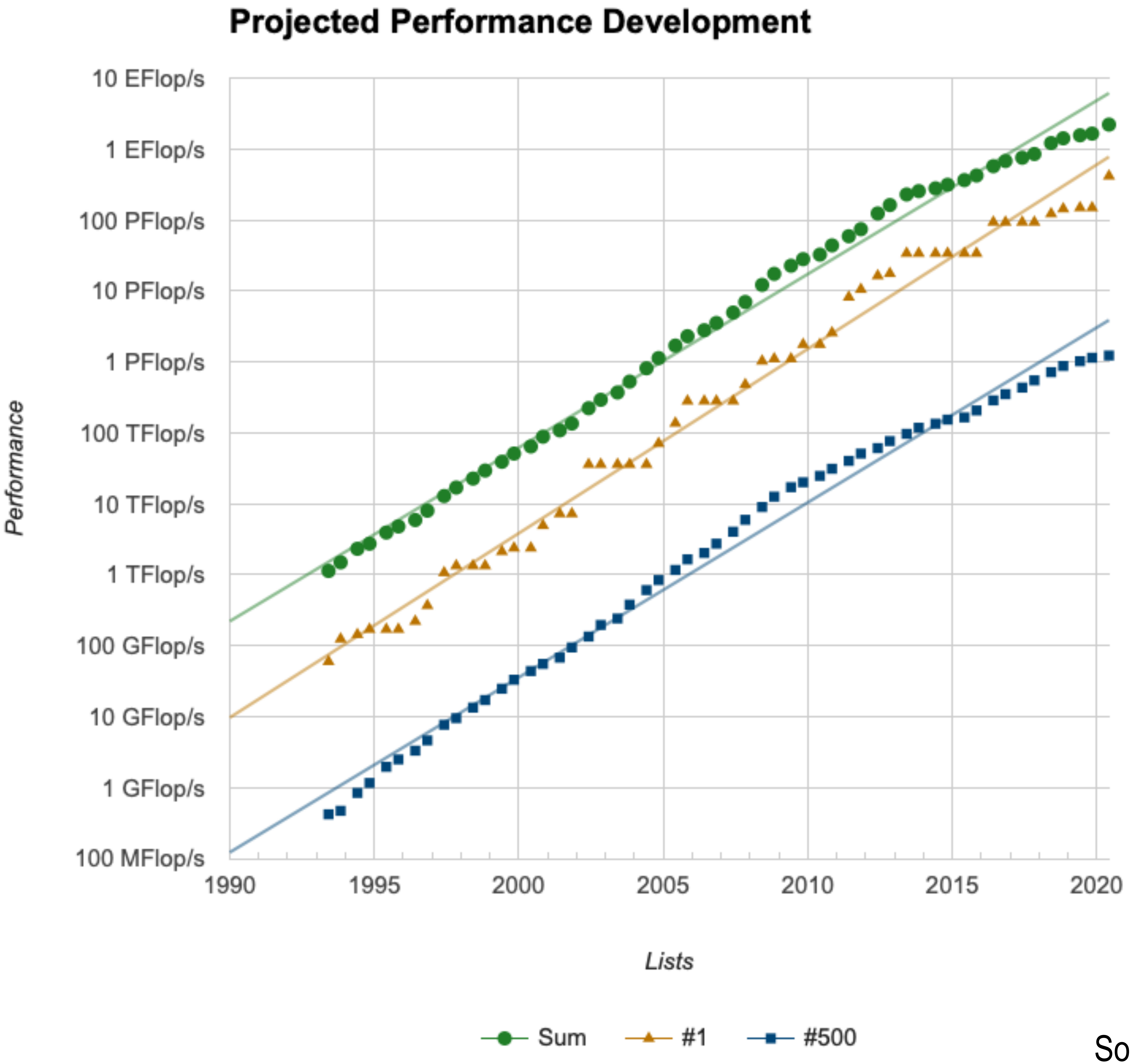
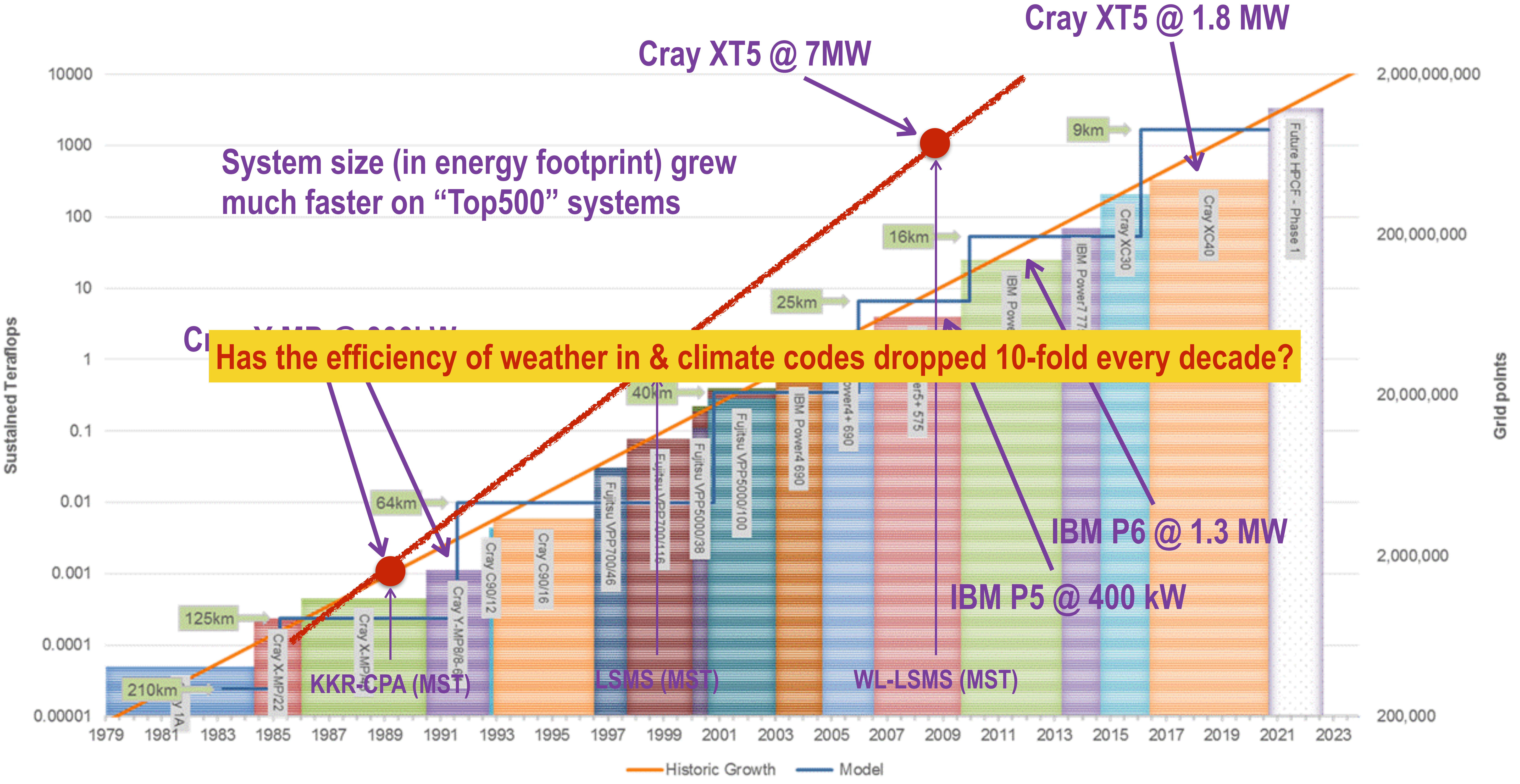
A useful definition of exascale computing for weather and climate modelling

Thomas C. Schulthess



Source: www.top500.org

“Only” 100-fold performance improvement in climate codes



Cray XT5 @ 7MW

Cray XT5 @ 1.8 MW

System size (in energy footprint) grew much faster on “Top500” systems

Cray Y-MP @ 300kW

Has the efficiency of weather in & climate codes dropped 10-fold every decade?

IBM P6 @ 1.3 MW

IBM P5 @ 400 kW

KKR-CPA (MST)

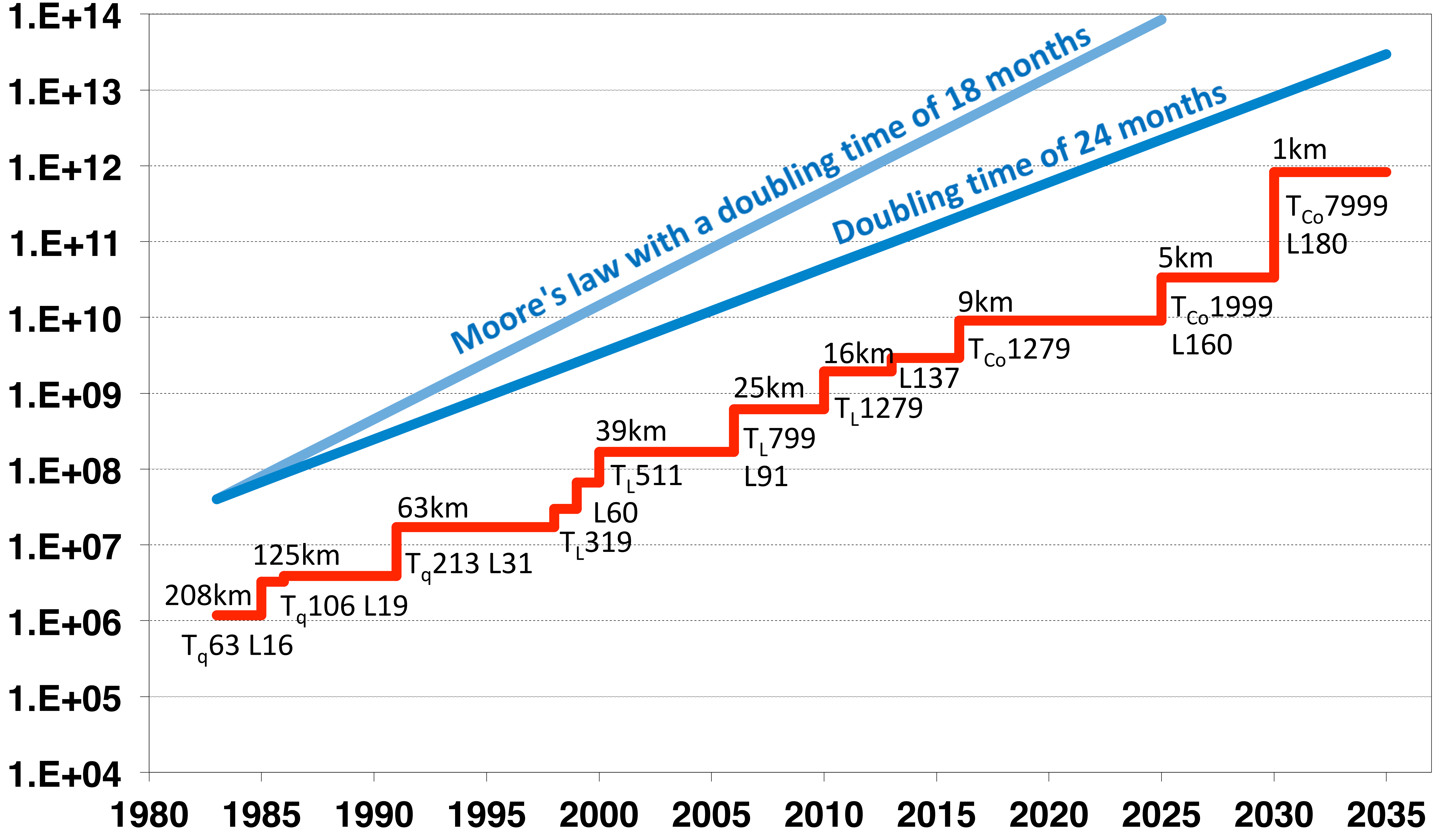
LSMS (MST)

WL-LSMS (MST)

Floating point efficiency dropped from 50% on Cray Y-MP to 5% on today’s Cray XC (10x in 2.5 decades)

Source: Peter Bauer, ECMWF

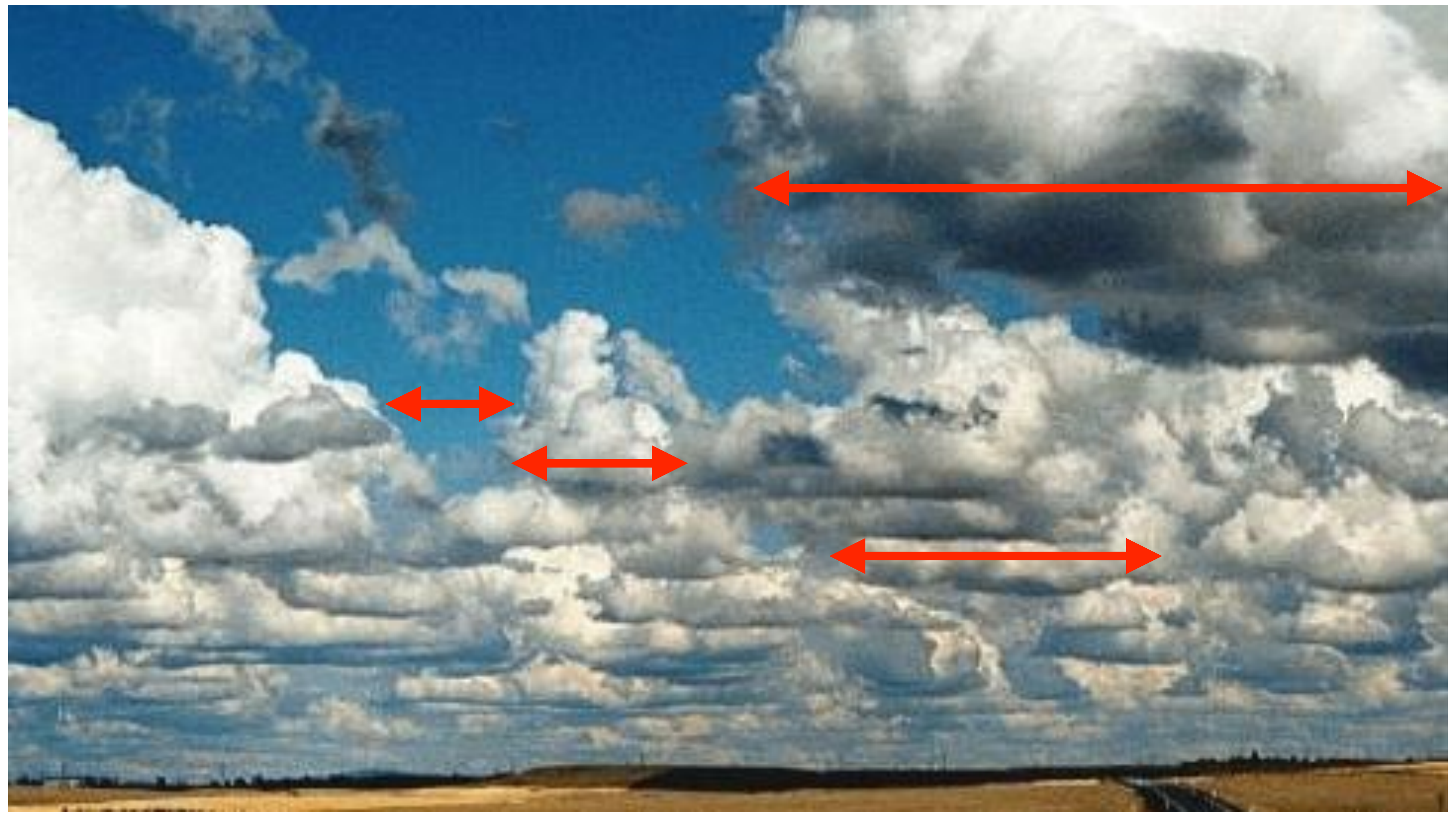
Computational power drives spatial resolution



Source: Christoph Schär, ETH Zurich, & Nils Wedi, ECMWF
Schulthess et al., 2019

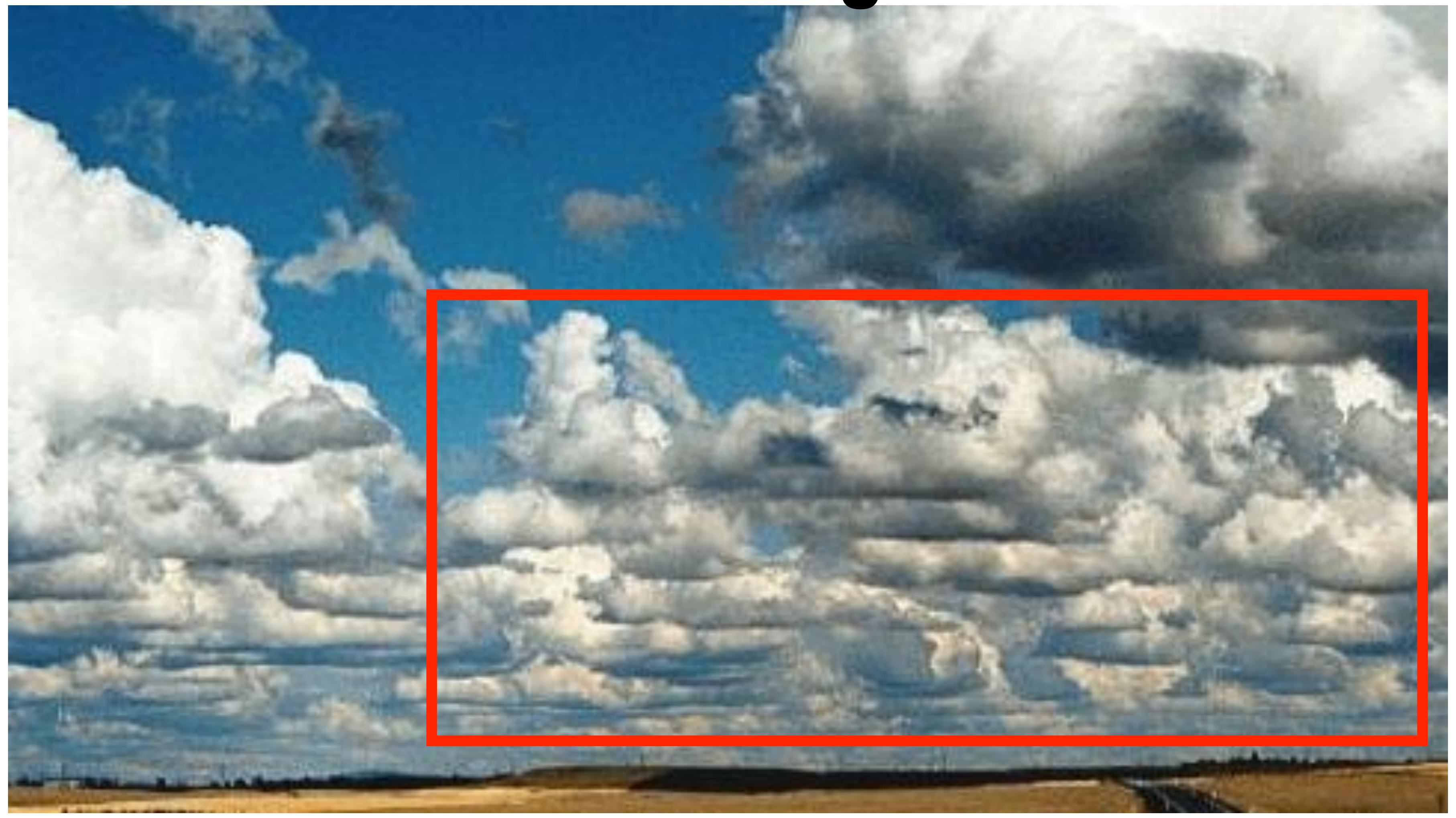
Resolving convective clouds (convergence?)

Structural convergence



Statistics of cloud ensemble:
E.g., spacing and size of convective clouds

Bulk convergence

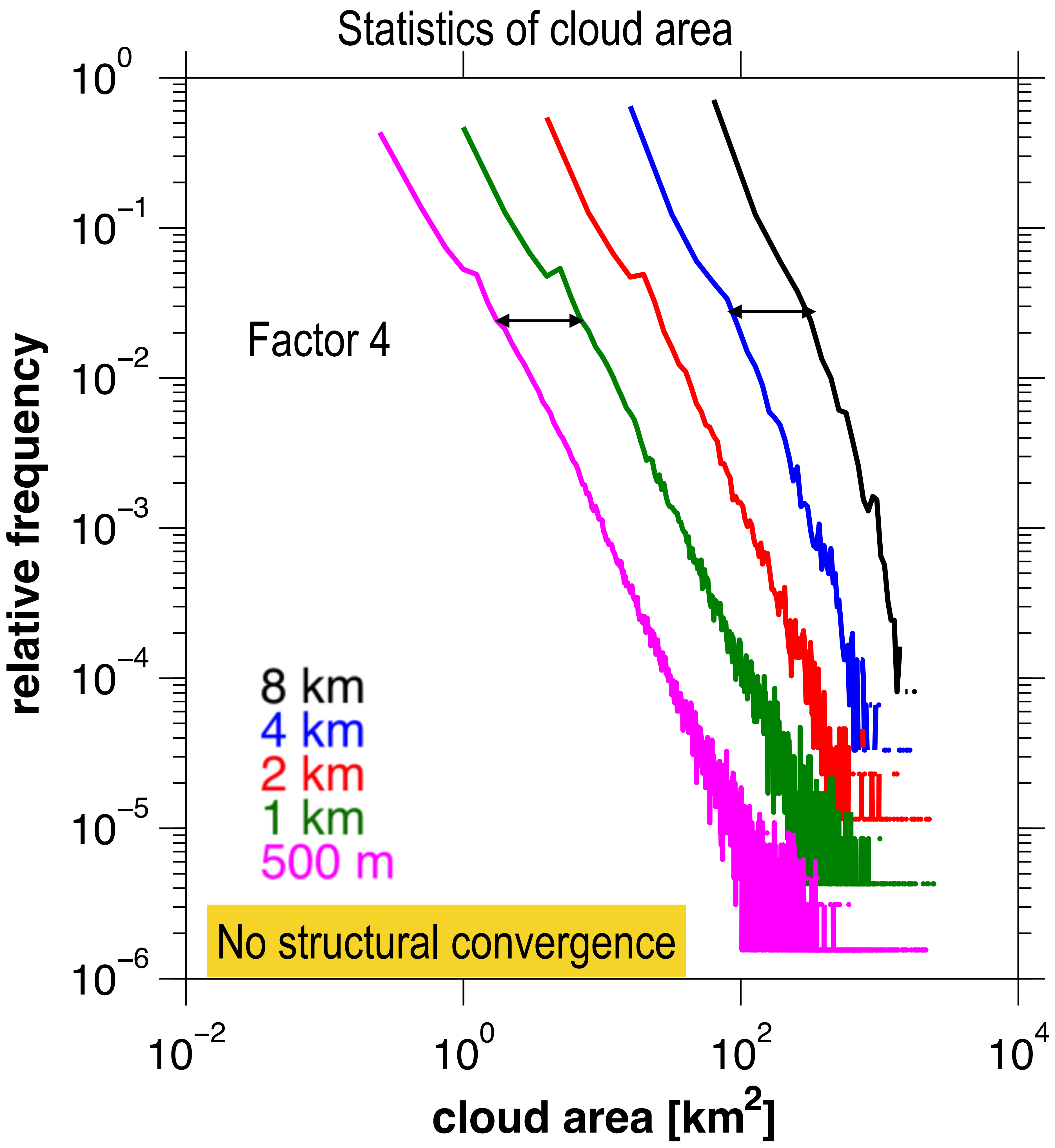


Area-averaged bulk effects upon ambient flow:
E.g., heating and moistening of cloud layer

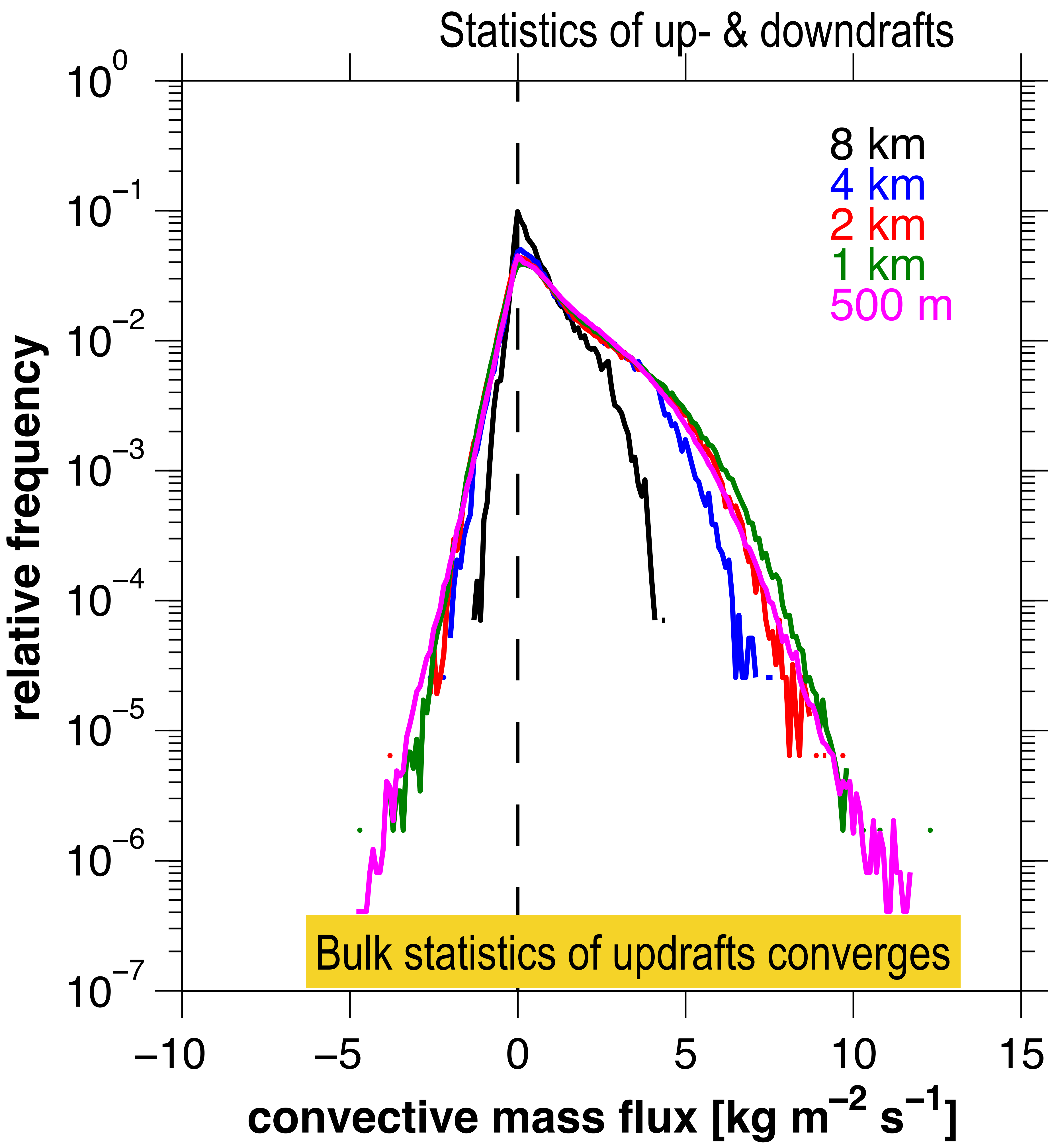
Source: Christoph Schär, ETH Zurich

Structural and bulk convergence

(Panosetti et al. 2018)



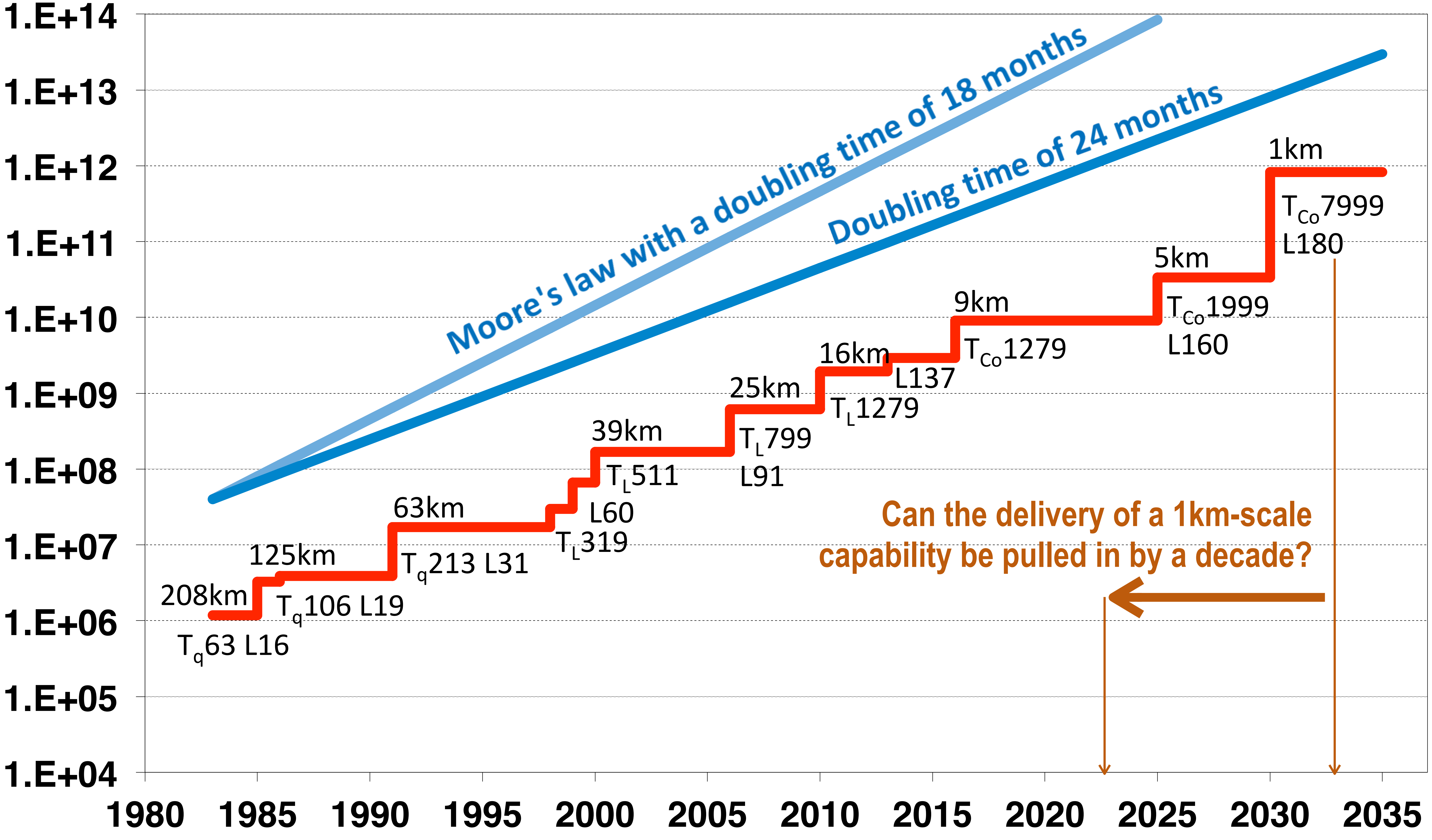
No structural convergence



Bulk statistics of updrafts converges

Source: Christoph Schär, ETH Zurich

Computational power drives spatial resolution



Can the delivery of a 1km-scale capability be pulled in by a decade?

Source: Christoph Schär, ETH Zurich, & Nils Wedi, ECMWF
Schulthess et al., 2019

Our “exascale” goal for 2022

Horizontal resolution	1 km (globally quasi-uniform)
Vertical resolution	180 levels (surface to ~100 km)
Time resolution	Less than 1 minute
Coupled	Land-surface/ocean/ocean-waves/sea-ice
Atmosphere	Non-hydrostatic
Precision	Single (32bit) or mixed precision
Compute rate	1 SYPD (simulated year wall-clock day)

Running COSMO 5.0 & IFS (“the European Model”) at global scale on Piz Daint

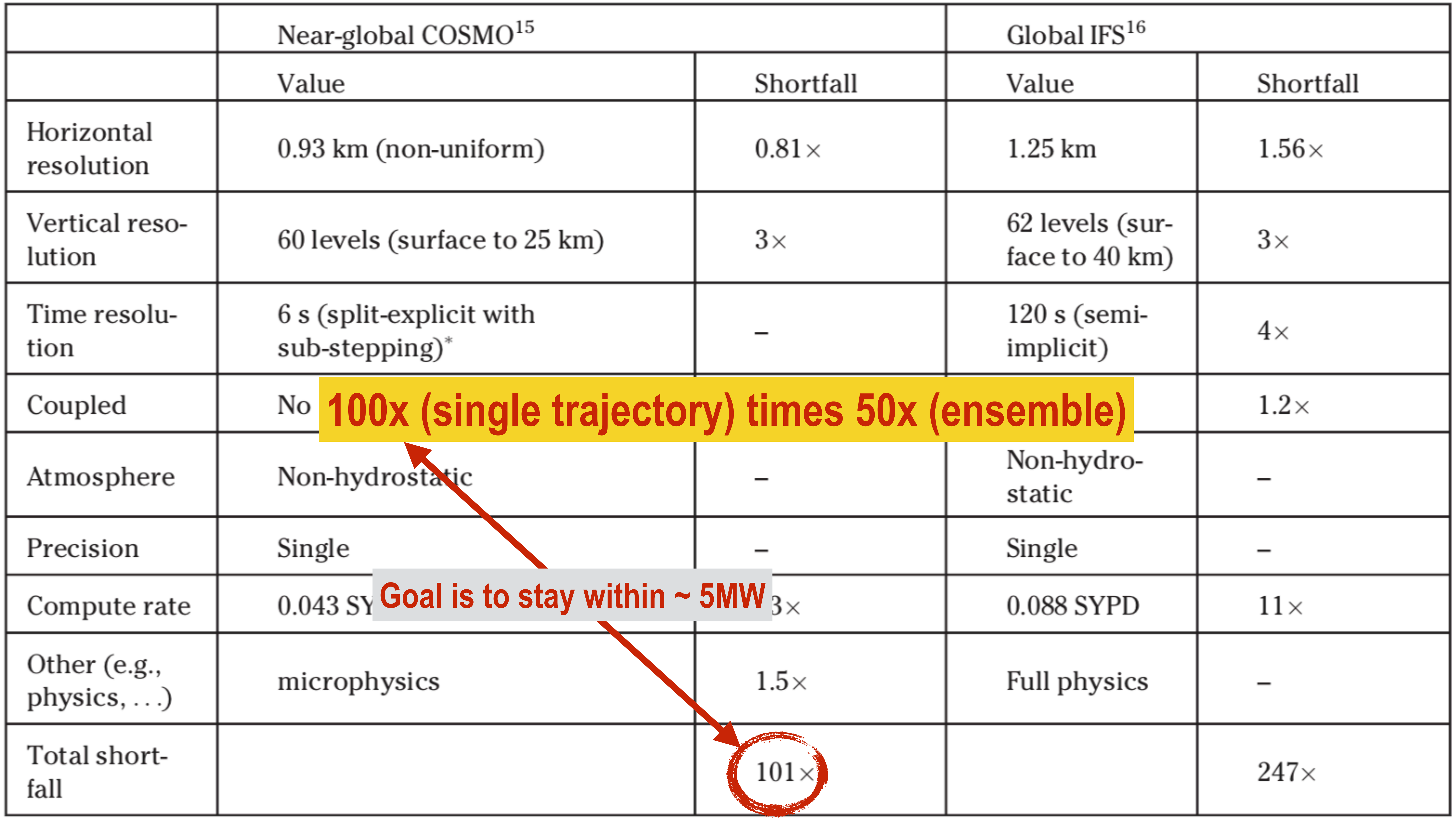
Scaling to full system size: ~5300 GPU accelerate nodes available

Running a near-global (±80º covering 97% of Earths surface) COSMO 5.0 simulation & IFS

> Either on the hosts processors: Intel Xeon E5 2690v3 (Haswell 12c).

> Or on the GPU accelerator: PCIe version of NVIDIA GP100 (Pascal) GPU

The baseline for COSMO-global and IFS



	Near-global COSMO[15]		Global IFS[16]	
	Value	Shortfall	Value	Shortfall
Horizontal resolution	0.93 km (non-uniform)	0.81×	1.25 km	1.56×
Vertical resolution	60 levels (surface to 25 km)	3×	62 levels (surface to 40 km)	3×
Time resolution	6 s (split-explicit with sub-stepping)*	–	120 s (semi-implicit)	4×
Coupled	No	[illegible]	[illegible]	1.2×
Atmosphere	Non-hydrostatic	–	Non-hydrostatic	–
Precision	Single	–	Single	–
Compute rate	0.043 SY	3×	0.088 SYPD	11×
Other (e.g., physics, . . .)	microphysics	1.5×	Full physics	–
Total shortfall		101×		247×

100x (single trajectory) times 50x (ensemble)

Goal is to stay within ~ 5MW

Memory use efficiency

Fuhrer et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-230, published 2018

Necessary data transfers

Actual data transfers

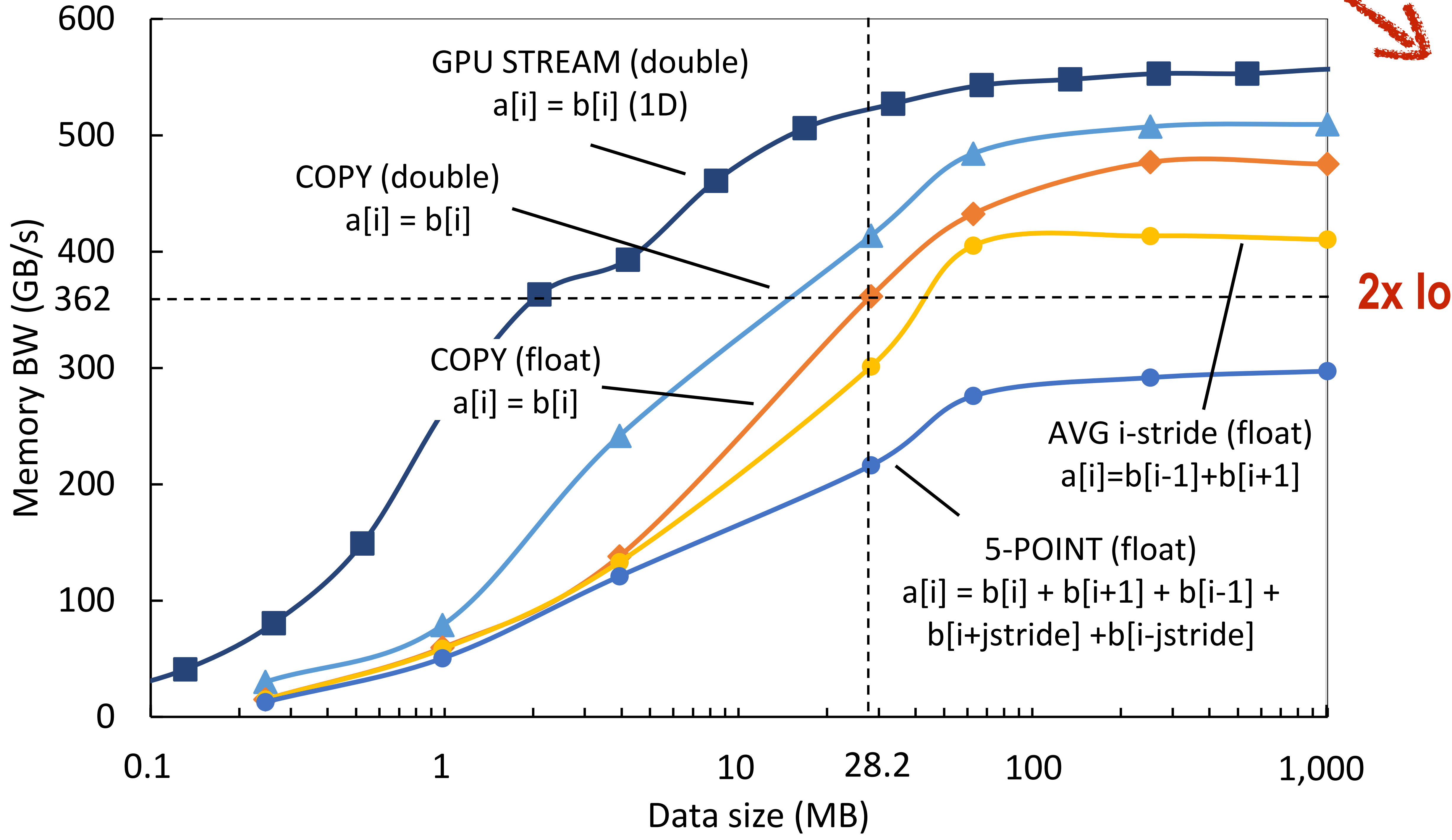
0.88

$$MUE = \text{I/O efficiency} \cdot \text{BW efficiency} = \frac{Q}{D}\,\frac{B}{\hat{B}} = 0.67$$

0.76

Achieved BW

Max achievable BW (STREAM)



0.55 w. regard to peak BW

2x lower than peak BW

Can the 100x shortfall of a grid-based implementation like COSMO-global be overcome?

1. Icosahedral/octahedral grid (ICON/IFS) vs. Lat-long/Cartesian grid (COSMO)

   2x fewer grid-columns

   Time step of 10 ms instead of 5 ms

   4x

2. Improving BW efficiency

   Improve BW efficiency and peak BW

   (results on Volta show this is realistic)

   2x

3. Strong scaling

   4x possible in COSMO, but we reduced available parallelism by factor 1.33

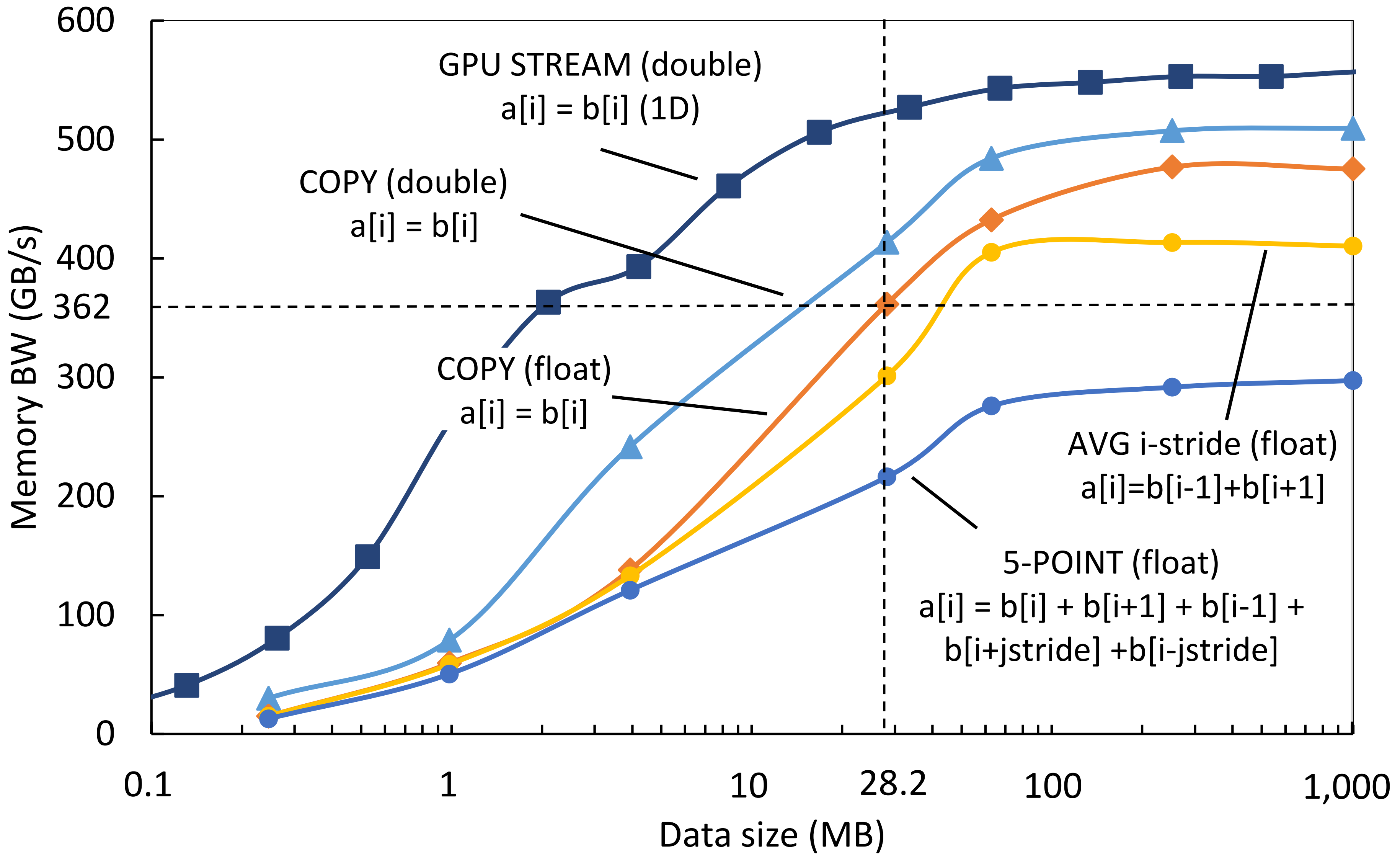
   3x

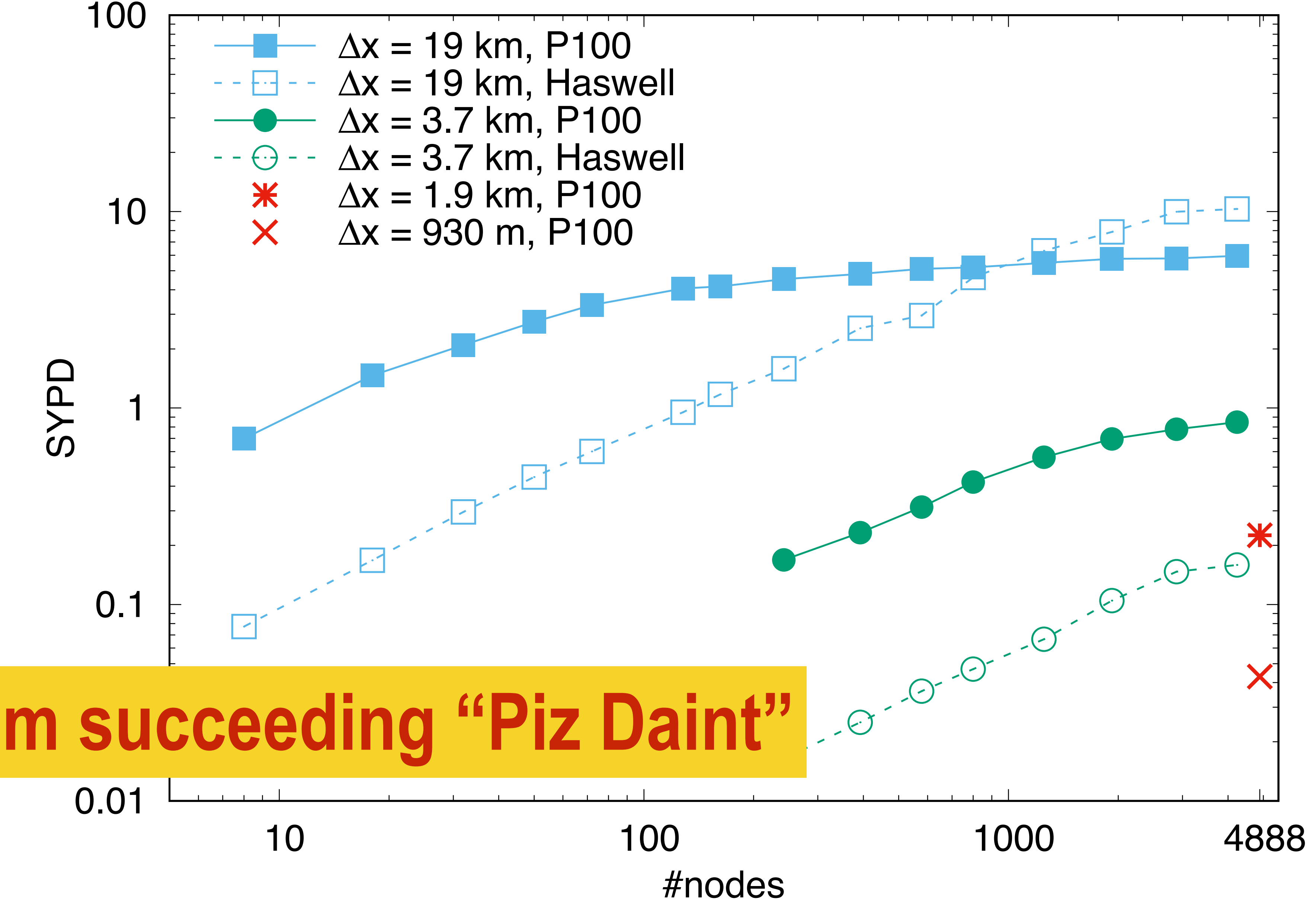
4. Remaining reduction in shortfall

   4x

   Numerical algorithms (larger time steps)

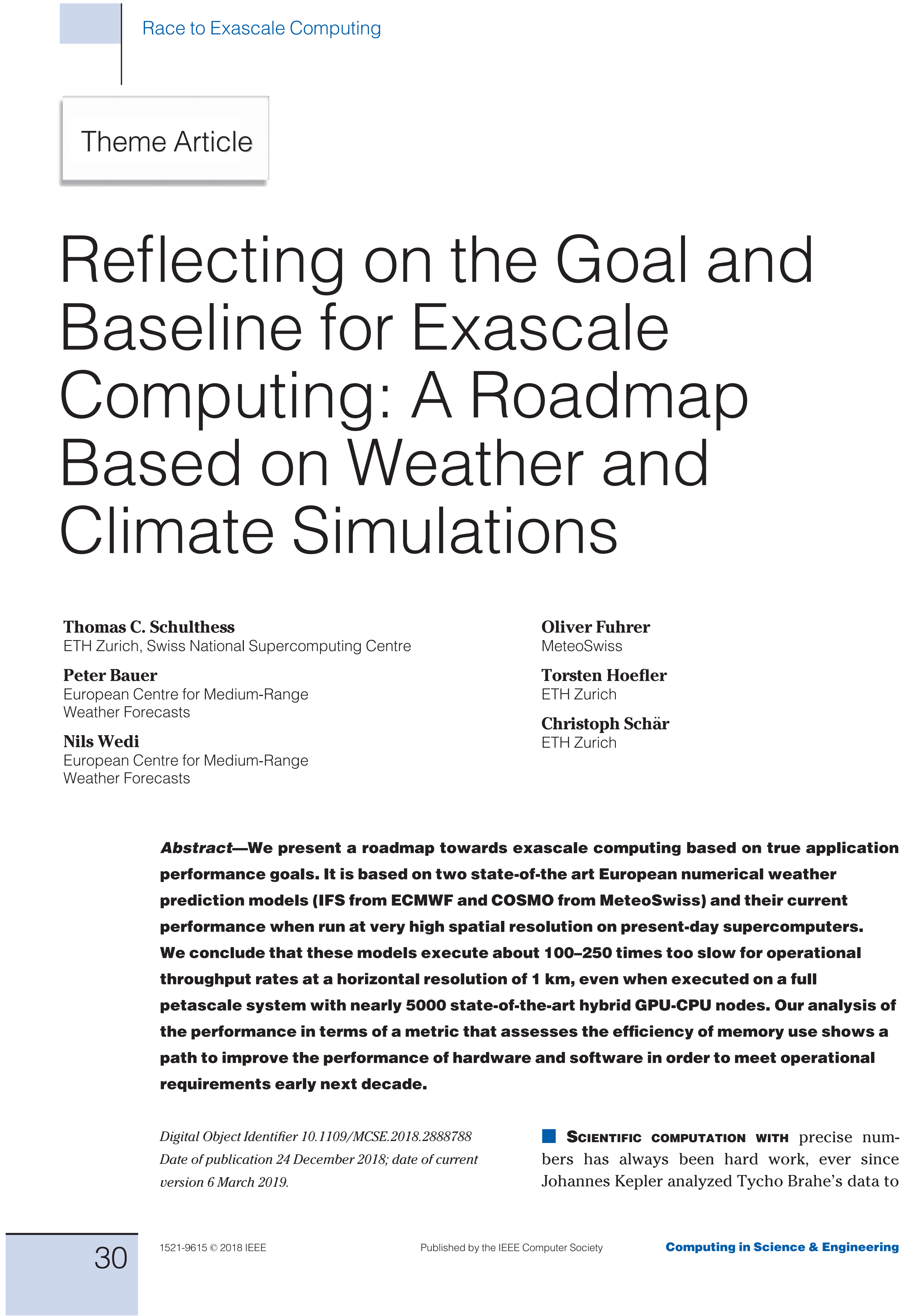
   Further improved processors / memory





But we don't want to increase the footprint of the 2022 system succeeding "Piz Daint"

Much of the data present here was from this article



Theme Article

Reflecting on the Goal and Baseline for Exascale Computing: A Roadmap Based on Weather and Climate Simulations

Thomas C. Schulthess
ETH Zurich, Swiss National Supercomputing Centre

Peter Bauer
European Centre for Medium-Range Weather Forecasts

Nils Wedi
European Centre for Medium-Range Weather Forecasts

Oliver Fuhrer
MeteoSwiss

Torsten Hoefler
ETH Zurich

Christoph Schär
ETH Zurich

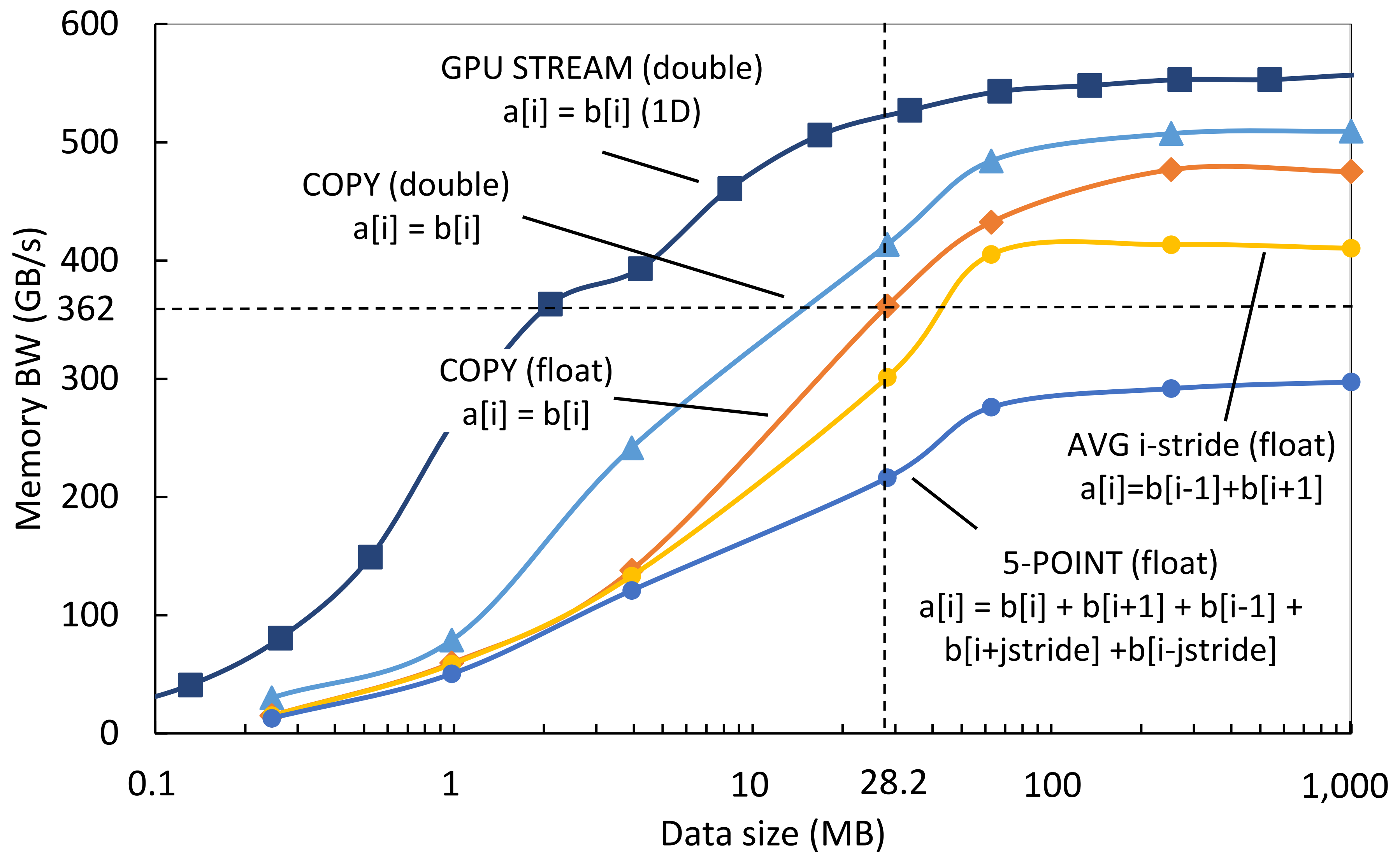
***Abstract*—We present a roadmap towards exascale computing based on true application performance goals. It is based on two state-of-the art European numerical weather prediction models (IFS from ECMWF and COSMO from MeteoSwiss) and their current performance when run at very high spatial resolution on present-day supercomputers. We conclude that these models execute about 100–250 times too slow for operational throughput rates at a horizontal resolution of 1 km, even when executed on a full petascale system with nearly 5000 state-of-the-art hybrid GPU-CPU nodes. Our analysis of the performance in terms of a metric that assesses the efficiency of memory use shows a path to improve the performance of hardware and software in order to meet operational requirements early next decade.**

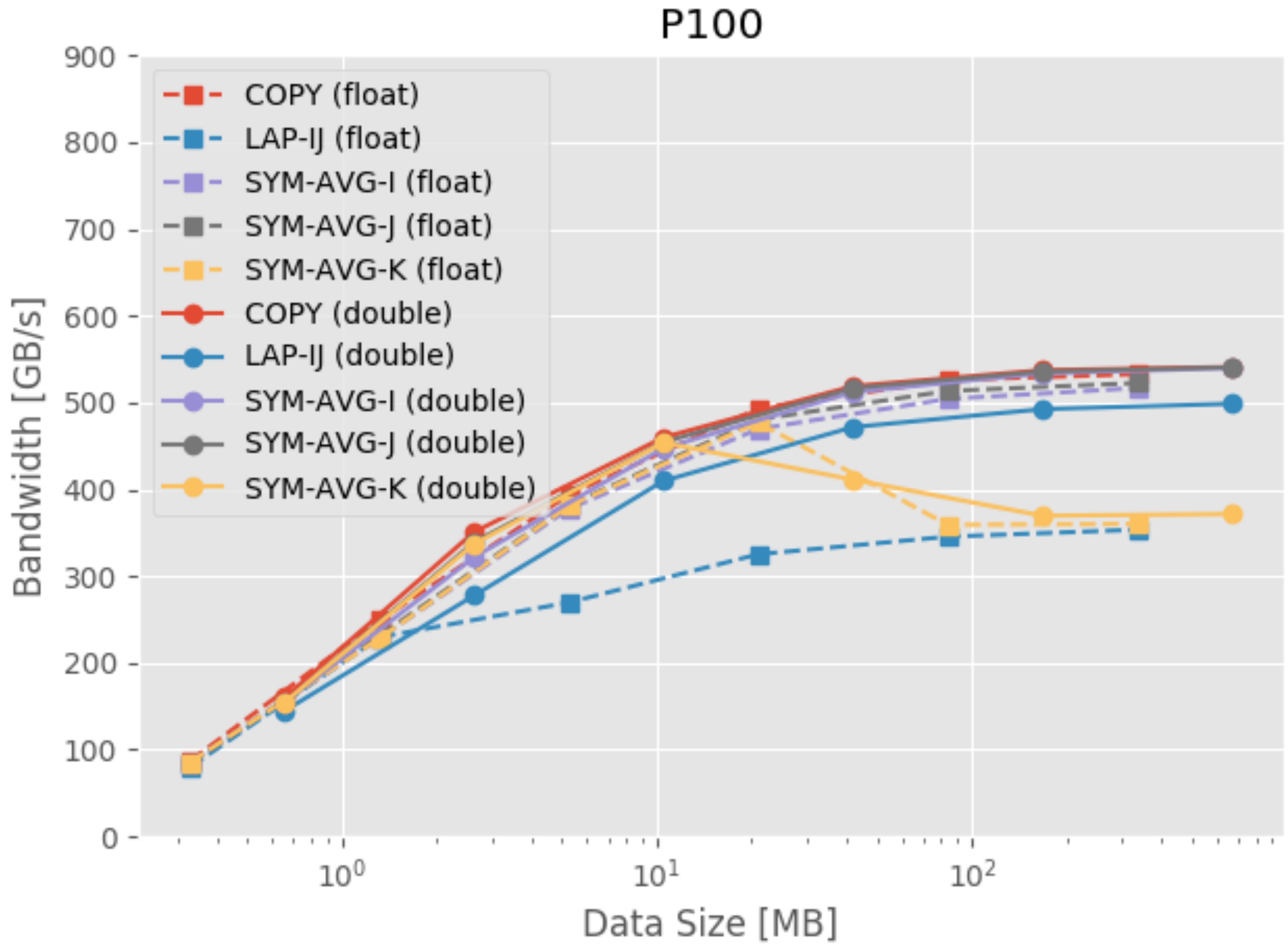
Digital Object Identifier 10.1109/MCSE.2018.2888788
Date of publication 24 December 2018; date of current version 6 March 2019.

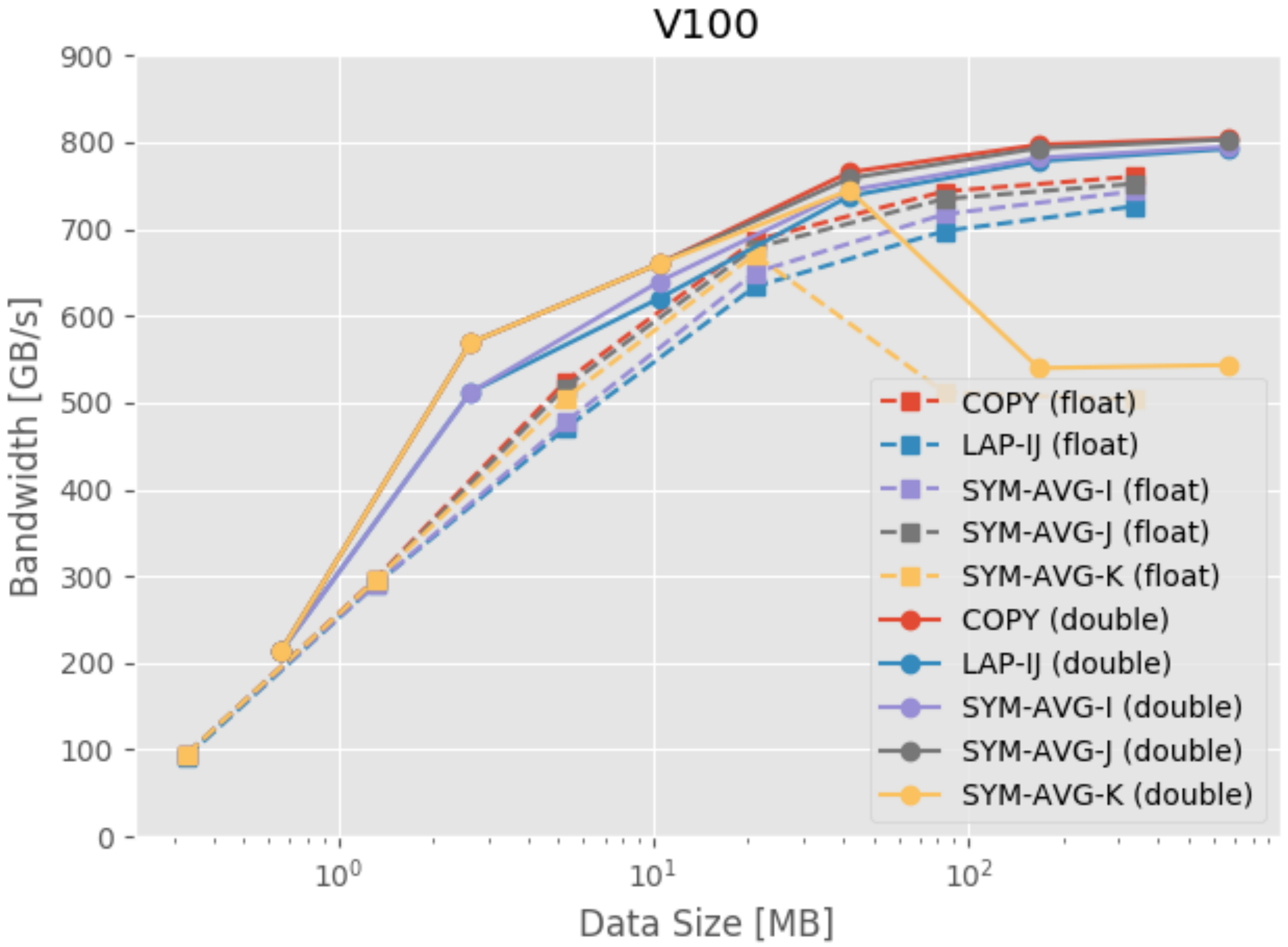
■ SCIENTIFIC COMPUTATION WITH precise numbers has always been hard work, ever since Johannes Kepler analyzed Tycho Brahe's data to

The good news:

memory performance is improving!

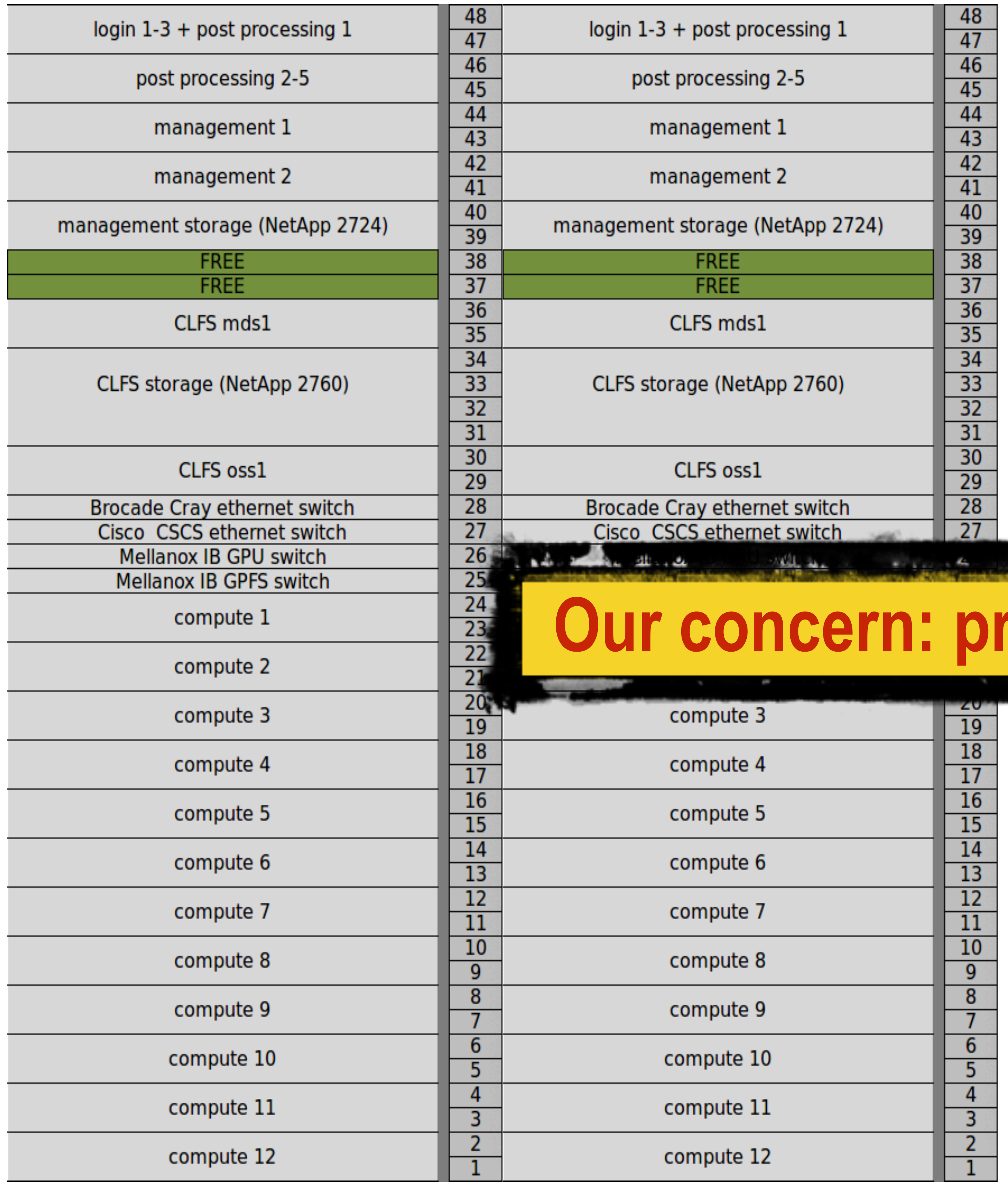






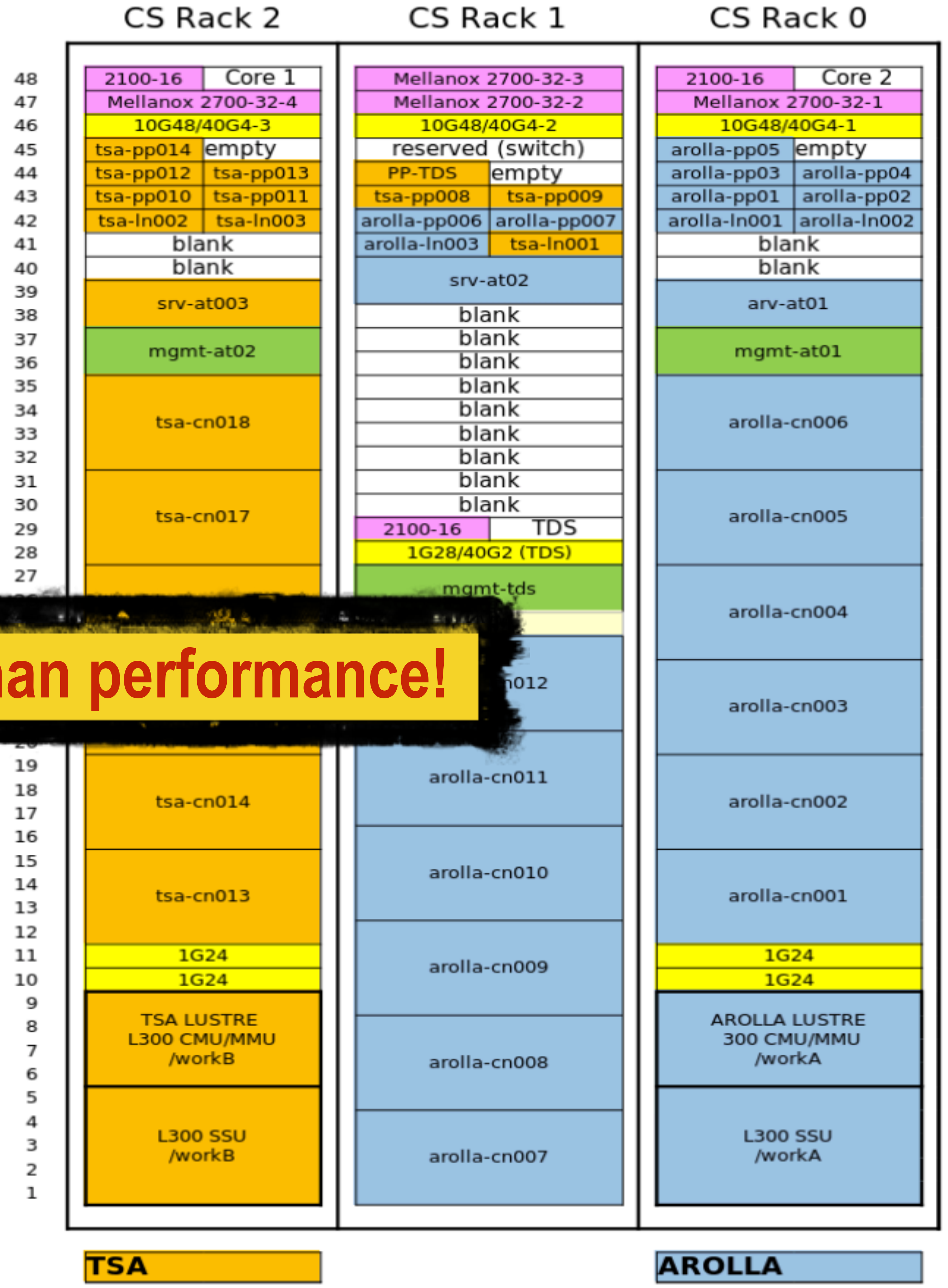
MeteoSwiss systems: Escha/Kesch (2015) vs. Arolla/Tsa (2019)

Two identical systems with 96 K80s (@480 GB/s) each



Escha Kesch

One system with two partitions of 96 and 48 V100 (@900 GB/s) each



Our concern: price increased faster than performance!

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Overcoming the 100x performance gap

- Our current estimates (modified/updated from Schulthess et al. 2019)
  - ~12x from improvements in software
  - >2x from improvements in memory performance
  - “only” factor 3-4 necessary from methods, algorithms, etc
- The real challenge will be data!
  - PRACE Tier 0 project based on 1 year allocation and 2.8 km horizontal resolution: 11.5 PB of data
  - Will Tier 0 projects that run at 1km horizontal resolution require 27x more online storage, or ~300 PB of data p.a.?

Use the Tier 0 project of MPI-M as opportunity to address challenge with data services – CSCS is willing to take on the challenge with partners

Collaborators

Tim Palmer (U. of Oxford)

Bjorn Stevens (MPI-M)

Peter Bauer (ECMWF)

Oliver Fuhrer (MeteoSwiss)

Nils Wedi (ECMWF)

Torsten Hoefler (ETH Zurich)

Christoph Schar (ETH Zurich)

Thank you!