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

Published by the IEEE Computer Society

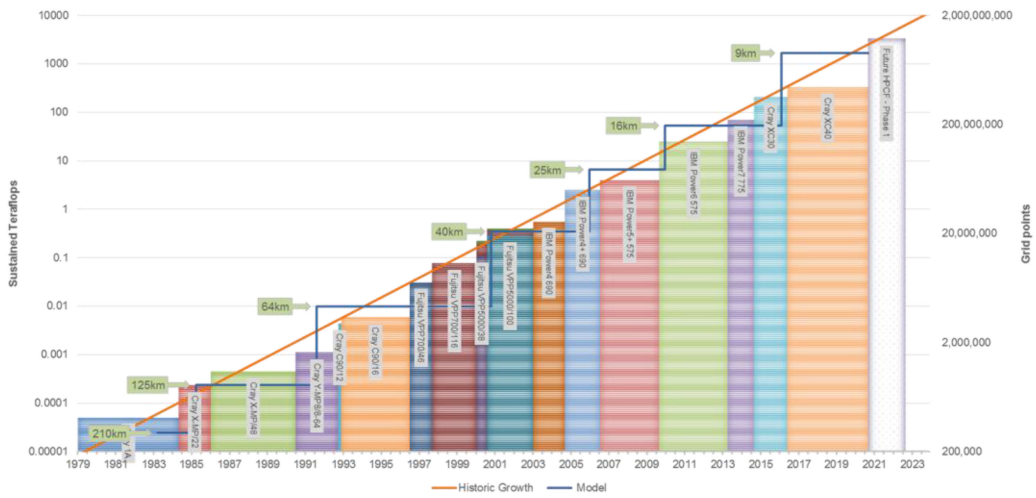

Figure 1. Growth of operational numerical weather prediction (NWP) model sustained performance achieved at ECMWF since 1979 (bars). Orange curve denotes linear (in log-space) fit to sustained performance; blue curve denotes corresponding spatial resolution upgrades translated to model grid point increase (courtesy M Hawkins, ECMWF).

derive the laws of planetary orbits that carry his name. This illustrates why scientists and the general public have been mesmerized by the performance of electronic computing machines in executing floating point operations. The fascination started soon after John von Neumann wrote his seminal report on the developments of Eckert and Mauchly. He made early supercomputers known to physicists and applied mathematicians working at numerous laboratories in the USA during the years after World War II. The invention of the transistor followed by the development of the complementary metal oxide semiconductor (CMOS) fabrication technology paves the way to modern supercomputing, giving modeling and simulations a seminal role in almost all areas of science.

SUSTAINED EXPONENTIAL GROWTH: PLANNING EXASCALE COMPUTING

CMOS allowed performance of microelectronic circuits to grow exponentially at constant cost. Known as Moore's Law, this development persisted for many decades and only recently began to taper off. Interestingly, from the late 1980 to about ten years ago, the performance of supercomputers measured in floating point operations per second (flop/s) grew faster than Moore's Law. This is, for example, documented with the sustained performance of first principles electronic structure codes used in materials science.

In 1990, a team of material scientists from Oak Ridge National Laboratory (ORNL) received a Cray Gigaflops Performance Award for sustaining 1.5 gigaflop/s on an 8-(vector)processor Cray Y-MP in a first principles electronic structure calculation based on multiple scattering theory.[1] During the 1990s, the same ORNL team developed a linearly scalable multiple scattering (LSMS) algorithm[1] that would map onto massively parallel processing arrays and sustained a teraflop/s (1000 gigaflop/s) in 1998 on a 1500 processor Cray T3E.[2] In 2009, a simulation based on the LSMS algorithm sustained more than 1 petaflop/s (one million gigaflop/s).[3]

Thus, in just two decades from 1990 to 2009, we have seen a million-fold performance increase in electronic structure computations of materials. This immediately benefited the simulation of nano-scale electronic devices. But what about other domains of science?

The European Centre for Medium-Range Weather Forecasts (ECMWF) has been tracking the performance development of the IFS model (www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation) on production supercomputers since the late seventies. When measured in sustained floating-point operations, the capability

of these simulations increased 100-fold per decade as shown in Figure 1 — a factor 10 less than simulation performance grew in materials science. While the architecture of these systems has been optimized differently than the supercomputers at ORNL, the supercomputer systems in Figure 1 nevertheless serve as calibration points for a direct comparison. One is thus led to wonder whether the efficiency of simulations used for NWPs decreased by an order of magnitude every decade since the late 1980s. In fact, when comparing relative performance for the IFS model over the past decades, 40%–50% of peak was sustained on the 1980s Cray X-MP and Y-MP while this rate dropped down to 5% on today's multicore Cray XC systems. This shows that any analysis centered around floating-point performance alone is too simplistic.

Another important aspect that often gets overlooked in the discussion of performance developments of the fastest supercomputers is their cost, which is well illustrated in terms of the energy footprint of the systems. The Cray Y-MP supercomputer that sustained a gigaflop/s on a material science computation in 1990 consumed about 300 kW. Jaguar, the Cray XT5 system at ORNL that sustained a petaflop/s in 2009 on a similar computation example consumed 7 MW. The million-fold increase in performance of these material science computations over 20 years includes a factor 23 in energy footprint. The energy footprint of the systems at ECMWF increased "only" by a factor 6 over the same time period.

This sobering expansion of energy footprint represents a significant increase in operations cost. The trend towards larger systems will likely continue with a global race in supercomputing that will see the deployment of first peak exaflop/s supercomputers consuming 30–40 MW in the early 2020s. While in some parts of the world the race towards exaflop/s-scale computing has strategic value that can justify almost any cost, the supercomputers used today in European open science have to earn their purpose in the context of the broader scientific enterprise in competition with other research infrastructures. Hence, the scientific performance and costs play an important role in determining whether a research infrastructure will be built and receives adequate support for operations (https://www.symmetrymagazine.org/article/the-value-of-basic-research).

Recently, the European Commission and (at time of this writing) 22 countries in Europe have declared their intent to join efforts in developing HPC technologies and infrastructures. While the goals and precise content of this initiative called EuroHPC (see https://ec.europa.eu/digital-single-market/en/blogposts/eurohpc-joint-undertaking-looking-ahead-2019-2020-and-beyond) are still being defined, a short-term ambition to deploy two "pre-exascale" systems by 2020, supercomputers with several hundred petaflop/s peak performance, is firm. It is thus safe to assume that, similar to other regions, Europe will have an exaflop/s-scale computing program in the 2020s. But it is equally clear that Europe will continue to have strong bottom-up programs in HPC that are funded by national initiatives, such as the German Gauss Centre or the Swiss High-Performance Computing and Networking (HPCN) initiative. Furthermore, Europe is playing a leading role in application development for extreme-scale computing. For example, ECMWF's IFS model, which in the US media is commonly dubbed the "European Model," attracts a lot of interest during the Hurricane season. Europe occupies the lead in global medium-range weather forecasting since 40 years (see https://mashable.com/2017/09/14/hurricane-irma-weather-forecast-models-gfs-vs-european/#iG20.FcLMOq3in).

Rather than developing a plan for European exaflop/s-capable supercomputing, our contribution to this special issue on the "Race to Exascale Computing" is to lay out an ambitious roadmap for developing next-generation extreme-scale computing systems for weather and climate simulations. Choosing a single domain is comparable to the approach IBM took with the development of the Blue Gene line of supercomputers that were deployed between 2004 and 2012. The resulting architecture will be usable in many domains but focusing on one enables us to set clear goals for ambitious design targets and to define consistent success metrics. Here, we begin by discussing the science goals and the implied performance targets for simulation rates.

In order to define an implementation roadmap, we study the baseline for running these types of simulations today on some of the

most performant petaflop/s-scale supercomputers that were available in 2016–2017. This will give us a clear picture of the shortfall of today's systems for reaching our science goals. We discuss a performance metric and analysis of these simulations and show that the purely floating-point-based analysis commonly used in supercomputing is not adequate for the problem at hand. Instead we propose a roadmap based on an analysis of data movements for the implementation of weather and climate models. This analysis is also representative of other simulation domains that rely on grid-based solvers of partial differential equations, such as seismology and geophysics, combustion and fluid dynamics in general, solid mechanics, materials science, or quantum chromodynamics. Thus, we argue that our techniques naturally address the most important performance challenges of modern computing.

AMBITIOUS GOALS FOR WEATHER AND CLIMATE

Over the past decades, weather prediction has reached unprecedented skill levels. For the medium range of five to ten days, there is consensus that the predictive skill of forecasting low-pressure systems and fronts, but also surface temperature and wind as well as precipitation has improved by about one day per decade.[4] This means for example that today's five-day forecast is as accurate as the four-day forecast a decade ago. This skill improvement is remarkable as it allows the reliable prediction of weather extremes, enabling governments, public and private sectors to plan and protect, and making billions of dollars of savings every year.

This skill trend is the combined effect of improvements in the representation of dynamical and physical processes, the more accurate provision of initial conditions, and significant investments in supercomputing. Today's top prediction centers run simulations at 10–15 km spatial resolution globally. Since its foundation over 40 years ago, the ECMWF has led global medium-range prediction, which is the result of centralizing national intellectual and financial resources from 34 member and cooperating states (www.ecmwf.int). Its integrated HPC center, of which 50% is allocated to research, has contributed to the effective transition of new methodologies and technologies into operations.

Climate and weather models use similar atmospheric components, but the consideration of extended time periods implies consideration of the ocean, sea-ice, and land-surface components. In addition, the impacts of natural and anthropogenic drivers, such as volcanic eruptions, variations in solar luminosity, as well as emissions of greenhouse gases and aerosols need to be accounted for, together with an improved representation of the energy, water and biogeochemical cycles. Sufficient throughput rates can only be realized in a cost-effective manner through a decrease in spatial resolution. While the most recent generation of global climate simulations (CMIP5) had an average resolution of about 200 km,[5] the next generation (CMIP6) is expected to reach 25–100 km, but it is not evident whether this will resolve the key uncertainties in climate projections.

The recent IPCC report shows overwhelming evidence that anthropogenic climate change is already happening, yet projections remain difficult and uncertain. Improved climate models are urgently needed both for mitigation and adaptation purposes, in order to reduce global warming and protect against its impacts. Current uncertainties are phrased in terms of the equilibrium climate sensitivity (ECS), which is the equilibrium global-mean surface warming from a doubling of atmospheric CO_2 concentrations. In the influential Charney report[6] of 1979, the ECS was estimated to be between 1.5 and 4.5 K. This wide uncertainty range has neither significantly shifted nor narrowed since then.[7] How should the Paris 2 °C target be achieved in the midst of this uncertainty?

Among the main causes behind these large uncertainties is the representation of clouds in climate models,[8,9] especially convective clouds (i.e., thunderstorms, rain showers, and shallow maritime clouds). Clouds may act as a positive or negative feedback to anthropogenic greenhouse gas emissions, depending upon whether they reflect more or less sunlight as the climate warms.

Both, weather and climate communities have the common long-term goal of pushing the horizontal resolution of simulations to scales of ∼100 m at which ocean eddies are resolved, and

Table 1. Ambitious target configuration for global weather and climate simulations with km-scale horizontal resolution accounting for physical Earth-system processes, and with today's computational throughput rate.

Horizontal resolution	1 km (globally quasi-uniform)
Vertical resolution	180 levels (surface to ~100 km)
Time resolution	0.5 min
Coupled	Land-surface/ocean/ocean-waves/sea-ice
Atmosphere	Non-hydrostatic
Precision	Single or mixed precision
Compute rate	1 SYPD (simulated years per wall-clock day)

the dynamics of convective clouds can largely be resolved based on first principles. But this goal is far out (probably a factor 100 000 or more in computational cost) from what today's simulations can accomplish. However, there is an intermediate goal at a horizontal resolution around 1 km. Experience from limited-area models indicates that at this scale convective systems reach "bulk" convergence,[10] meaning that the feedbacks may become adequately represented. Models of this resolution can currently be run for weeks to months on global scales,[11,12] or for decades on continental scales,[13,14] but global climate change simulations over decades would require a speedup of at least a factor 100 to reach the required timescale in the same wall-clock time.

Notwithstanding that there will be remaining uncertainties in the models, aiming for global km-scale frontier weather and climate simulations now will trigger a science and technology development that will lead to much more rapid improvement of todays' weather and climate models than would otherwise be achieved with incremental improvements of today's simulation systems. Thus, while the technology development for extreme-scale computing in weather and climate should not lose sight of the long-term goal with ~100 m resolutions, it seems like computing capabilities that allow global simulations at 1 km with reasonable throughput would represent a significant step towards making a qualitative difference, and we propose this as a goal post for developments of exascale computing systems in the coming decade. A more precise definition of this goal is given in Table 1.

Before we discuss whether this goal is achievable we revisit the current approach of adopting conventional flop/s-centric supercomputing systems to weather and climate. In Figure 2, we plot the advancement in simulation capability over the decades at ECMWF. These are the same simulations and computing systems used for the data in Figure 1, but rather than reporting the sustained flop/s, we report the increase of atmospheric degrees of freedom of the simulation. This measures the development of the computational complexity needed for a 10-day forecast (single forecast) over the decades.

From this plot one can conclude that general purpose supercomputing platforms that have been developed with a flop/s-centric metric, simply adopted for the purpose of weather and climate simulations, will deliver systems fit for simulations with 1 km horizontal resolution in the mid to late 2030s. Not a very encouraging perspective, in particular in view of additional challenges due to increased complexity in the description of the Earth-system.

BASELINE, WHERE WE ARE TODAY

Only a handful of global km-scale landmark simulations have already been performed.[11,12] In this section, we will focus on simulations executed with the COSMO model (www.cosmo-model.org) and the IFS model in order to establish a baseline of what can be achieved today on some of the largest supercomputers available. These simulations will serve as a baseline to estimate how far away we are from the exascale goal put forward in the previous section.

As already pointed out, the IFS model is a state-of-the-art model used for global weather and climate simulations. In contrast, COSMO is a regional weather and climate model that cannot readily be used for global simulations. But COSMO is, to our best knowledge, currently the only weather and climate model that has been systematically adapted for hybrid, GPU-accelerated HPC architectures,[15] and that is used

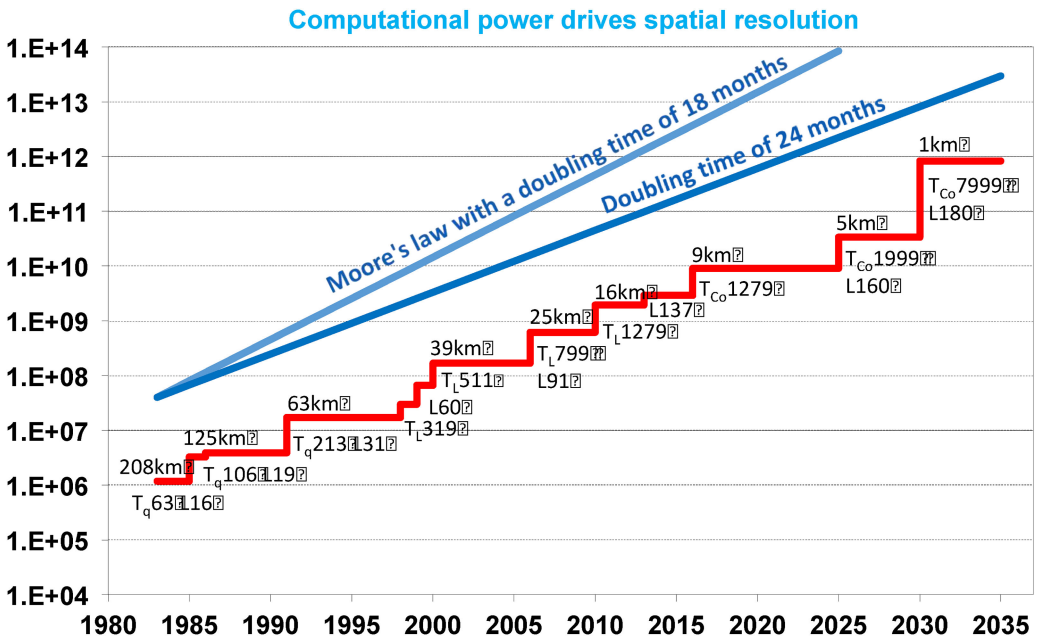

Figure 2. "Understanding grows only logarithmically with the number of floating point operations." (John P. Boyd). The progress in the degrees of freedom (vertical levels × grid columns × prognostic variables) of the ECMWF operational global atmosphere model in comparison to Moore's law, adjusted by ($P^{3/4}$) to account for the change in time-resolution also required in a three-dimensional dynamical system when increasing spatial resolution. ECMWF's actual progress up to 2018 doubles performance every 24 month. As illustrated, the ambitious goal of reaching a 1-km horizontal resolution with 180 levels in 2030 requires a faster progress. The numbers indicate the average grid-point distance in kilometers and the corresponding spectral resolution and levels used.

operationally for both weather and climate production simulations on such systems.

The performance results we use here have been conducted using a near-global setup of COSMO (covering 98.4% of the Earth's surface area) on the Piz Daint supercomputer.[15] The simulations were performed on the hybrid partition using a total of 4888 Cray XC50 nodes (almost the full system). These hybrid nodes are equipped with an Intel E5-2690 v3 CPU (code name Haswell) and a PCIe version of the NVIDIA Tesla P100 GPU (code name Pascal) with 16-GiB second-generation high-bandwidth memory (HBM2). The nodes are interconnected in a single fabric based on Cray's Aries technology. The simulation setup was using a standard benchmark setup for global atmospheric models. At a grid spacing of 0.93 km (1.9 km), a simulation throughput of 0.043 SYPD (0.23 SYPD) was achieved. The simulations had an energy cost of 596 MWh/SY (97.8 MWh/SY).

The performance results from the IFS are based on a global simulation at 1.25-km horizontal resolution with 62 vertical levels using the nonhydrostatic model variant,[16] on the Piz Daint supercomputer. The simulations were performed using a hybrid MPI/OpenMP configuration with 9776 tasks x 6 threads, utilizing only the CPUs on 4888 nodes of the Cray XC50 partition of Piz Daint. For these runs, each node is thus equipped with 12 Intel E5-2690v3 "Haswell" cores and 64 GiB of memory. The simulation used the full model at IFS cycle 43r3 including a realistic orography derived at 1-km resolution, a full suite of physical parametrizations with cloud microphysics, radiation, shallow convection, vertical diffusion boundary layer turbulence scheme, and the land-surface interaction. The model was run predominantly in single precision (some computations such as spectral transform pre-computations used double precision). With this configuration (without I/O) we achieved a throughput of 0.088 SYPD. The simulation had an energy cost of 191.7 MW·h/SY.

Table 2 compares the COSMO and IFS simulations with our goal of the previous section and estimates a shortfall in terms of the most relevant

Table 2. Estimation of shortfall of what is achievable today as compared to the ambitious goal proposed in this article (see Table 1). Shortfall factors for selected missing components (e.g., coupling, full physics) have been estimated from relative contributions in current lower-resolution simulations. (*) Smaller than necessary time step due to irregular grid (see text) and explicit integration.

	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×	No	1.2×
Atmosphere	Non-hydrostatic	–	Non-hydrostatic	–
Precision	Single	–	Single	–
Compute rate	0.043 SYPD	23×	0.088 SYPD	11×
Other (e.g., physics, . . .)	microphysics	1.5×	Full physics	–
Total shortfall		101×		247×

parameters. It is important to note that some shortfall estimates are not straightforward and are based on rough estimations based on the authors' expert knowledge. For example, the requirement for time step length depends on a number of factors, for example the time integration scheme, and has a large impact on total throughput.

These estimated shortfalls look intimidating, if they have to be overcome in the exascale timeframe early next decade. Nevertheless, we need to remind ourselves that neither model implementation has been optimized for this type of problem. In contrast to COSMO, IFS is a global model but the baseline runs were performed on 960 nodes of a supercomputer optimized for 10-km horizontal resolution. These runs could probably be accelerated by a factor of two to four with high parallel efficiency through optimization. COSMO is a limited-area model with a Cartesian grid that is inefficient when mapped onto the globe. Furthermore, the Piz Daint system has been designed in 2012 and was not optimized for these types of runs. For example, on the dedicated system for MeteoSwiss, COSMO runs on nodes that have a much higher GPU to CPU ratio. Nevertheless, the baseline runs we summarize in Table 2 show that an exascale system will have to perform 100–250 times better on weather and climate codes than the currently available supercomputers. However, for Earth-system models that include aerosols, trace gases and multimoment cloud schemes, the number of prognostic variables will be much larger leading to a corresponding increase of these factors.

WHAT IS A USEFUL METRIC OF PERFORMANCE?

In Figures 1 and 2, we have represented the performance evolution of the IFS model running on the ECMWF supercomputers in two different ways. The flop/s metric of Figure 1 is common in discussions of supercomputer performance, because it is used in the High-Performance Linpack (HPL) and the High-Performance Conjugant Gradient (HPCG) benchmarks of the Top500 list (www.top500.org). The atmospheric degrees of freedom we used in Figure 2, on the other hand, are a direct measure of the complexity of the computation that relates to the time to solution. ECMWF always required a 10-day forecast to run

Published by the IEEE Computer Society

in less than 1 h. The complexity of the simulation increases with improving capability, which is the horizontal and vertical resolution and the number of prognostic variables in our case. Figure 2 clearly demonstrates the exponential increase in performance of the ECMWF systems over the years, but with a doubling period of 24 months rather than 18 suggested by the flop/s metric in Figure 1, which coincidently corresponds to Moore's Law.

A common starting point to reason about performance in computational science is to consider the ratio between floating point operations required in the computation and the number of main memory accesses required to fill the operands. This flop per main-memory access metric is called the arithmetic intensity of the algorithm.

For instance, in order to multiply two dense N by N matrices, we execute $O(N^3)$ floating point operations on $O(N^2)$ operands, thus the arithmetic intensity for this class or algorithms is O(N). It increases with the size of the matrix and very quickly saturates the floating-point performance of the underlying computer. Since the materials science codes we discussed above rely mostly on this type of dense matrix operations, we can expect that the algorithmic implementation has to reach a very high rate of floating point operations, and it thus makes sense to use flop/s as the primary metric for performance. The same analysis applies to the HPL benchmark, and it is not surprising that the performance evolution of material science codes almost perfectly traces that of the Top500 list of fastest supercomputers measured in terms of the HPL benchmark.

However, weather and climate codes rely on entirely different algorithmic motifs. COSMO is prototypical for many grid-based codes that use low-order stencils to approximate differential operators with finite difference or finite volume methods. The resulting algorithms have an arithmetic intensity that is lower than one operation per byte of memory moved. In other domains of fluid dynamics, this intensity can be increased by choosing higher order methods. However, increasing the arithmetic intensity this way in climate models does not pay off, since the order of the time integration is dominated by first-order physics parametrization schemes. As a consequence, weather and climate simulations have a low arithmetic intensity and are dominated by data movement.

In order to assess the performance of a weather and climate simulation, we need to consider a metric that directly relates to data movement. In particular, we need to assess how efficiently a weather and climate model uses the memory subsystem of a particular machine. In a recent publication,[15] we have proposed a metric that compares the actual number of data transfers (D) executed in the computation with the necessary data transfer (Q) of the algorithm. The latter is a theoretical lower bound of the number of memory accesses required to implement the chosen algorithm on a given hardware architecture. We called the ratio Q/D the *memory I/O efficiency* or simply I/O efficiency[15] (not to be confused with I/O to storage), which measures how well our implementation reaches the memory movement lower bound of the algorithm. The I/O efficiency reaches a value of 1 when the implementation is optimal for the particular computing system.

The second element of our performance analysis is how well the memory bandwidth (BW) on the physical device is used for the data layout and memory access pattern chosen in the code. We call this the *bandwidth (BW) efficiency*[15] that is given by the ratio $B/\hat{B}$ of the achieved bandwidth B by the implementation on the system with the maximum achievable bandwidth $\hat{B}$.

Taken together, we now have a metric that directly relates to data movements. We call this the *Memory Use Efficiency* (MUE), which is the product of the memory I/O and BW efficiencies:

$$\begin{aligned}\text{MUE} &= \text{memory I/O efficiency} \times \text{BW efficiency} \\ &= Q/D \times B/\hat{B}.\end{aligned}$$

Our analysis in terms of MUE[15] of the near global COSMO runs we discussed in the baseline section above shows that the MUE of the optimized code is 0.67, with memory I/O and BW efficiencies of 0.88 and 0.75, respectively. Thus, the implementation of the COSMO model is not far from optimal for the GPU accelerated system we used. However, the BW efficiency of the underlying system is far from optimal for the memory access patterns of the implementation. Improving this is not in the hands of the application developers but needs the attention of the architects who

design the processor and implement the software that controls the memory subsystem.

Presently, we do not have a comparable performance analysis for the implementation of IFS at ECMWF. However, the parts of the numerical methods used in IFS that execute on the grid are similar to COSMO and our approach should still apply. The current implementation of IFS uses a spectral transform method for the dynamical core. If this will continue to be the method of choice at exascale, the performance analysis will have to cover spectral transforms as well and attention will have to be given to global communication.

REFLECTING ON THE ROADMAP TOWARD "EXASCALE COMPUTING"

The characteristics of computing systems have changed fundamentally during the five decades when CMOS-based technology (Moore's Law) was the engine of performance growth. In the past, floating point operations were expensive and memory access relatively fast, such that latencies of data retrieval could be neglected. Today, floating point units are inexpensive and fast compared to the access time of memory. A compute cycle takes less than a nanosecond, the time it takes light to travel 30 cm. It takes hundreds of cycles to retrieve an operand from main memory; moving it into a register will cost about 1000 times the energy it takes to execute a floating point operation with this operand. Thus, while in the past minimizing floating point operations in an algorithm and maximizing the usage of arithmetic units of the computer was the method of choice, today the focus has to be on economizing data movements, while at the same time making optimal use of the memory subsystem. Our proposed MUE metric allows measuring this efficiency.

The need to focus on data movement when designing next generation supercomputing systems is what motivated us to choose a science problem that is dominated by algorithms with low arithmetic intensity. A supercomputer that allows applications with low arithmetic intensity to use the memory subsystem effectively will work also for arithmetically dense algorithms. The opposite is not true, as we know from the current generation of multipetaflop/s-scale supercomputers. Choosing one representative science problem and a clear goal has another significant advantage for the design of systems. It forces us to look at other key challenges that are not directly related to supercomputing performance.

For instance, let us assume that a reference architecture that meets the 1 km-scale goal will be available in five years and all weather and climate simulations move to this architecture. When the climate community ran CMIP5 at 200-km horizontal resolution it produced about 2 petabytes of data. Assuming that in 2024 this community could suddenly run at 1-km horizontal resolution without however changing their workflow, we would have to plan for storage system in the range of 80 exabytes with appropriate I/O capabilities. That would be about 10 times the capacity anticipated by CERN after they have upgraded the large hadron collider to high luminosity in the same timeframe—the ATLAS and CMS experiments at CERN today produce about 800 petabytes of data. Upgrading the compute performance of supercomputing systems to meet the goal of 1 km-scale weather and climate simulation would lead to an unmanageable data problem unless the community is supported to change the workflows as well. Rather than writing out data to disk for postsimulation analysis, the state of the simulation (check-point) will have to systematically recorded and data will have to be analyzed in-situ while (parts of) the simulations are rerun. This approach will reduce I/O to storage media by orders of magnitude but requires the ability to recreate the model trajectories. A prototype of this proposed new workflow for exascale weather and climate simulations will be presented in a forthcoming publication.

The big question that remains to be answered is whether an affordable system that will deliver on our goals can be built in the coming five to six years. To answer this, we return to the baseline and estimated shortfalls presented in Table 2. We discuss COSMO since the performances of these runs have been analyzed in detail.[15] There are at least three aspects to the performance shortfalls of the current system that should be considered.

The first aspect is software, since COSMO is implemented on a Cartesian grid that was mapped onto an anisotropic latitude-longitude grid. Compared to a homogeneous grid the

near-global COSMO runs used twice as many grid points. As a consequence of this anisotropic grid the time step had to be reduced by a factor two for stability reasons—the COSMO-based limited-area model running at 1-km horizontal resolution at MeteoSwiss uses a timestep of 10 s rather than the 6 s used for the near-global runs.[15] Thus, by replacing COSMO with a truly global grid-based model that uses the same software technology, the shortfall of Table 2 will be reduced by a factor four to 25×.

The second aspect follows from the MUE analysis of the previous section. The BW efficiency of these runs was measured to be 0.75, where the peak BW was taken to be equal to that sustained on the STREAM triad (https://www.cs.virginia.edu/stream/). However, on the NVIDIA P100 GPU that used an early version of second generation High-BW Memory (HBM2), the STREAM triad sustains only about 70% of theoretical peak BW, and theoretical peak was lower than expected. It is realistic to assume that the memory subsystem could be improved to enhance this performance by a factor two. Preliminary test with NVIDIA's new generation V100 GPU shows that the BW efficiency will increase. Together with the higher peak bandwidth for HBM2 on NVIDIA's V100, the performance of COSMO improves by a factor 2. This would bring down the shortfall to 13×.

And the third aspect is scaling. Inspection of the strong scaling behavior of the COSMO-global runs[15] leads us to believe that there is enough remaining parallelism at 1-km horizontal resolution to reduce the time to solution by a factor four through strong scaling. This scaling potential will have to be reduced in an implementation with a near-isotropic grid since it will reduce the number of grid points by 12%. Thus, a factor three can be achieved within the same energy footprint by increasing the GPU to CPU ratio and using the latest generation of GPU or equivalent accelerators in the early 2020s.

The remaining shortfall after these straightforward performance enhancement possibilities is just 4×. This may not be trivial but it is within reach of a new model implementation based on an Icosahedral, octahedral or cube-sphere grid. A similar performance study is under way for IFS. While this model is already optimized to run on the globe, none of the opportunities used on COSMO to run on nonconventional architectures have been exploited yet. We are thus confident that by early next decade, both a grid-based and a spectral model, which can run with a throughput of 1 SYPD at a nominal horizontal resolution of about 1 km, will exist, assuming the number of prognostic variables remains at today's level.

CONCLUSION

Key to this development will be an appropriate domain specific software framework[17] into which the Swiss HPCN initiative has been investing since 2010 (see www.hp2c.ch and www.pasc-ch.org) when the refactoring of COSMO was started. A first version of a domain specific library (DSL) STELLA[18] was released in 2012 and is used operationally by MeteoSwiss since 2016. A generalization of STELLA to grids suitable for global models will be integrated into the GridTools framework, the successor to STELLA. The first public release of GridTools is scheduled in 2019 when MeteoSwiss will upgrade its COSMO model to run operationally on this DSL. CSCS and MeteoSwiss, along with its close partners at ECMWF in the UK and the Max Planck Institute of Meteorology in Germany, are exploring ways to bring this software technology to the broader weather and climate community.

On this basis, we will develop a reference architecture in the coming years that will enable global weather and climate models to run at a horizontal resolution of 1 km with a throughput of 1 SYPD. Our goal is for next generation systems, such as the one that will replace Piz Daint in the early 2020s, to support such simulations without significantly higher power consumption. We believe this is possible with a dedicated codesign effort based on useful performance metrics and simple but ambitious goals. The reference architecture developed here can be integrated into the EuroHPC exaflop/s-scale computing systems that are planned for the 2023–2024 timeframe. This would mitigate the risk of our codesign approach not delivering performance within the energy footprint of current multipetaflop/s scale systems.

REFERENCES

1. Y. Wang, G. M. Stocks, W. A. Shelton, D. M. C. Nicholson, Z. Szotek, and W. M. Temmerman, "Order-*N* multiple scattering approach to electronic structure calculations," *Phys. Rev. Lett.*, vol. 75, pp. 2867–2870, 1995.
2. B. Ujfalussy *et al.*, "First principles calculation of a unit cell (512 atoms) model of non-collinear magnetic arrangements for metallic magnetism using a variation of the locally self-consistent multiple scattering method," in *Proc. ACM SC*, 1998.
3. M. Eisenbach, C.-G. Zhou, D. M. Nicholson, G. Brown, J. Larkin, and T. C. Schulthess, "A scalable method for a *ab initio* computation of free energies in nanoscale systems," in *Proc. Conf. High Perform. Comput. Netw., Storage Anal.*, 2009, pp. 64:1–64:8.
4. P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, pp. 47–55, 2015.
5. K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of CMIP5 and the experiment design," *Bull. Amer. Meteorol. Soc.*, vol. 93, pp. 485–498, 2012 doi:10.1175/BAMS-D-11-00094.1.
6. J. Charney *et al.*, "Carbon dioxide and climate: A scientific assessment," Report of an Ad Hoc Study Group on Carbon Dioxide and Climate, 1979. Available at: www.nap.edu/catalog/12181.html.
7. IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate change, 2013. Available at: http://www.climatechange2013.org
8. B. Stevens and S. Bony, "Water in the atmosphere," *Phys. Today*, vol. 66, no. 6, pp. 29–34, 2013.
9. T. Schneider *et al.*, "Climate goals and computing the future of clouds," *Nature Clim. Change*, vol. 7, pp. 3–5, 2017.
10. D. Panosetti, L. Schlemmer, and C. Schär, "Convergence behavior of convection-resolving simulations of summertime deep moist convection over land," *Clim. Dyn.*, to be published. doi:10.1007/s00382-018-4229-9.
11. C. S. Bretherton and M. F. Khairoutdinov, "Convective self- aggregation feedbacks in near-global cloud-resolving simulations of an aquaplanet," *J. Adv. Model Earth Syst.*, vol. 7, pp. 1765–1787, 2015.
12. H. Yashiro *et al.*, "Performance analysis and optimization of nonhydrostatic icosahedral atmospheric model (NICAM) on the K computer and TSUBAME2.5," in *Proc. Platform Adv. Sci. Comput. Conf.*, 2016. doi:10.1145/2929908.2929911.
13. D. Leutwyler *et al.*, "Evaluation of the convection-resolving climate modeling approach on continental scales," *J. Geophys. Res. Atmos.*, vol. 122, no. 10, pp. 5237–5258, 2017.
14. A. Prein *et al.*, "The future intensification of hourly precipitation extremes," *Nature Clim. Change*, vol. 7, no. 1, pp. 48–52, 2017.
15. O. Fuhrer *et al.*, "Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0," *Geosci. Model Dev.*, vol. 11, pp. 1665–1681, 2018. doi:10.5194/gmd-11-1665-2018.
16. N. P. Wedi and P. K. Smolarkiewicz, "A framework for testing global non-hydrostatic models," *Quart. J. Roy. Meteorol. Soc.*, vol. 135, no. 639, Part B, pp. 469–484, 2009. doi:10.1002/qj.377.
17. T. C. Schulthess, "Programming revisited," *Nature Phys.*, vol. 11, pp. 369–373, 2015.
18. T. Gysi *et al.*, "STELLA: A domain-specific tool for structured grid methods in weather and climate models," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2015, pp. 1–12.

Thomas C. Schulthess is a Professor of Computational Physics with ETH Zurich, Zurich, Switzerland, who directs the Swiss National Supercomputing Centre (a.k.a. CSCS). He leads the Swiss High-Performance Computing and Networking initiative that promotes application software development for extreme scale computing and develops the Swiss supercomputing roadmap. Contact him at schultho@ethz.ch.

Peter Bauer is the Deputy Director of the Research Department with the European Centre for Medium-Range Weather Forecasts, Reading, U.K. He leads the ECMWF Scalability Program that prepares the forecasting system for future high-performance computing and big data handling challenges. He also coordinates European Commission research projects in support of the Scalability Program. Contact him at peter.bauer@ecmwf.int.

Nils Wedi leads ECMWF's Earth System Modelling section that addresses all aspects of scientific and computational performance relating to ECMWF's numerical weather prediction model and its ensemble forecasting system. Contact him at nils.wedi@ecmwf.int.

Oliver Fuhrer leads the weather model development team at MeteoSwiss, Zurich, Switzerland, and holds a Lecturer position with ETH Zurich, Zurich, Switzerland. He led the project that built the first operational weather prediction model running on a GPU-accelerated high-performance computer. Contact him at oliver.fuhrer@meteoswiss.ch.

Torsten Hoefler is a Professor of Computer Science at ETH Zurich, Zurich, Switzerland. His main research interests include application-centric implementation and optimization of high-performance computing systems. He directs the Scalable Parallel Computing Laboratory performing research at the intersection of performance-centric algorithms, programming models, middleware, and hardware systems. Contact him at torsten.hoefler@inf.ethz.ch.

Christoph Schär is a Professor at ETH Zurich and active in high-resolution climate modeling. He is leading a Swiss initiative to develop and exploit a km-resolution continental-scale climate-modeling capability with the goal to improve the simulation of the water cycle. Contact him at schaer@env.ehz.ch.