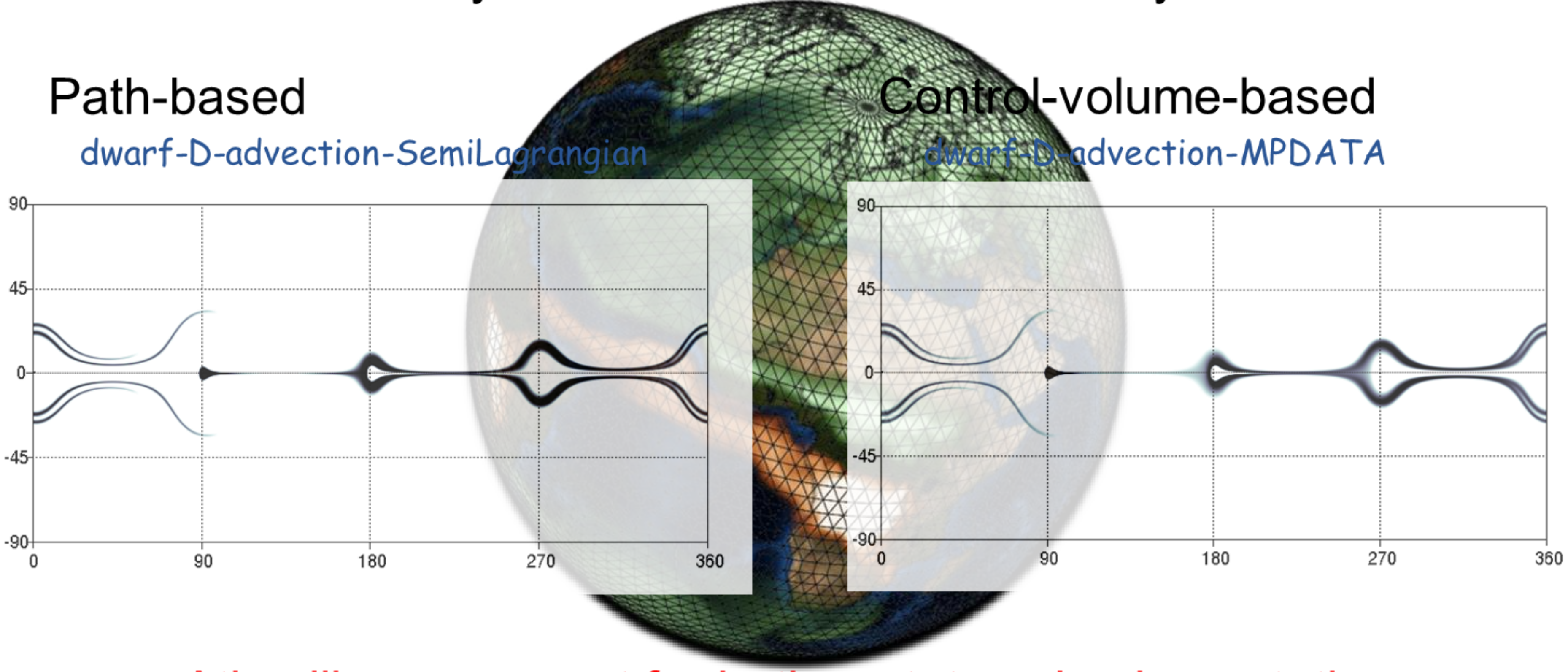
Rossby-Haurwitz test case after 7 days

Path-based

dwarf-D-advection-SemiLagrangian

Control-volume-based

dwarf-D-advection-MPDATA

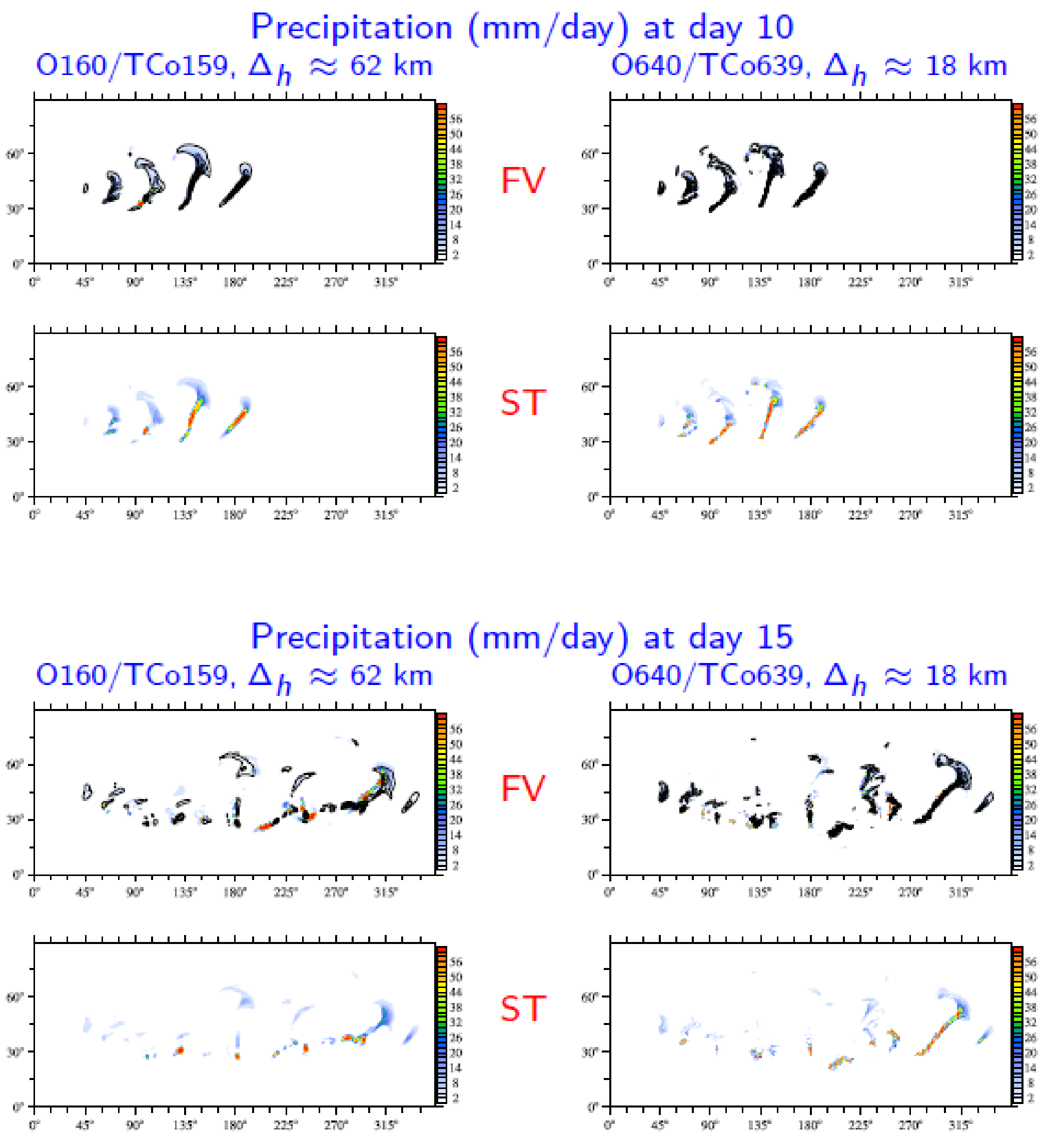


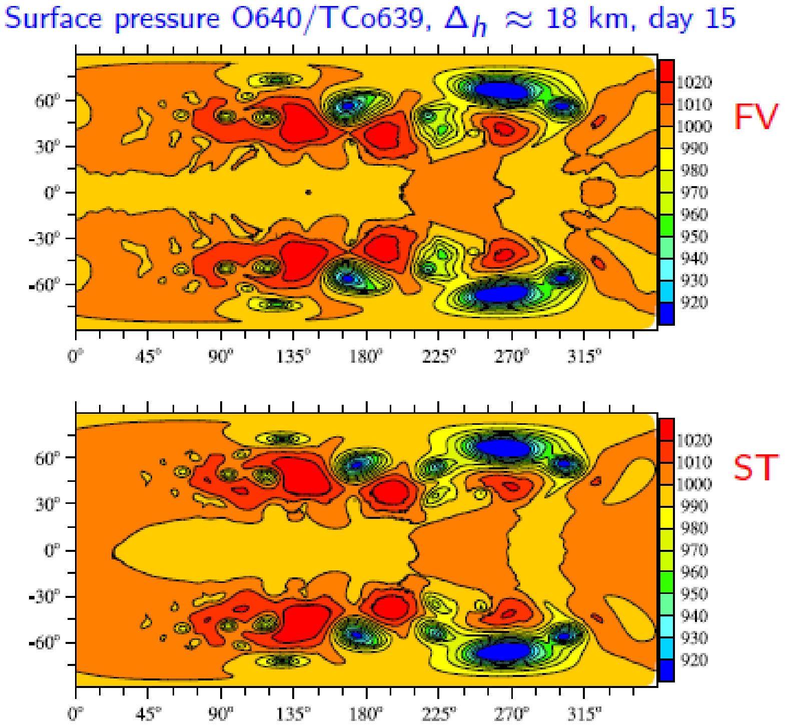
Atlas library support for both prototype implementations

Moist baroclinic instability with FVM and spectral-transform IFS (ST) using large-scale condensation and diagnostic precipitation:

Alternative dynamical core choices on the same grid with the same physics!

Christian Kuehnlein





Finite volume

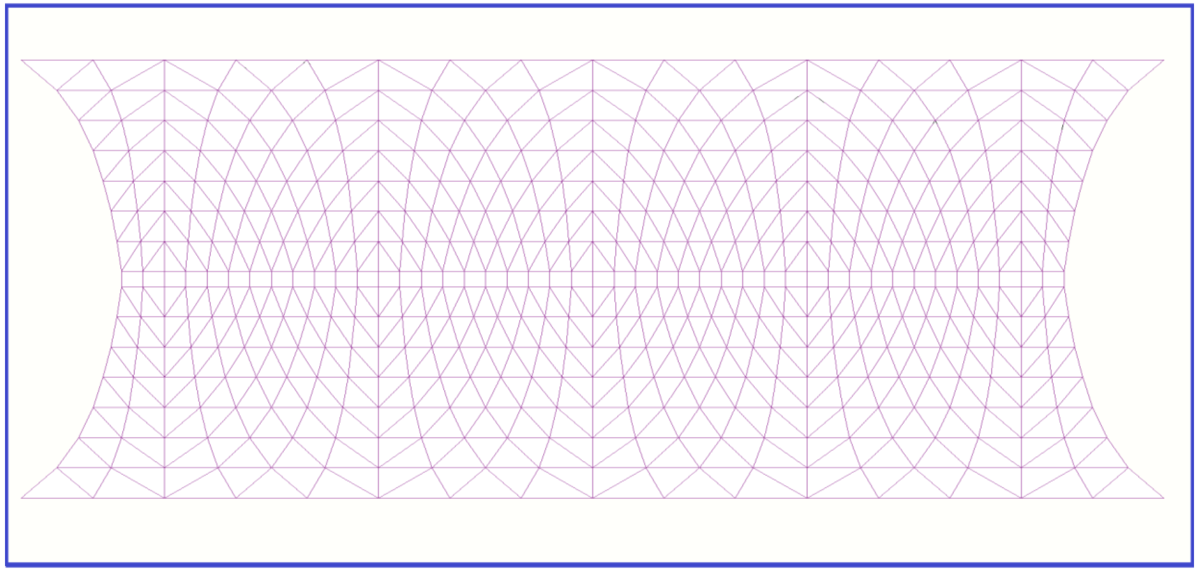
Spectral transform

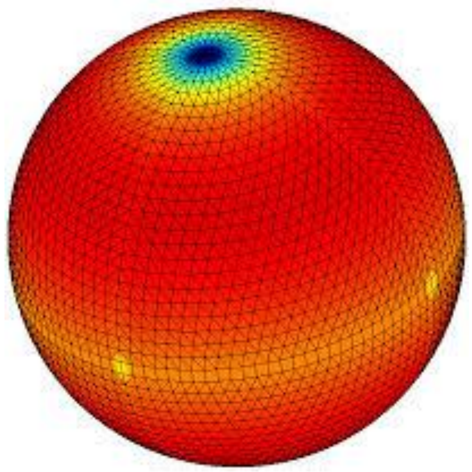
Another alternative: higher-order finite-difference development by *M. Glinton & P. Bénard*

Bespoke Krylov solvers

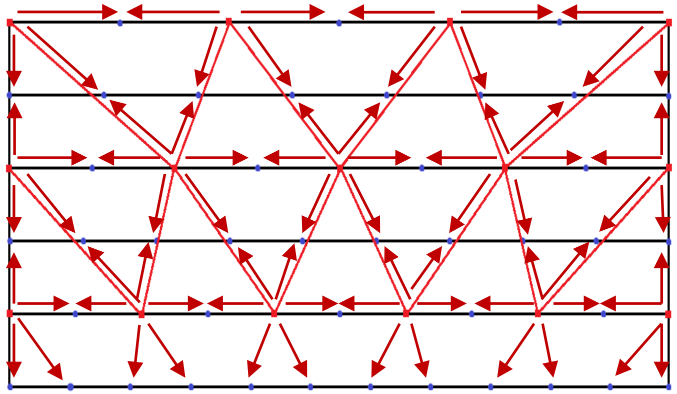
Funded by the European Union

Parallel restriction, prolongation and Atlas mesh generation



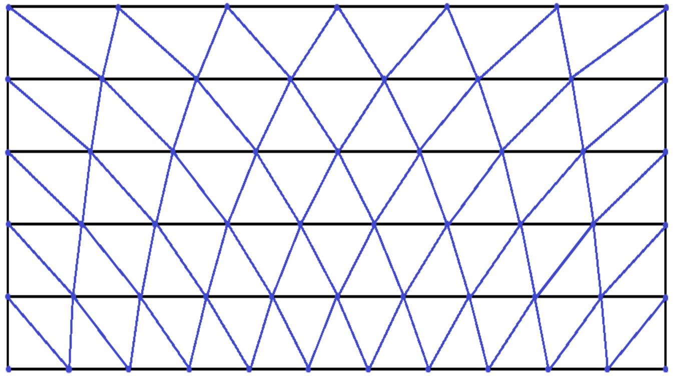


Distributed & Atlas



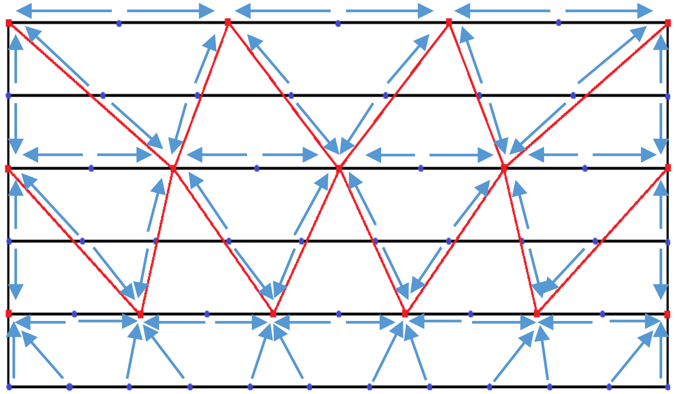
Coarse mesh

interpolation



Fine mesh

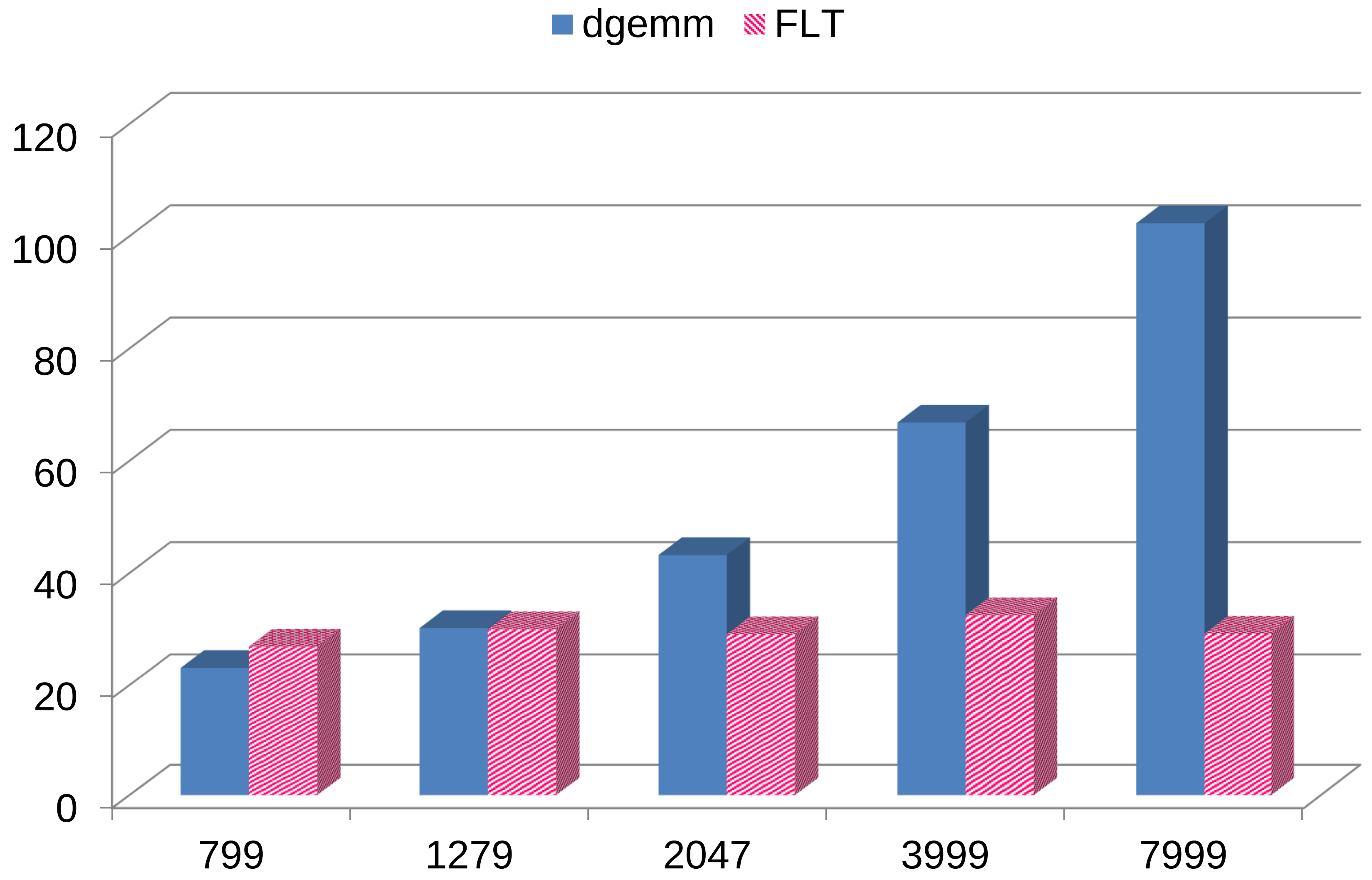
restriction



Coarse mesh

Fast Legendre Transform

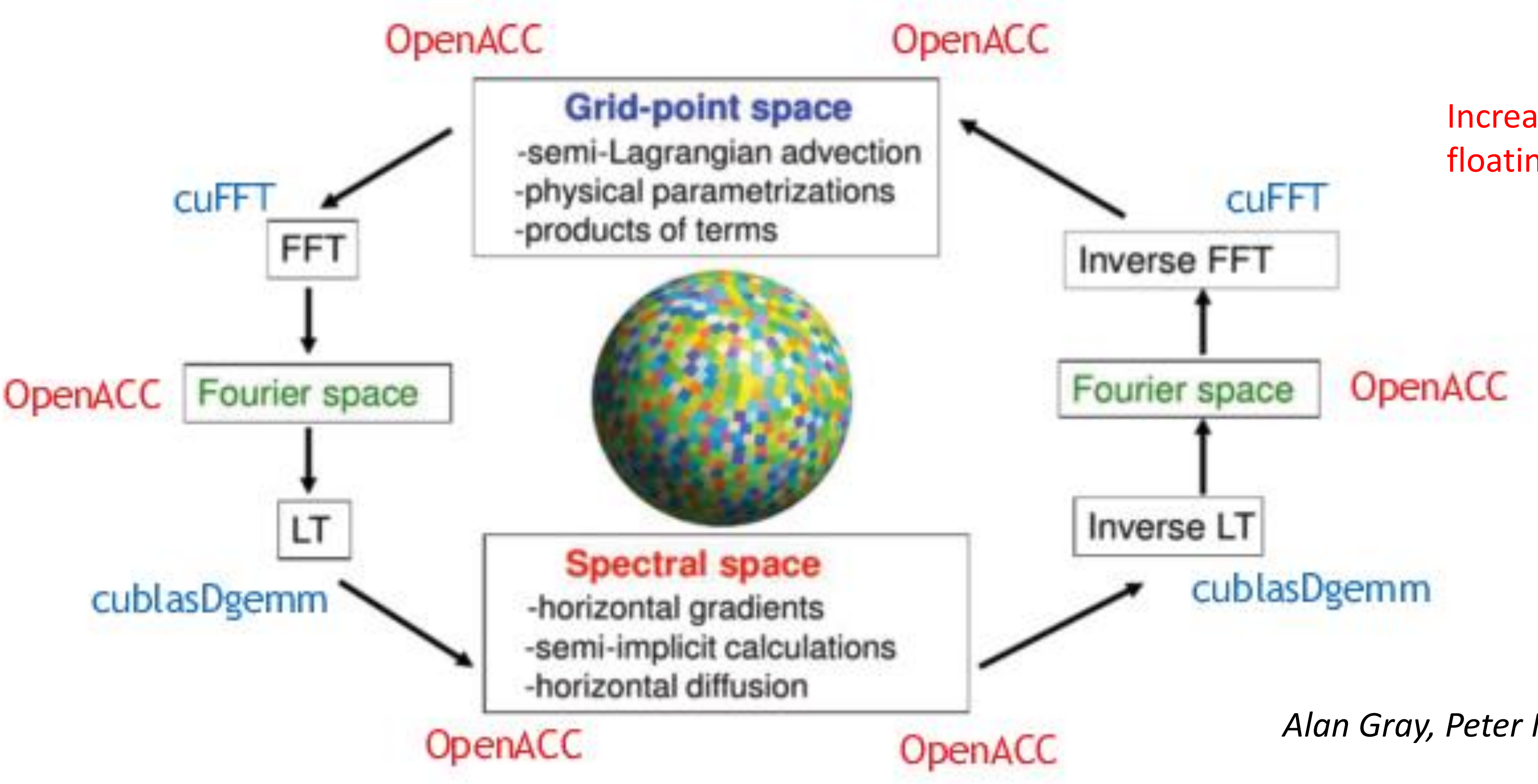
Reduces number of floating point operations



Number of floating point operations for direct or inverse spectral transforms of a single field, scaled by $N^2 \log^3 N$

(Wedi et al, 2013)

Schematic description of the *spectral transform method* in the ECMWF IFS model



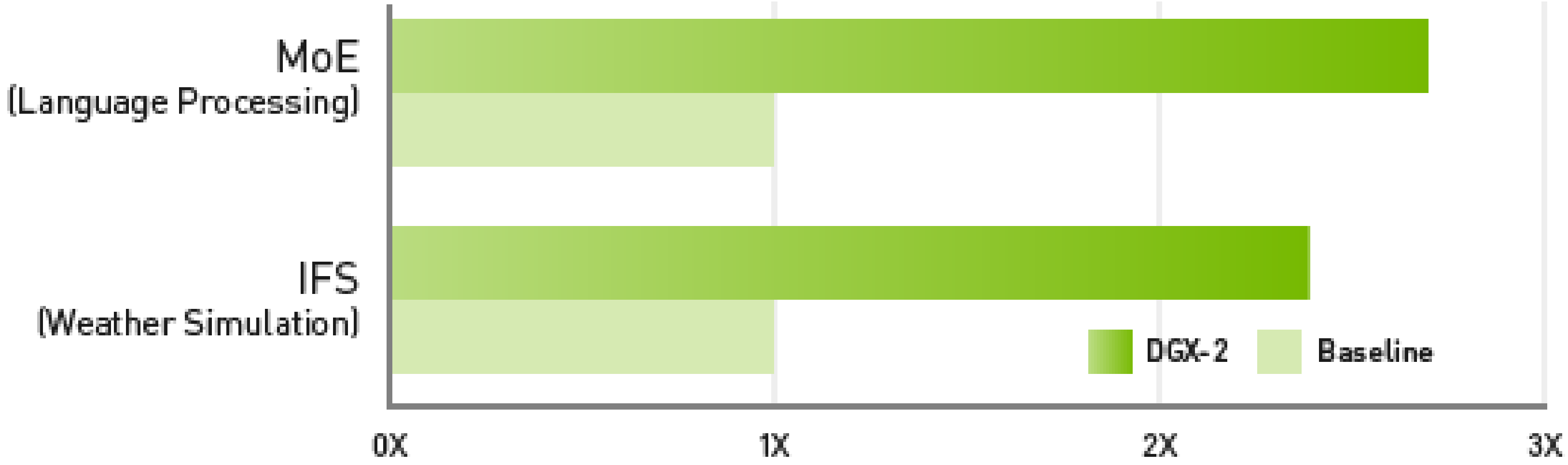
Increases number of floating point operations

Alan Gray, Peter Messmer, NVIDIA

Funded by the European Union

Will Deep Learning influence algorithmic choices for weather & climate ?

NVSwitch Delivers a >2X Speedup for Deep Learning and HPC*



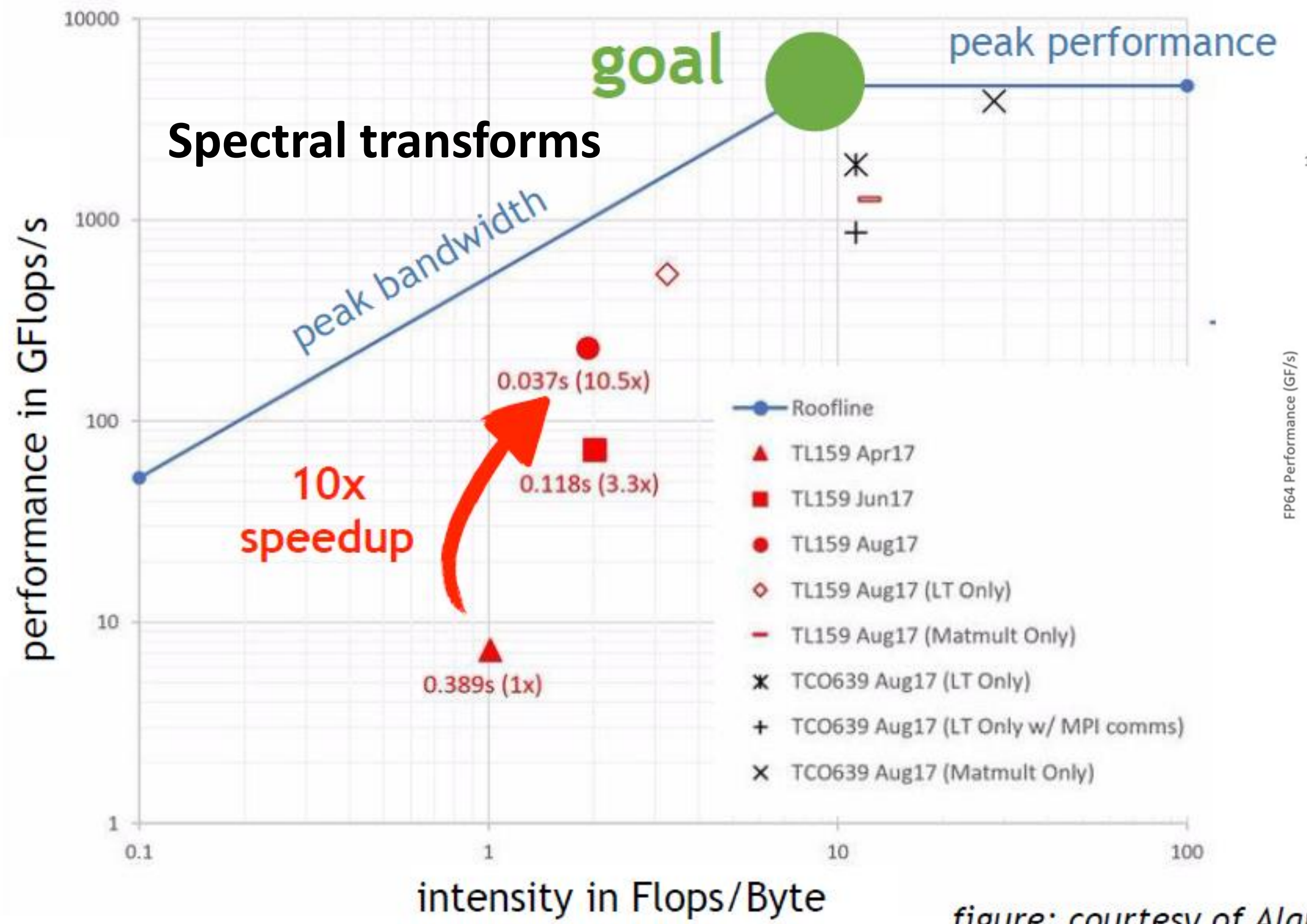
System Configs: Each of the two DGX-1 servers have dual-socket Xeon E5 2690v4 Processor, 8 x V100 GPUs; servers connected via a 4 EDR (100Gb) InfiniBand connections. DGX-2 server has dual-socket Xeon Scalable Processor Platinum 8168 Processors, 16 x Tesla V100 GPUs.

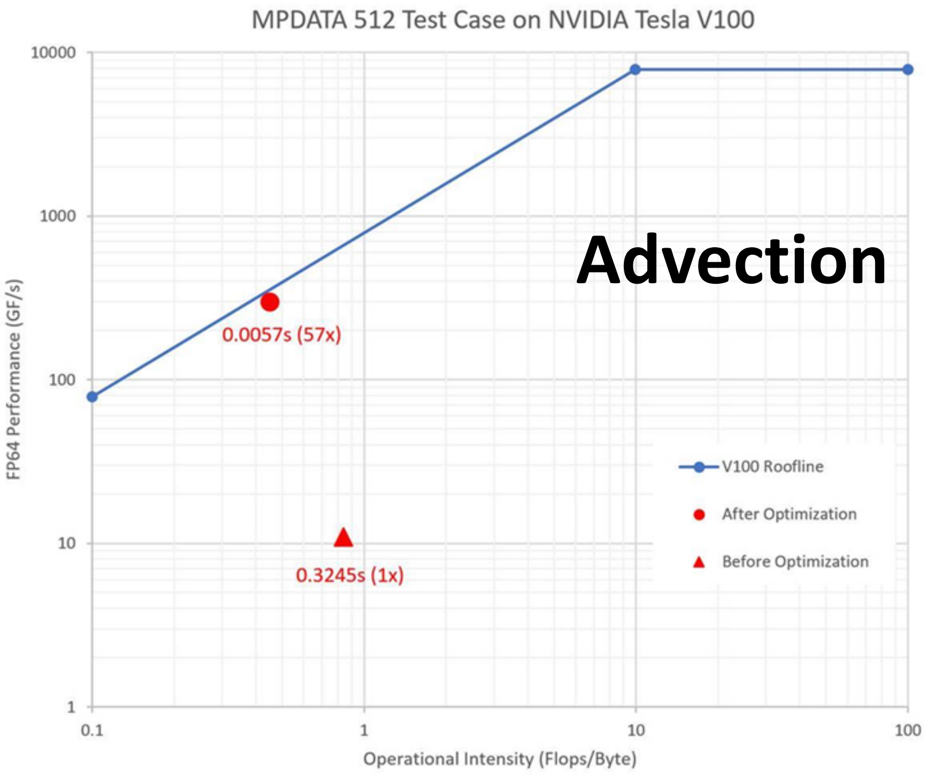
See talk by A. Gray

https://news.developer.nvidia.com/nvswitch-leveraging-nvlink-to-maximum-effect/

Analysing performance with roofline plots

Funded by the European Union





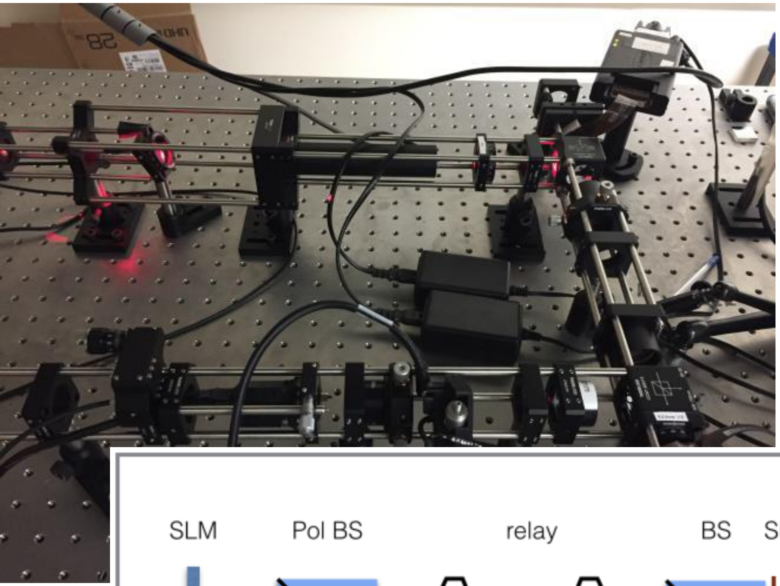
GPU speed-up

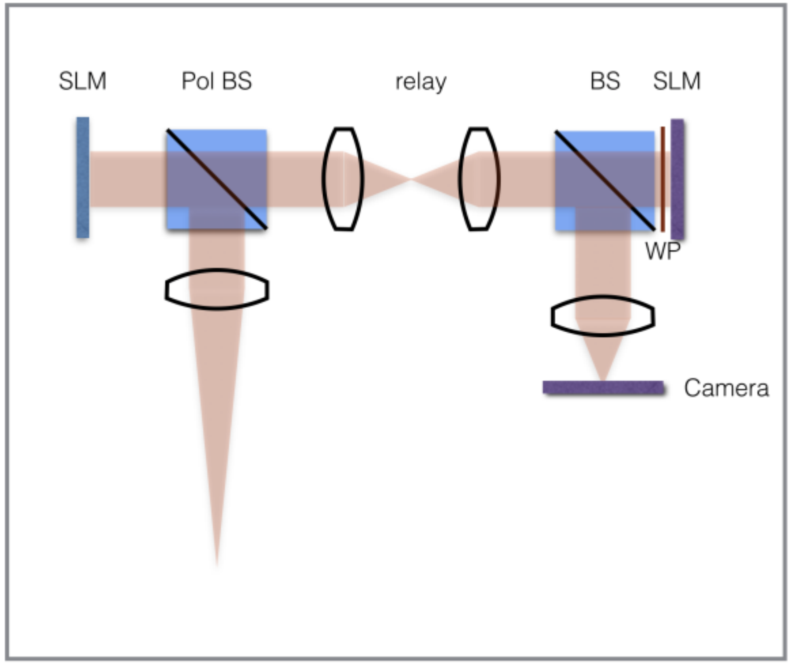
figure: courtesy of Alan Gray, Peter Messmer (NVIDIA)

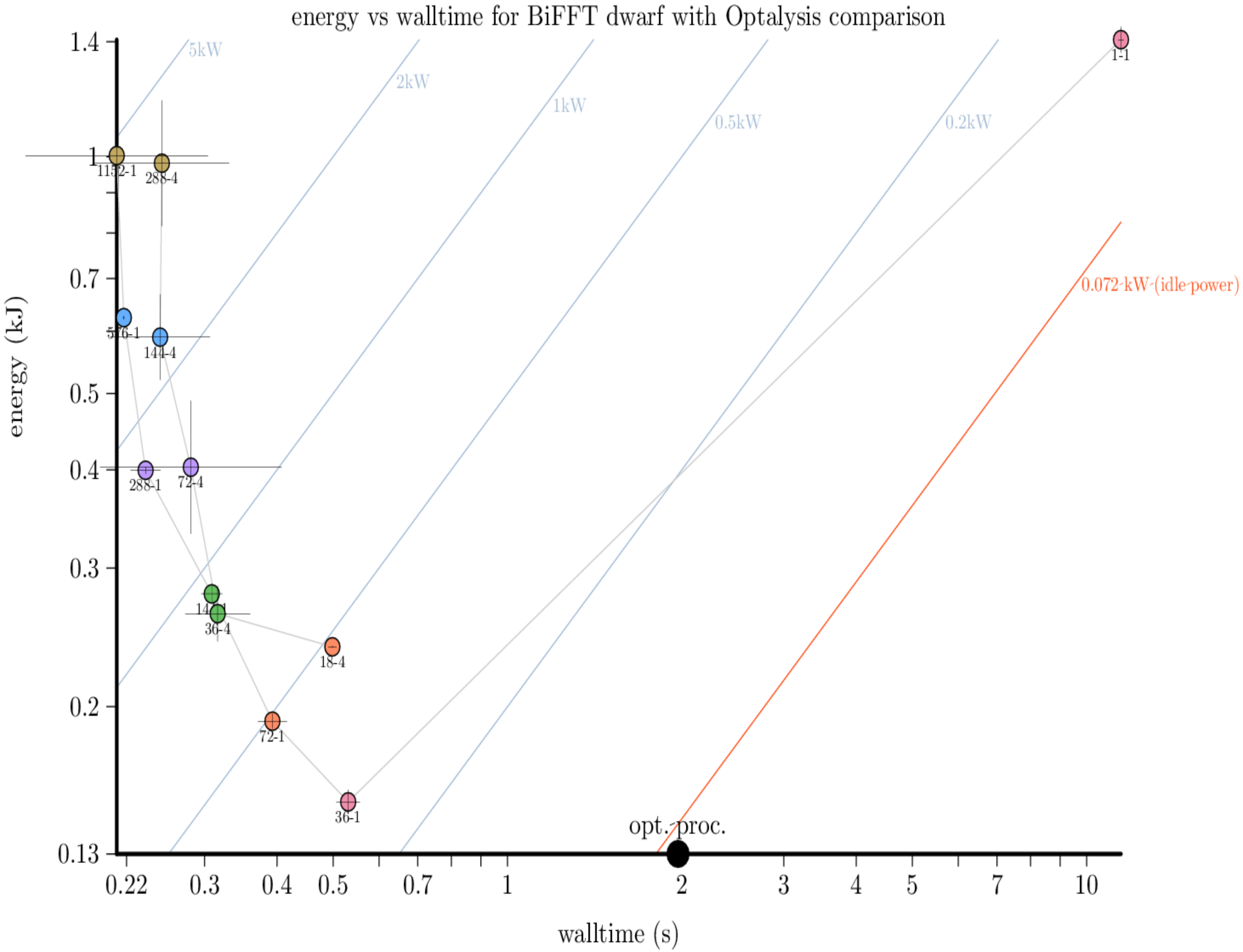
ECMWF MeteoSwiss DMI RMI NVIDIA METEO FRANCE Toujours un temps d'avance ICHEC Deutscher Wetterdienst Wetter und Klima aus einer Hand Optalysys Bull PSNC Loughborough University

Funded by the European Union

Optalysys: optical processor for spectral transform (biFFT and spherical harmonics) at speed of light ?

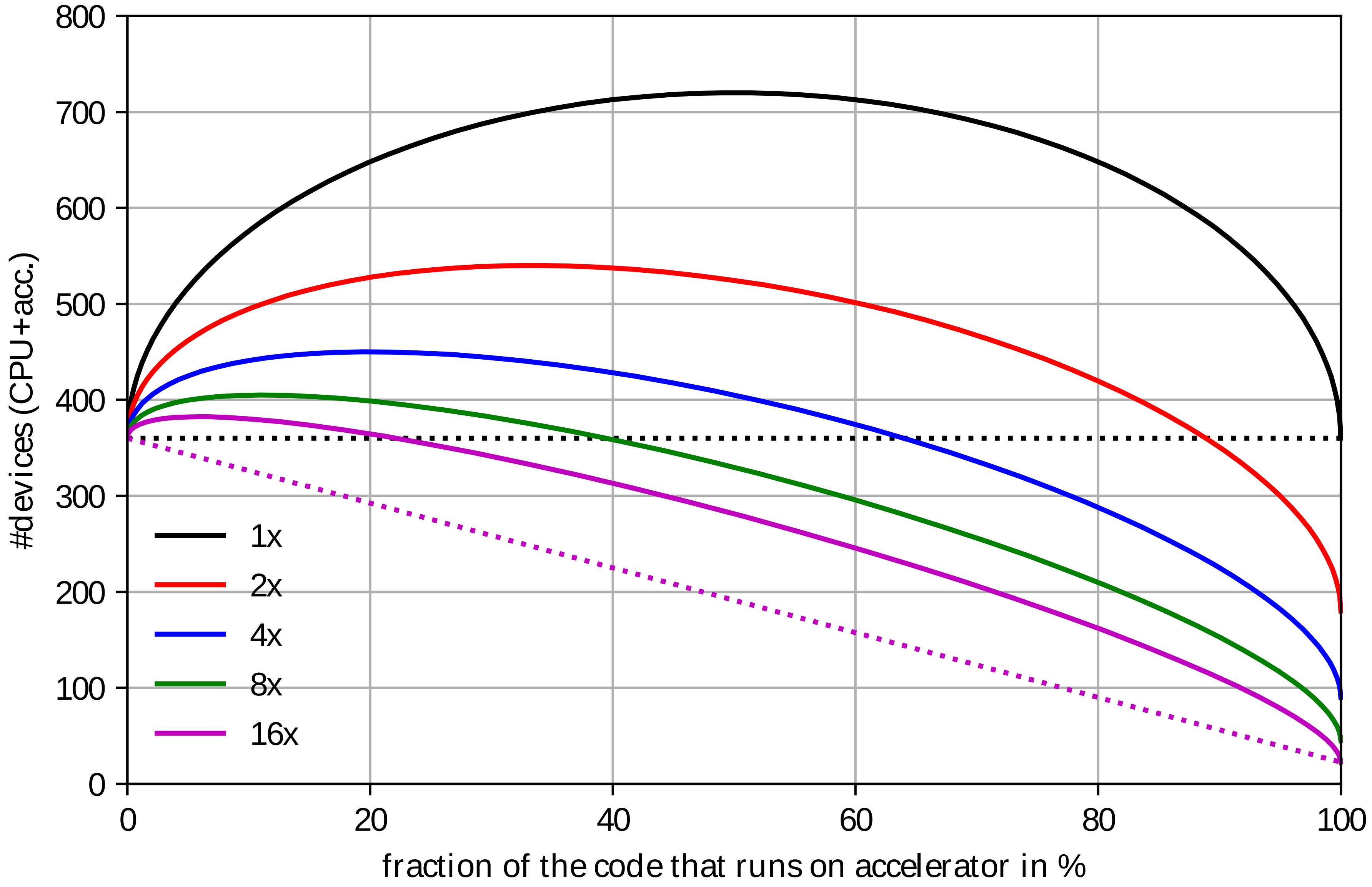






Funded by the European Union

Benefit of accelerators – theoretical model number of devices (acc.+CPU) at MétéoFrance



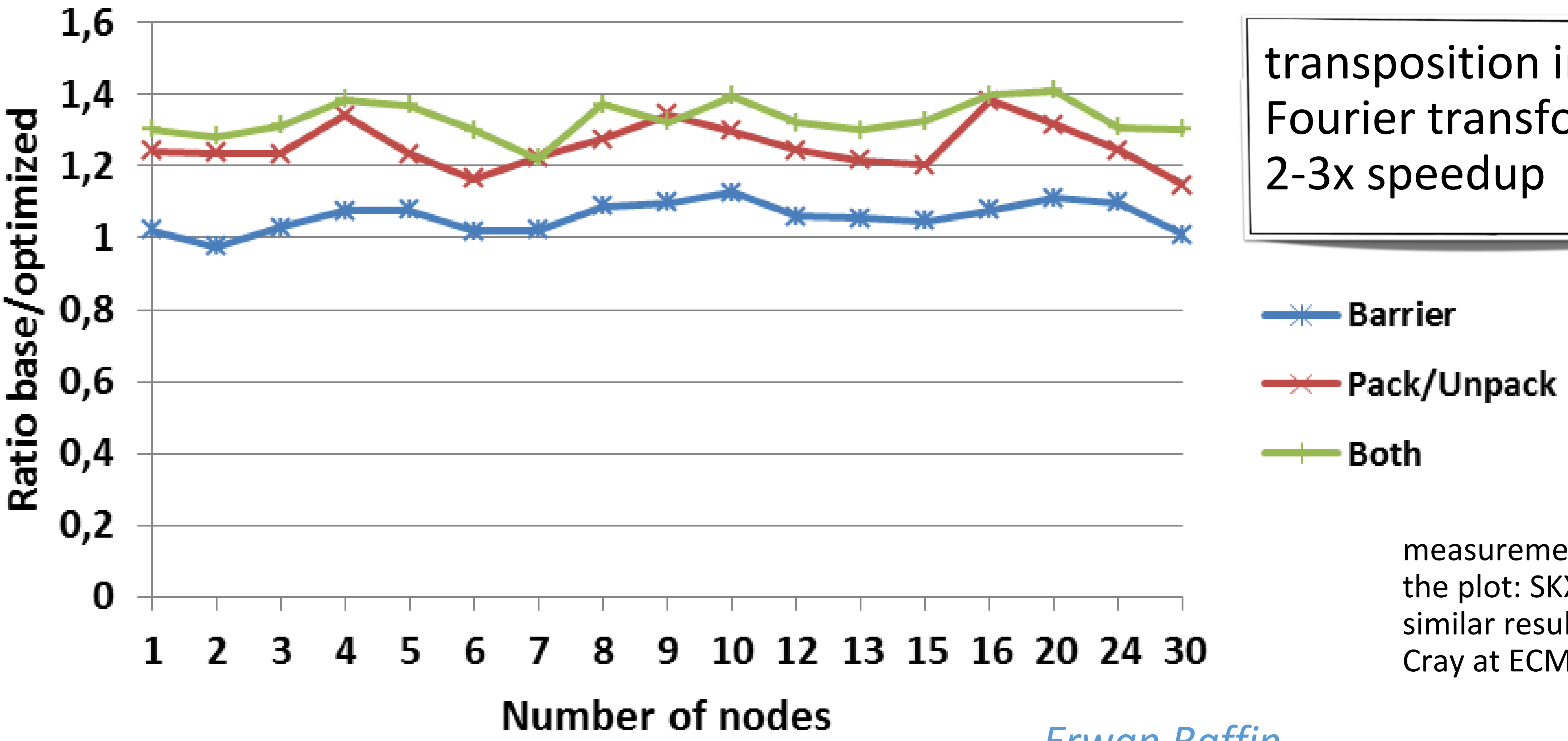
Assumes:
sequential execution
perfect scalability

Philippe Marginaud, Andreas Mueller

Funded by the European Union

Spectral transform optimisation by Atos/Bull

ESCAPE



transposition in Fourier transform: 2-3x speedup

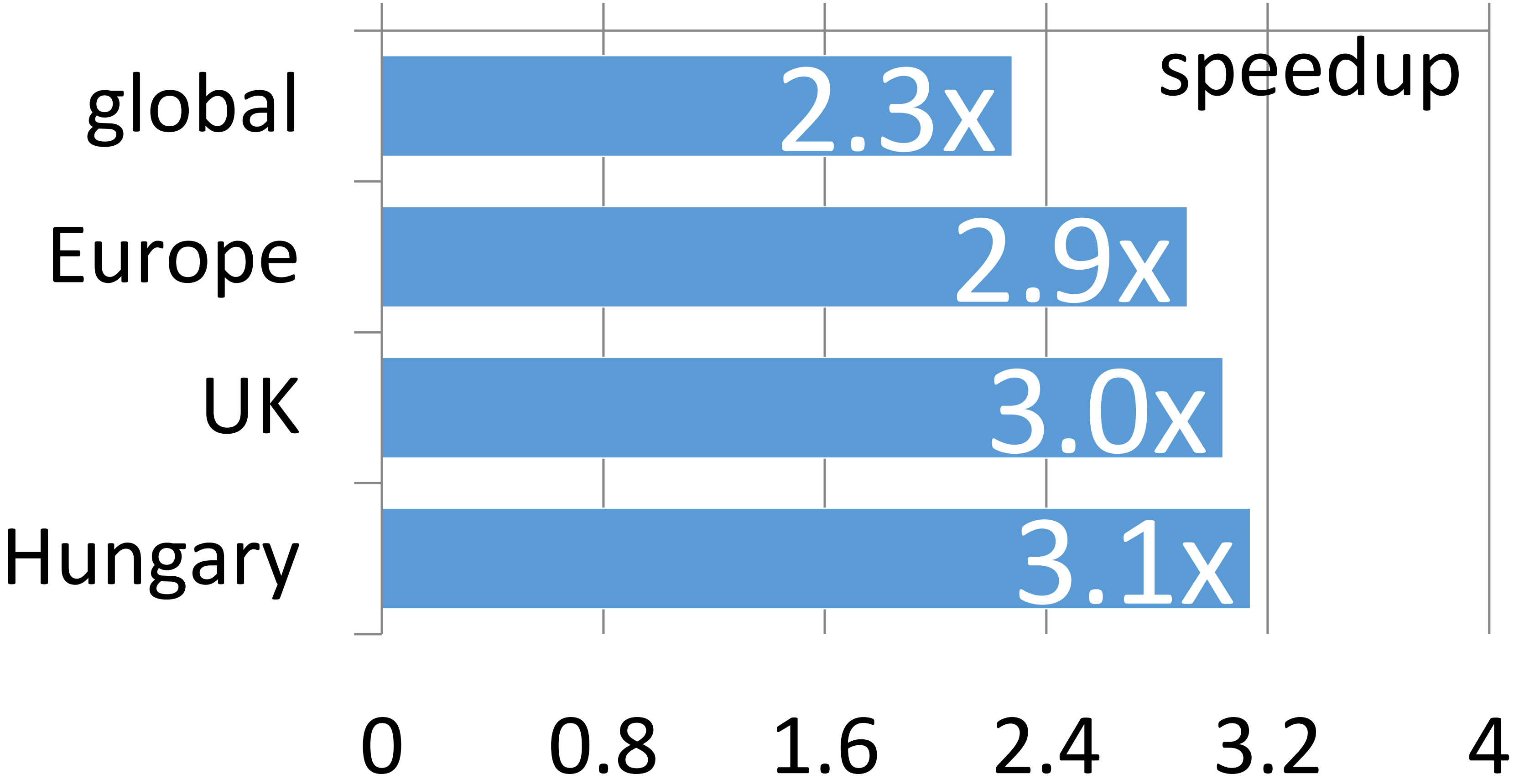
measurements for the plot: SKX, similar results on Cray at ECMWF

Erwan Raffin

Funded by the European Union

Optimisations in IFS on CPUs postprocessing of spectral data at 9km resolution

Single CPU



speedup compared to current operational transform used for postprocessing

Andreas Mueller

Geosci. Model Dev., 11, 1665–1681, 2018
https://doi.org/10.5194/gmd-11-1665-2018

Geoscientific Model Development
Open Access
EGU

Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0

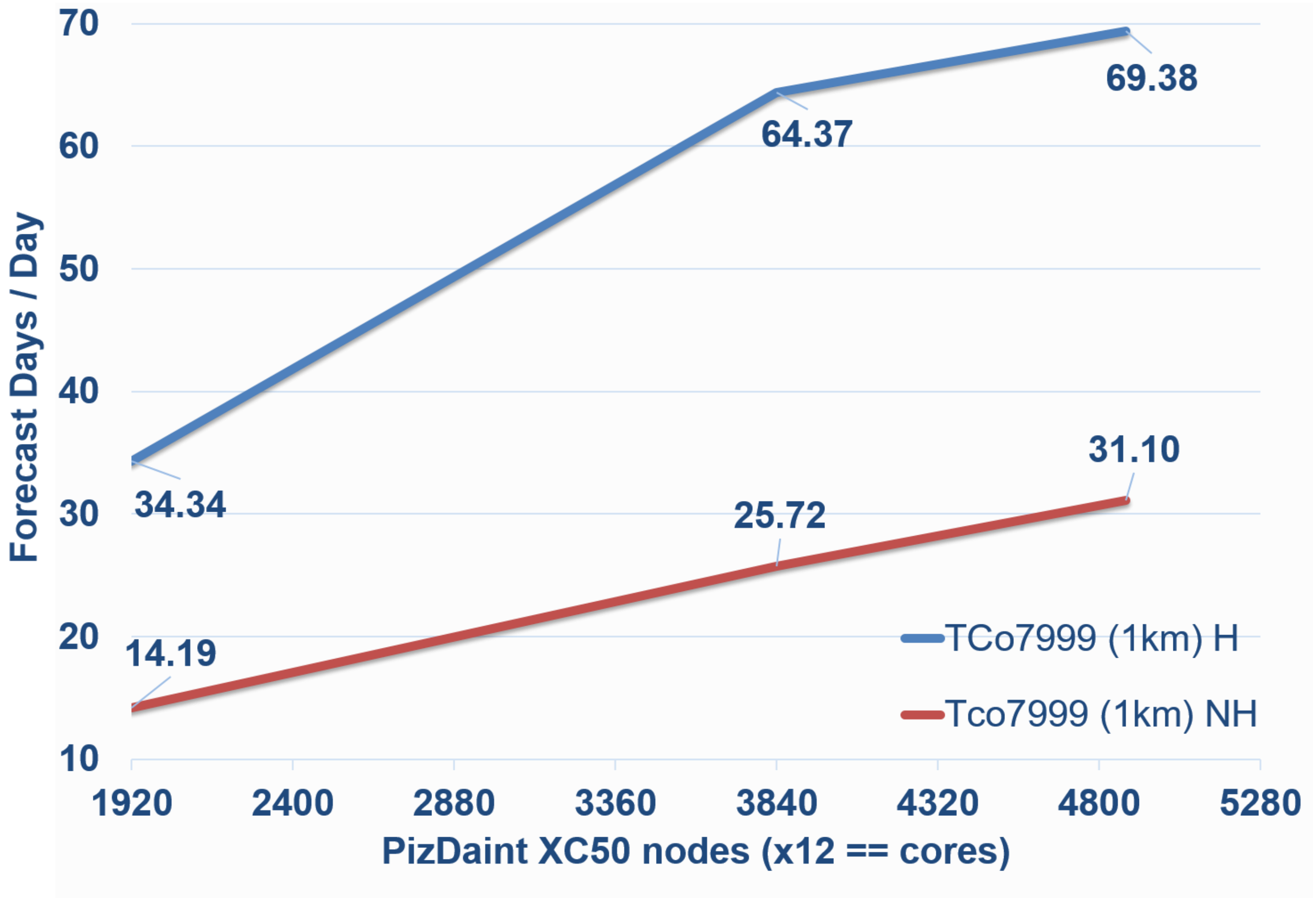
Oliver Fuhrer[1], Tarun Chadha[2], Torsten Hoefler[3], Grzegorz Kwasniewski[3], Xavier Lapillonne[1], David Leutwyler[4], Daniel Lüthi[4], Carlos Osuna[1], Christoph Schär[4], Thomas C. Schulthess[5,6], and Hannes Vogt[6]

[1]Federal Institute of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland
[2]ITS Research Informatics, ETH Zurich, Switzerland
[3]Scalable Parallel Computing Lab, ETH Zurich, Switzerland
[4]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland
[5]Institute for Theoretical Physics, ETH Zurich, Switzerland
[6]Swiss National Supercomputing Centre, CSCS, Lugano, Switzerland

Correspondence: Oliver Fuhrer (oliver.fuhrer@meteoswiss.ch)

Received: 16 September 2017 – Discussion started: 5 October 2017
Revised: 7 February 2018 – Accepted: 8 February 2018 – Published: 2 May 2018

IFS 1km: strong scaling on PizDaint

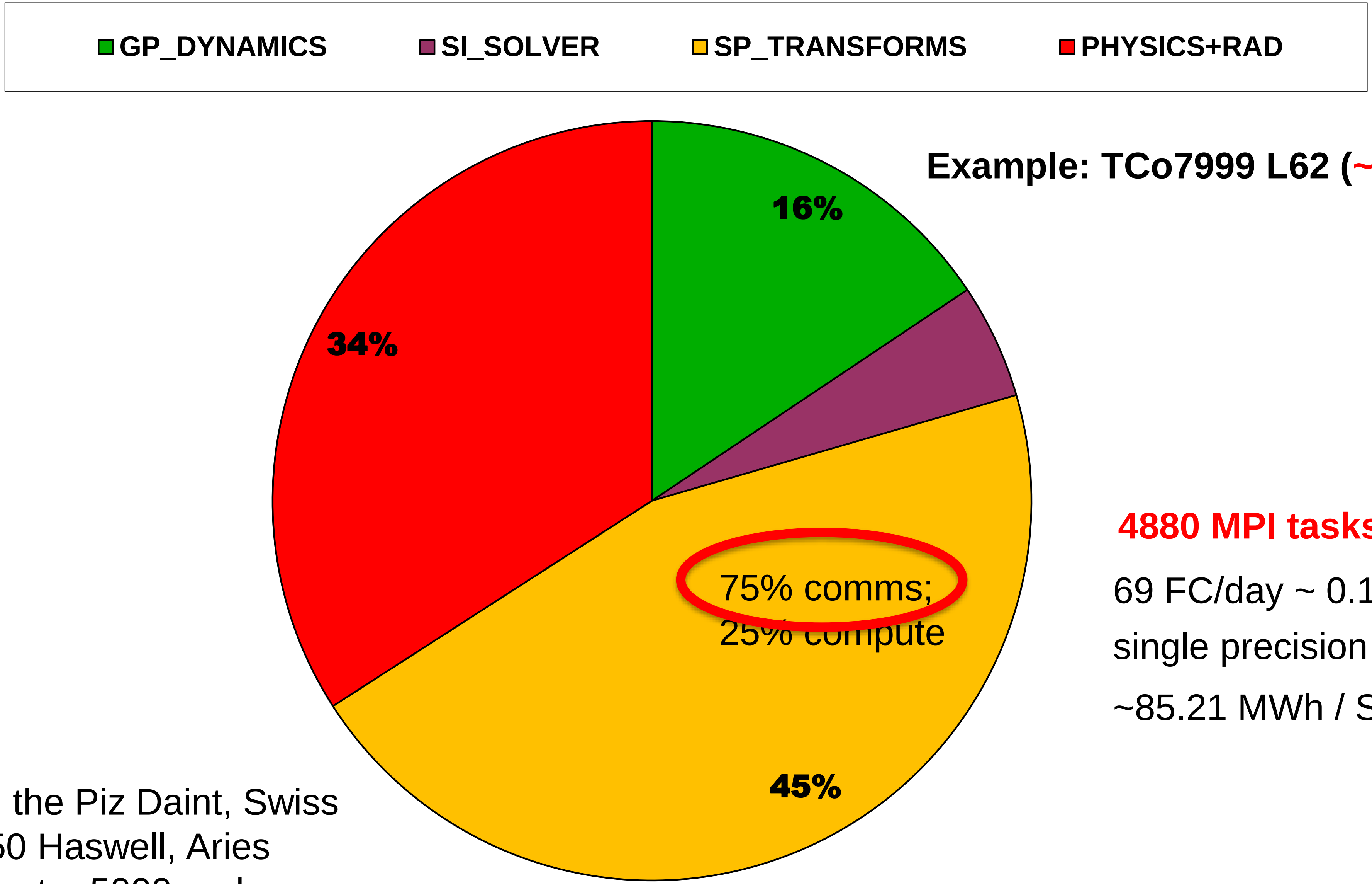


Goal ~1 year / Day

Result of algorithmic changes and single precision

Many thanks to Thomas Schulthess & Maria Grazia Giuffreda !

The cost profile of a 1.25km IFS atmosphere simulation on Piz Daint (CPU only)



Example: TCo7999 L62 (~1.25km)

4880 MPI tasks x 12 threads

69 FC/day ~ 0.19 SYPD

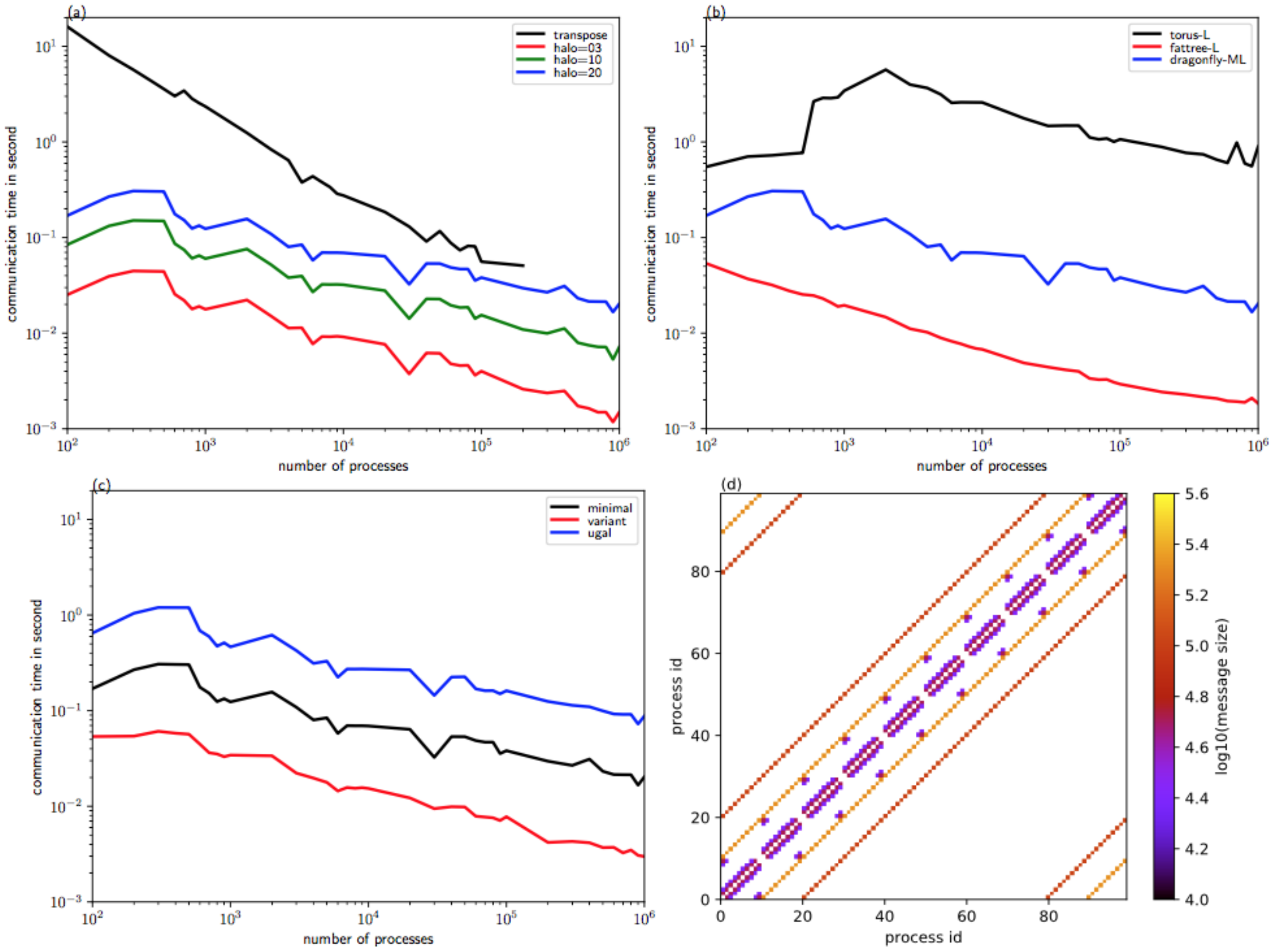
single precision / FLT

~85.21 MWh / SY

Based on the Piz Daint, Swiss Cray XC50 Haswell, Aries interconnect, ~5000 nodes total

Funded by the European Union

Simulating performance and scalability of MPI communications

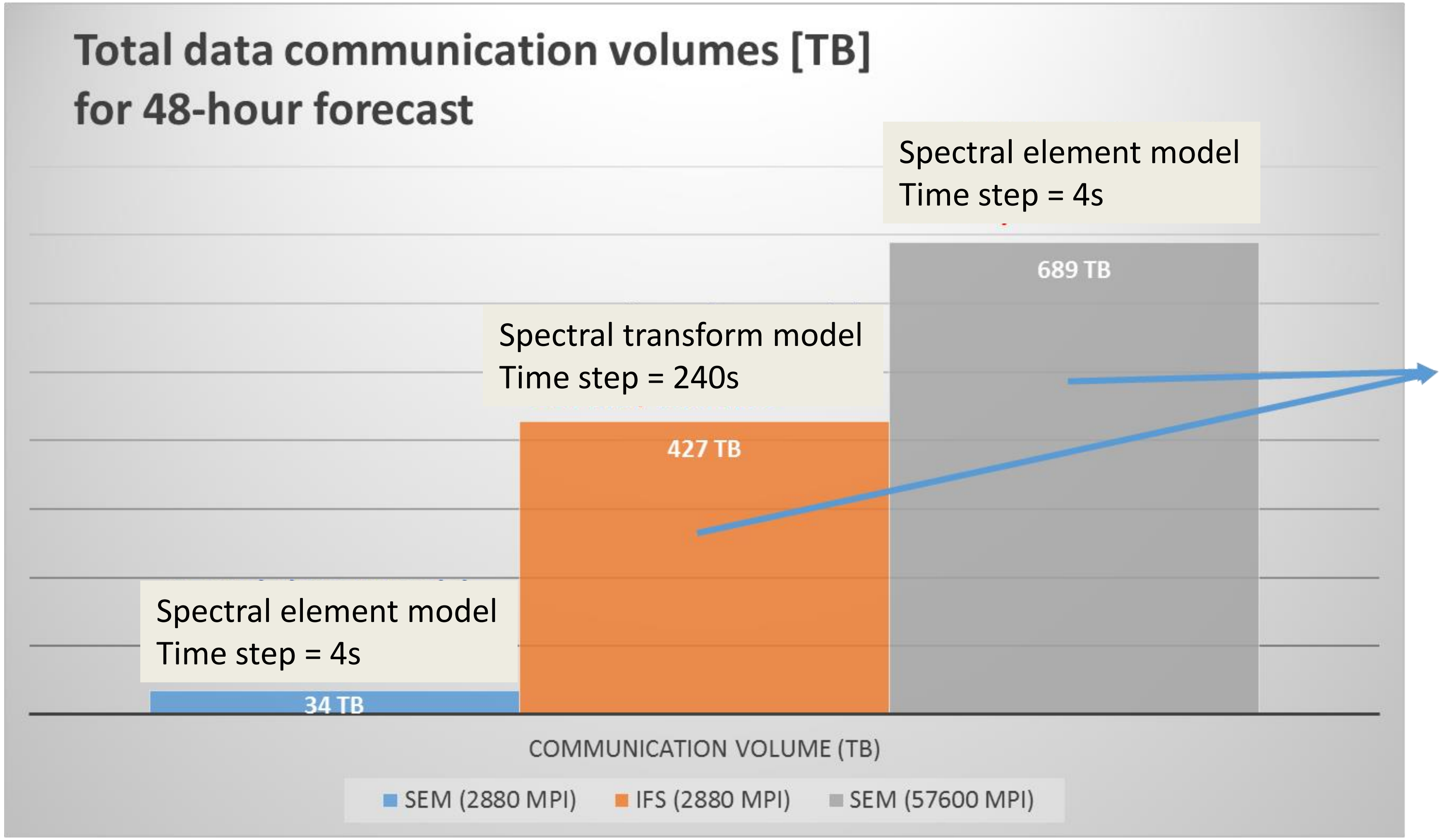


Communication time as a function of halo size, topology & routing algorithm, compared to spectral transpositions

Zheng and Marginaud (GMD, 2018)

MPI collectives at least across a subset of nodes are required in NWP & climate!

Communication is bad – small time steps are worse



Same time to solution!
Energy efficiency?

Data movement x100 (x1000) more expensive than computations in time (energy)!

[Shalf et al. 2011]

Funded by the European Union

Co-ordinated by ECMWF

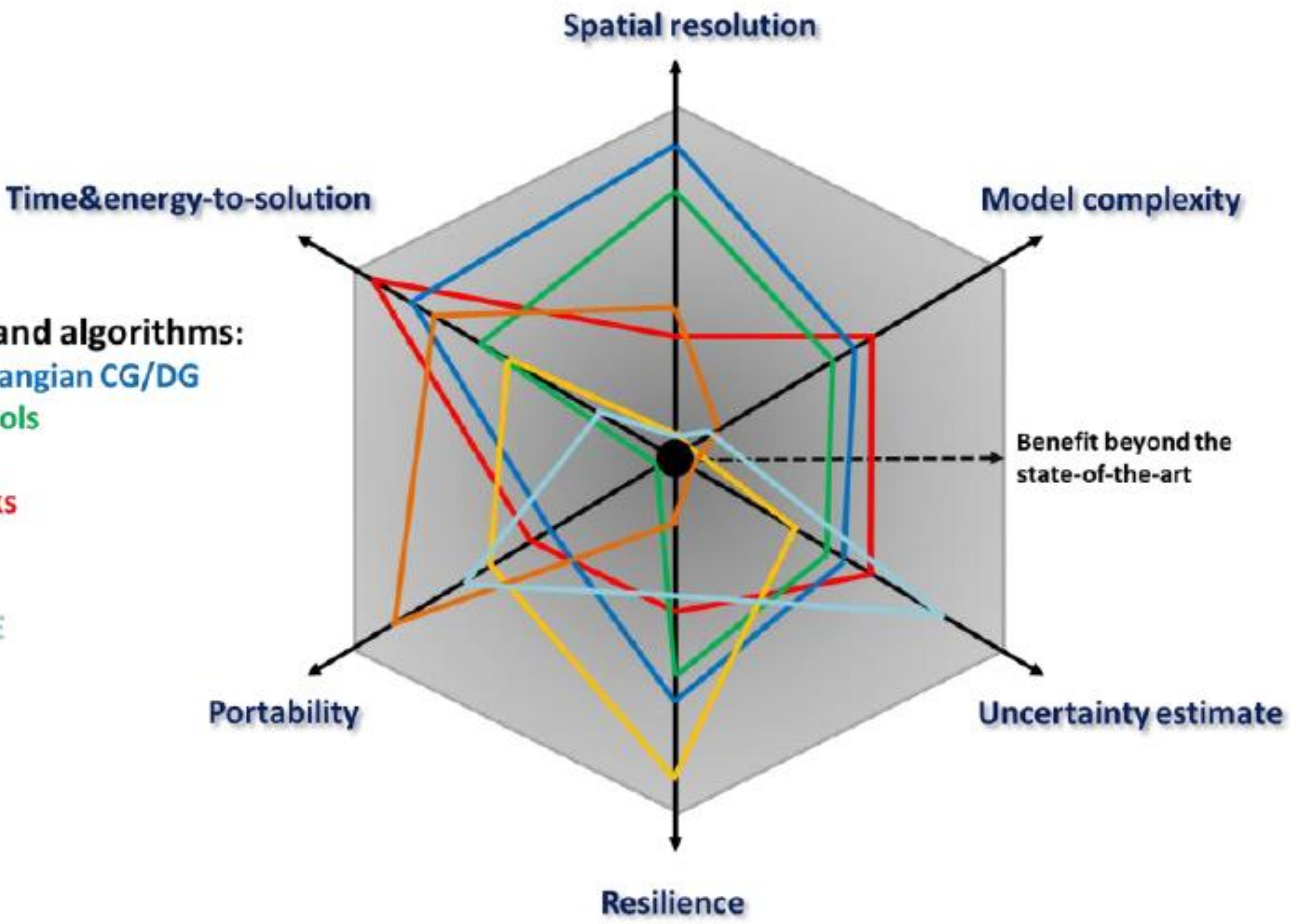
ESCAPE 2

ECMWF | DKRZ Deutsches Klimarechenzentrum | Max-Planck-Institut für Meteorologie | MeteoSwiss | Barcelona Supercomputing Center – Centro Nacional de Supercomputación | cea

Loughborough University | RMI | Politecnico Milano 1863 | MOX | DMI | cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici | Bull atos technologies

Mathematical methods and algorithms:
- Semi-implicit, semi-Lagrangian CG/DG
- Hierarchical multigrid tools
- Fault resilient solver
- Artificial neural networks

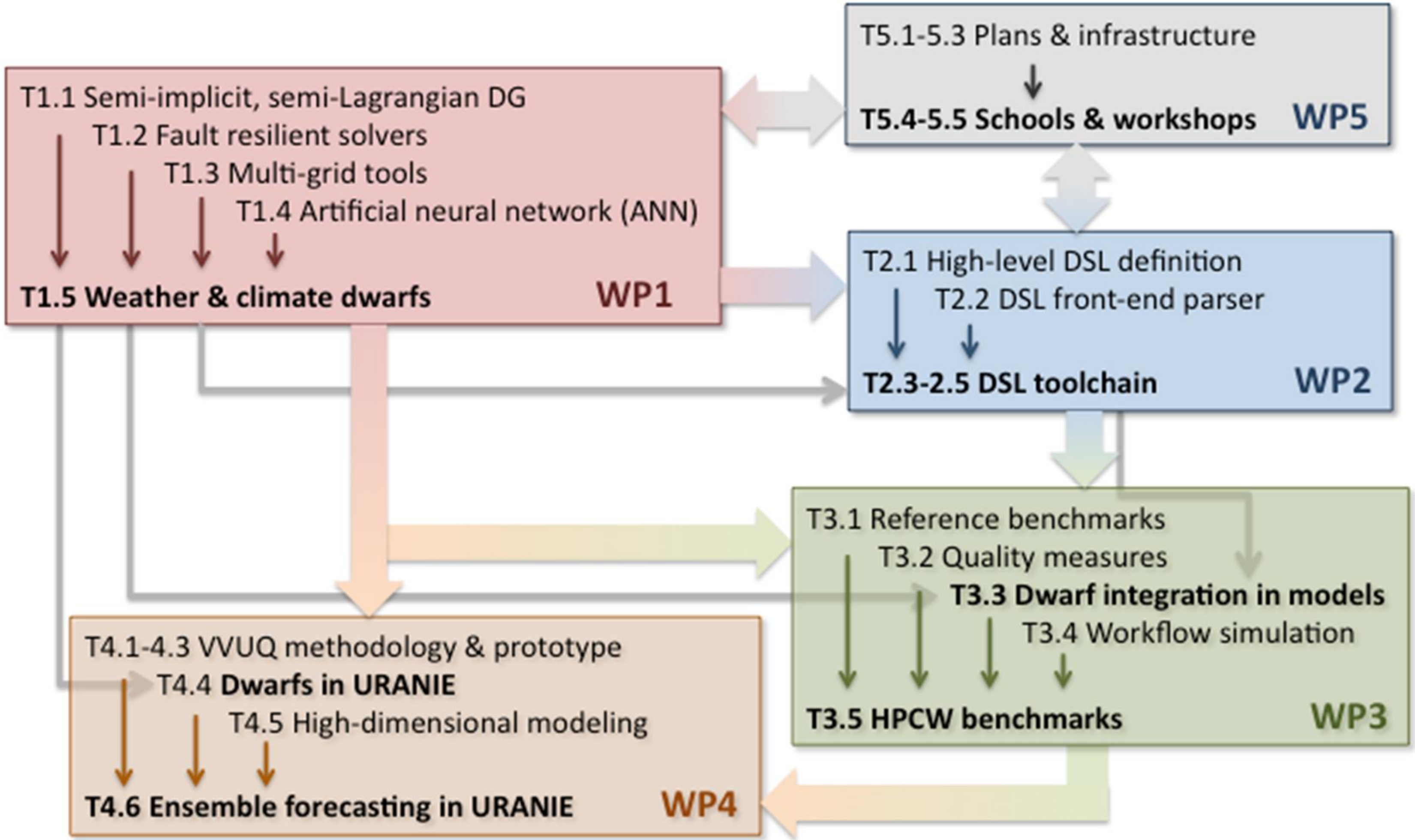
and:
- DSL toolchain
- Ensemble based URANIE
- State-of-the-art



ESCAPE 2

starting October 2018

Funded by the European Union

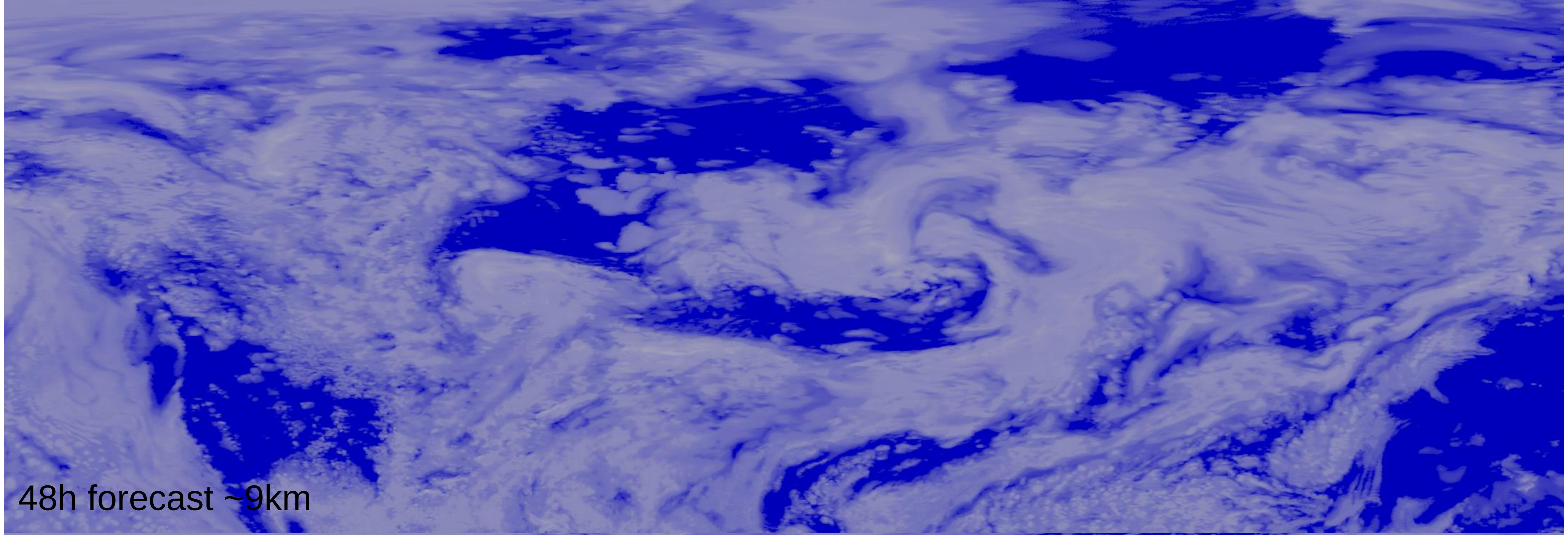


ECMWF · DKRZ · Max-Planck-Institut für Meteorologie · MeteoSwiss · Barcelona Supercomputing Center · cea · Loughborough University · RMI · Politecnico Milano · MOX · DMI · cmcc · Bull atos technologies

Summary

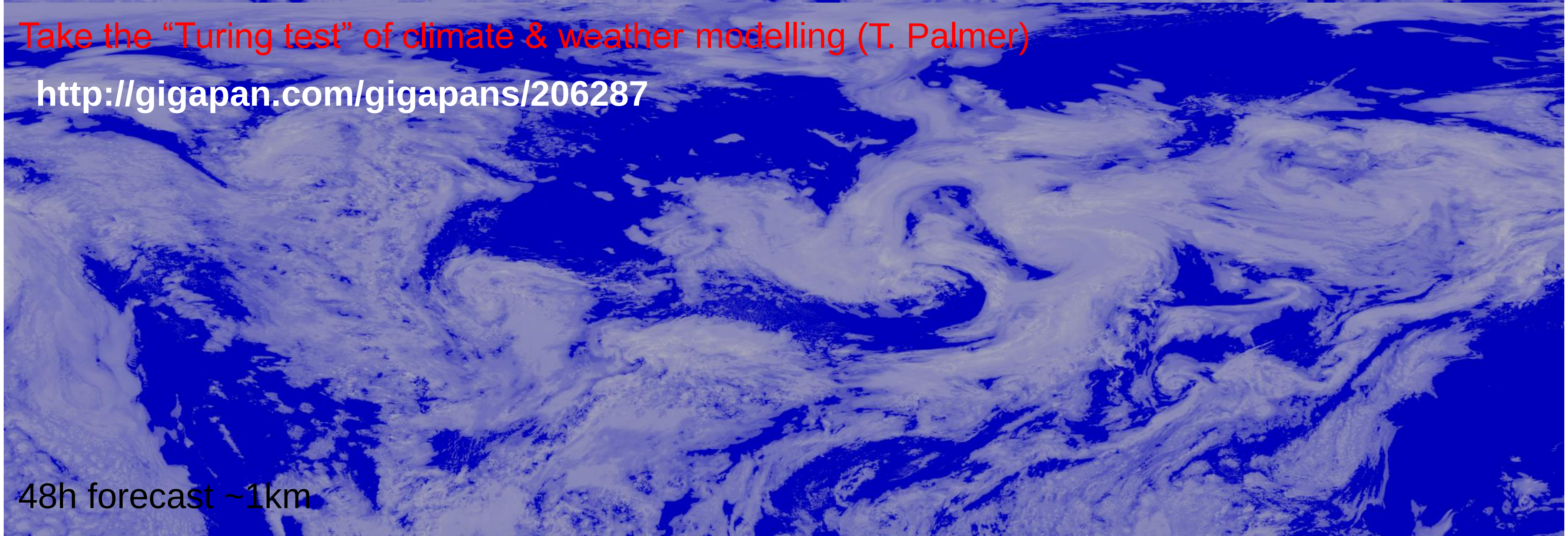
Funded by the European Union

- Numerical weather prediction & climate needs sustained efforts to evolve together with emerging computing opportunities
- *ESCAPE and ESCAPE-2 will deliver*
  - Pioneering approaches for refactoring society critical legacy codes
    - Weather & climate dwarfs
  - Energy-efficient accelerator use in global weather & climate prediction
    - Scrutiny of the need for precision
  - Co-development of novel mathematical algorithms & hardware adaptation
  - Pioneering mathematical algorithm development with hardware adaptation using DSL toolchains
  - A HPCW benchmark and cross-disciplinary Verification, Validation, and Uncertainty Quantification (VVUQ)
  - Application resilience



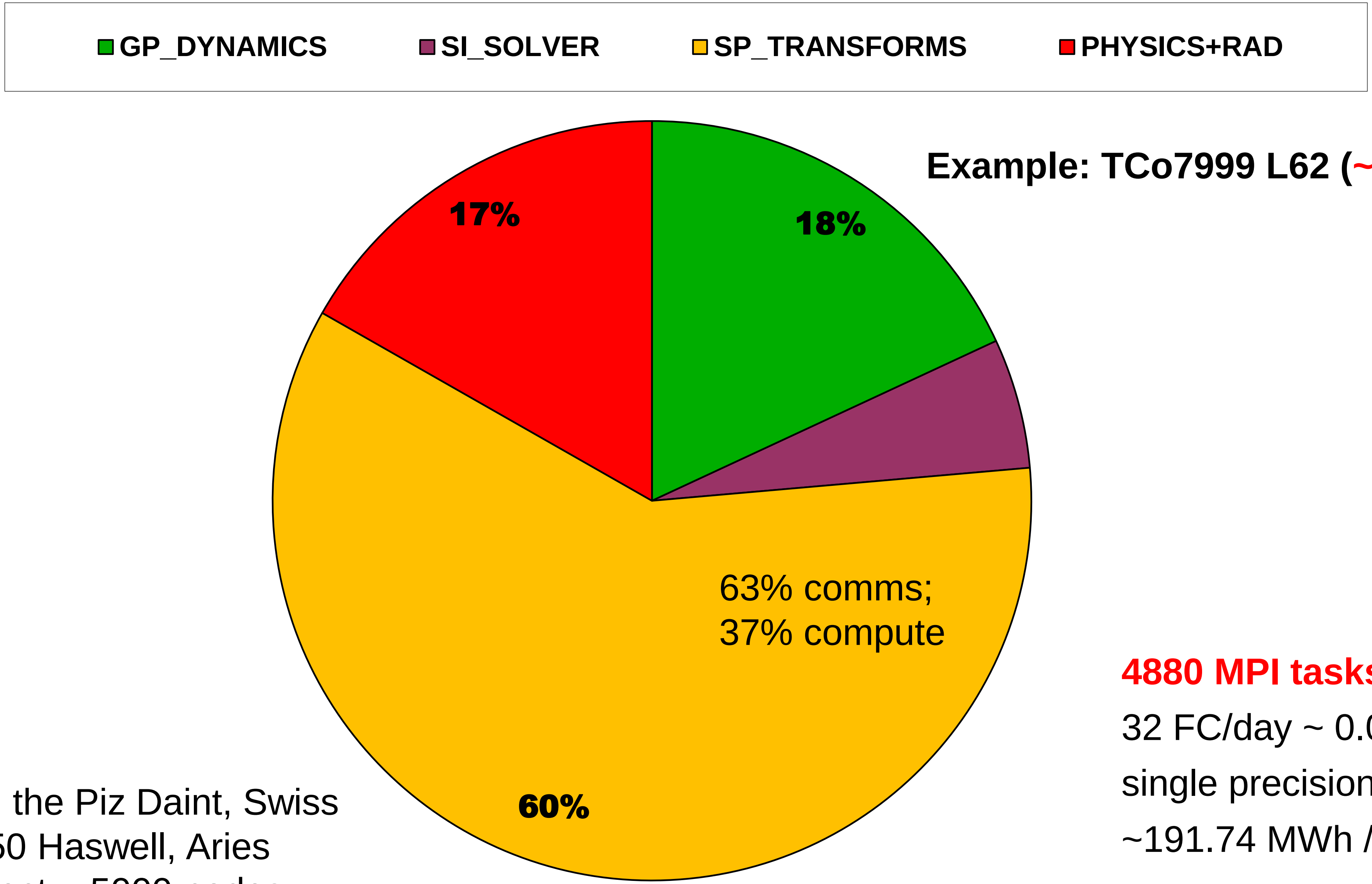
Take the “Turing test” of climate & weather modelling (T. Palmer)

http://gigapan.com/gigapans/206287



Additional slides

The cost profile of a 1.25km (non-hydrostatic) IFS atmosphere simulation Piz Daint



Example: TCo7999 L62 (~1.25km)

4880 MPI tasks x 12 threads

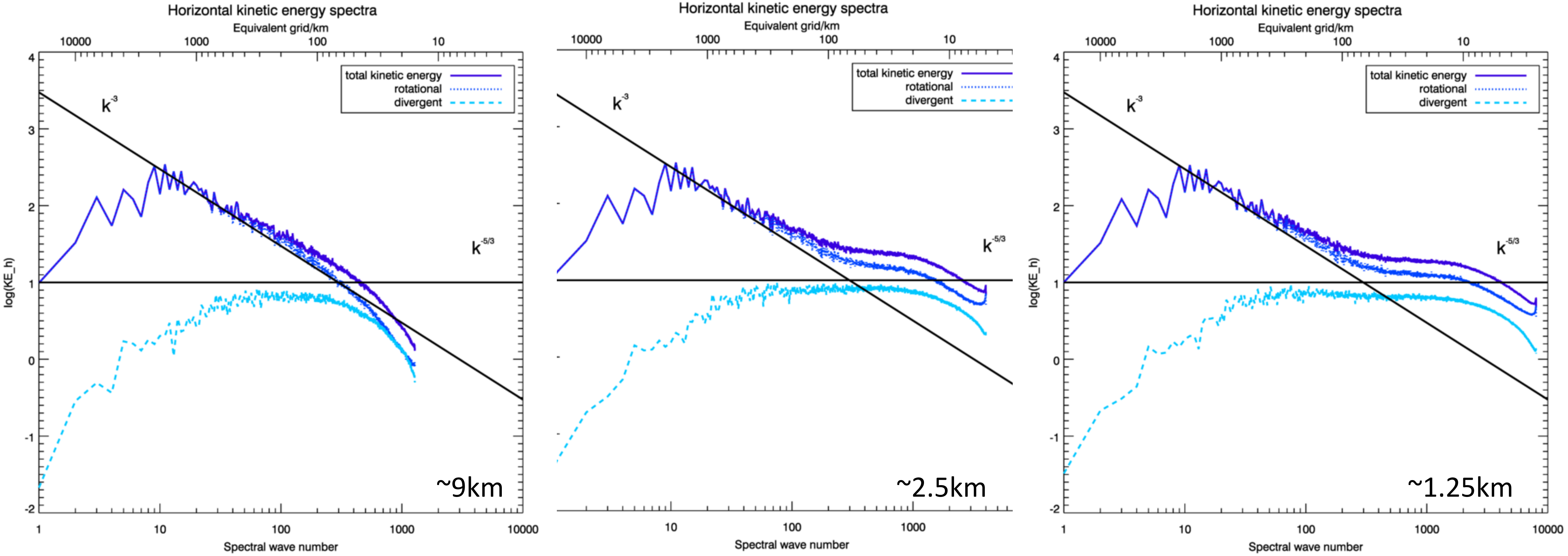
32 FC/day ~ 0.088 SYPD

single precision / FLT

~191.74 MWh / SY

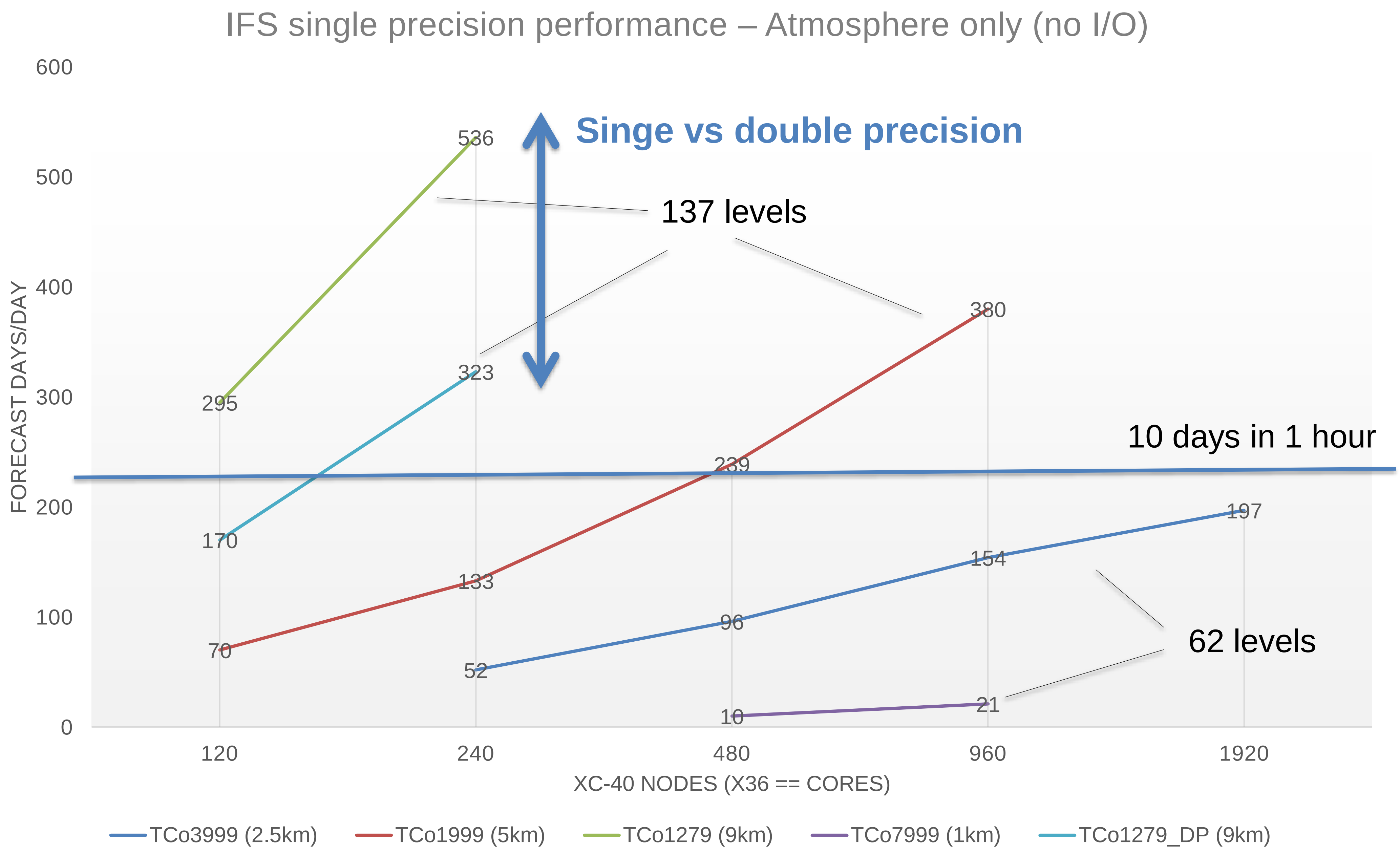
Based on the Piz Daint, Swiss Cray XC50 Haswell, Aries interconnect, ~5000 nodes total

Global KE - Spectra ~500hPa



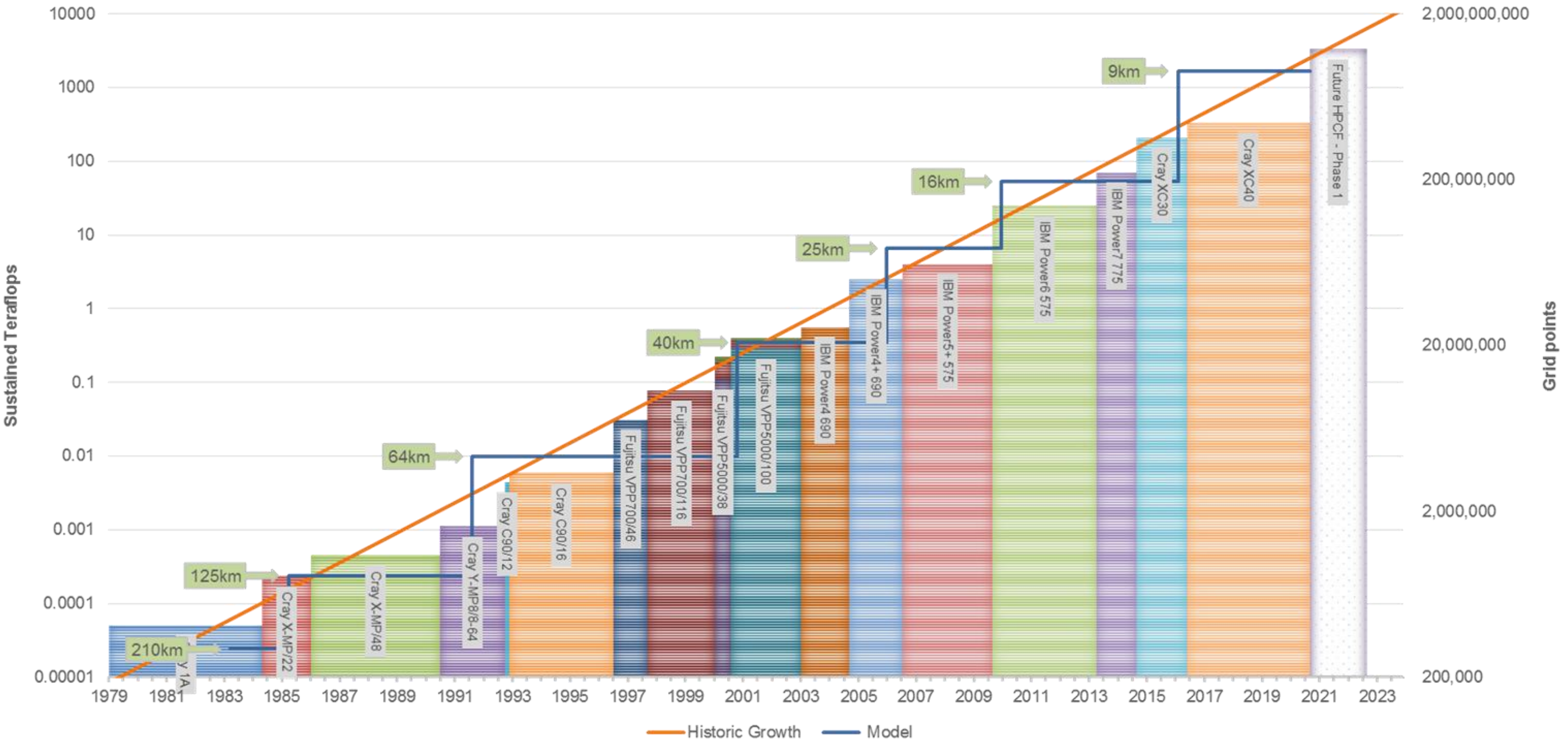
Resolve rather than parametrize much of the crucial vertical transport of momentum and heat

ECMWF

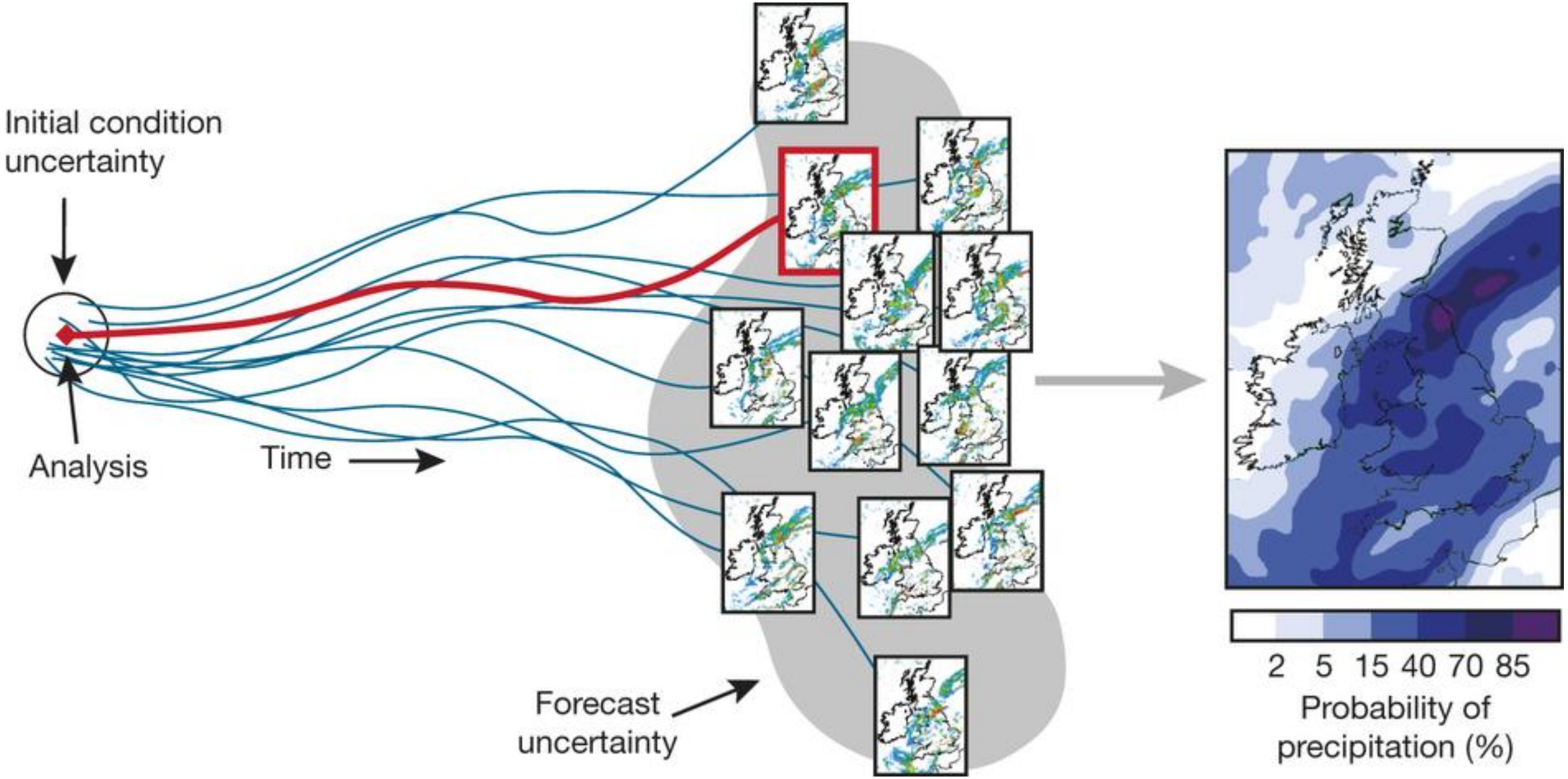


(Vana, Dueben et al 2017)

Sustained HPC performance



Ensemble of assimilations and forecasts



(Bauer et al, 2015)