From Pascaline to “Piz Daint” in the Alps infrastructure: A Modern-Day View of Computing in Science

Thomas C. Schulthess

“Piz Daint,” CSCS’ current flagship supercomputer

Introduced in 2013 and since 2017 features 5,704 NVIDIA P100 GPU accelerators, dubbed “Pascal”

World’s Most Powerful AI-Capable Supercomputer?

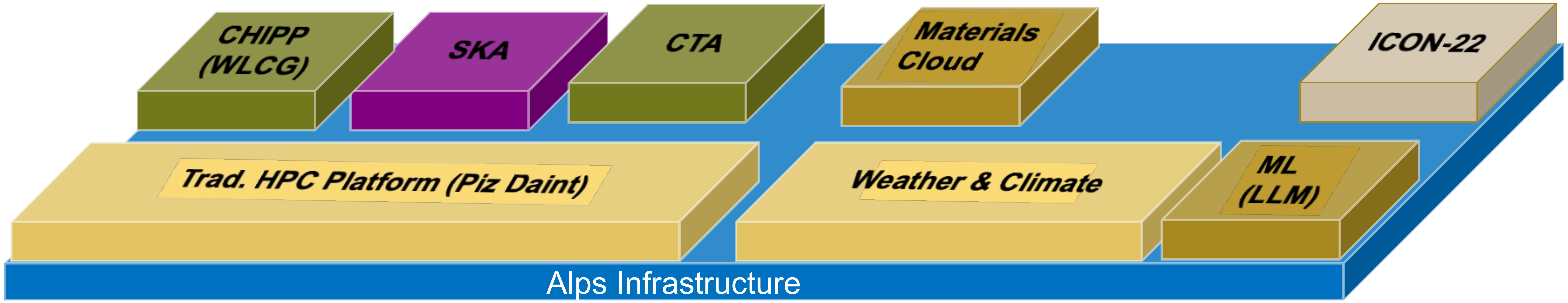
CSCS, Hewlett Packard Enterprise and NVIDIA Announce World’s Most...

12.04.2021

"Alps" system to advance research across climate, physics, life sciences with 7x more powerful AI capabilities than...

“Piz Daint” in the “Alps” Infrastructure

To a particular community, a platform will look like a dedicated supercomputer



CHIPP (WLCG)
SKA
CTA
Materials Cloud
ICON-22
Trad. HPC Platform (Piz Daint)
Weather & Climate
ML (LLM)
Alps Infrastructure

vClusters

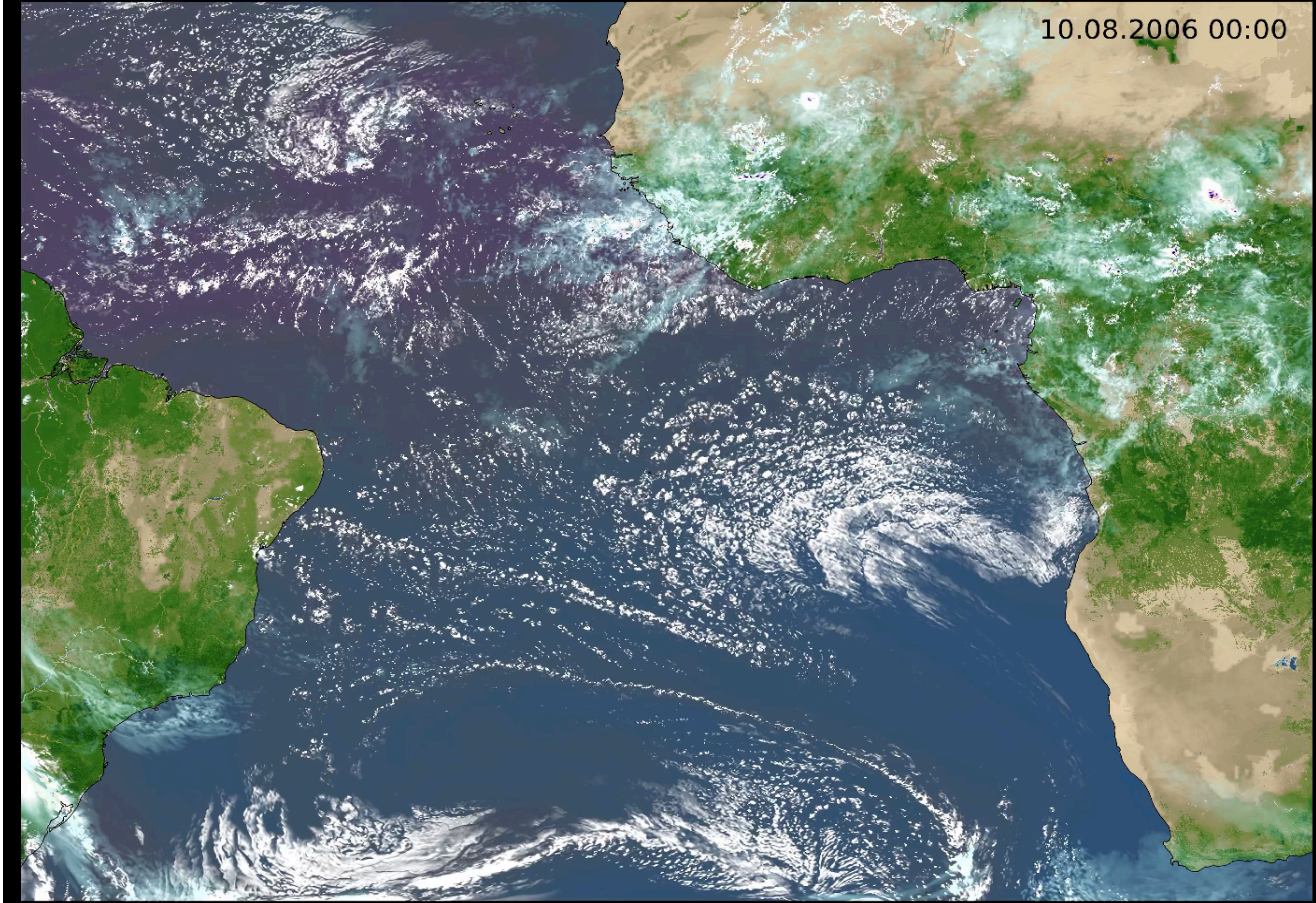
FIRECREST

SARUS

Slingshot network
Cray System Management software (µ-service arch.)

> “Supercomputers are by definition the fastest and most powerful general-purpose scientific computing systems available at any given time.”

–Dongarra et al. in “Numerical Linear Algebra for High-Performance Computers,”
SIAM 1998.



About 300×10^6 grid points

Christoph Heim, ETH Zürich

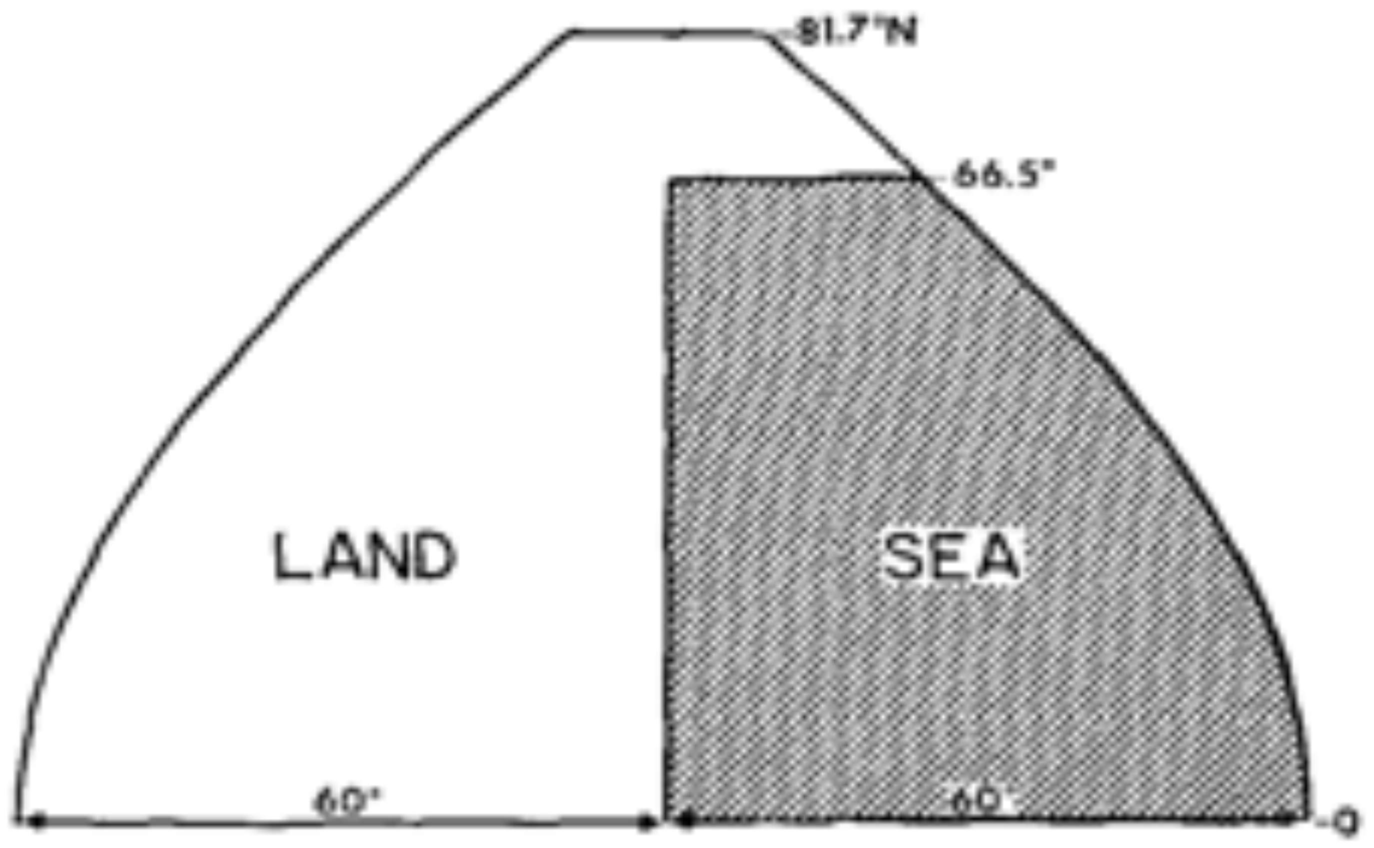
Frist 2 x CO2 general circulation model experiment

Manabe and Wetherald (1975): The Effects of Doubling the CO_2 Concentration on the Climate of a General Circulation Model. J. Atmos. Sci., Vol. 32(1), 3-15

Nobel Prize 2021
Hasselmann, Manabe & Parisi

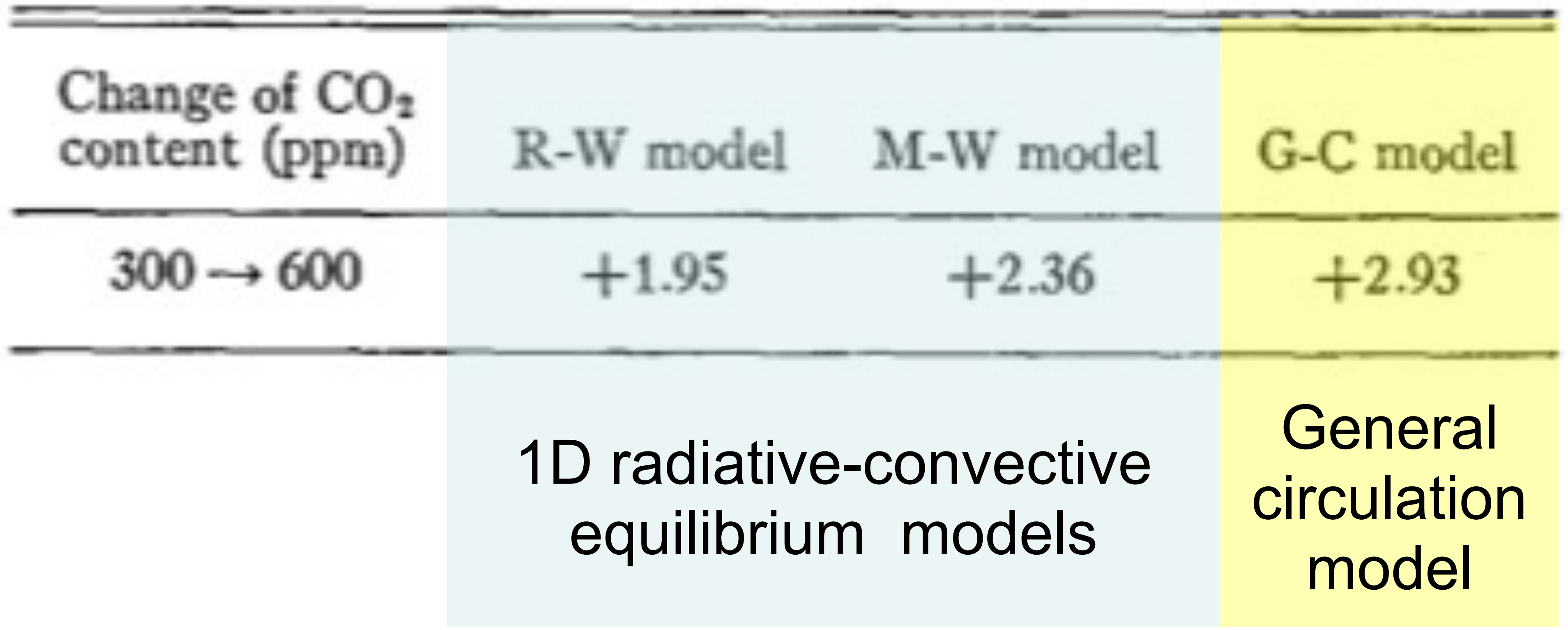
Main simplications:

- Atmosphere only
- Idealised distribution of land and sea, not global (only 120° longitude & periodic)
- Lateral resolution about 450 km with 9 layers (about 20 x 34 x 9 = 5220 grid points)
- No seasonal cycle, no diurnal cycle
- Prognostic water vapour and snow, bucket model over land, specified clouds
- Integrated for only a few 100 days



Notable results:

- Equilibrium climate sensitivity of **2.9°C**



Change of CO_2 content (ppm)	R-W model	M-W model	G-C model
300 → 600	+1.95	+2.36	+2.93
	1D radiative-convective equilibrium models		General circulation model

- Polar amplification
- Intensification of hydrological cycle
- Weakening of extratropical storm tracks

Scaling of computing and model performance

IBM Stretch 7030 (1961-1982)

About 1 MFLOPs = 10^6 ops / s

Piz Daint (CSCS, Lugano, 2013-2024)

20 PFLOPs = 20×10^{15} ops / s (2017-2024)

Manabe & Weatherald GCM
Δx = 450 km
9 layers

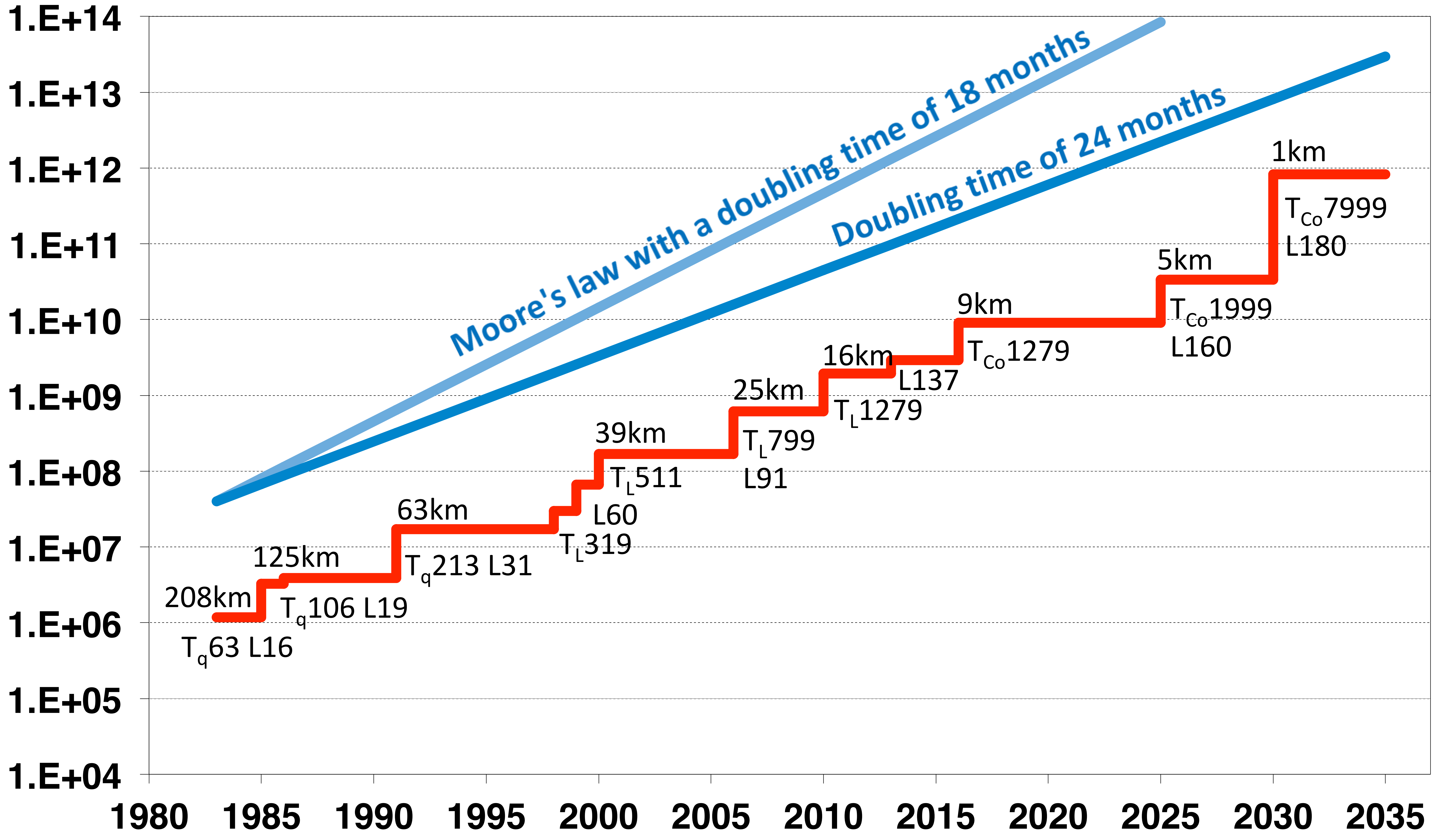
Today's runs?
Δx = 500 m
150 layers

Scale to today's computers
assuming optimal use of hardware

ETH zürich

Evolution of computing systems and model capability at ECMWF

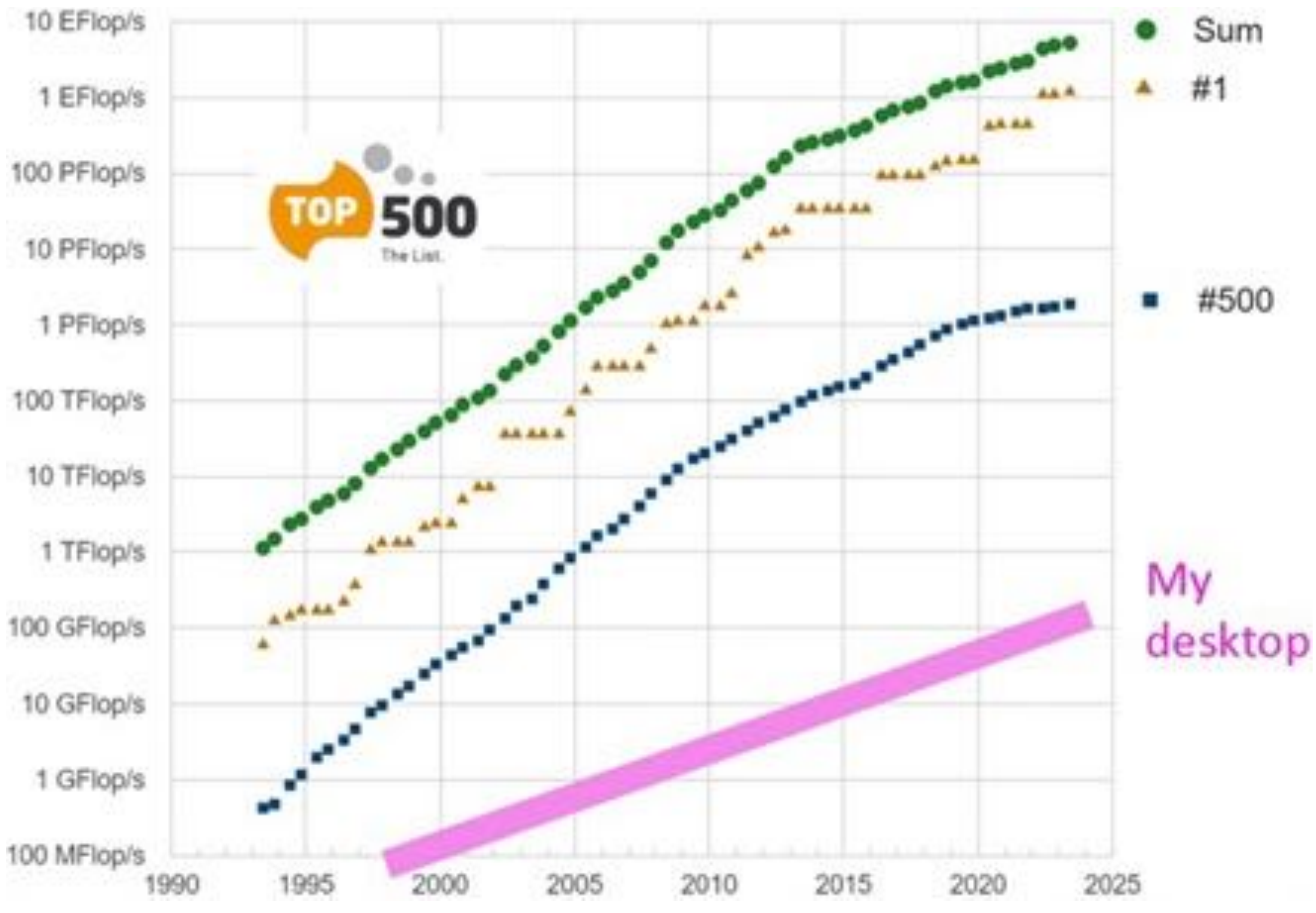
Computational power drives spatial resolution



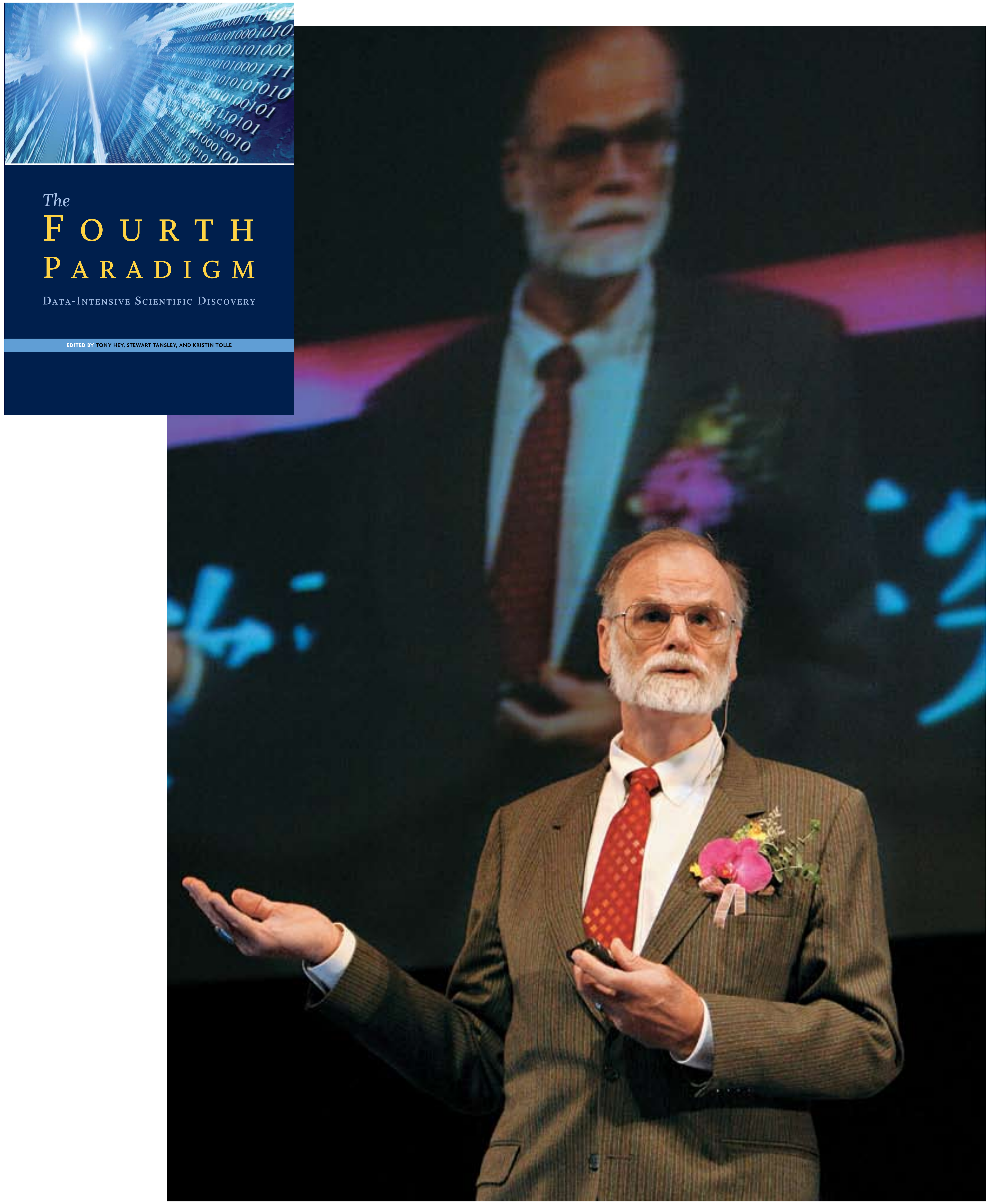
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Schulthess et al., Comp. Sci. Eng. 21 (1), 31-40 (2018)

A Bifurcation in Moore's Law?

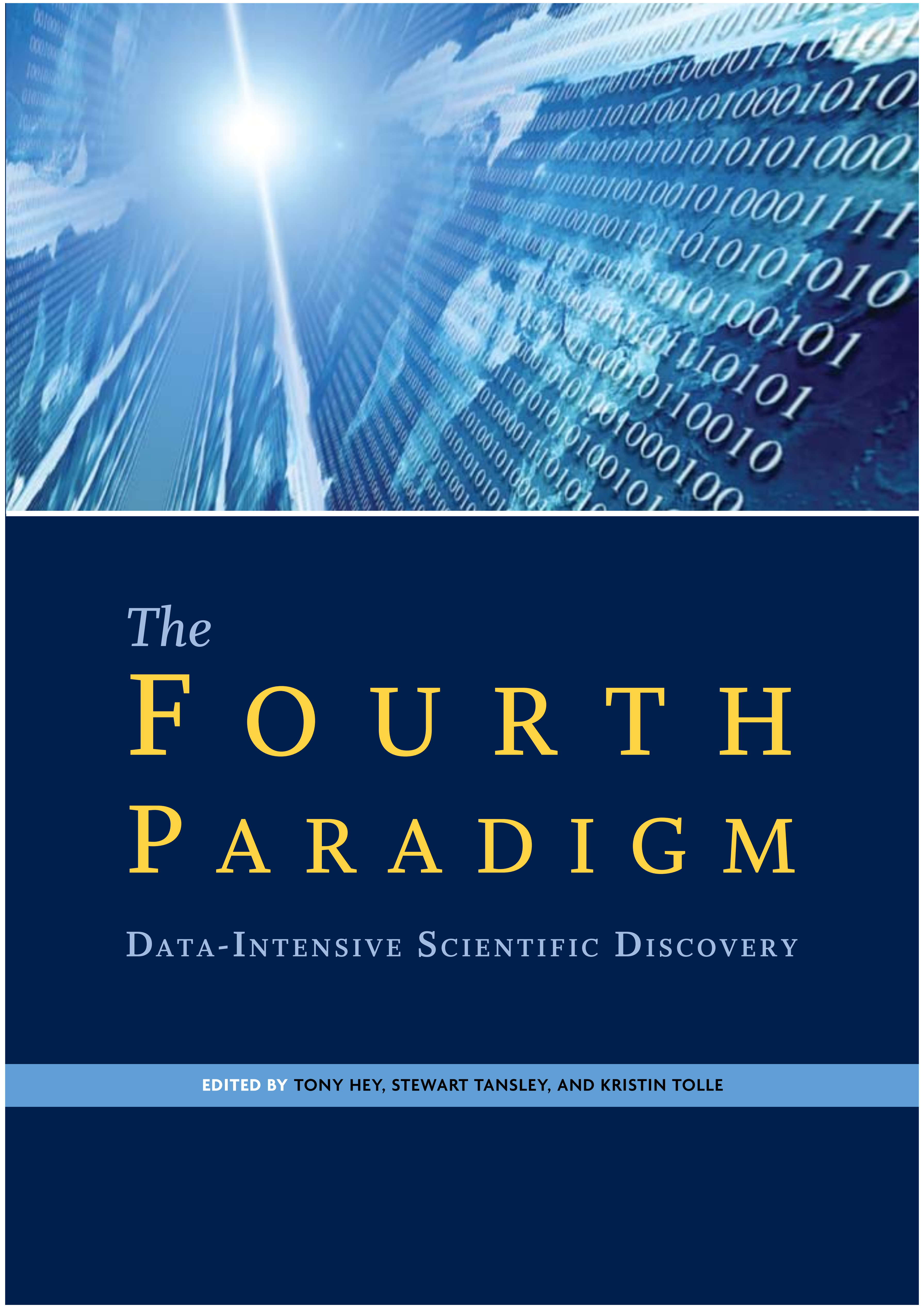


Nick Trefethen, SIAM News, September 01, 2023



Jim Gray on eScience:
A Transformed Scientific Method

National Research Council’s Computer Science and Telecommunications Board, Jan. 11, 2007

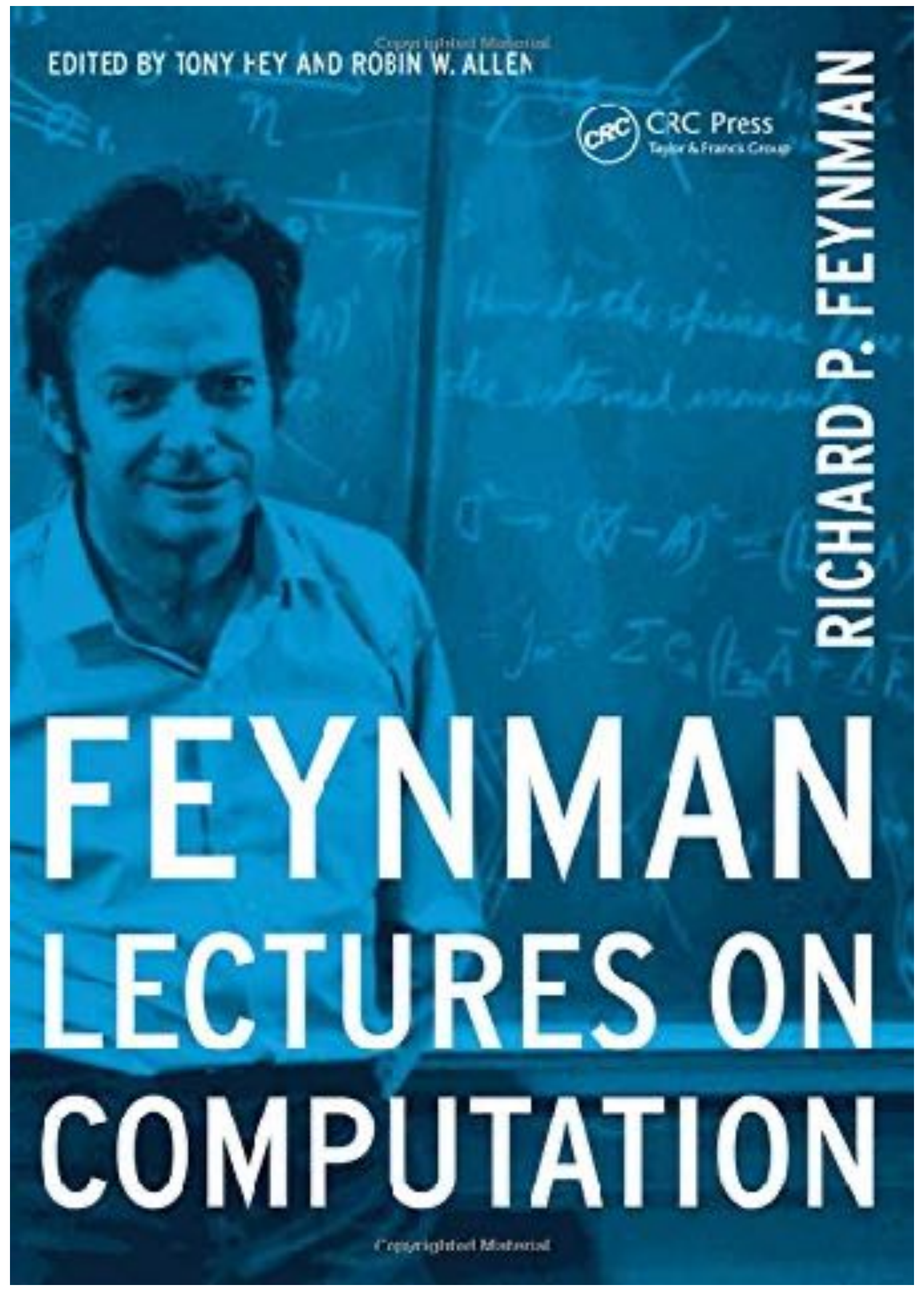


The

FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



Copyrighted Material

EDITED BY TONY HEY AND ROBIN W. ALLEN

CRC Press
Taylor & Francis Group

RICHARD P. FEYNMAN

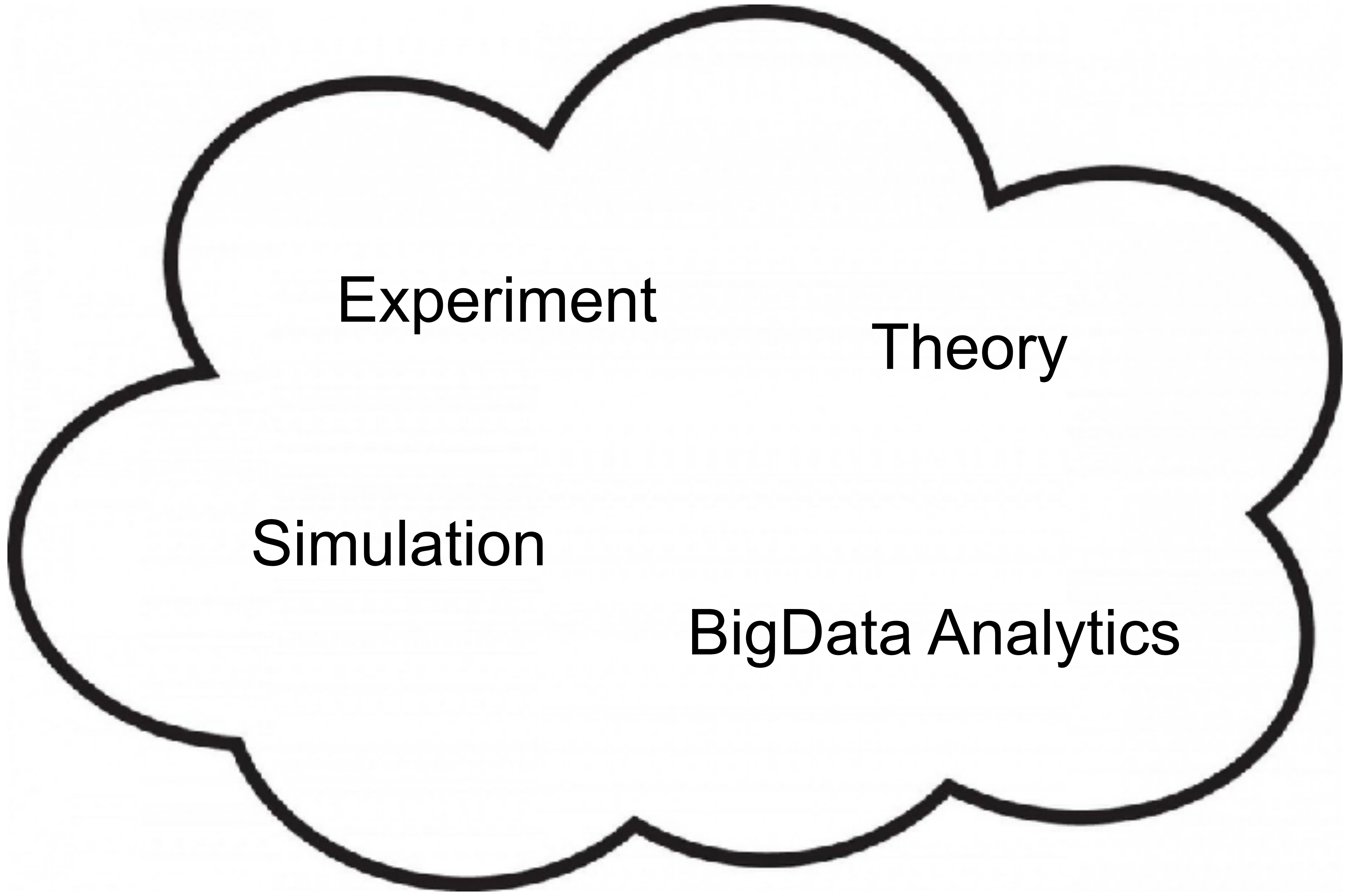
FEYNMAN LECTURES ON COMPUTATION

Copyrighted Material

Characteristics of BigData Analytics

Important considerations when dealing with digital data:

1. Velocity
2. Volume
3. Variety
4. Veracity
5. Value

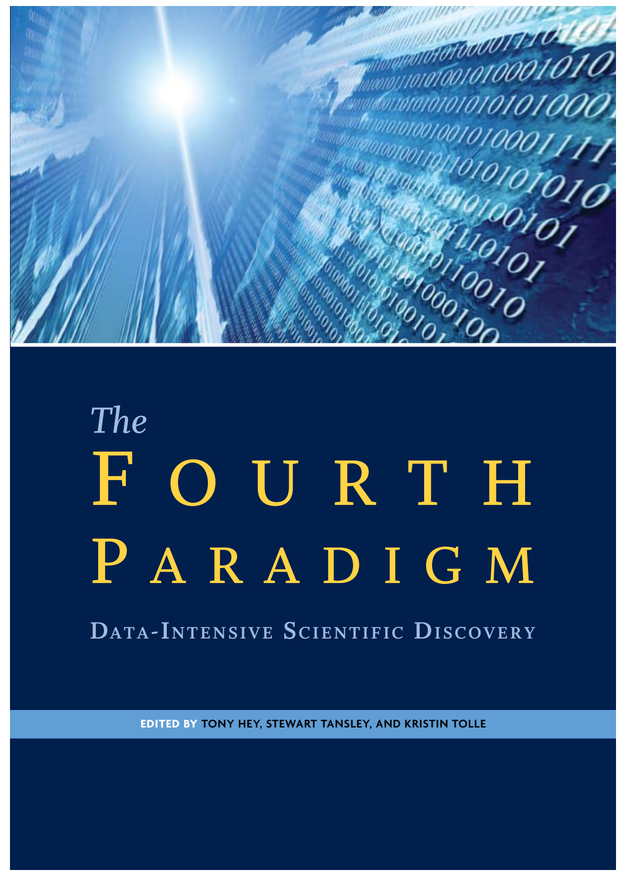


Experiment

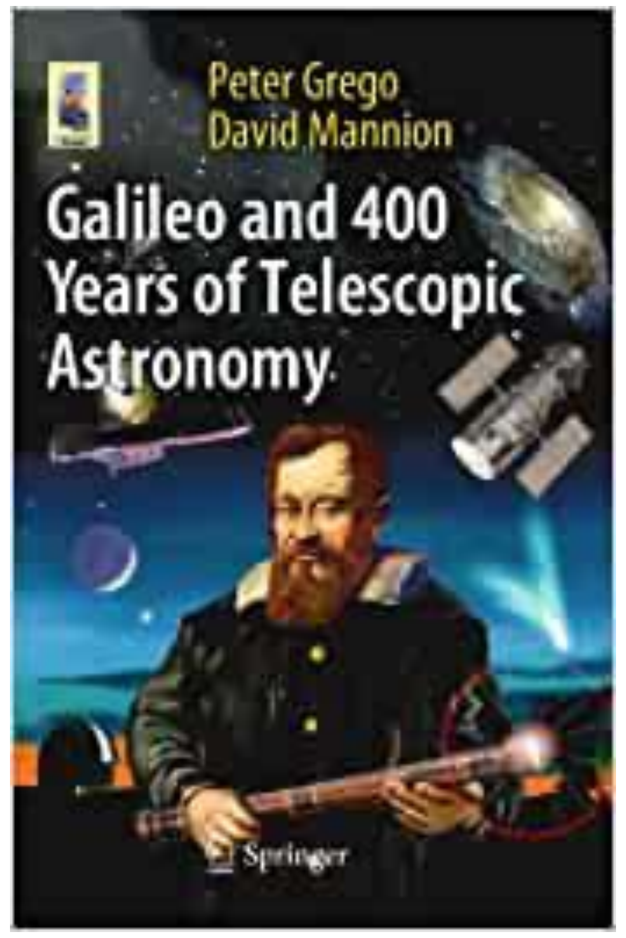
Theory

Simulation

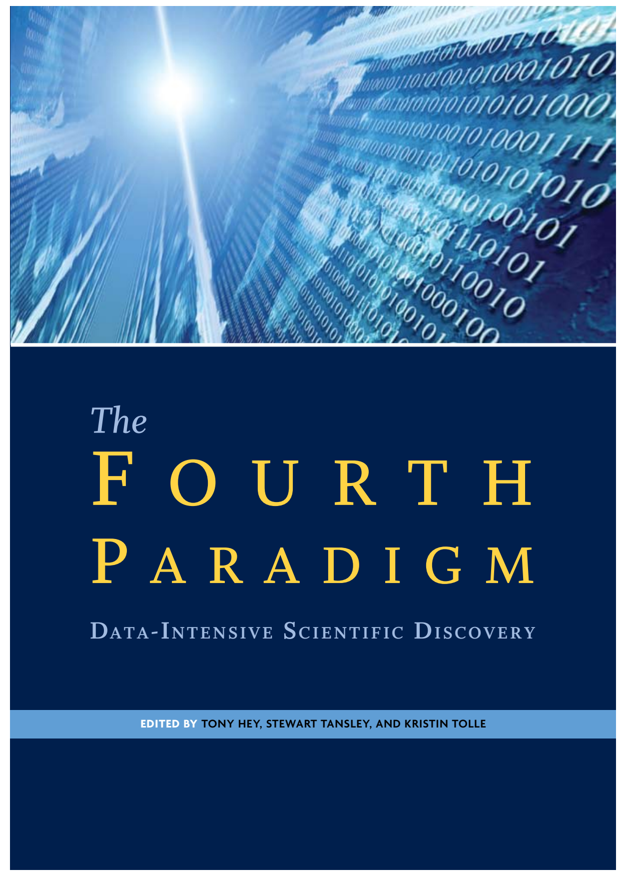
BigData Analytics



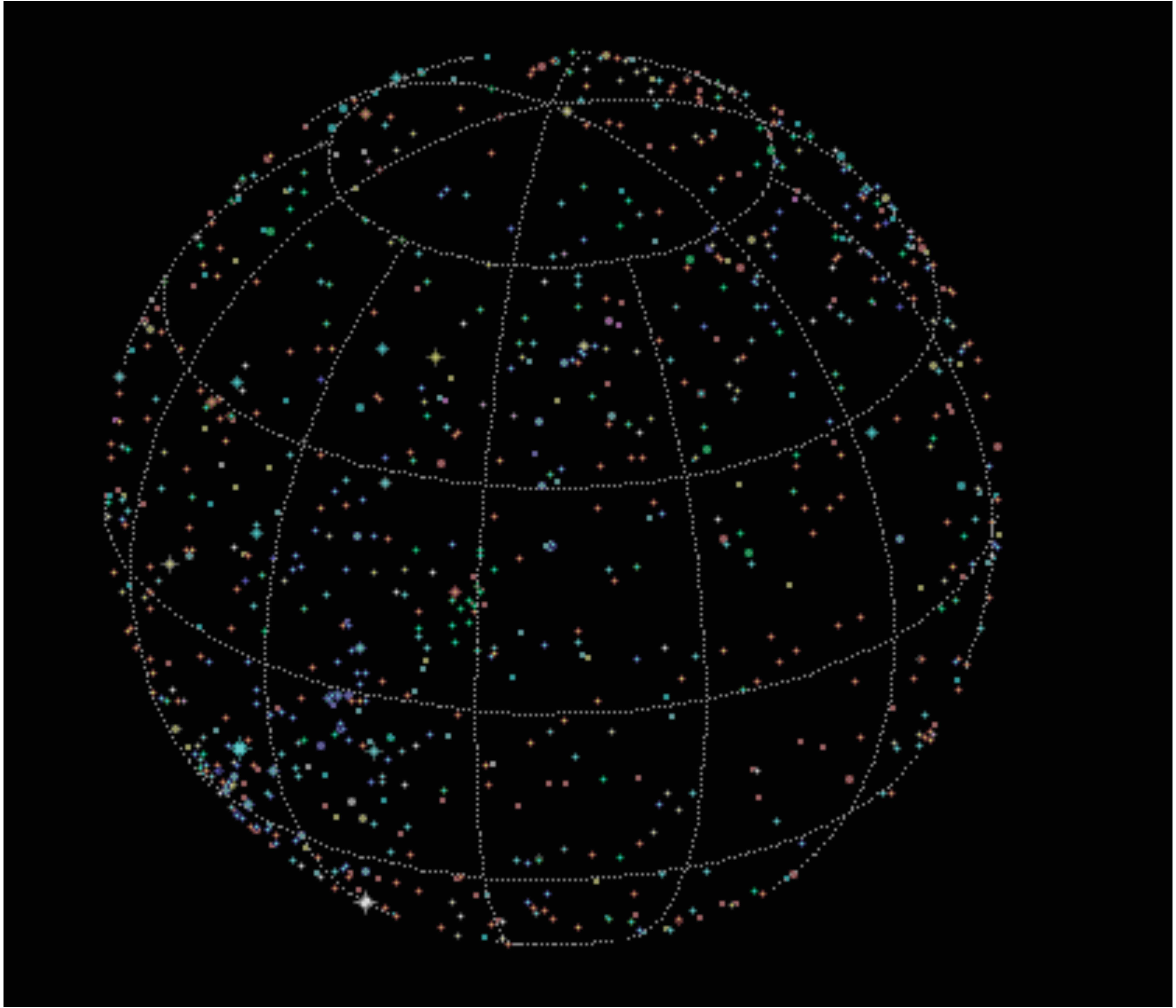
> “I’ve studies all available charts of the planets and stars and none of them match the others. There are just as many measurements and methods as there are astronomers and all of them disagree. What’s needed is a long term project with the aim of mapping the heavens conducted from a single location over a period of several years.”
>
> –Tycho Brahe, 1563



Experiment (observation), Veracity and Variety

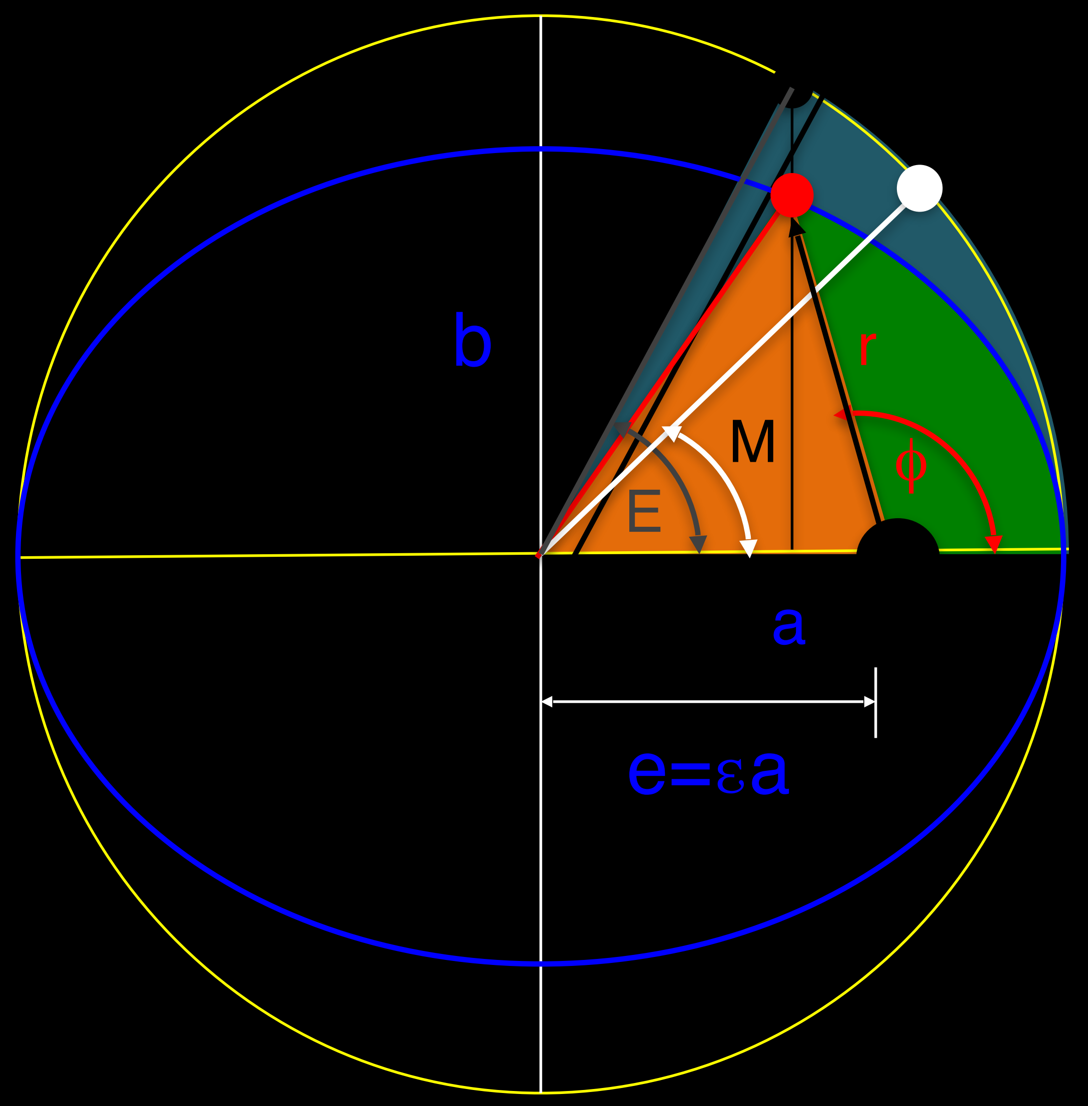


The
FOURTH
PARADIGM
Data-Intensive Scientific Discovery
EDITED BY: TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



source: www.pafko.com/tycho/

source: www.pafko.com/tycho/

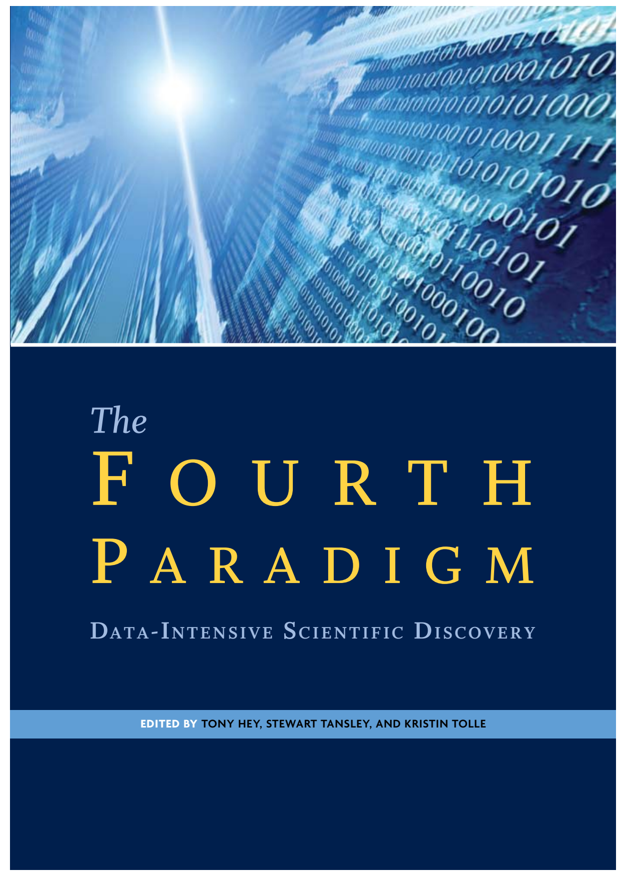


$$M = E - \varepsilon \sin E$$

1. Solve $E(M)$ (Numerics)
2. Solve $\phi(E)$ (Geometry)

source: www.pafko.com/tycho/

Data Analytics

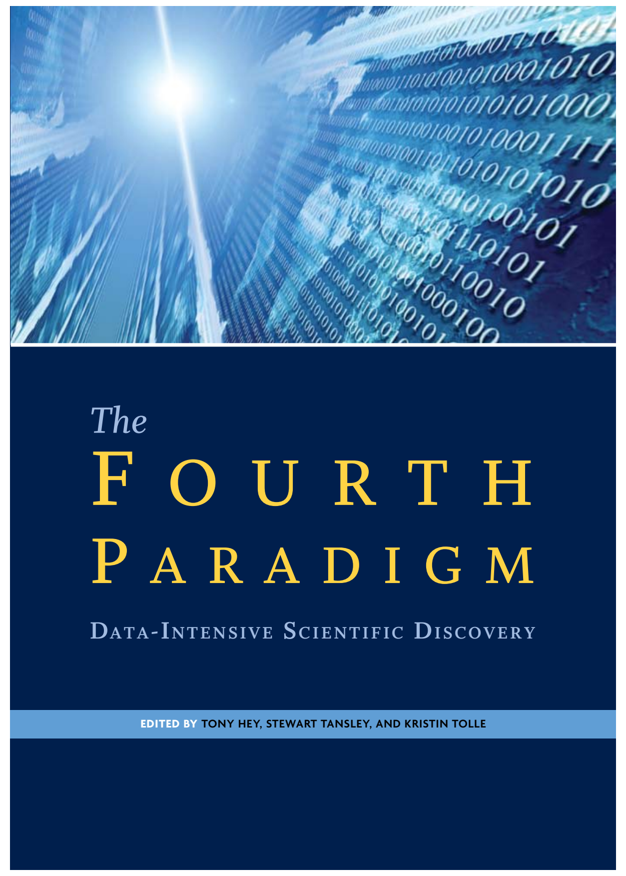


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

source: www.pafko.com/tycho/

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Theory



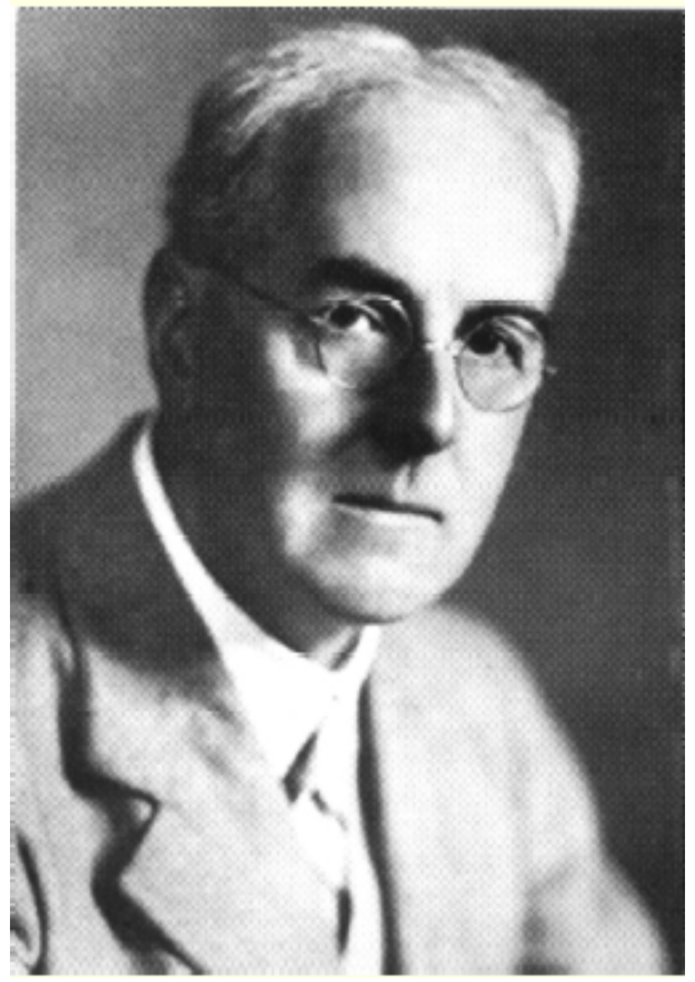
The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Jean Joseph Le Verrier predicts existence and position of Neptune to within 1º, confirmed by Johann Galle on 09/23/1846

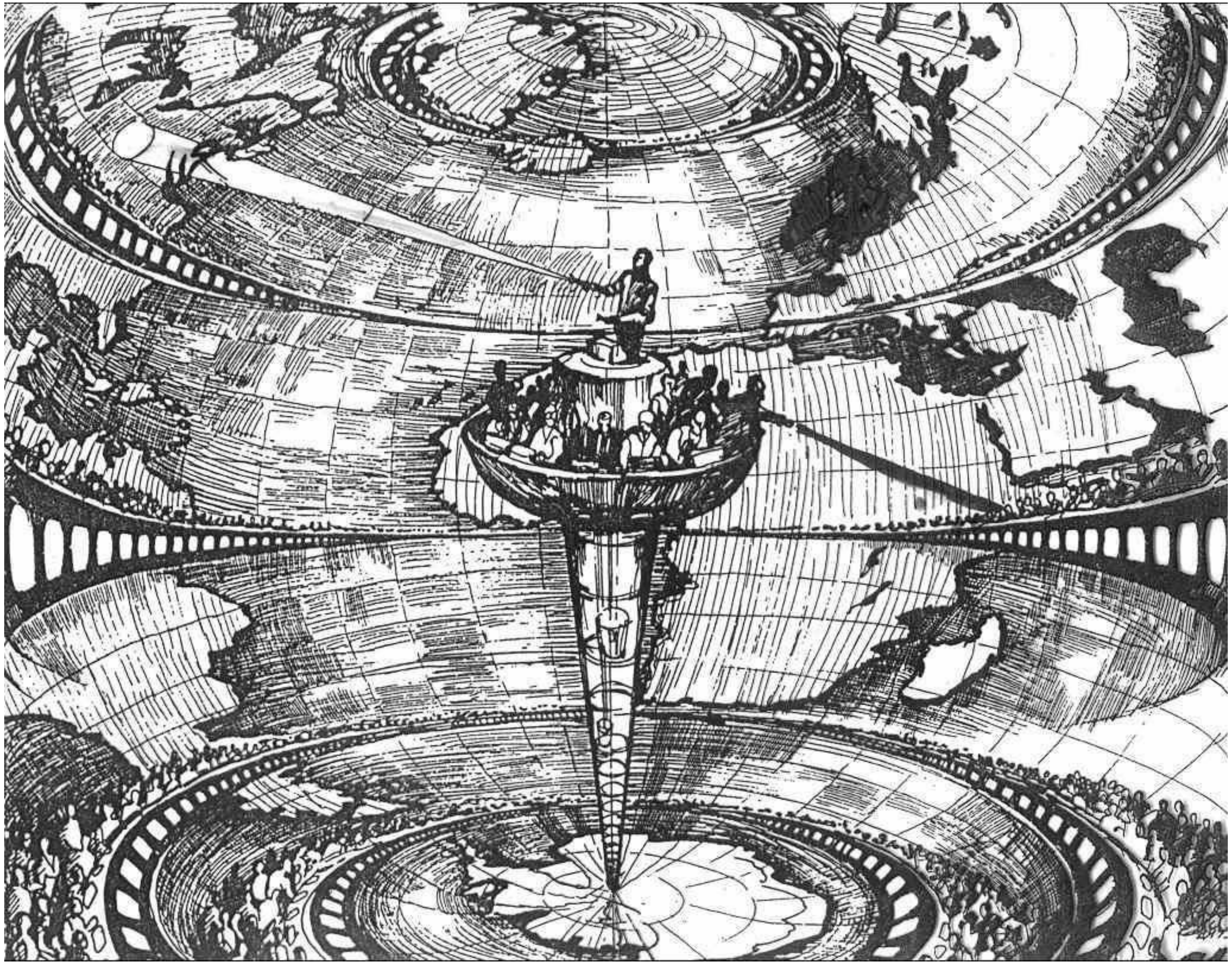
Similar predictions around the same time by John Couch Adams

Simulation?

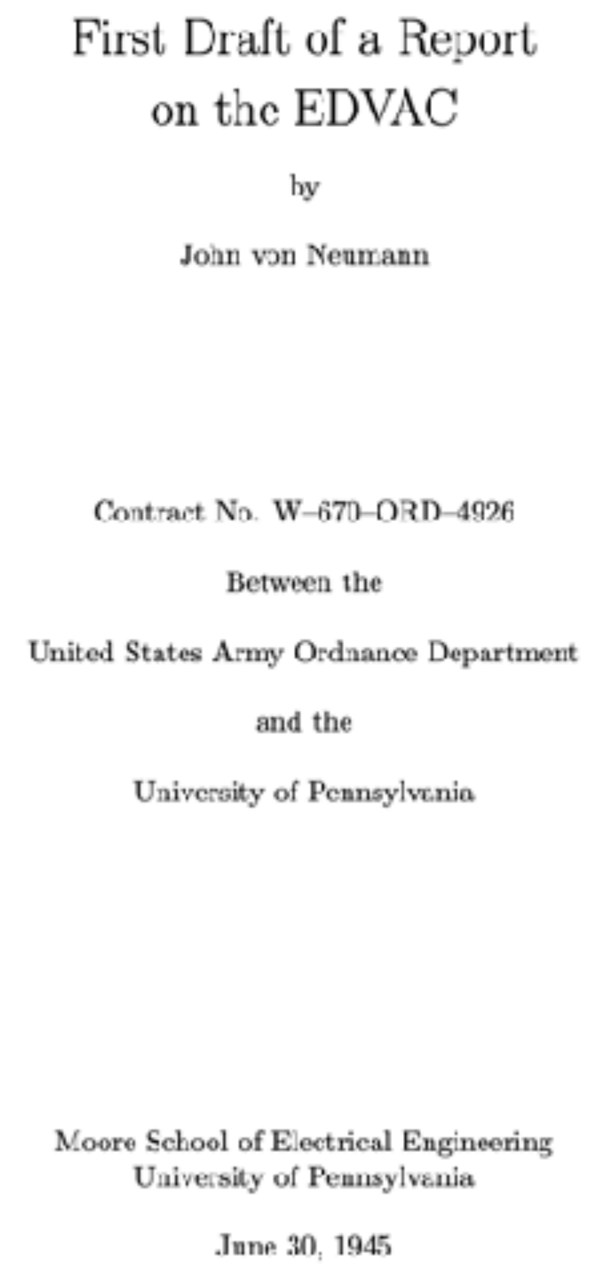
Richardson’s forecast factory (1922)



Lewis Fry Richardson:
Weather Prediction by Numerical Process



Bulk synchronous parallel (BSP) computing model



First Draft of a Report
on the EDVAC

by

John von Neumann

Contract No. W–670–ORD–4926

Between the

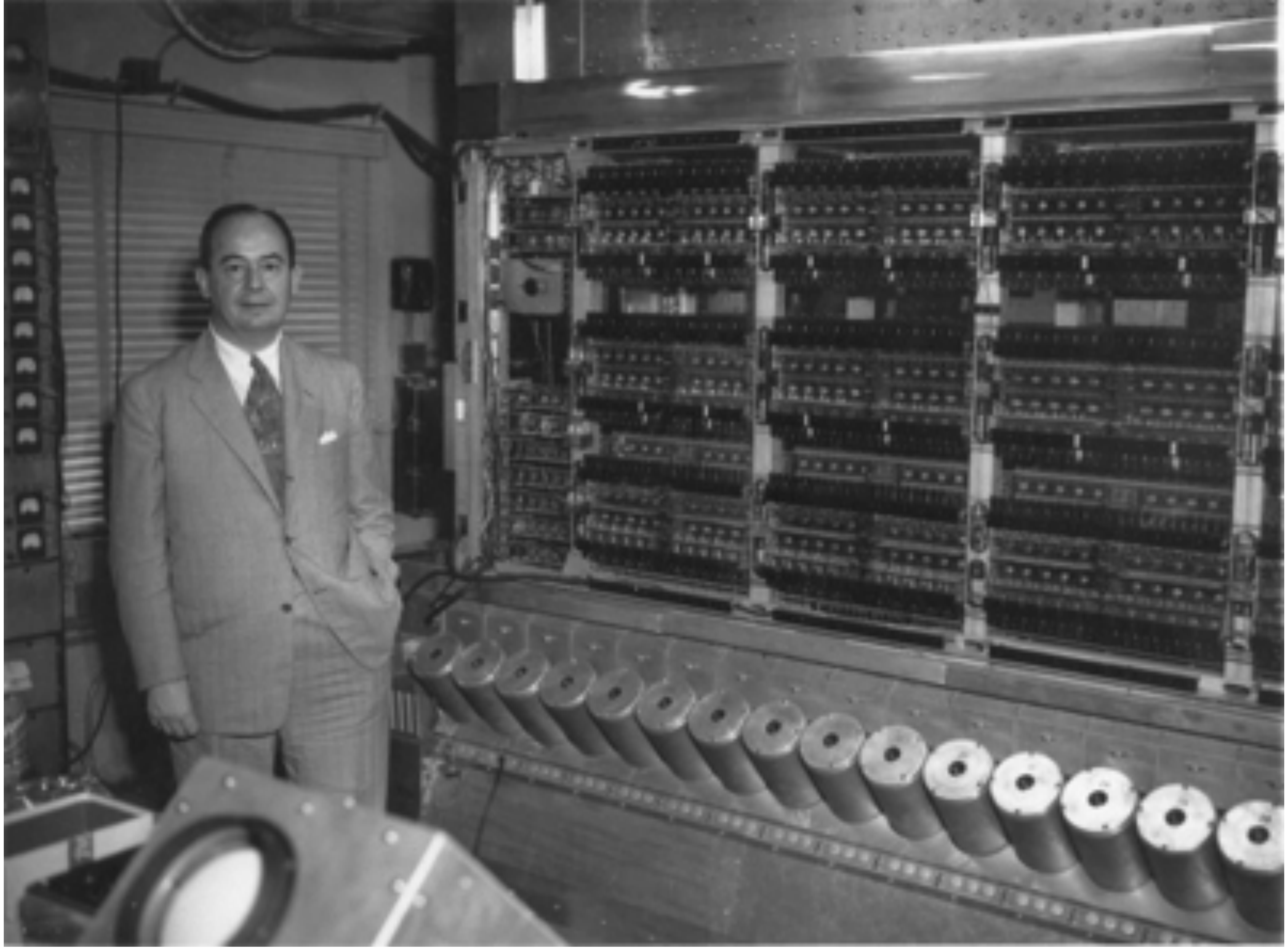
United States Army Ordnance Department

and the

University of Pennsylvania

Moore School of Electrical Engineering
University of Pennsylvania

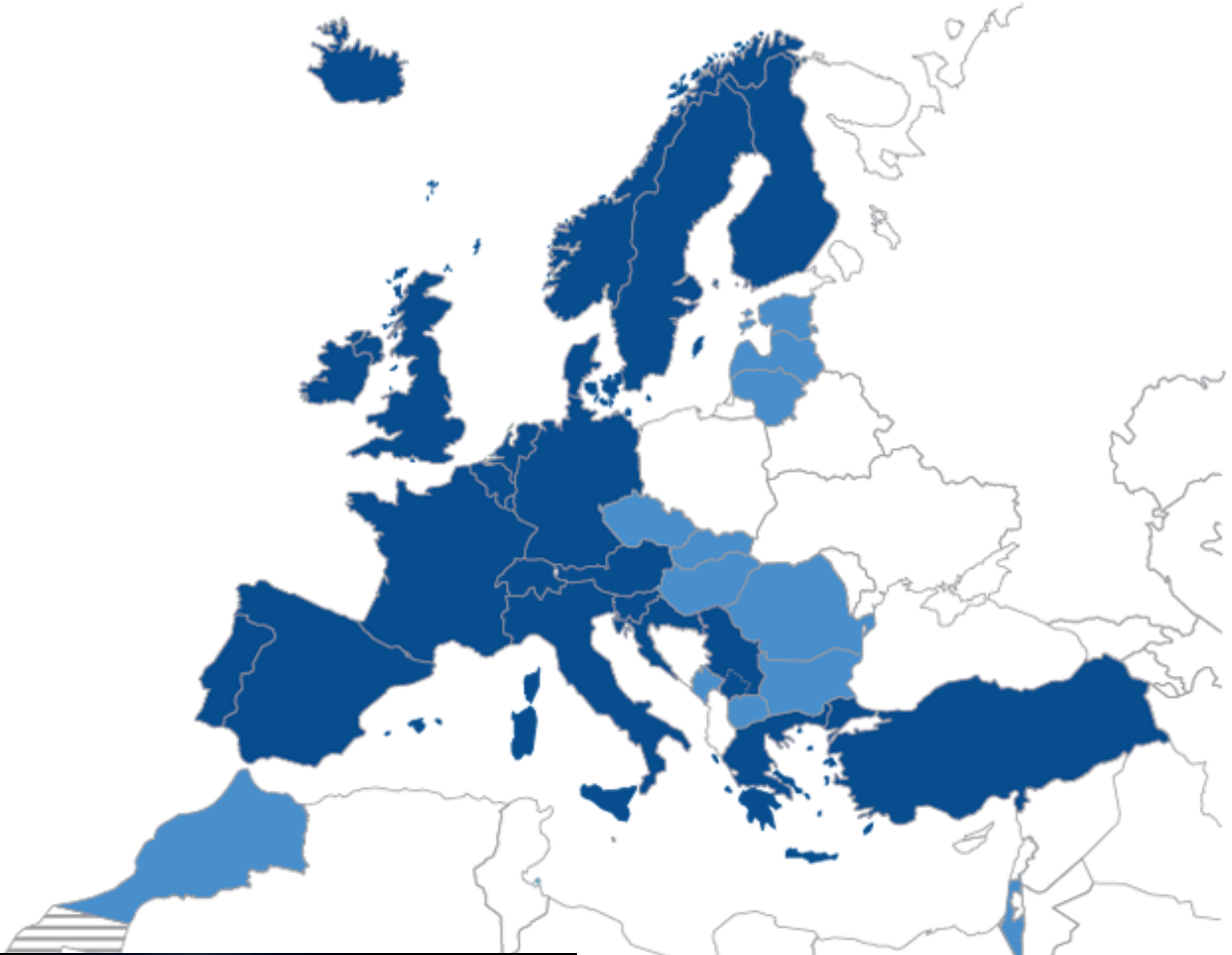
June 30, 1945



John von Neumann with first “electronic computing instrument” that was built at Princeton’s IAS between 1946 and 1952

Source: https://conifer.rhizome.org/mudd/turing/20180328150956/https://www.princeton.edu/turing//alan/history-of-computing-at-p/

European Center for Medium-Range Weather Forecasts



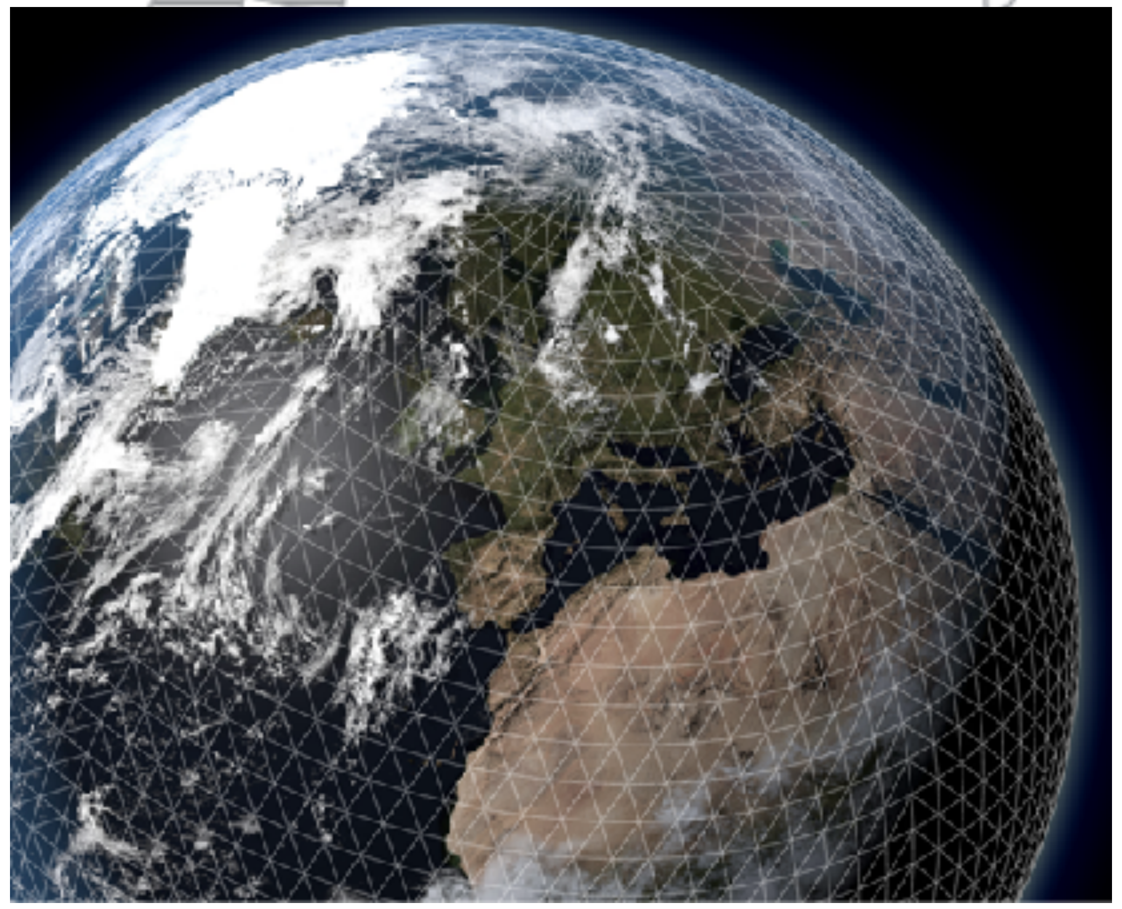
ECMWF

An independent intergovernmental organization established in 1975

Switzerland was founding member of ECMWF among 18 countries

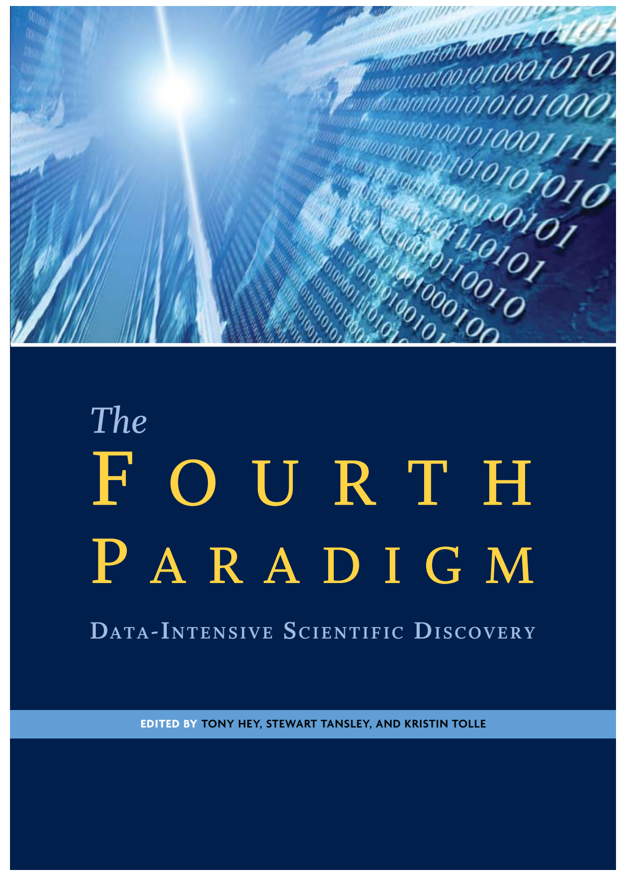
Today the worldwide leading numerical weather prediction center

Provides input data for the weather predictions of MeteoSwiss



adapted from Schär, ETH Zürich

All of the above + Velocity, Volume and Value

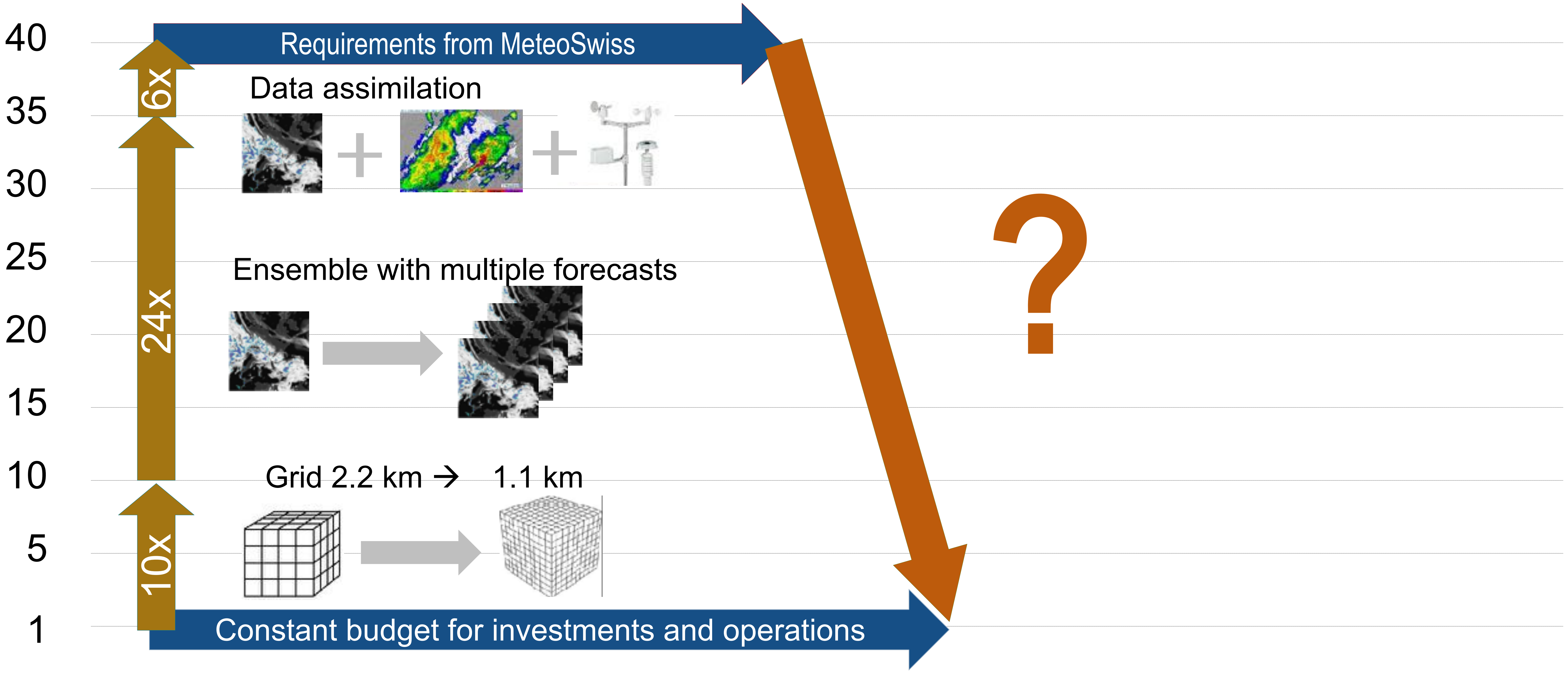


The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

MeteoSwiss' Performance Ambitions in 2013



Requirements from MeteoSwiss

Data assimilation

Ensemble with multiple forecasts

Grid 2.2 km → 1.1 km

Constant budget for investments and operations

We need a 40x improvement between 2012 and 2015 at constant cost

Porting codes to GPUs, Xeon, ARM, etc.

CUDA (C / C++ / Fortran)

```
__global__ void add_pw_ekin_gpu_kernel(int num_gvec__,
                                       double alpha__,
                                       double const* pw_ekin__,
                                       cuDoubleComplex const* phi__,
                                       cuDoubleComplex const* vphi__,
                                       cuDoubleComplex* hphi__)
{
    int ig = blockIdx.x * blockDim.x + threadIdx.x;
    if (ig < num_gvec__) {
        cuDoubleComplex z1 = cuCadd(vphi__[ig], make_cuDoubleComplex(alpha__
                                                                     alpha__
        hphi__[ig] = cuCadd(hphi__[ig], z1);
    }
}
```

OpenCL

```
__kernel void vector_add(const int n, __global float *a, __global float *b, __global float *c) {
  const int i = get_global_id(0);

                      + b[i];
```

OpenACC

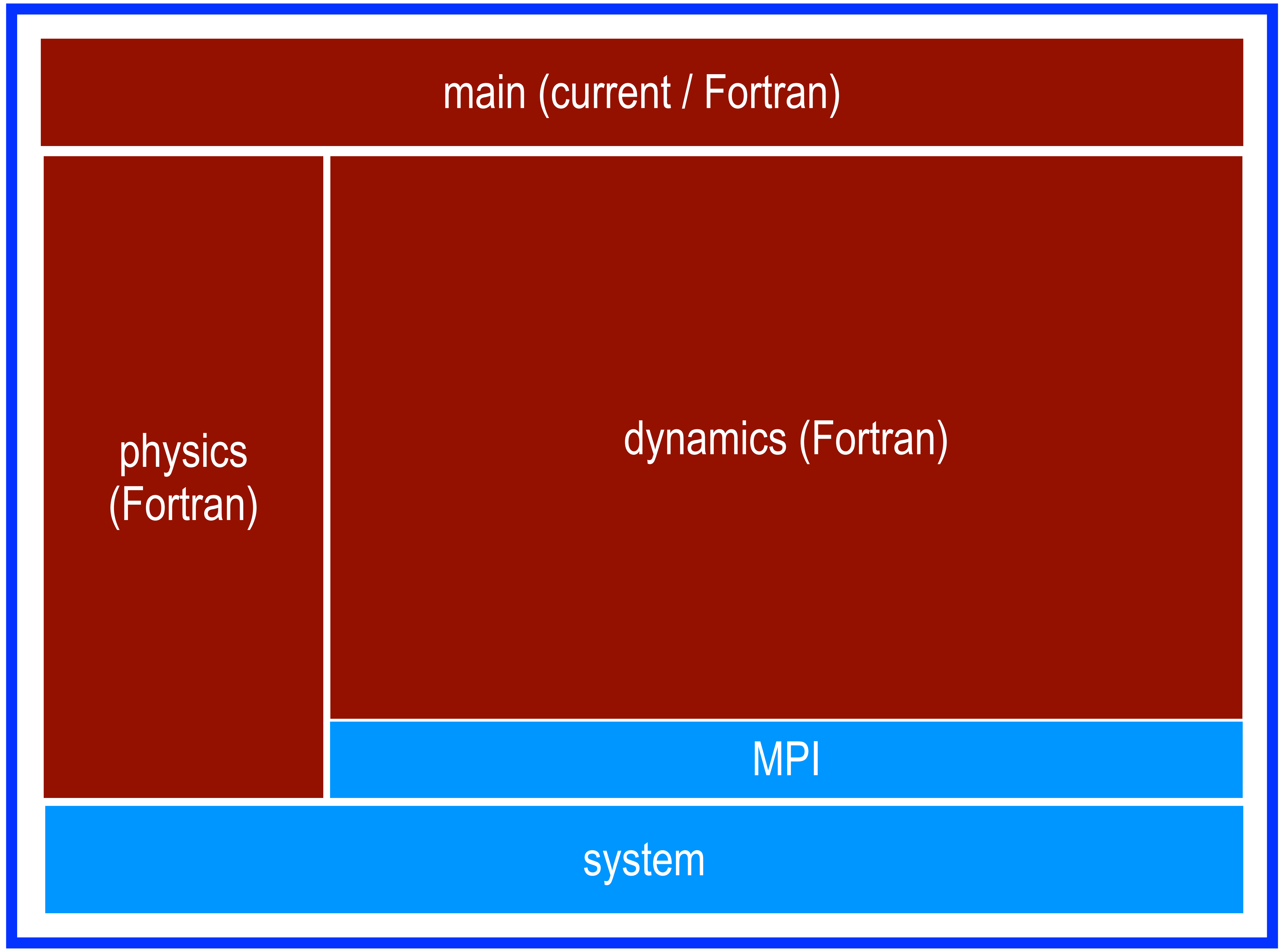
```
    acc = 0
    !$acc parallel present(x)
    !$acc loop reduction(+:acc)
    do i = 1, N
       acc = acc + x(i) * x(i)
    enddo
    !$acc end parallel
    call mpi_allreduce(acc, accglobal, 1, MPI_DOUBLE, MPI_SUM, MPI_COMM_WORLD, err)
```

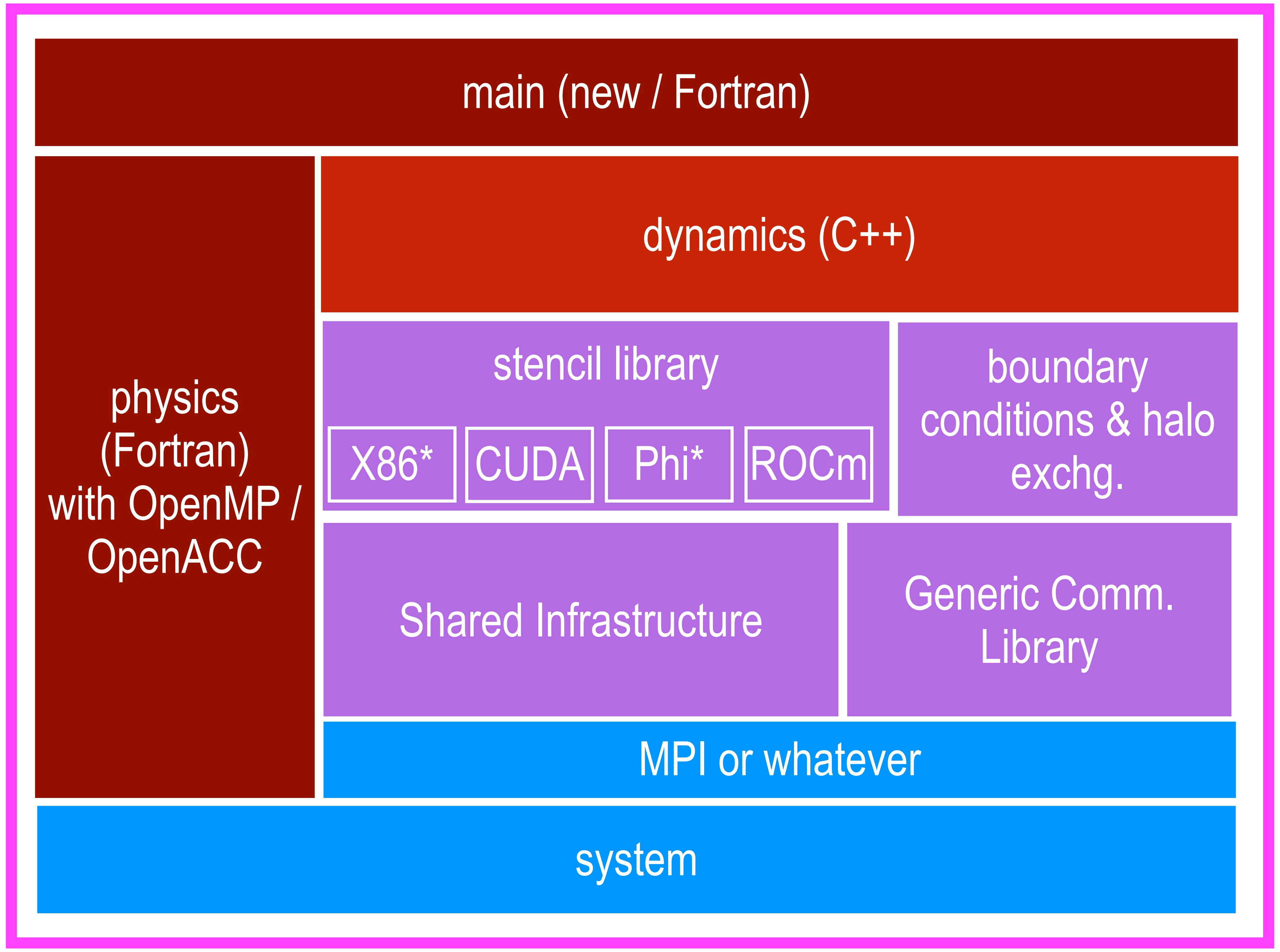
OpenMP 4.x

```
        omp target data map(tofrom: x[0:n],y[0:n])

    #pragma omp target
    #pragma omp for
    for (int i = 0; i < n; i++)
        y[i] += a * x[i];
}
```

COSMO: old and new (refactored) implementation

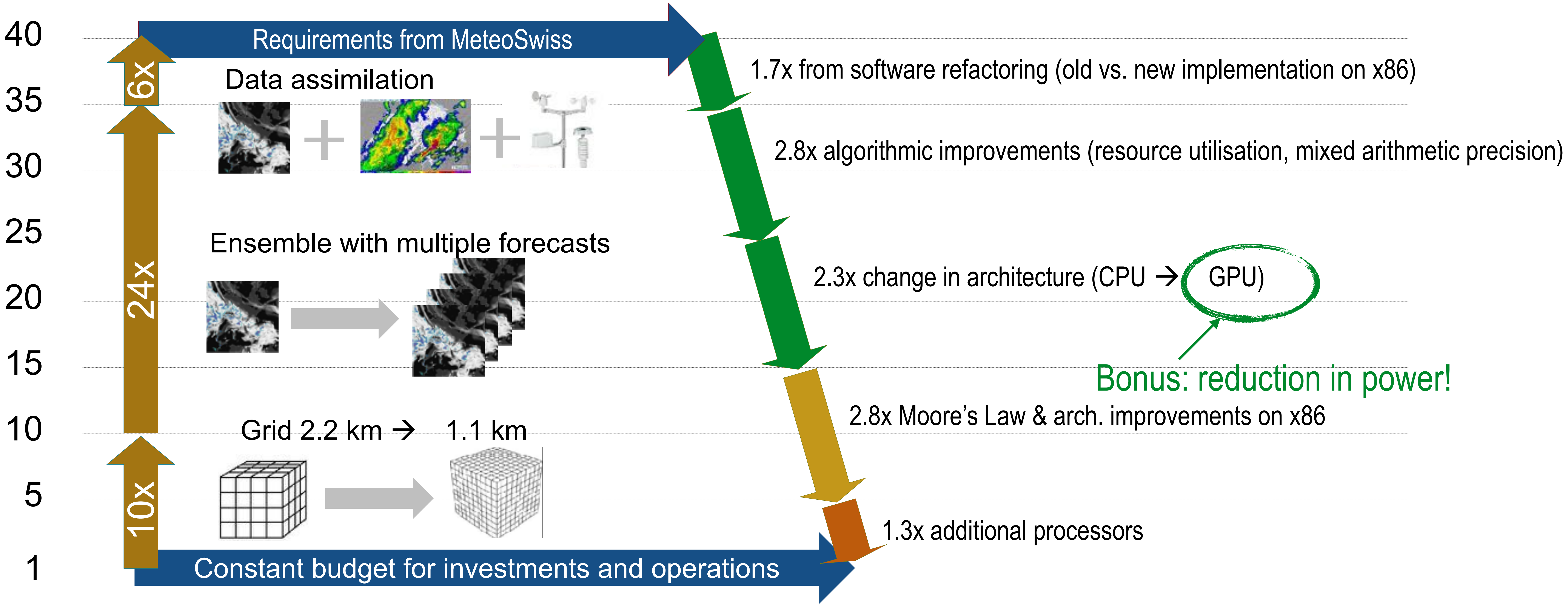




* two different OpenMP backends

Where the factor 40 improvement came from

Investment in software allowed mathematical improvements and change in architecture



Requirements from MeteoSwiss

Data assimilation

Ensemble with multiple forecasts

Grid 2.2 km → 1.1 km

Constant budget for investments and operations

1.7x from software refactoring (old vs. new implementation on x86)

2.8x algorithmic improvements (resource utilisation, mixed arithmetic precision)

2.3x change in architecture (CPU → GPU)

Bonus: reduction in power!

2.8x Moore's Law & arch. improvements on x86

1.3x additional processors

There is no silver bullet!

Setting a new baseline for atmospheric simulations

The state-of the art implementation of COSMO running at most weather services on multi-core hardware.

↓ **~10x**

The refactored version of COSMO running at MeteoSwiss on multi-core or GPU accelerated hardware.


Platin Sponsor
Swiss ICT Award 2015
SIX
Gold Sponsoren
Swiss ICT Award 2015
winterthur
trivadis
swiss ICT

COSMO at Δx = 50 km
189 x 142 x 60 = 1.6×10^6 grid points

COSMO at Δx = 3 km
2750 x 2065 x 60 = 340×10^6 grid points

Christoph Heim, ETH Zürich

“Exascale:” our goal for 2024-2026 climate applications runs

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)

Schulthess, P. Bauer, N. Wedi, O. Fuhrer, Th. Hoefler, Ch. Schär, Comp. Sci. Eng. 21 (1), 31-40 (2018)

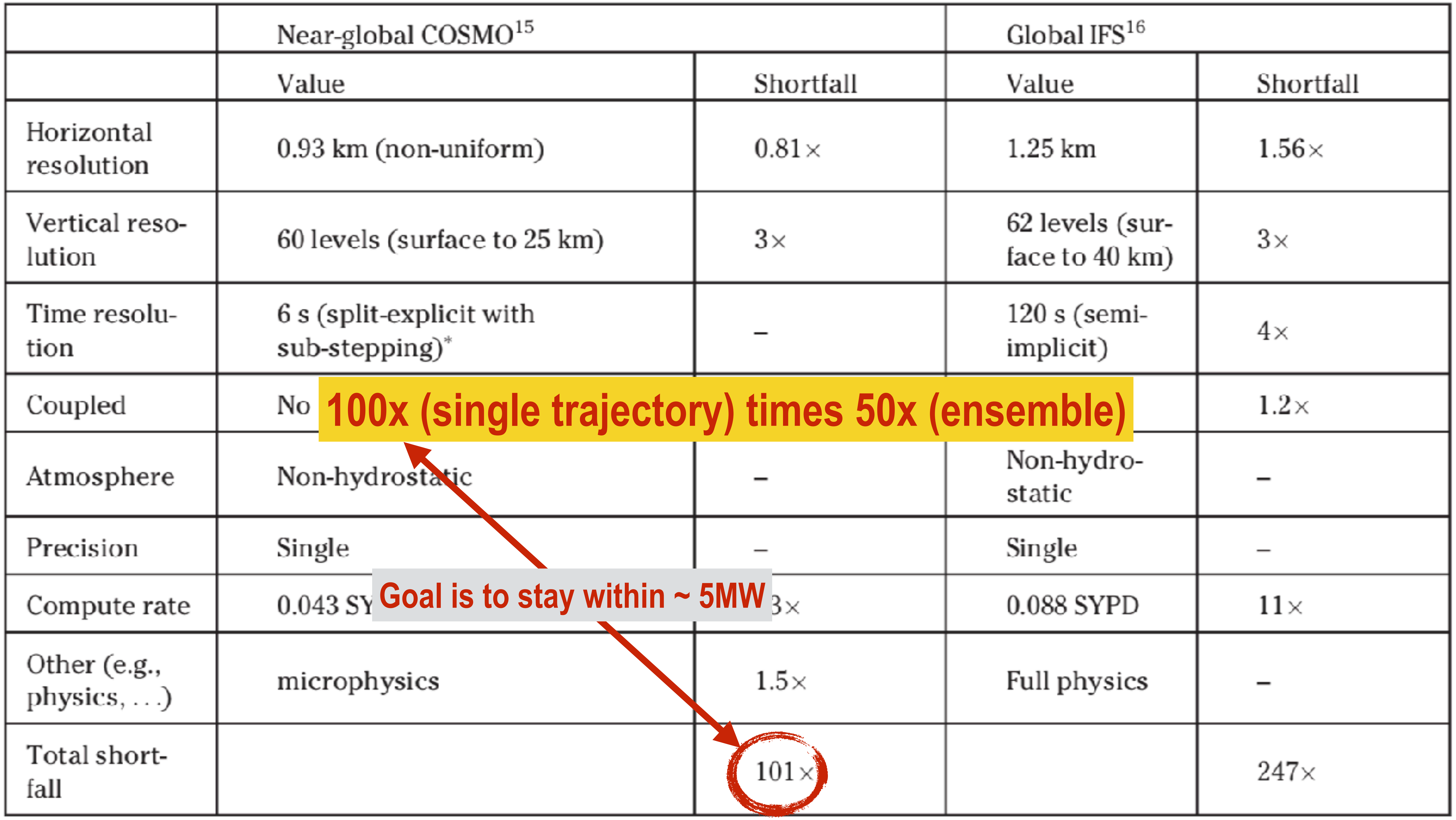
Baseline in 2018: Running COSMO & IFS (“the European Model”) at global scale on “Piz Daint”

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 near-global and IFS



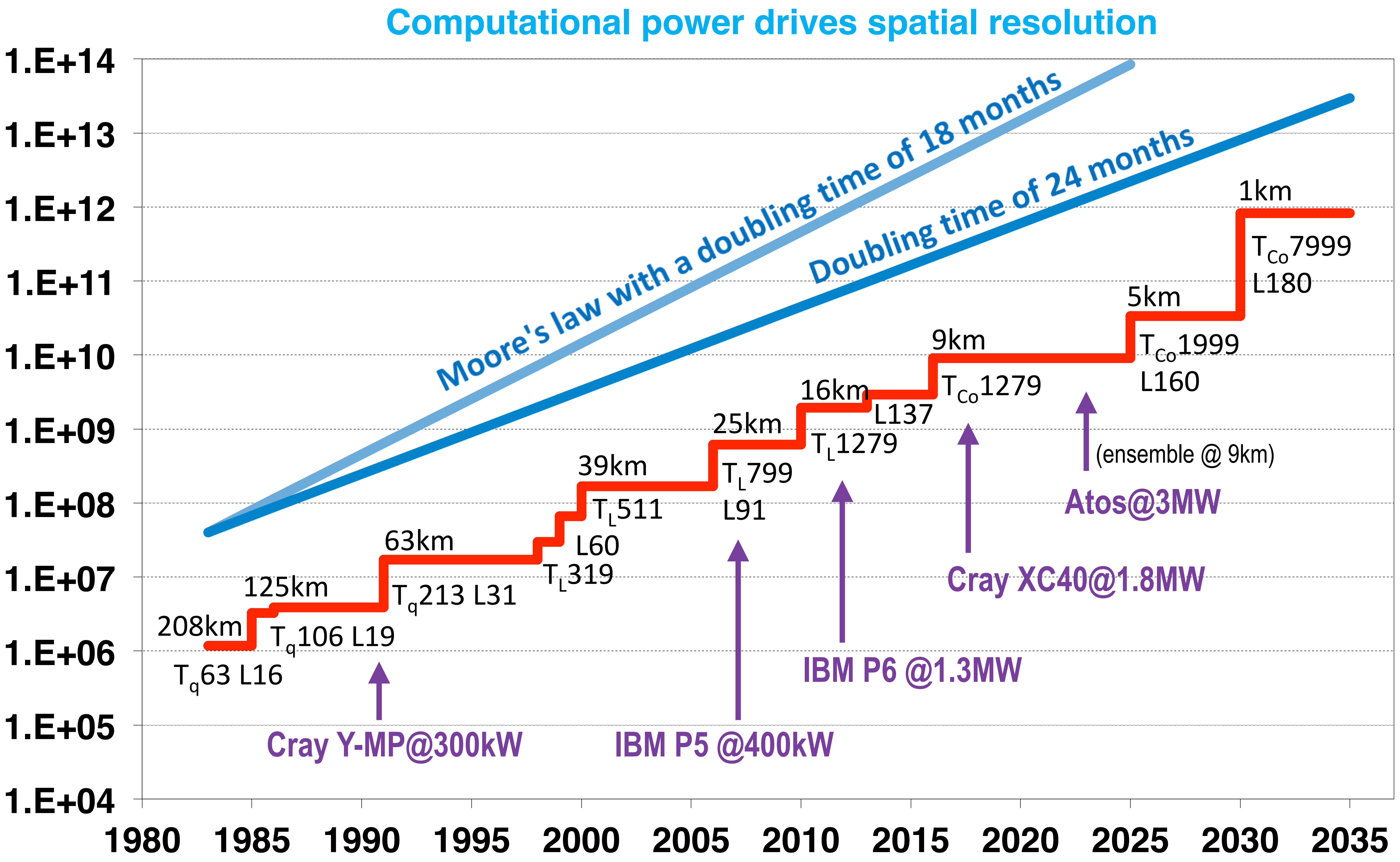
	Near-global COSMO[15]		Global IFS[16]	
	Value	Shortfall	Value	Shortfall
Horizontal resolution	0.93 km (non-uniform)	0.81×	1.25 km	1.56×
Vertical resolution	60 levels (surface to 25 km)	3×	62 levels (surface to 40 km)	3×
Time resolution	6 s (split-explicit with sub-stepping)*	–	120 s (semi-implicit)	4×
Coupled	No			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

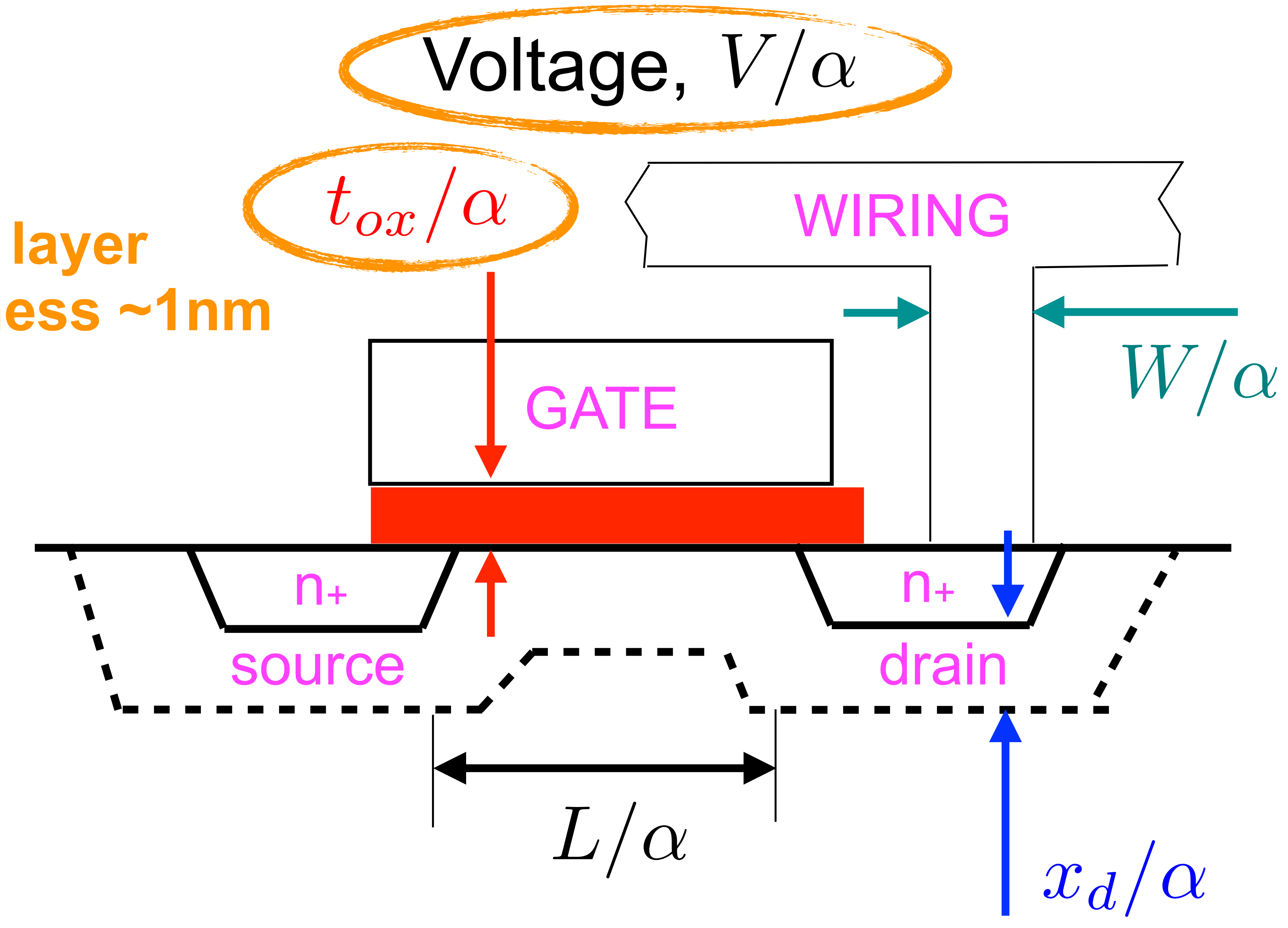
Schulthess et al., Comp. Sci. Eng. 21 (1), 31-40 (2018)

Evolution of computing system and model capability at ECMWF



The end of Dennard Scaling

Robert H. Dennard (1974)



Oxide layer thickness ~1nm

p substrate, doping $\alpha^* N_A$

SCALING

~~Voltage: V/α~~

~~Oxide: t_{ox}/α~~

Wire width: W/α

Gate Width: L/α

Diffusion: x_d/α

Substrate: $\alpha^* N_A$

CONSEQUENCE:

Higher density: $\sim \alpha^2$

Higher speed: $\sim \alpha$

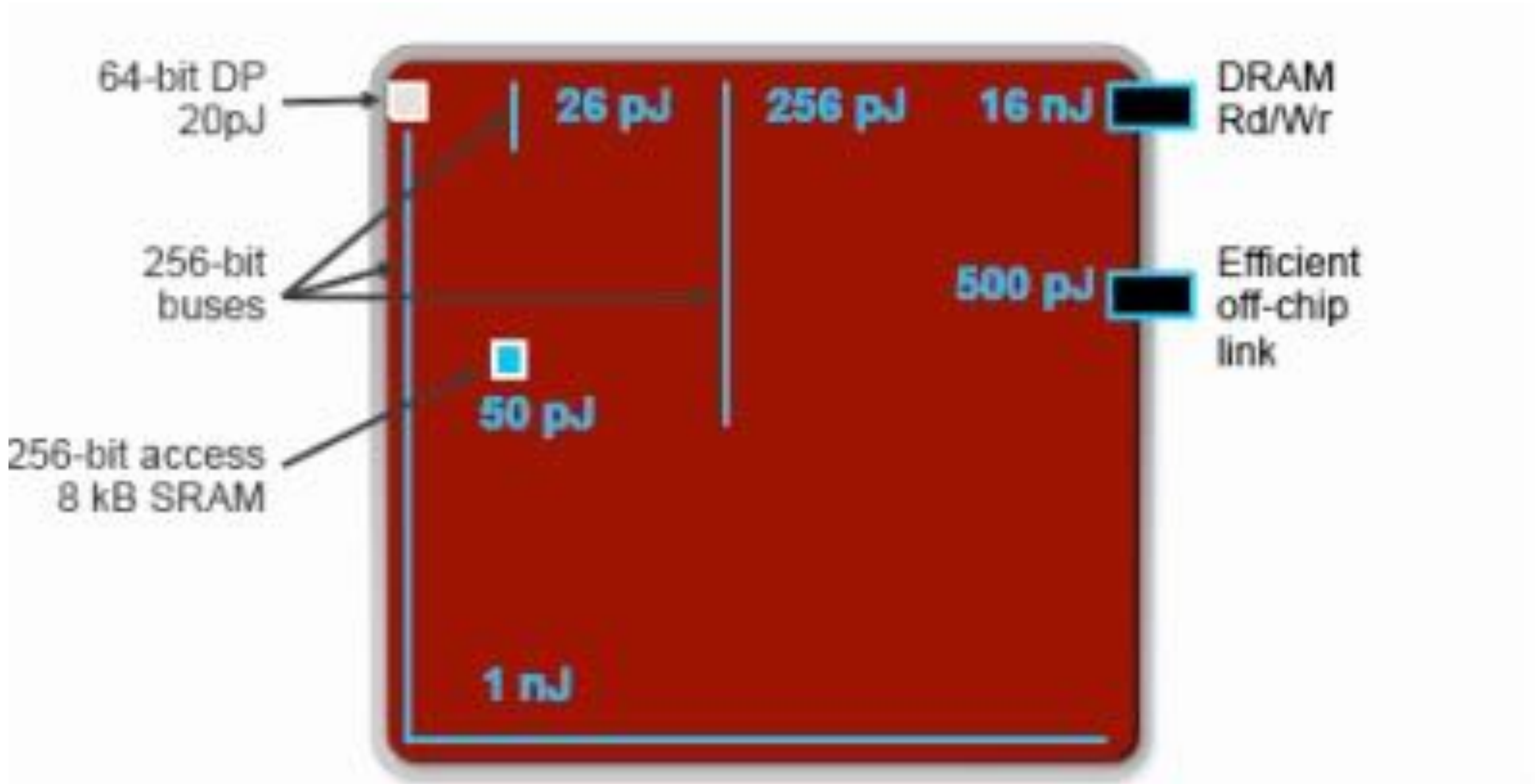
Power/ckt: $\sim 1/\alpha^2$

~~Power density: $\sim$ constant~~

Source: Ronald Luijten, IBM-ZRL ret.

Who consumes how much energy (28nm)

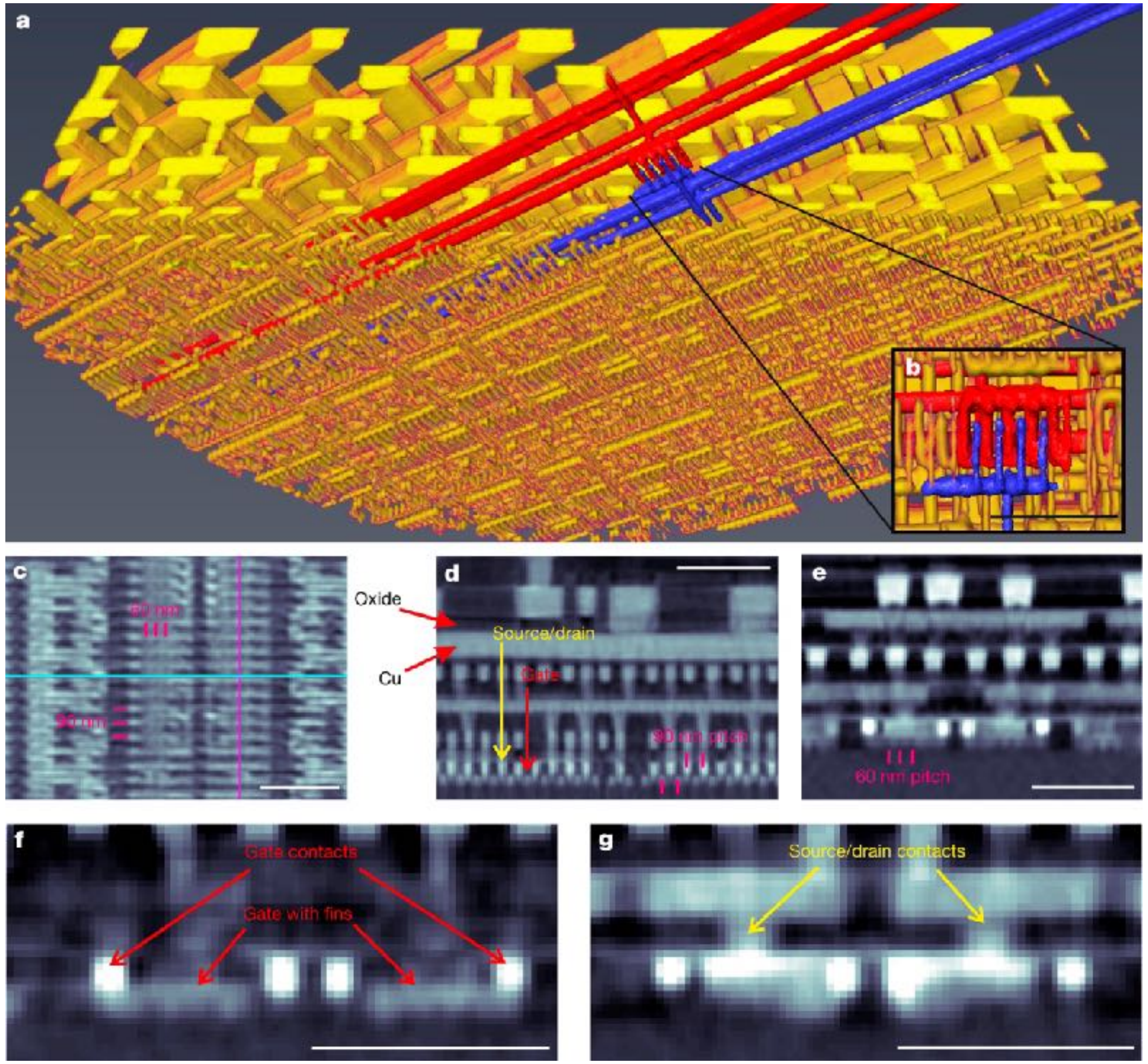
- 64 bit floating point unit: 20 pJ
- 256-bit access 8kB SRAM: 50 pJ
- 256-bit bus across die: 1,000 pJ
- Read/write to DRAM: 16,000 pJ



Source: Bill Dally, 2011

By a wide margin, most energy is spend in moving data on the die and to memory

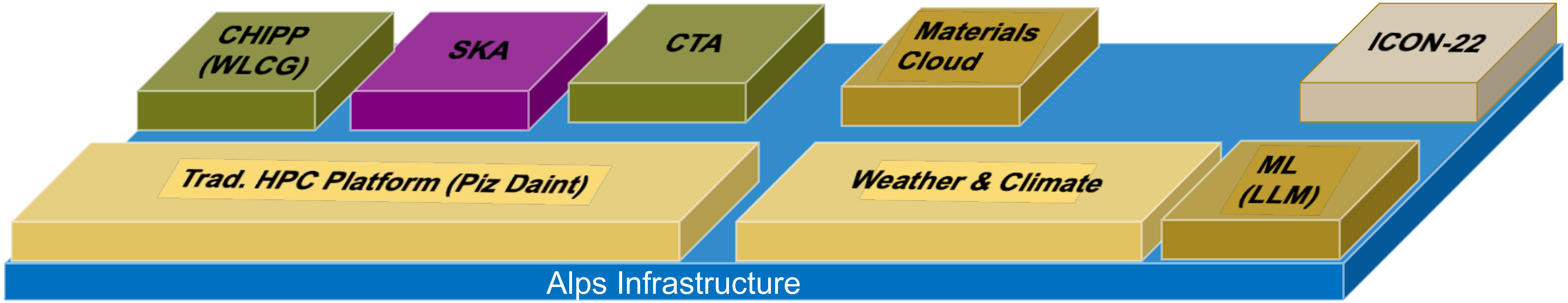
Developing algorithms that maximise data locality should be THE TOP PRIORITY



M Holler *et al.* *Nature* **543,** 402–406 (2017) doi:10.1038/nature21698

“Piz Daint” in the “Alps” Infrastructure

To a particular community, a platform will look like a dedicated supercomputer



CHIPP (WLCG)

SKA

CTA

Materials Cloud

ICON-22

Trad. HPC Platform (Piz Daint)

Weather & Climate

ML (LLM)

Alps Infrastructure

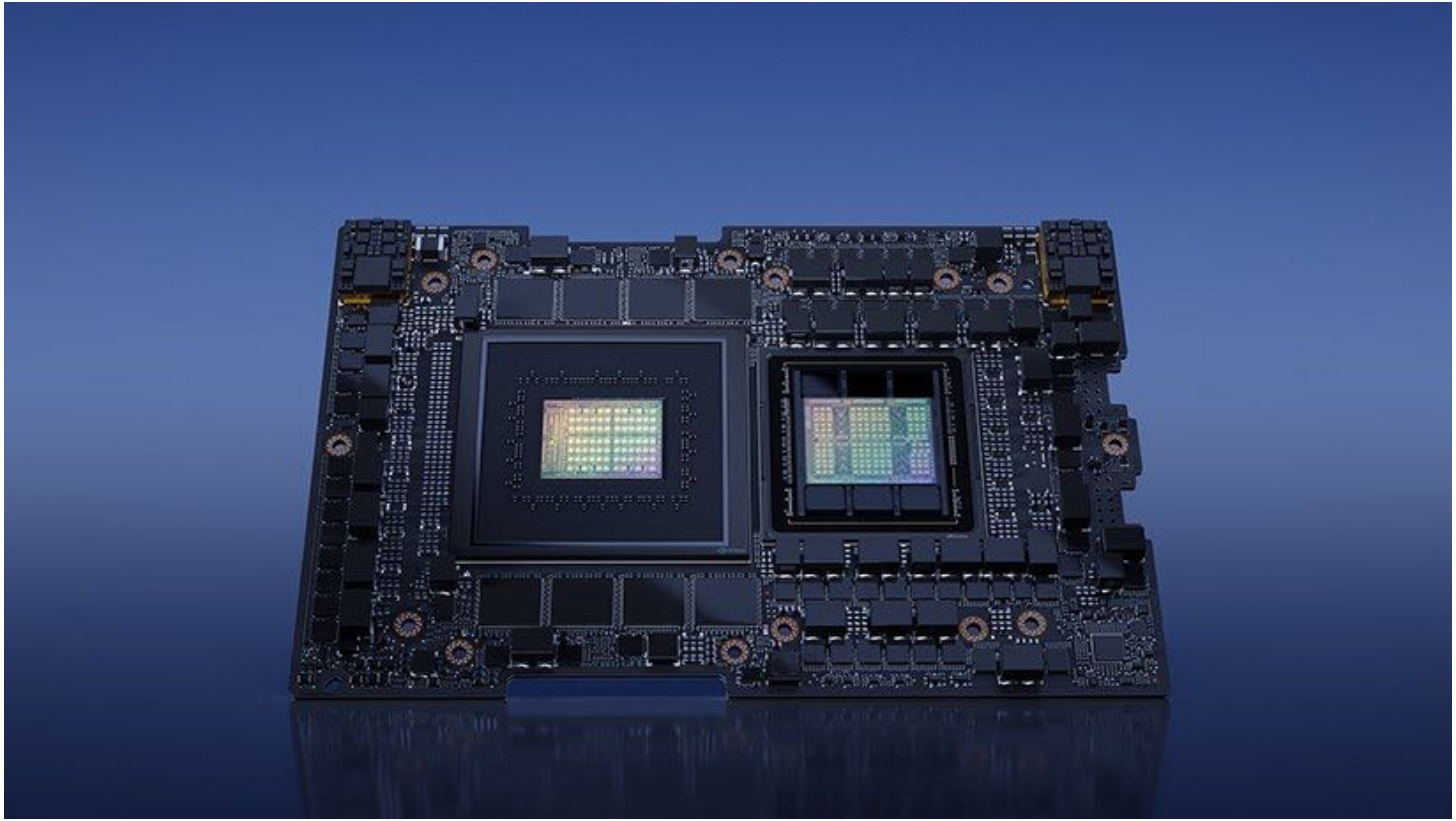
vClusters

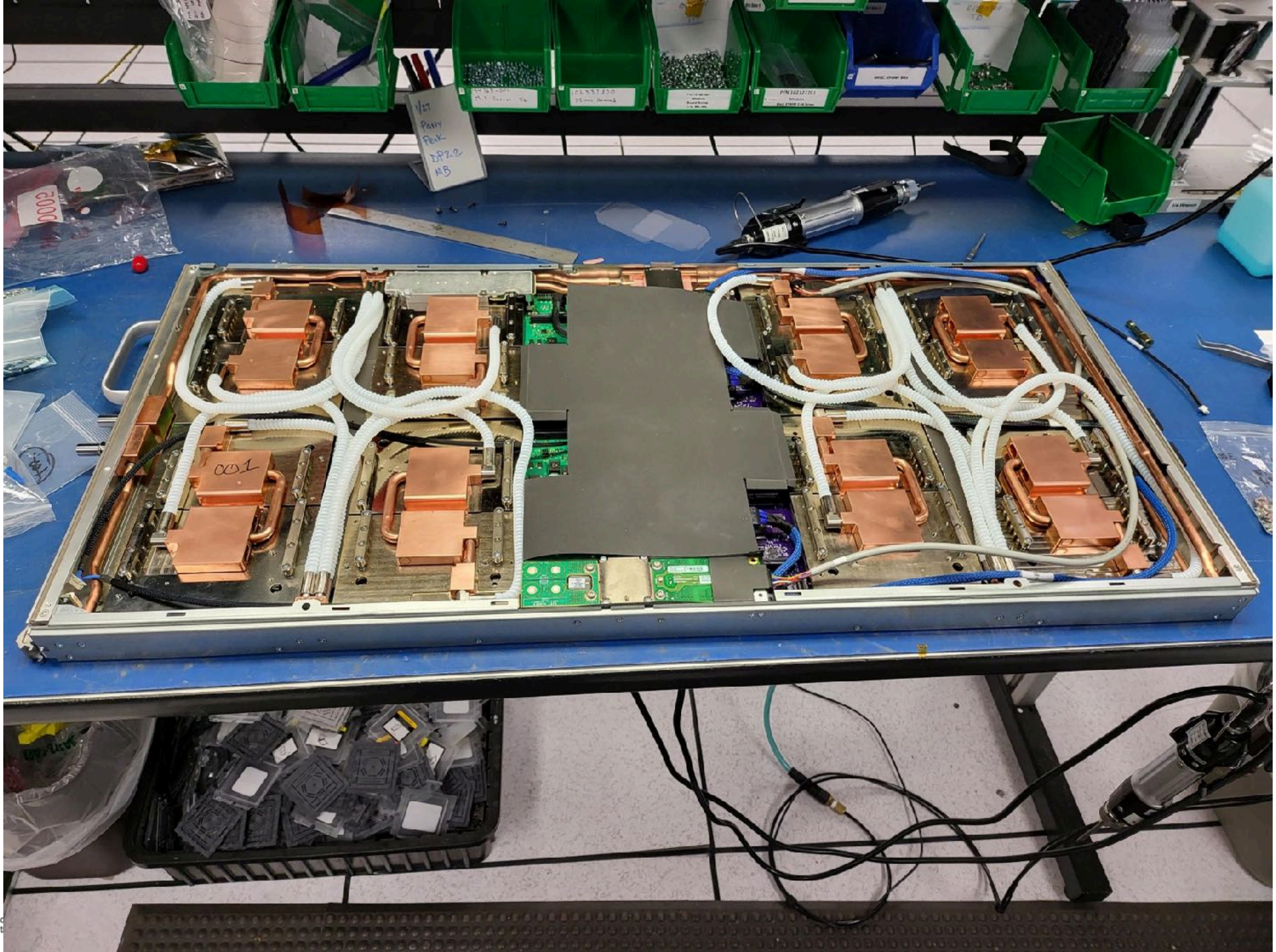
FIRECREST

SARUS

Slingshot network

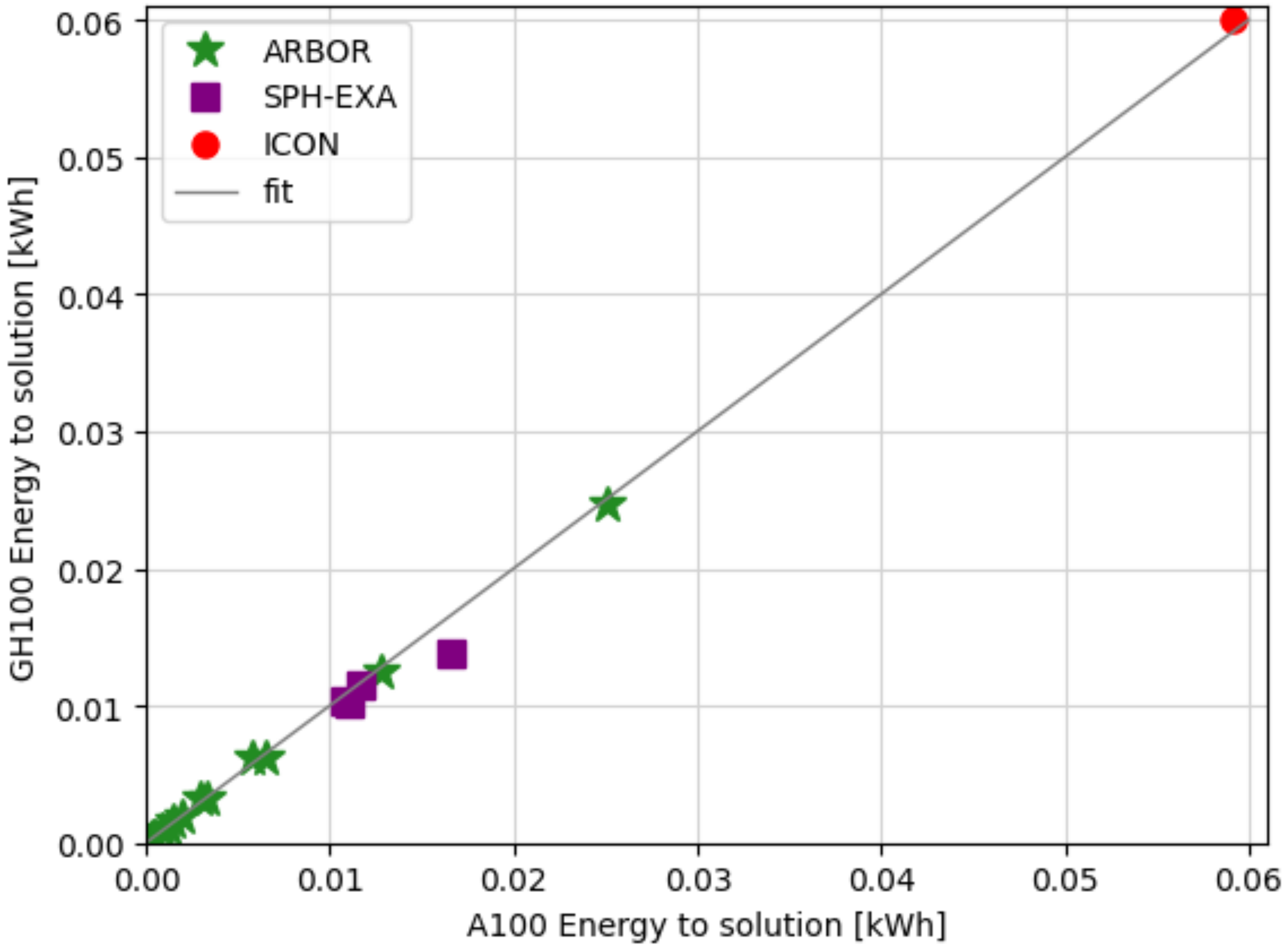
Cray System Management software (μ-service arch.)




001

Preliminary comparison A100 vs. GH200(*)

(*) A02 engineering samples



ICON benchmark

	time (s)	power (W)
A100	2196	388
GH200	1518	570
	1.45	1.47

B. Cumming and W. Sawyer (CSCS)

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Baseline running on "Piz Daint"

	Near-global (COSMO)		ICON global	
	Value	Shortfall	Value	Shortfall
Horizontal resolution	0.93 km (non uniform)	0.81x	5 km (uniform)	25x
Vertical resolution	60 levels	3x	90 levels	2x
Time resolution	6s (split-explicit with sub-stepping)	—	40s (split-explicit with sub stepping)	5x
Couple	No	1.2x	No	1.2x
Atmosphere	Non-hydrostatic	—	Non-hydrostatic	—
Precision	Single	—	?	—
Simulation rate	0.043 SYPD	23x	0.4 SYPD	2.5x
Other (e.g. physics, …)	Microphysics	1.5x	Full physics	—
Adjusted to 5300 nodes	5300 nodes	—	1000 nodes	0.19x
		101		190

Conclusions

- Science can greatly benefit from fully embracing digitalisation
- Moore’s Law is fading: computing and data research infrastructures have their cost
- A multitude of computer architectures is the consequence
- Continued investments in algorithms and software development is essential
- Software engineering has to become a first class citizen

Thank you to CSCS, partners such as MeteoSwiss, HPE/Cray, NVIDIA, as well as many colleagues and collaborators

Tim Palmer (U. of Oxford)

Bjorn Stevens (MPI-M)

Peter Bauer (ECMWF)

Oliver Fuhrer (MeteoSwiss)

Nils Wedi (ECMWF)

Sadaf Alam (U. of Bristol)

Torsten Hoefler (ETH Zurich)

Christoph Schar (ETH Zurich)

Thank you