PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Sustainability in astroparticle physics

V. Grinberg,[a,*] **K. Jahnke,**[b,*] **V. Lindenstruth,**[c,*] **C. Markou,**[d,*] **S. Funk,**[e] **U. Katz**[e] **and M. Roth**[f]

[a] *European Space Agency (ESA), European Space Research and Technology Centre (ESTEC), Keplerlaan 1, 2201 AZ Noordwijk, the Netherlands*

[b] *Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*

[c] *Goethe-University Frankfurt, Institute for Advanced Studies, Max von Laue Street 12 , 60438 Frankfurt, Germany*

[d] *Institute of Nuclear and Particle Physics, NCSR Demokitos, 27 Neapoleos Str., Agia Paraskevi Attikis, 15341 Greece*

[e] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen Centre for Astroparticle Physics, Erwin-Rommel-Str. 1, D-91058 Erlangen, Germany*

[f] *Institute for Astroparticle Physics (IAP), Karlsruhe Institute of Technology, POB 3640 D-76021 Karlsruhe, Germany*

The topic of sustainability is becoming increasingly important in research activities in astroparticle physics, both in existing and also in future instruments. At this years International cosmic ray conference (ICRC 2021) one session was dedicated to this topic. This publication will summarise the findings of this well-attended online session.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)1401

1. Introduction: doing science in a developing climate crisis (Knud Jahnke)

Humanity is confronted with a global, existential, and anthropogenic climate crisis. For decades its impacts have already been developing and felt in many particularly vulnerable regions – but the effects now tangibly also started hitting the Global North: droughts in California, floods in Germany, more frequent local temperature records everywhere, increased extent of wildfires in Australia, USA, Canada, and Russia. Global average surface temperatures have already increased by more than 1°C compared to pre-industrial times, locally often much more than that. So beyond rising sea levels as the main yet abstract threat of the past, the climate crisis now has arrived for most of humanity.

The situation can be summarised like this: Yes, the climate crisis is real; yes, it's made by us; yes, we can still limit it; but we will have to change our behaviour [1]. Most people actually tend to agree with this – yet, one of the questions we frequently hear being raised after this statement is: "But why does that relate to my work in research? The main emissions come from countries X and Y / the coal industry / industrial consumption / meat production / the housing sector?" Qualitatively, that seems like a well-founded argument. But the answer is: With >200 countries, many fields of industry, many different fossil fuel companies, many sectors of energy consumption, that there is no single factor, not a single country dominating CO_2-emissions, hence there is no culprit that is "mainly" responsible. At least if we look at this issue from the consumer side – in the end the total is the sum of the parts. Our science is one of these parts.

The measure of how relevant this part could be comes in shape of the 'permissible' carbon emission per person. The Paris Agreement [2] provides a scientific calculation for the globally remaining CO_2 emission budgets that limit global warming to a maximum of +1.5°C, with 50% probability. This budget is a global remaining emission of ~410 Gt CO_2 from 2022 onward [3]. Under the assumption of global climate neutrality by 2050 and a (controversial) equal distribution across 7 billion humans, this would permit each one of us to emit ~60 tCO_2 total until 2050, corresponding to about 2 tCO_2 per year, when starting in 2022.

In contrast, the current emission per person in e.g. Germany is currently ~10 tCO_2. This is a total emission for every citizen *on average*. But how high are our science work related emissions in comparison? Until recently we did not have good data about this. This changed in the past years for the field of astronomy, where assessments have calculated the total and per-scientist CO_2-emissions for the Australian astronomy community [4], the Max Planck Institute for Astronomy (MPIA) [5], the CFHT telescope [6, 7], in-person vs. online conferencing [8], and more recently the European Southern Observatory [9], as well as planning for the GRAND experiment in China [10]. This ties in with the wider-context Labos1point5 initiative of research labs in France [11].

These assessment reports showed us that e.g. the MPIA in 2018 had about 18 tCO_2 emissions per scientist, the Australian community even beyond 40 tCO_2 – and these are both solely the work-related emissions in addition to all "private" emissions from food, housing, mobility, and general consumption. In both cases the dominating factors of science-related emissions are business flights and electricity use, mainly for computing. At the level of MPIA the remaining 60 tCO_2 per-person emission budget according to the – internationally binding and ratified – Paris Agreement would be used up in 3 years, at the level of the Australian astro community in less than two.

This is obviously not sustainable – and there are two very different views on why we should take this knowledge as the incentive to drastically reduce our science-related emissions:

1) *The 'moral' view:* we emit way beyond our remaining budgets and have a moral obligations for the livelihood of future generations to drastically reduce this. If along others we do not, then we are keeping Earth on a path towards +3–4°C, with catastrophic consequences.

2) *The 'selfish' view:* under the assumption that society as a whole will quickly move to reducing carbon emissions, this will e.g. increase the cost of CO_2-emissions, as well as reduce the social acceptance of large emitters. If our current mode of work is so drastically dependent on CO_2-emissions through frequent flying e.g. for communication and carbon-intensive electricity for computing, then we quickly need to reduce this dependency. Else our research in the future will become more expensive and socially less accepted – with detrimental consequences.

This means: we have to change the way we work – and by 'we' in this case specifically means astronomy and astroparticle physics. We have to identify the specific instances and reasons for e.g. why we fly, which emissions this entails, how we plan energy- and data-intensive experiments, supercomputing, and how we can in general decarbonise our communication, our data, and our way we create new fundamental knowledge. Solutions for this will be very diverse: we first need to identify the reasons and paths of emissions, and then identify solutions. Some of these solutions might lie on the personal level, but more often will require combining aspects on the level of institutions, our community, or even society. Only by setting clear decarbonisation goals for our work, and only by carefully quantifying emission sources we will be able to search for and find solutions, and implement them one by one. This will not be an event, but a process. With the remaining budgets vs. the current science-related emissions this decarbonisation process has to be started by us, and has to be started now.

The following sections will describe the situation and initial, partially already scalable solutions for our science emissions, in the fields of conferencing and travel, green computing, and green experiments.

2. Conferences and travel (Victoria Grinberg)

Academic travel takes on many shapes, ranging from collaboration meetings and conferences to instrument construction and commissioning, from research visits to job interviews, from PhD students attending summer schools to senior researchers participating in grant committees (Gokus, Jahnke et al., in prep.). The expectation of international mobility gives rise to a secondary layer of travel many will undertake privately to be able to see family and/or friends.

It is thus of no surprise that travel-related emission is one of the major contributions to the CO_2 budget of astrophysical research, reaching up to ~45% of the total emission of an individual institutes and/or communities, depending on their location and scientific focus [4, 5, 12], with similar estimates for whole collaborations such as GRAND [10]. Addressing academic travel is thus paramount when addressing the academic CO_2 footprint. As the whole field of academic travel-related emission cannot be addressed adequately here, we aim to provide an overview of some of the ongoing efforts, especially focused on conferences, and pointers towards useful publications.

The most sustainable travel is no travel at all. But full abstinence from travel is not a personal decision that can be easily made by an individual in current astro(particle)research without

significant impact on their career. This is especially the case for early career researchers, including PhD students, postdocs and others in untenured positions where travel enables networking that is seen as essential to reach tenure. A systemic change is thus necessary – while individual contributions as possible triggers for change are not to be neglected (see [13] and references therein).

First step towards changing our approach to travel is quantifying it. Where such estimates have been made, the absolute amount of flight emission is, for both individual institutes or communities [5, 12] and conferences [8], dominated by long-haul flights that cannot be easily replaced by, e.g., trains. To reduce travel-based emission other approaches than alternative modes of transport are thus necessary. This may include alternatives to in person meetings (remote or hybrid events), combining multiple trips in one (including a flexible enough financial framework to enable combined trips and possibly pricing in the environmental impact of additional flights), or re-consideration whether certain trips are necessary at all.

A simple consideration is the comparison, as discussed in [8], of the 2019 European Week of Astronomy and Space Science in Lyon, with a total emission from travel only of over 200 000 kg CO_2 equivalent, which does not yet take into account hotel- or venue-related emission, with the 2020 edition of the same conference that had to take place online due to the covid pandemic. Taking into account laptop, network- and zoom-server related emissions, the total for the remote meeting is below 1000 kg CO_2 equivalent. [14] provide a tool[1] to easily estimate the carbon footprint of a given meeting; note in particular that different emission factors and estimates may lead to results that differ by a factor of a few. [14] further use the example of the Athena X-ray Integral Field Unit Consortium to discuss how the travel footprint associated with the instrument and its consortium can be reduced, in particular by reduction of the number of the consortium meetings and by transitioning working group meetings from face-to-face to video conferences, without detrimental impact on the overall project.

While remote meetings are clearly superior for environmental sustainability reasons, they are often cited as less conducive to new collaborations and alienating, especially for early career researchers. On the other hand, they are also more accessible to those who face barriers attending face-to-face meetings due to, for example, financial constrains, care responsibility, health constraints, disabilities, or teaching responsibilities; they are, thus, in many regard, more inclusive [15]. Additionally, online meetings do not require travel visas that are hard to impossible to obtain for researchers from many developing countries[2]. They do, however, require a good and stable internet connection, which can be challenging in some locations.

The ICRC meetings themselves are a good example for the increased reach of remote meetings: the 2019 face-to-face meeting in Madison, WI (USA) had 857 participants from 39 countries who have submitted 1062 abstracts. Researchers from certain countries have not been granted visas to attend the meeting, excluding whole communities. The 2021 online meeting had 1601 registered participants from 54 countries who have submitted 1400 abstract, i.e. increasing the number of participants almost twofold. In particular, a session on sustainability with the 2021 line-up, including speakers from outside the cosmic ray community, would have likely not have

[1] https://travel-footprint-calculator.irap.omp.eu/
[2] https://www.nature.com/articles/d41586-018-06750-1

been possible in a face-to-face meeting, as external speakers may have been hesitant to travel.

As an additional complication, simply transferring a face-to-face conference into an online or hybrid setting, without re-thinking the conference structure, will rarely work. Monitor fatigue, challenging time-zones, lack of direct feedback, feelings of isolation are just some of the problems. Successful online meetings require re-thinking of conferences and collaborations that most conference organizers had no time to do when the COVID-19 pandemic forced many events online. But in a way, the pandemic only accelerated a development that has been predicted and ongoing, albeit much slower, for years. Today, systematic approaches to defining better online conference format are being discussed and published (see esp. [16] for an in-detail discussion and [17] for a short summary) and first how-to guides describing more- or less successful online meetings appear [18].

We may not be able – and perhaps not willing – to give up travelling as part of the academic profession. But both personal decision and systemic changes in how we approach travel and conferences will help us to at least reduce the environmental footprint of work-related travel.

3. Green computing (Volker Lindenstruth)

Despite rising server efficiency, the increasing IT demand is overcompensating the efficiency savings, resulting in a constant surge of IT energy usage.

Green Computing is to be understood as efficient computing, delivering the same results for less energy consumption and therefore less environmental footprint. There are three major areas of concern here. First the data centers hosting the supercomputers typically require significant amounts of energy for the cooling of the systems and the provision of uninterruptable power, second there is the computer architecture and third there are the algorithms themselves, where different implementations can vary in power consumption and execution speed by several orders of magnitude. In the following we will present examples of all three domains.

3.1 Data Center Architecture

The average power consumption of data centers in Germany alone in 2020 was 2 GW [19]. Assuming the conventional energy mix in 2020, this corresponds to 6.4 Mt CO_2. The average PUE of the data centers was 1.63, which corresponds to a power usage of the data center alone of 773 MW. This amount of power is mostly required for the cooling infrastructure but also for redundant power, such as battery backups. One of the driving factors for this high cooling power requirement is the choice of the cooling technology. For instance, if data centers are cooled with air, very high air flow rates are typically required and also large temperature differences are required between the cold air supply and the hot air leaving the servers. Also, moving large amounts of air volume requires additional energy.

We have developed an alternative technology, which has proven to be much more efficient [20][21]. It is based on the concept of transferring the waste heat of the IT equipment to cooling water as early as reasonably possible. Since the thermal capacity of water is a factor 4000 larger than air the corresponding flow rates and temperature differences are equivalently lower.

There are several approaches for cooling with liquids. On the one hand there are systems with direct water cooling implementing different kinds of heat sinks cooled by a liquid which is pumped through. This technology has the disadvantage that the appropriate heat sinks have to be specifically

developed for each and every server architecture, connecting to all devices, which produce heat. This means, however, that there are additional lead times in the development and that components cannot easily be swapped out against other similar devices. In addition, this technology typically does not remove 100% of the generated heat, leaving the requirement for additional air cooling in the data center. Another approach includes different implementations of immersion cooling where the IT equipment is subjected to a liquid, which directly removes the heat. There are so called open bath or enclosed chassis implementations available.

However, when considering the cooling of IT equipment, the cost of the said equipment must be taken into account, in particular if the chosen cooling technology has an impact on the available market of that technology. The mass market provides in general the largest competition, the fastest time to market, and therefore typically the best cost efficiency. This cost factor is of particular importance as computers are usually replaced after 5 years of operation, therefore reiterating any cost overhead every 5 years. In comparison, data centers are operated typically for more than 20 years. To date, the above-mentioned cooling technologies are more specialized, limiting the market of available IT hardware and therefore bear the risk of higher cost and later the availability of new hardware (CPUs, GPUs, etc.).

Therefore, we have focused on the general IT and server architecture, which is air cooled. However, our technology does not rule out direct water cooling of the servers. When considering air cooled COTS servers, there is the additional power, required by the fans inside the server box. Noting that the efficiency of a fan is inversely proportional to the square of its rotational speed, it is obvious that larger fans with lower RPM should be preferred over smaller fans with correspondingly higher rotational speed. Therefore, very small servers, like the 1U servers, which as a rule implement batteries of counter rotating fans in series, are very inefficient. Taking into account that racks are commonly filled with servers, there is no need for such small enclosures. In our experience, 2U or larger servers provide quite efficient fan cooling where the fan power is below 7% of the total server power at maximum server power consumption. Given that the average operating power of a HPC server is generally at 60% of maximum power consumption, the typical fan power is reduced to below 2% of the server power.

One very efficient way to cool the hot air leaving the servers are heat exchangers, which are mounted in the rear door of the rack. Such heat exchangers are commercially available and are built such that they do not present a significant back pressure to the air leaving the servers. For example, such a rack operating at 30 kW IT power would have a heat exchanger back pressure at an air flow of 4500 m^3/h of below 30 Pa. Another important aspect is the air velocity inside the rack. Depending on the particular configuration air flow rates are normally between 1ms and 2 ms. This means that the hot air leaving the IT equipment needs less than 0.2 s before it hits the heat exchanger. Therefore, any vertical effects can be neglected. It is possible to combine high power compute servers with low power file servers inside the same rack without these two very different systems affecting each other. There is also no strong requirement to seal the rack. Customarily there are some small openings for the cable feed throughs. Given the small over pressure behind the heat exchanger the amount of hot air leaving the rack is negligible. We have been operating many different servers since 2010 according to those principles without any issues.

The cooling water circulating the heat exchangers has to be cooled back. Figure 1 shows a sketch of standard cooling architecture. There are two cooling loops. The secondary circuit is

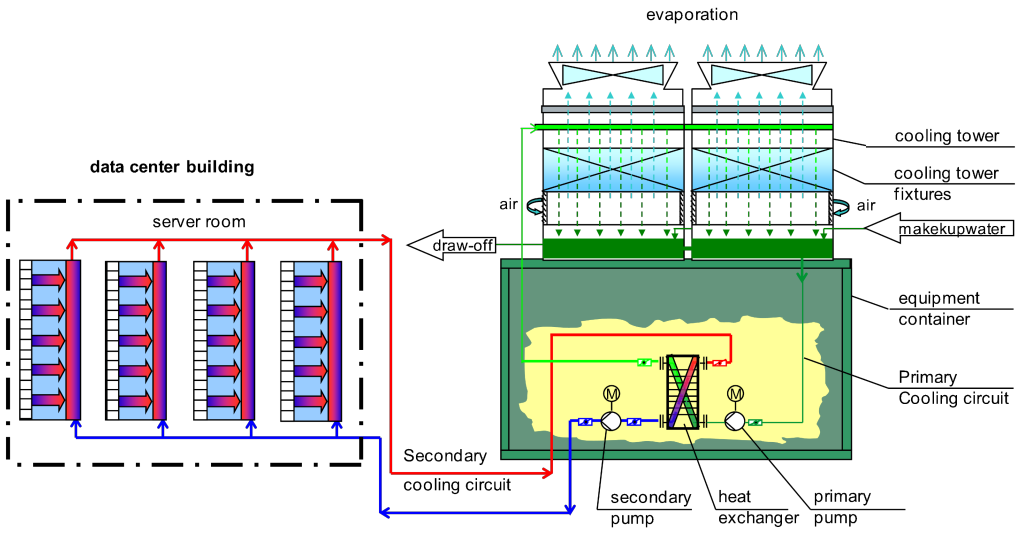

Figure 1: The Green Cube cooling scheme

closed and circulates water between the rear door heat exchangers and a heat exchanger, coupling the secondary circuit to the primary circuit. This water is clean. The cooling water in the primary circuit is cooled most efficiently with open-loop evaporative coolers. Basically, water is evaporated inside the cooler, which provides the cooling effect to the rest of the water in the primary circuit. As a rule of thumb there are two cubic meters of water required per hour and MW of cooling power. The cost of evaporated water is negligible compared to the energy cost. One of our sites uses the water of a nearby river for evaporation. There are two sets of redundant pumps moving the water in the two cooling circuits. The reason for decoupling the cooling loop inside the white space from the primary circuit is that the open-loop coolers are subjected to the environment and for instance collect pollen during spring time. There are sand filters in the primary loop but it did not seem advisable to pump such water directly through the heat exchangers in the data center.

The important feature of this cooling system is that it can cool down to the wet bulb temperature, which in Germany remains below 22° C and is typically above 20° C for about 100 hours per year. During operation we measure small temperature differences between the primary and the secondary circuit, the primary cold-water supply is slightly warmer than the wet bulb temperature. We operate the secondary circuit at full power at a design ΔT of 3° K at full cooling power. The required water flow rates do not present a high demand on pumping power. It should also be noted that it is not required to regulate the water flow rate of the individual heat exchangers. In the worst case the water flow rate through a heat exchanger of a rack with little power consumption would be higher than absolutely necessary but given the low power requirement of the water pumps this is negligible. The air temperature leaving the heat exchanger is slightly higher than the cooling water return temperature. This, however, is a design parameter of the particular heat exchanger.

All in all we have shown that the room temperature can be kept below 8° C above the wet bulb temperature and therefore at or below 30° C. This temperature is well within the ASHRAY allowable temperature envelope.

The concept of rear door heat exchangers confines the warm air to the rear part of the rack. Consequently, there is no requirement for hot or cold aisles or any air flow regulation in the data center. The racks can be arranged in any way. Figure 2 shows a photo of the fifth floor of the data

Figure 2: One Green Cube floor

center Green Cube at the GSI Helmholtz Center. It should be noted that there is no need for high ceilings or any air flow regulation. Therefore, the racks can be stacked as in a high bay warehouse. The Green Cube implements 6 floors. Currently the two upper floors are built. Each floor has a cooling capability of 2 MW with 2N redundant power.

The HPC data center of Goethe-University has been in operation since 2010, while the Green Cube has been operating since 2016. The Green Cube operates at about 15% of its design cooling capability and has demonstrated an overall average cooling efficiency of 7% of the IT power or a PUE of 1.07. Those numbers have been verified by the German TÜV. We have measured at 2.4 MW a PUE of below 1.03. The Green Cube at the GSI Helmholtz Center has carried the German eco-label "Blauer Engel" for several years now.

3.2 Computer Architecture

The computer architecture can also provide significant energy savings. In general there is a trade-off between the number of GPUs, CPU cores and memory in a server. Naturally different applications have different requirements. One metric here is the application memory requirement per CPU core. In our experience 4 GB/core is a reasonable low estimate. It is a good idea to have a few dedicated nodes in a cluster with significantly higher memory installations for special applications. It should be noted that the cost of the memory in a server can easily become a significant fraction of the overall cost. Another metric is the ratio between CPU cores and GPUs in a server. Generally there are some tasks better performed on CPUs, which are often used to orchestrate the processing on the GPUs. If choices are not optimal here either GPUs or CPUs are left underutilized, wasting both money and also energy. One of our latest installations at the CERN ALICE experiment implements 8 CPUs per server GPU and 8 GPUs in one server, which has 512 GB of main memory. There is a total of 250 of those servers, therefore implementing 2000 GPUs, 16000 physical CPU cores and a total of 125 TB of main memory. The memory installation here is extensive since there are rather large data buffers required by this application. Another aspect is the choice of processor and GPU. There are different energy efficiency levels available. However, this must be verified with the given applications at hand. In general the latest hardware implementing the latest silicon technology provides the best energy efficiency.

Nowadays the compute performance is limited by the cooling capability of the server. Both the CPUs and GPUs will throttle their internal clocks according to their utilization and temperature. Those adjustments usually happen very quickly. It is very important to understand this behavior, particularly with regard to the context of the applications. The result may turn out to be sobering. One very important aspect in this context is the use in parallel computers where many servers cooperate. This often requires all servers to complete one task (for example processing one tile of a matrix) before all can continue to the next step. In this context the slowest node determines the overall performance and consequently requires all other nodes to wait, wasting energy during that time. We have performed several tests of large batches of GPUs, measuring the spread in performance under comparable conditions. Running a DGEMM benchmark on GPUs with the exact same voltage and clock setting resulted in 15% variations of the compute performance. These performance variations are simply silicon process variations during production. When running the LINPACK algorithm for example this would mean that all GPUs would operate at the lowest performance limit and therefore all GPUs would be slowed down to the performance of the slowest one. This has obviously the energy inefficiency associated with it since the faster GPUs will remain in some waiting pattern. Therefore it is advisable to benchmark the efficiency performance of the GPUs and to adjust their clock rates and supply voltages to an optimum for the entire system. This optimization process typically requires a significant number of benchmarks. The DGEMM algorithm is quite useful here as it has a very high compute utilisation of the GPU. A system which is optimized in this way will provide also for applications a very efficient performance.

Servers often come with lots of extra features for the various potential application fields. Servers used in the HPC environment usually do not require a large fraction of thouse features. For instance unused USB subsystems can consume significant power and should be powered down. Same is true for DVD devices and the like. It is advised to measure the standby power of a server with no HPC applications running. This power should be optimized to a minimum. Usually this process involves some trial and error procedure. But considering the usually large number of servers, implemented in a HPC system this effort is worth trying.

The servers usually regulate their fans independently, according to load and environmental conditions. The fan power can become quite large, exceeding 100 W and needs to be monitored. Also here there is a trade-off between the server compute power and the additional cost for the fan cooling. A good operating point can be found with reasonable fan speed and low overall power consumption.

Several of our systems have scored high (positions 1, 2, 8) in the Green500 world ranking list of the most efficient computers [22][23]. All those systems have implemented fans and some, the first ranked in particular, were in competition with systems which used immersion cooling, where server fans are excluded and the pumps to move the cooling liquid is not accounted for.

3.3 Algorithmic Engineering

The availability of highly parallel manycore architectures, GPUs, wide-vector processors, and new memory technologies is leading to a paradigm shift in the design of algorithms and to a huge increase in efficiency, which can be several orders of magnitude. Efficiency increases can be directly transferred into a faster knowledge gain, while they usually directly translate in energy efficiency improvements of the same order, leading to “Green HPC”.

Nowadays, processors offer an increasing amount of vector instructions, while the supported feature sets differ between architectures. Today's processors implement 512-bit wide vector registers. This means that a program which does not use vector features operates only at $1/8^{th}$ at double precision or $1/16^{th}$ at single precision of the compute performance of the processor. Well vectorized code also works efficiently on GPUs. It should be noted that vectorization should be taken into account when the algorithm is developed. The data structures and the algorithm itself must be engineered properly. The existing auto vectorization features of the compilers cannot repair what is faulty at the concept level [24]. The vectorization package Vc was developed to enable portable software development, which allows the optimal use of particular vectorization feature sets while maintaining portability across platforms without overhead [24]. The standard ISO/IEC 19570:2018 is now based on Vc. Vectorizing existing code typically requires the refactoring of data structures, often by realigning them from arrays of structs to structs of arrays.

The next level of highly energy efficient algorithms use GPUs. These devices are designed to operate massively parallel like the processing of the pixels of an image. Therefore the algorithm has to have a high degree of parallelism in order to use GPUs well. In addition GPUs implement the fastest available memories exceeding 1 TB/s access rate. The latest GPUs implementing PCIe 4.0 have demonstrated the capability of transferring simultaneously reading and writing in excess of 50 GB/s.

Several large software packages have been ported to run on GPUs. Examples here are the development of an Open-CL lattice QCD program, which, after optimization, is 10 times faster than before and runs simultaneously on 4 GPUs with good scalability [25].

In the area of relativistic molecular dynamics, the UrQMD package has been rewritten and accelerated by a factor of 150 [26]. Track reconstruction in nuclear and particle physics must recognise even the most complex decay patterns. Complex algorithms for 4D event reconstruction (3D plus time) for various experiments at CERN [27][28] and FAIR [29] are indispensable for the operation of these experiments. Corresponding libraries are currently being further developed to provide the necessary functionality and performance. The first optimization step based on cellular automatons and Kalman filters has resulted in a speed increase of 10,000 times [29]. In the area of life sciences, the analysis of electron microscopic data with Bayesian inference could be accelerated 45 to 450 times [30]. All of these improvements have enabled these applications to run efficiently on GPUs. In addition, algorithms for very large datasets, especially graphs, with 1,000-fold speed increases have been developed.

In general the future of efficient computing is to be found in massively parallel computing using wherever possible vector or vector like instructions and data structures. The price performance and energy efficiency of GPUs outperforms CPUs. There is still a large amount of legacy software, which has not been programmed according to those paradigms and requires a rework. This often requires some effort but taking into account the already demonstrated benefits an overhaul should be undertaken rather sooner than later. Many of the modern experiments in particle physics would simply not be possible without these highly efficient algorithms. For example the on-line reconstruction software for the ALICE experiment at CERN has been adopted to run to more than 95% on GPUs. However, the software was written in HIP allowing the same source code to execute on CPUs and GPUs by different vendors. Direct comparisons have shown that an equivalent CPU only system would have increased the CAPEX cost by a factor of 7, which corresponds to a potential

cost increase of 36 Million Dollars.

4. Green experiments (Christos Markou)

Reducing the carbon emissions of experimental and research facilities is an important issue, not only in relation to the sustainability of the infrastructures, but also in relation to the overall "message" the scientific community can send to the rest of the world concerning actions against climate change. In the case of small to medium scale experiments, the possible solutions can range from the obvious choice of purchasing green energy directly from appropriate certified providers, all the way to the rather unusual choice of producing the required energy "in–house" using Renewable Energy Infrastructure (REI).

We are presenting a case study of such an approach in the context of KM3NeT [31], the underwater neutrino telescope which is being built in the Mediterranean Sea. We have investigated possible strategies and technical choices, legal issues, we performed a detailed technical study and considered the financial side of this endeavor, in order to answer the question whether it is feasible, whether it makes sense and what could be the best approach. The investigation was performed in the context of the KM3NeT–2.0 project [32], funded by H2020 and was carried out in the period 2018–2019.

KM3NeT is a distributed infrastructure with two active installation sites, in the South of France off the coast of Toulon and in Italy, off the South East coast of Sicily. A third site off the South West Peloponnese in Greece is a potential site for a later stage. The case study considered all three sites on an equal basis, and assumed that the REI in each case would serve approximately equal size detectors and shore stations. Only the energy requirements during operation were considered at the actual experimental sites. Energy requirements during R&D, construction and installation and the activities in individual institutions were not included.

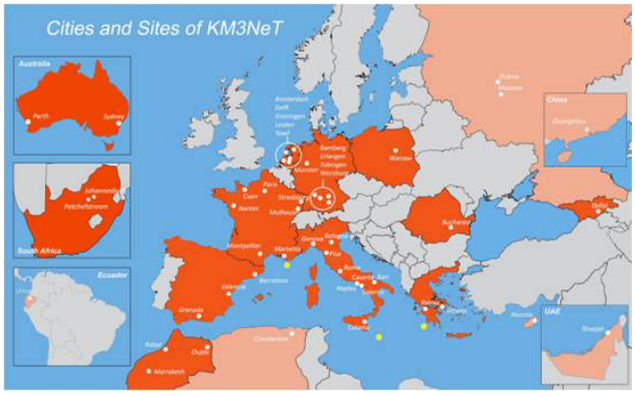

Figure 3: The KM3NeT collaboration map

During operation, the energy budget for each site, with full detector in operation, is estimated to be in the range 580 - 650 kW, corresponding to an average 615 kW or 5.4 GWh per year, equivalent to ~1330 tCO_2. Such energy requirements can be satisfied with a small–to–medium sized REI. To this end, three possible strategies have been identified and pursued:

1) Use *certified energy providers* over the grid. This is an obvious choice which although satisfies the requirement for zero carbo emissions, has no added value as its exposure to the society is rather limited and is difficult to communicate in an engaginf matter.

2) *Collaborate with REI producers* with the intention to add to their infrastructure.

3) *Establish our own REI*, provided it makes sense financially, also counting in the added value in terms of societal engagement.

It turns out that choice 1 above is the only working scenario for our French site due to the existence of a highly mature green energy provider market and the existence of complicated National legal procedures which can result in an extended period until completion for the other two choices. Choice 2 is not feasible as the scale of our porject is far too small for the typical comercial REI size, thus making the addtion of said infrastructure a complicated issue with little financial iterest for the private sector. Choice 3 is an attractive solution for both the Italian and the Greek sites, especially in view of the strong interest expressed by the respective local authorities and communities for collaboration.

4.1 Technology choices

In order to have viable solutions, off-the-shelf mature technologies should be chosen, optimally suited to the installation sites. Sicily and Peloponnese have similar climates, with typical Mediterranean conditions involving a lot of solar irradiation all year round, as well as relatively strong wind patterns. Both Photovoltaic (PV) panels and land-based wind turbines are well tested and mature off-the-shelf technologies with costs which have decreased in the recent years. Extensive know-how in installation and maintenance exists widely in most countries, and certainly in the two regions under consideration. Other possible solutions like geothermal, wave, tidal, off shore wind, floating wind, OTEC and other technologies are either ill suited to the specific sites or not mature enough to be commercially viable yet.

Our working model is to produce the energy required in each site, provide it to the local grid and then purchase it back under appropriate agreements. This model allows for the opportunistic character of the two chosen renewable energy technologies and eliminates the need for energy storage solutions which are prohibitively expensive. The key point in our scenario is the use of two kinds of REI: a large scale facility to generate the bulk of the energy required, supplemented with small scale REI of *high aesthetic quality* designed to be installed in an urban environment. For PV panels, the large scale facility comes in the usual form of panels installed inthe countryside, while specially designed PVpanels can be installed on the vertical surfaces of buildings inside the urban web. Similarly for wind, the usual Horizontal Axis Wind Turbines (HAWT) can be installed in appropriate places with high wind patterns, while inside the cities, smaller scale Vertical Axis Wind Turbines (VAWT) can be installed in public open areas, like parks, playgrounds, seafront jetties etc. In this way, the local authorites and communities will be active partners in the project providing the neccessary real estate for the REI installation, benefiting from the upgrade of the urban web and the provision of surplus electricity to local schools, hospitals, public buildings etc. In both sites, the local authorities have been largey supportive of this scheme.

4.2 Legal issues

A detailed study of both National and European legislature has identified no major legal issues or obstacles to the realisation of such a project by the corresponding legal entites of the research institutions responsible for the installation sites of KM3NeT.

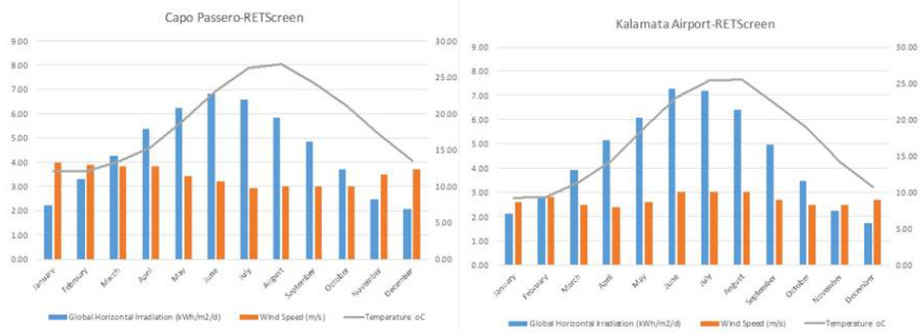

Figure 4: Weather data for Italy (left) and Greece (right) over a 12 month period

4.3 Does it make sense financially?

The success of implementation and the long term viability of such a project relies heavily on the financial issues involved. In order to study the financial health of our working scenario we went through the following steps:

1) Various configurations of REI were defined with full specs for each of them according to the manufacturers.

2) Detailed, realistic simulations of energy production were performed for these configurations.

3) Proper calculations of the actual cost of energy production with based on real market data and the real costs for REI installation, running costs and maintenance, taking into account different inflation scenaria and infrastructure lifetime.

4) A detailed comparison with normal grid energy (non-green) costs was performed.

Weather data from several databases (PVGIS, PVGISCM-SAF, RETScreen, etc) provide hourly data with spatial resolution of the order of 3 km over land. These data are the product of land based stations, sattelite observations and extrapolation results over several years. A compilation of data from RETScreen is shown in figure 4. The interesting quantities include ambient temperature, atmospheric pressure, solar irradiation, wind speed and direction. These, together with information on the terrain and surrounding landscape in Capo Passero, Sicily and Kalamata, Peloponnese, as well as the detailed REI configurations and their specs were fed to simulation programs like PVsyst, SAM and HOMER to calculate the energy production over the lifetime of the project. An extesive list of losses were considered, including variations of irradiance levels, temperature variations, soiling, ohmic losses, power and voltage threshold losses, turbine performance degradation, etc.

Taking into account the local REI market in both Italy and Greece and the differences in the wind and solar weather patterns, the following configurations were defined:

For Italy, 1 HAWT at 3 MW installed capacity, 6 VAWT at 60 kW for the urban installation, and PV panels of 140 kW total installed capacity, including 40 kW of PV facades.

For Greece, 1 HAWT at 2.3 MW installed capacity, 6 VAWT at 60 kW for the urban installation, and PV panels of 440 kW total installed capacity, including 40 kW of PV facades.

The simulations produced the Specific Energy Production for each component of the proposed REI, as shown in figure 5. The overall performance of the system can be assessed by the *Levelized Cost of Energy (LCOE)* which is the ratio of the *Total lifetime cost* over the *Total lifetime Energy*

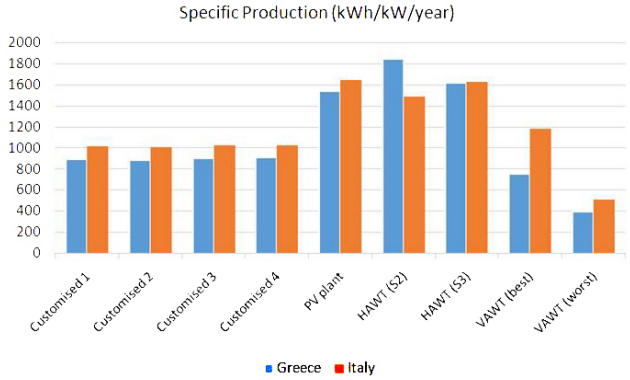

Figure 5: Specific Energy Production for various REI configurations (Customised 1-4 refer to different PV facade models)

Figure 6: LCOE for large and small scale REIs in Italy and Greece compared to wholesale energy costs in December 2019

Production. In the Total lifetime cost, the material and installation costs, maintenance costs, margin for component replacements and inflation estimates were included. As far as inflation is concerned, we considered a scenario based on the average 10-year inflation in each country and a scenario with double this average.

The Total Lifetime Energy Production is the Specific Energy production for the complete installed system over the expected lifetime. For this end, we considered two cases, spanning 15 and 25 years. The results are shown in figure 6 for the various REI configurations in each country, with the 2 inflation and the 2 lifetime scenaria. The LCOE should be compared to the actual costs of purchasing energy over the grid. The solid line on the figures corresponds to the average *Wholesale* cost of electricity in the European markets in December 2019. Although this is a volatile quantity, it is evident that the cost involved in satifying the energy requirements of KM3NeT in *all* installation

sites under the scenario of creating our own REI is far more attractive financially than purchasing the necessary energy from the grid in the normal way.

This study has shown that a medium sized research infrastructure can become a zero carbon footprint by opting for an unusual implementation, one which involves the local communitites who will benefit from the creation of the said REI, while at the same time, the projected costs over the lifetime of the project make it an attractive solution.

Final remark

The ways of sustainable approach outlined here may serve as a seed for future developments in our field. But also beyond that, we hope to radiate into other areas of science and even more into everyday life by leading the way.

References

[1] IPCC, *Summary for Policymakers*, in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, V. Masson-Delmotte et al., eds., Cambridge University Press, in press.

[2] United Nations, "Paris Agreement." https://treaties.un.org/pages/ViewDetails.aspx?chapter=27&clang=_en&mtdsg_no=XXVII-7-d&src=TREATY, 12, 2015.

[3] German Advisory Council on the Environment, *Using the CO_2 budget to meet the Paris climate targets, Environmental Report 2020, Chapter 2*, 2020.

[4] A.R.H. Stevens, S. Bellstedt, P.J. Elahi and M.T. Murphy, *The imperative to reduce carbon emissions in astronomy*, *Nature Astronomy* **4** (2020) 843 [1912.05834].

[5] K. Jahnke, C. Fendt, M. Fouesneau, I. Georgiev, T. Herbst, M. Kaasinen et al., *An astronomical institute's perspective on meeting the challenges of the climate crisis*, *Nature Astronomy* **4** (2020) 812 [2009.11307].

[6] N. Flagey, K. Thronas, A. Petric, K. Withington and M.J. Seidel, *Measuring carbon emissions at the Canada-France-Hawaii Telescope*, *Nature Astronomy* **4** (2020) 816.

[7] N. Flagey, K. Thronas, A.O. Petric, K. Withington and M.J. Seidel, *Estimating carbon emissions at CFHT: a first step toward a more sustainable observatory*, *Journal of Astronomical Telescopes, Instruments, and Systems* **7** (2021) 017001.

[8] L. Burtscher, D. Barret, A.P. Borkar, V. Grinberg, K. Jahnke, S. Kendrew et al., *The carbon footprint of large astronomy meetings*, *Nature Astronomy* **4** (2020) 823 [2009.11344].

[9] European Southern Observatory, *Annual report 2020*, 2021.

[10] C. Aujoux, K. Kotera and O. Blanchard, *Estimating the carbon footprint of the GRAND project, a multi-decade astrophysics experiment*, *Astroparticle Physics* **131** (2021) 102587 [2101.02049].

[11] O. Berné, *Labos 1point5: a collective action to reduce the carbon footprint of research at the scale of France*, in *Astronomy for Planet Earth: Forging a Sustainable Future*, p. 3, July, 2021, DOI.

[12] F. van der Tak, G. Nelemans, S. Bloemen, L. Burtscher, S. Portegies Zwart, R. Wijnands et al., *The carbon footprint of NL astronomy in 2019*, in *Astronomy for Planet Earth: Forging a Sustainable Future*, p. 6, July, 2021, DOI.

[13] "L. Hackel and G. Sparkman: Reducing Your Carbon Footprint Still Matters." https://slate.com/technology/2018/10/carbon-footprint-climate-change-personal-action-collective-action.html.

[14] D. Barret, *Estimating, monitoring and minimizing the travel footprint associated with the development of the Athena X-ray Integral Field Unit*, *Experimental Astronomy* **49** (2020) 183 [2004.05603].

[15] S. White, *An environment for everyone*, in *Astronomy for Planet Earth: Forging a Sustainable Future*, p. 5, July, 2021, DOI.

[16] *The Future of Meetings: Outcomes and Recommendations*, Zenodo, Dec., 2020. 10.5281/zenodo.4345562.

[17] V.A. Moss, M. Adcock, A.W. Hotan, R. Kobayashi, G.A. Rees, C. Siégel et al., *Forging a path to a better normal for conferences and collaboration*, *Nature Astronomy* **5** (2021) 213.

[18] H.M. Günther, J.R.A. Davenport, S.J. Wolk and S. Gallagher, *How to organize an online conference - Lessons learned from Cool Stars 20.5 (virtually cool)*, in *The 20.5th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun (CS20.5)*, Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun, p. 333, Mar., 2021, DOI [2105.08795].

[19] D.R.H. Borderstep Institute, "Rechenzentren 2020, energiebedarf der rechenzentren steigt trotz corona weiter an."

[20] V. Lindenstruth and H. Stöcker, "Building for a computer centre with devices for efficient cooling."

[21] V. Lindenstruth and H. Stöcker, "Methods and apparatus for temperature control of computer racks and computer data centres."

[22] D. Rohr, G. Neskovic and V. Lindenstruth, *The L-CSC cluster: Optimizing power efficiency to become the greenest supercomputer in the world in the green500 list of november 2014*, *CoRR* **abs/1811.11475** (2018) [1811.11475].

[23] D. Rohr, M. Bach, G.N. and Volker Lindenstruth, C. Pinke and O. Philipsen, *Lattice-csc: Optimizing and building an efficient supercomputer for lattice-qcd and to achieve first place in green500*, in *High Performance Computing - 30th International Conference (ISC), Frankfurt, Germany, July 12-16,*, pp. 179–196, 2015.

[24] M. Kretz, *Extending C++ for explicit data-parallel programming via SIMD vector types*, Ph.D. thesis, Goethe University Frankfurt am Main, 2015. 10.13140/RG.2.1.2355.4323.

[25] M. Bach, V. Lindenstruth, O. Philipsen and C. Pinke, *Lattice QCD based on OpenCL*, *Comput. Phys. Commun.* **184** (2013) 2042.

[26] J. Gerhard, V. Lindenstruth and M. Bleicher, *Relativistic hydrodynamics on graphic cards*, *Comput. Phys. Commun.* **184** (2013) 311.

[27] S. Gorbunov, D. Rohr, K. Aamodt et al., *Alice hlt high speed tracking on gpu*, *Nuclear Science, IEEE Transactions on* **58** (2011) 1845.

[28] ALICE Collaboration, *Real-time data processing in the ALICE high level trigger at the LHC*, *Computer Physics Communications* **242** (2019) 25.

[29] S. Gorbunov, U. Kebschull, I. Kisel, V. Lindenstruth and W.F. Müller, *Fast simdized kalman filter based track fit*, *Comput. Phys. Commun.* **178** (2008) 374.

[30] P. Cossio, D. Rohr, F. Baruffa, M. Rampp, V. Lindenstruth and G. Hummer, *BioEM: GPU-accelerated computing of bayesian inference of electron microscopy images*, *Comput. Phys. Commun.* **210** (2017) 163.

[31] "KM3NeT web portal." https://www.km3net.org.

[32] "The KM3NeT-2.0 H2020 project web portal and documents therein." https://www.km3net.org/km3net-infradev/low-carbon-footprint-with-renewables/.