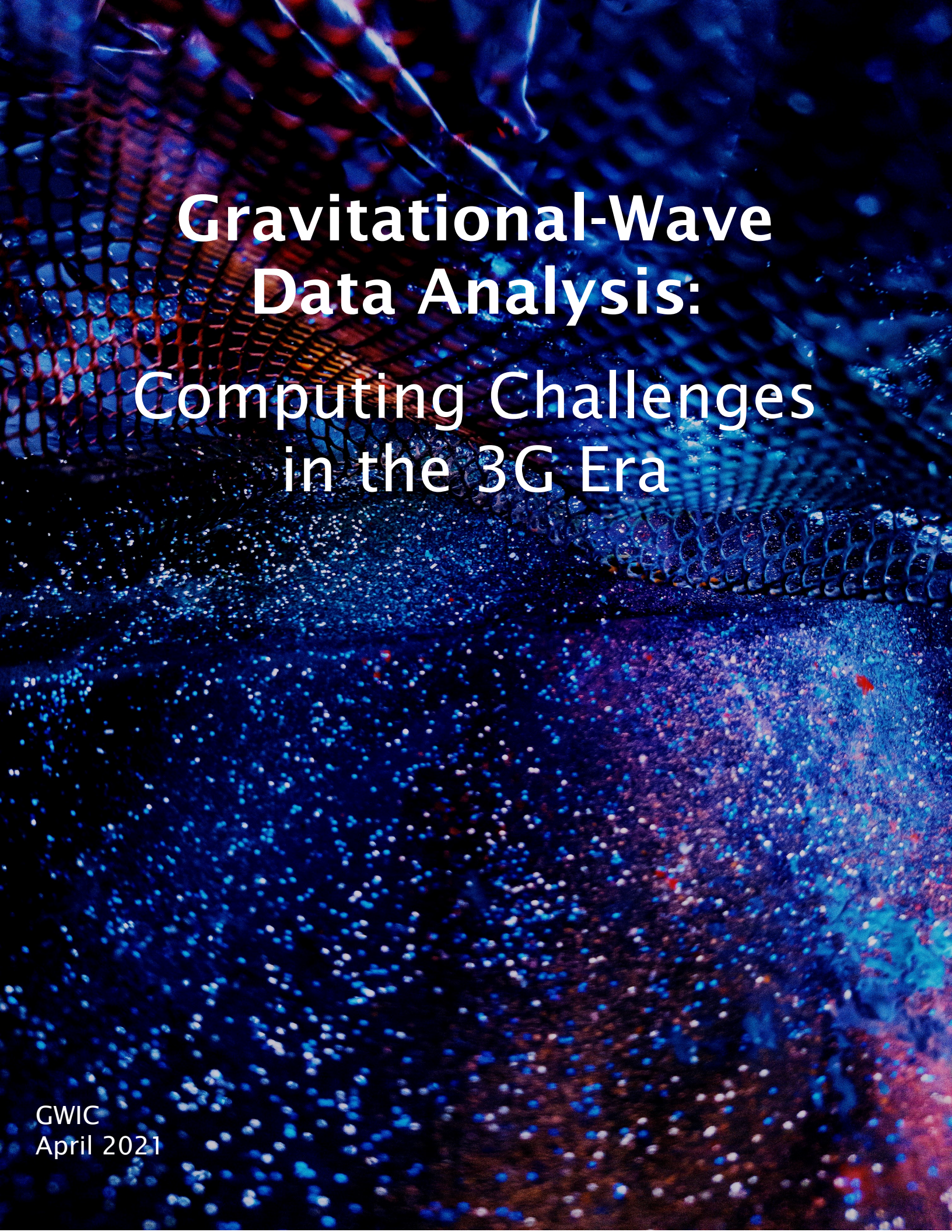


Gravitational-Wave Data Analysis:

Computing Challenges in the 3G Era

GWIC
April 2021

This document was produced by the GWIC 3G Subcommittee and the GWIC 3G Data Analysis Computing Challenges Subcommittee

Final release, April 2021

Cover: Andre Moura

Contents

1. Introduction

To frame and motivate this report, we begin with background on the existing 2G detector scientific collaborations and an overview of the computing models and methods currently employed. For additional background on the science driving the computational needs, please refer to the 3G Science Case Report.[1]

The Advanced LIGO/Advanced Virgo collaboration (LVC) is composed of three Gravitational Wave (GW) interferometers located in Hanford (WA), Livingston (LA) and Pisa (Italy). In September 2015, the LVC began a series of advanced era detector runs, with the nomenclature "O#". O1 ran from September 2015 to January 2016, and as well as the first ever detection of GWs, the run ended with the detection of three binary black hole (BBH) mergers. O2 ran from December 2016 until the end of August 2017. As well as the detection of a number of other BBH mergers, O2 saw the first ever detection of a merger of two neutron stars (BNS). O3 began on April 1st 2019, and, due to lockdown imposed by the COVID-19 pandemics, was terminated earlier than scheduled on Mar 27th, 2020. It is further expected that the Japanese interferometer KAGRA will join the upcoming O4 run.

From a data analysis computing perspective, a primary challenge in the transition from O1 to O2 was the increased computational power needed for both the search and parameter estimation phases. In the search (detection) phase, the waveform template banks increased in size to accommodate a larger range of masses. In the parameter estimation phase, while the computational cost of each run remained almost the same as in O1, the greatly increased number of GW sources, as well as the number of exploratory runs that were needed for the BNS merger discovery, caused the computational cost to explode. In addition, the discoveries provided an opportunity to perform unforeseen and computationally-intensive analyses to measure the Hubble-Lemaitre constant H0, test the validity of GR and to constrain the internal physics of neutron stars.

For its third observing run (O3), the LIGO-Virgo collaboration estimates its ongoing data analysis computing requirements at 700 million CPU core-hours[1] per year, to perform some 80 astrophysical searches, follow-up activities, and detector characterization activities. The 10 most computationally intensive of these analyses comprise about 90% of the demand, with a long tail of the remaining 70. Most of this computing consists of pleasingly parallel High Throughput Computing (HTC) for "deep" offline searches; 10% is for low-latency data analyses needed to generate rapid alerts for multi-messenger (electromagnetic, neutrino) followup. Very little high-performance parallel computing is required, with the exception of Numerical Relativity simulations, which are not included in this assessment.

While during O1 the vast majority of the computing power was provided by dedicated LIGO-Virgo clusters, either on-site or in large computing centres, during O2 and O3 more and more use was made of external shared computing resources. This growth of shared, external computing resources prompted the development of a distributed computing model, similar to the ones used for example by large LHC collaborations. Furthermore, the Virgo, LIGO and KAGRA collaborations are joining efforts moving from partially interoperable owned computing resources to a wholly shared common computing infrastructure

[1] On a reference Intel Xeon E5-2670 2.6Ghz CPU

under the IGWN (International Gravitational Wave observatories Network) rubric[2, 3], by adopting common tools and interfaces.

The LVC has carefully investigated the use of parallel GPU and MIC architectures for its most compute-intensive searches; CUDA GPUs have been the most successful and cost-effective, and were deployed at scale for the first time in O3.

Currently little use is made of commercial cloud resources for data analysis computing aside from service infrastructure, for example for the low-latency services that generate and distribute the public event candidate alerts. There are no major technical obstacles to performing HTC on the cloud (they are similar to other shared resources), however the logistics and cost-effectiveness of funding and managing metered computing are still being understood, while the use of academic cloud resources (such as the EGI FedCloud in Europe) are being investigated.

LIGO/Virgo generates $\sim$20TB per interferometer (IFO) per observing year of h(t) strain data used by most analyses, and 1PB per IFO per observing year of raw data (all channels, full sample rate). The scale of gravitational-wave data analysis is no longer "big data" by 2020 standards — but data management remains non-trivial in a distributed HTC environment nonetheless.

In the following, we will adopt the following nomenclature:

- Analysis pipelines: individual data analysis codes written by domain scientists, on top of Core software libraries and frameworks, to extract the scientific results from data;
- Core software: common data analysis applications used by scientists, scientific computing libraries, and experiment software frameworks;
- Cyberinfrastructure (US) or e-Infrastructure (EU): services used to run distributed, federated computing, and the corresponding software for data management, workload management, AAI, etc.; Computing facilities: the underlying computing resources (CPU, storage, networks, local support) supporting data analysis computing infrastructure and codes.

2. Challenges

The demand for data analysis computing in the 3G era will be driven by the high number of detections (up to hundreds per day), the expanded search parameter space (0.1 solar masses to 1000+ solar masses) and the subsequent parameter estimation (PE) follow-up required to extract the optimal science. Given that GW signals may be visible for hours or days in a 3G detector (as opposed to seconds to minutes in the 2G era), their waveforms may exceed physical memory limits on individual computing resources. On top of this, as the 3G network will be an international effort, there will be an increased need to develop appropriate and scalable computing cyberinfrastructure, including data access and transfer protocols, and storage and management of software tools, that have sustainable development, support and management processes.

Data storage should not be challenging, since GW data products are not expected to grow by many orders of magnitude in the 3G era; the challenge will be in the computational power needed to extract the science content from the data.

In terms of computational challenges for the analysis of gravitational-wave data, there appear two major aspects for the 3G era, relative to 2G: compact binary coalescence (CBC) detection and CBC parameter estimation.

2.1 CBC Detection

At present, the majority of GW detections are made using a matched-filter based template bank. This requires covering the parameter search space with many hundreds of thousands of theoretical waveform models (or templates). The cost of matched-filter searches for CBC signals grows dramatically with improvements in sensitivity as detector bandwidth is increased, so straightforward application of existing methods are unlikely to meet the science requirements of the 3G era. Specifically, both the number and duration of waveform templates needed increases substantially for the lower frequency limits afforded by the 3G detector designs. Thankfully, the dominant costs are independent of predicted rates and rely on parallel algorithms with independent tasks that are horizontally scalable. 3G detectors will observe vastly more gravitational waves than 2G detectors, including some with tremendous signal-to-noise ratios and some that remain in the detector's sensitive frequency band over much longer times. Given the increased duration and higher number of signals, extracting all of the science encoded in these observations would be prohibitively expensive with today's data analysis techniques; much work will be needed on core software and search pipeline R&D to reduce the overall computing requirements and overcome new challenges.

Assuming waveform template banks are still applied in the 3G era, a number of questions need to be answered:

- How many templates will be needed to cover 1-1000(+) solar mass events?
- Will sub-solar-mass events (0.1-1 solar masses) need to be included in the template bank?
- How will matched filtering perform if there are overlapping mergers in the data?
- How large must template banks be (i.e., how densely will waveforms need to be placed in parameter space) in order to maintain O2-era SNR targets; or inversely, how much SNR will be lost if template bank size is constrained?

- What are the science impacts of massively reducing the number of templates at the cost of sensitivity?

A major foreseeable problem is how to answer these questions when the ground-based GW data analysis community is fully occupied with development for 2G science runs. The Gravitational-Wave community is still relatively small, and will need to grow and/or divide its efforts between exploiting the existing 2G data and preparing for future 3G data.

2.2 CBC Parameter Estimation

The cost of CBC Parameter Estimation using Bayesian inference may present an even larger challenge. Improvements in detector sensitivity translate into orders of magnitude greater detection rates, and PE costs scale linearly with the number of detections. Furthermore, most PE techniques rely on inherently sequential algorithms (e.g., Monte Carlo Markov Chain, or "MCMC"), whose convergence properties vary from algorithm to algorithm. To make statements on the parameters of a system, a certain number of statistically independent samples from the posterior distribution are required. The current 2G runtime to achieve this number of samples varies not only from algorithm to algorithm, but also from source to source, with the fastest analysis taking days, and the longest taking weeks. This latency is unacceptable in the 3G era.

The development of new, more convergent (and thus more rapid and efficient) algorithms is a field of research in its own right, with the typical timescale for the development of a new algorithm usually being on the order of a decade. The development of such algorithms for 3G must be an interdisciplinary effort involving GW scientists, experts in statistical inference and professional programmers. The adaptation of advanced algorithms designed for other domains will almost certainly require extensive modification for GW astronomy applications. Again, this is something that would take a number of years to accomplish. Such a development will require a large investment of human effort on top of that required by the development of new detection algorithms.

Also, the scientific requirements for PE in the 3G era are not yet well-defined. In the 2G era, the LVC performed multiple, computationally-intensive PE analyses on every viable CBC candidate, and yet were constrained by available human effort as much as by computing. In addition, the computing costs of development, testing, and exploration of PE codes were greater than that of the final PE runs used for publication.

In an era of multiple signals per day, it is unclear if it will be feasible or even desirable to run a deep PE analysis on every CBC candidate. It might be that the science goals demand different degrees of PE investigation (and therefore computing resources) for different candidates. The answer will obviously be driven by the computing efficiency of PE codes (and the degree of automation possible in the PE process), but the range of scientific scenarios is not obvious due to the many remaining uncertainties: what is the lower limit of PE investigation needed to realize our most basic 3G scientific goals, and what is its present computing cost; what is the upper limit of useful PE investigation we can currently achieve simply by applying more computing power, given the expected signal quality; where between these two limits do we see the maximum scientific benefit as a function of cost; and do we see rapidly diminishing returns for additional computing after a certain degree of precision? Are there obvious "tiers" of investigation and cost which we should target for different kinds of candidate signals? How should we define those tiers, given the science we hope to extract from the data?

These questions need to be explored iteratively by GW scientists over the next decade in close collaboration with statistical inference and computing experts, as new computing efficiencies are realized and new scientific goals are identified and translated into PE requirements.

Increasing heterogeneity and complexity of computing platforms drives the need for i) better software engineering and testing; ii) additional organizational expertise and effort in optimization, distributed computing (architecture, engineering, support), and computing management; iii) better tools, services, and processes for sustainable optimization; iv) better education and consulting for scientists/developers who are not first and

foremost software engineers; v) automated testing for diverse hardware platforms and environments; and vi) more complex deployment, orchestration, instrumentation, and accounting of DA workflows.

2.3 Burst and continuous sources

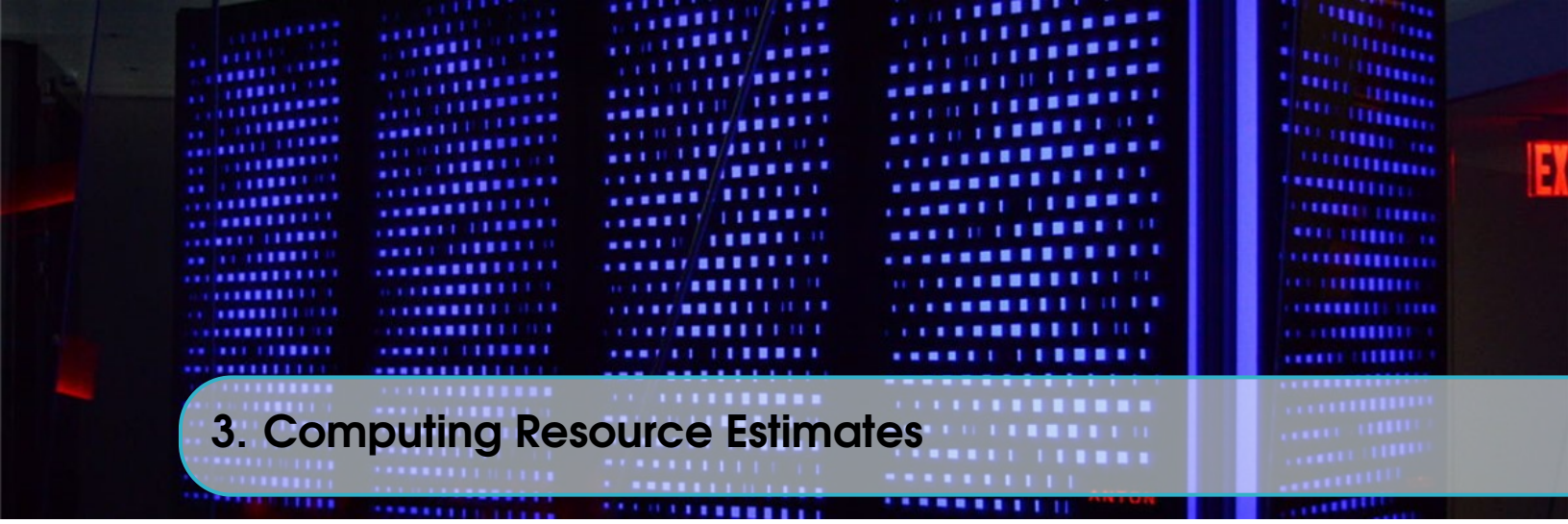
Unmodeled burst source searches, a major component of 2G gravitational-wave data analysis computing demand, should scale more or less linearly with the number of interferometers and their observing time; unlike CBC, there is no scaling with detector sensitivity or low-frequency cut-off. Thus, the computational demands of unmodeled burst source detection should be relatively modest in the 3G era.

Evaluating the computing resources needed for Continuous-Wave (CW) signal searches is a challenge in itself, however. In the 2G era, such searches were limited by many orders of magnitude less available computing power than optimal, and thus the increased cost of 3G detection will not pose a fundamentally new problem for CW searches.

All-sky searches, with current pipelines (see, for example, [4]), scale linearly with both the number of detectors and observation time, but the lower frequency regions do not contribute that much to the computation time, so the required computing power is not expected to explode like in the CBC case. However, aiming for a better intrinsic sensitivity of the searches or for a larger explored parameter space would very quickly increase the computation time.

Targeted searches, looking for signals from known or suspected pulsars, are again linear with both the number of interferometers and the amount of accumulated data, but need a very small fraction of the computing power needed for all-sky searches, even if the number of potential sources will grow.

Ultimately in the 3G era, all-sky searches sensitivity will still be limited by the available computing power (mitigated by algorithm developments and code optimisation), more or less by the same amount in the 2G era given the same scientific objectives.



3. Computing Resource Estimates

Approximately 1 million CPU core-hours of computing were used for parameter estimation of each signal in the LVC's second observing run (O2), compared to 10k CBC signals per year expected for 3G IFOs; a 3 orders-of-magnitude increase.

A naive scaling of current matched-filter CBC searches and PE to 3G design sensitivities will require many orders of magnitude more CPU and RAM than 2G. Moore's Law alone will not deliver the required performance increase, moreso because it is slowing already due to both cost management by a shrinking set of hardware producers and — barring a breakthrough — the physical limits of silicon transistors. CERN's outlook is a 10%/year performance/cost improvement at constant cost, with very large uncertainties [5]. This does not include the coprocessor/accelerator (GPU, FPGA, etc.) market, but there are no panaceas there; the performance improvements that are being delivered by both increased CPU parallelism and co-processors are becoming more complicated to exploit in software via task and/or data parallelism. Many existing codes won't benefit from additional CPU parallelism, and a large amount of effort will need to be dedicated to effectively exploiting such architectures.

Optimistically, more efficient techniques for background estimation may reduce CBC search costs by a factor of a few; machine learning-based searches may live up to their promise in culling signals from non-Gaussian noise. Even a diminished Moore's law could help improve the performance of sequential codes by a factor of a few over the next decade. And newer, more parallel algorithms will track Moore's Law closer than sequential algorithms.

However, something has to give: barring an unexpected breakthrough, we must prepare to be clever about how and where in parameter space our searches can be less sensitive, in order not to miss what we care about most. This could be done, e.g., via matched-filtering with sparser template banks and/or less faithful waveforms in the most expensive and/or less scientifically valuable regions of the parameter space, or via reduced-order modeling, waveform compression, greater reliance on unmodeled burst searches, and/or machine learning technologies. We don't yet know the right mix of approaches for 3G-scale analysis. We may need to identify new "sweet spots" of sensitivity vs. cost given our science goals.

If the GW science we can realize will indeed be limited by the efficiency of search and PE codes, targeted investments in data analysis development and optimization today, in advance of the 3G era, may pay outsized scientific dividends. These investments should be driven by first understanding the detection and PE requirements needed for specific discoveries, and by stressing that both the computing infrastructure (for more efficient exploitation of available resources) and core software R&D and quality management (for pipeline optimisation) are crucial 3G investments.

Even with all the possible innovations in the core software, and the mitigations that can be obtained by refining the scientific strategy, the large number of observations to process will require a fundamentally distributed data analysis system, rather than the dedicated Data Centers mostly used in 2G runs; a transition that already started during O2 and that will take advantage of the large shared cyberinfrastructures for scientific computing that we expect will be available in the 3G timescale.

4. Core Software (AI, Algorithms, etc.)

Scientific data analysis will necessarily rely on both cyberinfrastructure for distributing and managing those analyses, and the core software and libraries implementing the underlying algorithms. As highlighted in the challenges, and quantified in the resource estimates, the improved sensitivity and corresponding scientific potential of 3G detectors requires that the core software be well-optimized, yet also able to target the diverse hardware of a shared and distributed computing environment.

Much of the core software currently used to analyze gravitational waves is written and maintained by domain scientists themselves. Among this are examples of software leveraging GPU [6–9], CPU vectorization [10] and high-performance libraries [11]. However this individually-driven approach to performance already shows its limitations in second-generation analysis codes. Core software for 3G will require extensive optimization targeting diverse hardware, which needs both software engineering and domain science expertise. In addition to developing this software, careful benchmarking to feed into more detailed estimates of computational costs will be essential. All of these activities, to move beyond the ad-hoc approach of the 2G era, need both coordination and consistent staffing of scientists and developers with the requisite cross-disciplinary expertise.

Additionally, the resource estimates of this report are largely based on extending the data analysis techniques used for current detectors to 3G detectors. However, new data analysis techniques are also under constant development, and in particular the success of machine learning in several scientific disciplines has prompted investigations of those methods in gravitational-wave data analysis. We cannot adequately summarize those investigations here; a recent overview is in [12]. Examples of the applications considered are searches for compact binaries [13–18], searches for isolated pulsars [19, 20], parameter estimation for compact binaries [21–23], and detector characterization [24–26].

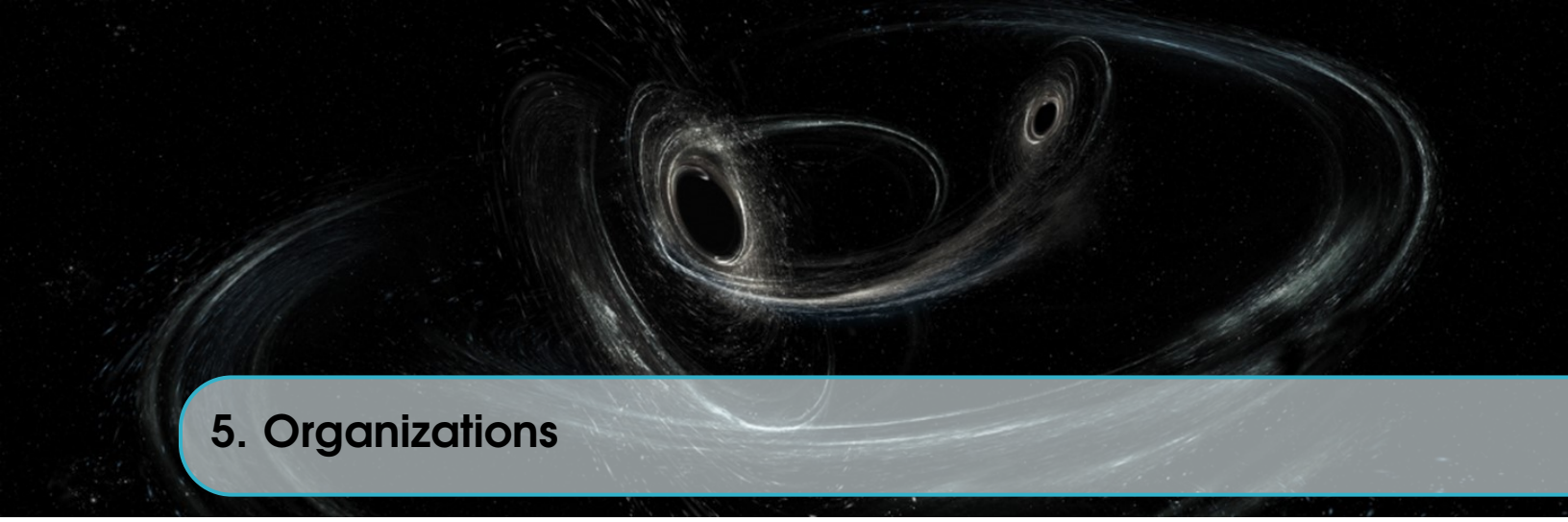
Some of the papers above find machine learning techniques to be extremely promising; others find those techniques to compare poorly to traditional approaches. More relevant will be the state of the art in the 3G era. Some of the most successful applications of machine learning—such as long short-term memory for speech recognition [27, 28], or convolutional neural networks for image classification [29, 30]—come from machine learning techniques that are adapted to particular problems. It is not unreasonable to expect that the full potential of machine learning to gravitational-wave data analysis will likewise benefit from similar domain-specific techniques.

Both hardware-targeted optimization of traditional techniques, and production-ready machine learning techniques, will need to be tested across realistic data sets. The most appropriate technique for each data analysis task, whether search or parameter estimation, will also depend on the time frame in which results are needed: “early warning” pre-merger detection, low-latency CBC detection and follow-up, or higher-precision but higher-latency “deep” analyses. Optimal hardware procurement will also need to be carefully benchmarked on validated and tuned production-ready core software.

Taken together, all of these considerations emphasize the need for both a stable organizational structure in which cross-disciplinary expertise can be coordinated, and formal mock-data challenges that rigorously compare different approaches, similar to those successfully used by LISA. A LISA challenge begins with

publicly released simulated noise, with simulated signals hidden inside. Anyone in the community is then welcome to build and apply computational tools that attempt to find and interpret the hidden signals. Such activities provide increasingly precise estimates of required computing resources, and inform scientific prioritization if resources must be constrained. As an example, the data challenges spurred the development of stochastic search techniques, because a template-bank approach would not be computationally feasible. In the 3G case, mock-data challenges can also be used to test the readiness and usability of the emerging distributed computing cyberinfrastructure, similarly to what the LHC experiments did in the early years of what later became the WLCG.

Recommendation: The community should organize mock-data challenges for 3G data analysis, similar to ones undertaken by the LISA community, to systematically test competing techniques and ensure that the best approaches are robust and production-ready.



5. Organizations

The GW community is rapidly moving from an era of single experiments managing their own computing needs, to a time of global collaboration between GW experiments, within a more and more complex international environment where it will be competing for resources with other sciences that are also exploring very large scale data sets, and other global collaborations. As it stands today the GW community does not have a mechanism for organizing a strong collaboration in computing. We believe that in preparation for the future it is timely to consider what a global computing collaboration for GW would look like and the advantages that would arise from having such a structure.

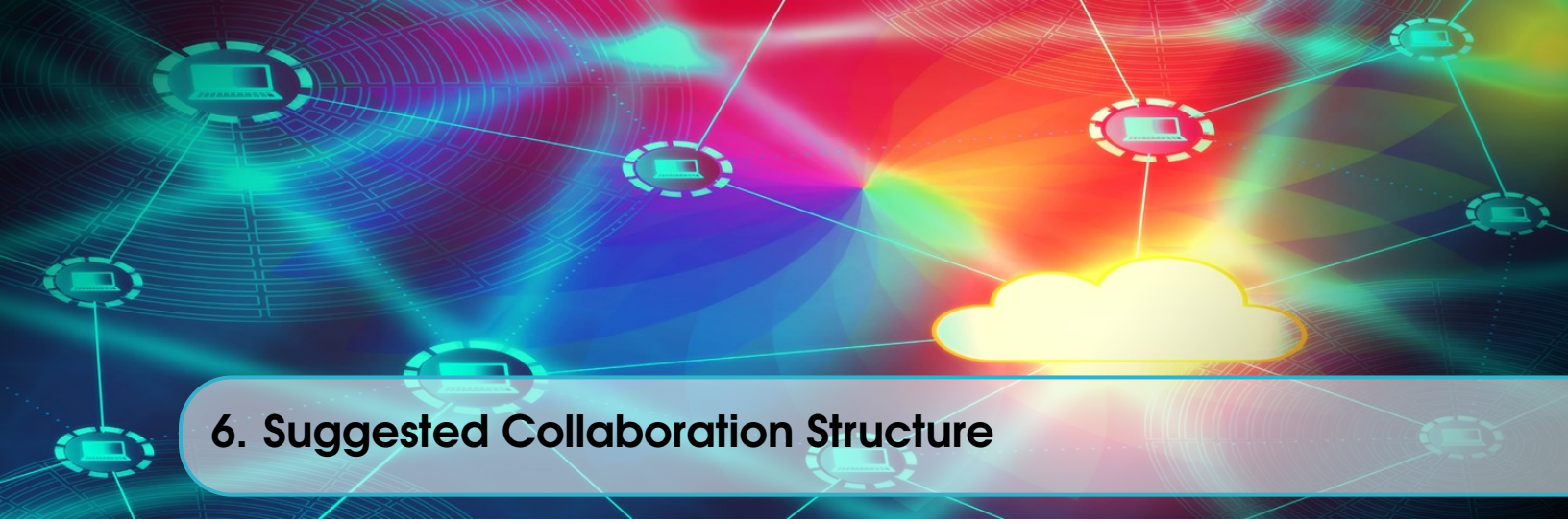
It will be important for the GW community to have a strong and recognised voice with the cyberinfrastructure communities who are advancing collaborative, federated, global computing services. Thus it will be very important to create a GW computing management structure to have a clear input in discussing and collaborating with other large research computing efforts as well as the cyberinfrastructure. This organization must gather and represent the needs of the global GW community, and to be able to negotiate and plan with a clear, recognizable voice. There are several reasons behind this:

- The GW community has challenges in terms of the nature of its workflows and resource needs, that may not be well represented by the existing cyberinfrastructure communities (for example large memory requirements, use of complex pipelines, specific CPU/GPU needs).
- In an era where GW does not expect to have owned, dedicated resources, a body is needed to address common challenges and explore synergies with others. Computing for GW will be co-located in data centres used by other large-scale sciences (LHC, SKA, etc.)
- Similarly, a body is needed to address other cyber infrastructure needs on international resources: facilities, networks, etc., either in collaboration with other sciences or in representing GW-specific needs.
- It is needed as a single political voice to face funding agencies, resources providers, etc.
  - This is essential to have to argue coherently internationally for people and hardware resources.
  - It represents a point of organization to collaborate on project proposals - and to be coherent between those across different Funding Agencies.
  - This is where the WLCG (in the case of High Energy Physics, HEP) has been very successful - the funding agencies recognise the global nature and organization and are open to discussion on how to collaborate on resources.
- A management body is needed to agree on responsibilities between various GW collaborations; at various levels - perhaps MOUs for services, data and other policies, resource provision, etc. This will include GW presence in open data policy and funding, open access activities and would be coherent with that in other Astronomy of physics domains.
- A global computing collaboration management body could also oversee and commission community-wide technical developments needed for progress, and to help integrate with the global infrastructures it wants to use; for example on Authentication and Authorization Infrastructure, or other areas. Such a body would also be able to oversee common operational support (across GW collaborations) in such

key areas as security coordination, in conjunction with other cyberinfrastructures and so on. It can provide a presence in relevant technical bodies (e.g. those of EOSC, OSG, ASCR, PRACE, etc.). Again, this organization can work internationally to make best use of all of the funding opportunities available, and to be coherent across them.

A global computing management could also address specific needs where good coordination is required:

- There is a management challenge in balancing between effort needed for support of the current observing runs and the effort needed for future developments and preparation. This could be addressed with a global view of effort across the collaborations.
- Given the length of time to develop new algorithms, the management body could also set up groups to tackle specific problems with experts drawn from across the experiments, and help to resource and fund these.
- Similarly working groups on "R&D" topics may be required for new search algorithms, PE, use of new computing architectures, etc. Again a common management will be key.



6. Suggested Collaboration Structure

Drawing lessons from other sciences operating at a global scale, we propose some strategies for management that seem to work well.

First, while it is essential that such a global management structure have the support of the experiment/detector/observatory managements, it should also not be too prescriptive and must have the buy-in of the technical teams. In the case of LHC and the WLCG, this mandate comes via a formal MoU between the experiments, the facilities, and the funding agencies. This MoU covers both the contributions of computing and storage resources, but also the agreements of how the collaboration will work, respective roles and responsibilities. In the LHC case this is a very formal MoU but implemented in a collaborative and non-confrontational way.

A simple management body (here referred to as a Management Board - MB) would oversee ongoing work and organise collaborative developments etc. Its members would be drawn from all of the stakeholders, the GW collaborations, the cyber infrastructures, and the computing facilities that are funded to provide resources to GW. It would call upon specific technical expertise as needed. It would meet regularly (e.g., monthly), and manage the ongoing operations, set up working groups for development, and provide the voice of the global GW collaboration in computing. The bullet points enumerated above would be its key tasks.

Such a body must report back to the stakeholders once or twice a year, for example. It should have scientific review from the experiments, but also report back to the funding agencies in some form. In this way, trust can be built, and new requests for resources will have a foundation that is already understood. It is also important to recognise that this feedback and review ensures that national interests are served. This body could in turn establish technical working groups that would report to it, focused on specific developments, software, operations, infrastructure, etc.

Once such a management structure is in place, a medium and long term plan for resource and funding needs should be constructed. This plan would be updated regularly as part of the stakeholder review, and while of course it will evolve and have many uncertainties, it will nevertheless provide guidance to the funding agencies and other resource providers. This long-term planning aspect is one of the important benefits of an organisation as suggested here.

This organization and initial plan should be put in place as early as possible in order to get buy-in, and if it is decided to have a formal agreement, the MoU process will take some time to implement. At the same time as this organisation is being set up, a first computing roadmap should be written, initially as a discussion document, but then leading into detailed technical plans. These initial steps will provide strong leadership in the computing activities for GW, and provide a foundation for long-term collaborative investment.

Recommendation:

- Set up now a computing collaboration for GW, with a structure of a management body and appropriate stakeholder review as discussed above. The collaboration could be bound by an MoU structure.
- Provide an initial resource plan (hardware and effort) with an outlook of a few years, to be updated yearly, and maintained as the current best outlook.

- Provide a computing technical strategy and roadmap, for discussion and later to become implementation plans.

7. Cyberinfrastructure, Resources, Facilities

Given that the hardware resources for 3G will be orders of magnitude larger than 2G, how will these resources need to be organized and managed? A centralized "walled garden" of homogeneous, dedicated GW clusters likely to be inefficient and prohibitively expensive. A diverse worldwide GW collaboration is unlikely to select a single external computing provider (commercial or otherwise). 3G computing optimization may require specialized hardware for different searches. The landscape of 3G computing resources will thus necessarily be both more heterogeneous and more distributed than that of the 2G era.

The larger organization of the GW community will also have a profound effect on the organization of computing. A highly-centralized collaboration with proprietary data requires and enables a more centralized computing organization, whereas a decentralized community of scientists with open data necessitates more loosely-coupled forms of organization. The implications of the larger GW community structure on computing management and infrastructure must be taken into account when considering the pros and cons of competing visions for future 3G GW communities.

An improved cyberinfrastructure will be needed to make a more complex, dynamic network of data analysis resources usable and robust for a distributed collaboration of scientists. This includes distributed scheduling, identity management and access control infrastructure, cybersecurity tools and services, data analysis software development and testing tools, resource accounting and reporting tools, low-latency alert and coordination infrastructure, as well as public data and code release infrastructure.

We expect increasing heterogeneity and complexity of computing platforms in the 3G era:

- of processing hardware; due to the opportunities for cost savings, data analyses may need to support multiple generations of CPUs, GPUs, MIC platforms and treat them each as distinct platforms. "Lowest common denominator" code capable of running on any platform may not be efficient enough.
- of providers — resources will include those internal to the project, partners & collaborators, institutional, regional/national, commercial, volunteer.
- of target operating systems and software environments. Containerization, etc. are tools to help mitigate this complexity but carry their own costs and are not a complete solution.
- of batch/queueing systems.
- of storage and network interfaces and capabilities.
- of policies for identity and access management, workflow prioritization.
- of accounting models and accounting systems
- of motivations and expectations — e.g., mutual scientific or strategic interest, public or scientific recognition, financial or other compensation, etc. — not all of which will be explicitly spelled out in MOUs, SLAs, or contracts.

8. People Resources

Two major challenges to computing in the 2G era have been uncertain and discontinuous funding streams for computing labor embedded within the collaboration groups outside the IFO laboratories, and a mismatch between existing job categories and new hybrid scientific computing roles. The need to professionalize software development and engineering, and support increasingly complex computing environments, demands more full-time professional computing expertise in side-by-side collaboration with domain scientists (as opposed to part-time volunteer/service work by scientists, or non-collaborative computing service providers).

LIGO's recent experience has shown that the return-on-investment of dedicated, "full-stack" computing optimization by a team of both GW science and computing experts is overwhelmingly positive, in terms of saved computing costs for expensive data analysis codes. The GW community should plan on making this investment for 3G. Data analysis efficiency improvements needed for 3G will require increased software development effort. This will require time from busy GW scientists and/or the hiring of more computing specialists. An investment in GW-fluent computing specialists can free up GW scientists to focus on GW science.

Long-term software development costs may also be higher for rapidly-evolving parallel hardware platforms (GPU, MIC, AVX512, FPGA, etc.) than for traditional CPUs, given that parallel programming interfaces are less stable targets. Most single-threaded CPU codes have run on new hardware with minimal modifications for more than 30 years. Data analysis on more distributed, non-dedicated computing grid/cloud platforms will require computing infrastructure investments. This will also require time from busy GW scientists (serving as a new "tax" on their research output) or the hiring of more computing specialists.

In total, 3G computing labor costs must go up by a factor of a few relative to existing 2G computing models. Notably, there may be potential and interest in broader collaboration in this area, e.g. with HL-LHC (and other HEP like DUNE) as well as SKA. Many of these are not traditional software development or IT roles that can be outsourced beyond the project — they are hybrids of domain science, research computing, consulting, software engineering, and distributed systems development and operations roles. These roles benefit enormously from institutional and project "memory" — large, distributed scientific projects pay dearly in time, money and quality when domain-specific experience and relationships are lost.

It is also difficult to recruit and retain career professionals on overlapping 1-3 year awards, and to find research funding for computing, which is not always "transformative" science in and of itself, but needed to enable transformative science. This is an old problem but will become more acute with the increasing need for these computing roles in the 3G era. Recognition and job security at the employer level (at universities and institutes) will be critical.

These concerns are common within other science communities, and could help to drive a common effort to recognise software and computing as scientific career tracks, as well as collaborate in a common lobby for more focused investment in software.

Recommendation: In concert with HEP and Astronomy communities, we recommend that the GW community strongly advocate to funding agencies to provide more stable funding streams and, along with

university and institute employers, define and recognize new roles for hybrid domain science / Computer Science staff dedicated to the computing challenges discussed in this report.

Recommendation: Long-term career opportunities must be provided for GW-fluent computing professionals to support domain scientists; the bulk of computing infrastructure and core software development effort cannot be effectively delivered in the 3G era by domain scientists or external (non-GW-fluent) computing professionals.

9. Global Collaboration, Other Sciences, Common Challenges

As the GW community prepares for the next generation of detectors and associated computing challenges, it is opportune that several other sciences in Astronomy and HEP are facing similar challenges. These include managing large scale global computing and data distribution, co-existing in scientific data centres with other sciences, and planning the evolution of computing infrastructures and software in order to benefit from evolving technology, whilst maintaining stability and reliability. In addition, new challenges of open data, data sharing and collaboration, and other policy-level concerns are coming to the forefront.

There are several dimensions to the potential for collaboration, with a number of opportunities:

- Collaboration between scientific collaborations on basic technology concerns: evolution of processors and the difficulty of obtaining the best performance, how to manage this in the long term with an outlook of continually changing hardware. This area includes also how to manage the change of technology (e.g. if tape were to disappear); the unpredictability of cost evolution of computing and storage, etc.; use of novel architectures and systems (HPC, HTC, etc.).
- Common software challenges, often in order to manage the previous point; here opportunities are many, with the HPC community internationally, and also with entities like the HSF, and software institutes in the USA and EU.
- Co-location of GW experiments in data centres together with other sciences. This can be very costly for the support in those centres unless there are some commonalities and synergy between the resources and services required by all of the experiments supported. In some cases the centre may mandate that the experiment use a certain type of resource.
- Collaboration with other Astronomy (CTA, Vera Rubin Observatory, SKA, etc.) and similar particle physics collaborations. Engagement across the board to explore synergies and common interests is useful and potentially hugely beneficial in driving towards common solutions.
- Collaboration with international cyberinfrastructures (OSG, EGI, EOSC, etc.) and how to benefit from the resources and services they provide, but also how to benefit from the funding opportunities that may come via those infrastructures. Often such funding will be expecting broad synergies and will not fund a specific experiment alone. Also such cyberinfrastructures can provide key collaboration points on policy, cyber security (policy and reaction), agreement on global coherence of federated Authentication and Authorization Infrastructure systems, and so on.
- Policy development and interaction with funding agencies from a broad scientific consensus. Common statements from a wide range of (similar) sciences (e.g. astronomy, astroparticle, GW, HEP, NP) could have a very big impact with funding agencies on common concerns. This would also help in a common strategy for provision of network bandwidth and capability, and areas such as large scale use of HPC, which may require a different access and usage policy for our communities compared to their traditional user base.
- Opportunity to engage with global collaboration on scientific computing infrastructure (see below).

In many of the above, bottom-up collaboration on topics of common interest is often the most direct and beneficial. However, a management view is needed to ensure coherence and to avoid duplicate or competing efforts, especially with limited resources. Thus the GW community would be well advised to work together with their scientific and international peers in order to exploit such opportunities. The management organisation proposed in an earlier section would take the lead on organising this.

The HEP community and others, has a working group looking into the evolution of the cost of hardware (CPU, disk, tape) in preparation for the HL-LHC upgrades on the 2026 timescale. Lessons from that study are relevant also to GW. CERN and other institutes have tracked costs over more than 15 years, and use recent trends to set expectations for the evolution over the coming years. Simple "Moore's law" predictions no longer seem to be valid and are outweighed by market forces and cost management by the relatively few hardware producers. It is a fact that there are now only a handful of chip fabrication companies, only two HDD manufacturers and that tape drive technology is a monopoly with only IBM building drives and only two media producers.

For the past several years the HEP community has used as guidance a 20% per year improvement in performance for CPU and disk storage capacity for a constant cost, to predict the affordable capacity of complete server systems (not the raw disk space or raw CPU performance). For CPU this applies to standard x86-like processors. For tape storage capacity the capacity increase per year was estimated to be somewhat more, around 25%/yr.

In the last 2 years however we have observed a worrying trend, that there are large fluctuations in the market prices of key components (in particular RAM), and shortages of supply, and that the effective growth is now only around 10% per year for constant cost. In addition all of these markets are now in the hands of very few large manufacturers, who control the markets, and thus prices. The upshot of this is that it is currently very difficult to make a realistic estimate of the price/performance evolution for the coming years. This is a significant concern when trying to understand the overall cost of computing and storage for future experiments.

Recommendation: Identify (and if necessary, establish) the right forums for the GW community, HEP, and Astro communities to regularly and systematically discuss the overlap in our computing problems, share information, and plan specific technical collaborations where useful. For example:

- Coordination of Distributed Scheduling, Computing Security, and Identity and Access Management efforts between HL-LHC and LIGO/Virgo.
- Computing optimization approaches, techniques and tools.
- Software engineering tools and technologies.

Some of this interaction is already happening (e.g., the ESCAPE project), but the GW community should expand those efforts and have a stronger voice in forums where these matters are discussed. The 3G gravitational-wave community cannot afford to reinvent wheels.

The high-luminosity LHC experiments face similarly-daunting computational scaling problems over the same timeframe, but they have an order of magnitude larger starting requirements — they are helping to blaze a trail in optimization and distributed computing infrastructure which we can and should follow and collaborate on. Being explored are Exabyte-scale capabilities for data management, able to serve data in an efficient way to heterogeneous and globally distributed compute resources including HTC clusters, HPC clusters, and commercial clouds. We foresee the requirement to support a variety of processors ranging from CPU, accelerators (e.g., GPU, MIC), FPGA and other innovative architectures.

CERN/WLCG is basing its strategy on 3 areas of investment:

- A federated data cloud ("data lake") capable of the long term curation and serving of exabyte-scale data, complemented with tools and services to allow experiments to manage that data and content delivery tools to efficiently serve the data to remote compute resources. This data lake would also provide an effective mechanism for hosting and serving data to the broader community including public

open access, potentially with the inclusion of commercial interests.

- A significant investment in application software skills. The ability to make use of the full set of available resources in coming years requires development of software skills and the ability to move applications to appropriate compute architectures in an agile way as they evolve. The infrastructure will also depend on a clearinghouse of software tools and services for science communities. The HEP community has initiated the HEP Software Foundation (HSF) as the vehicle through which to address the long term software investment and also to provide the locus for a software tools clearinghouse. It is not necessarily specific to HEP, and could also be strengthened with the inclusion of other science communities.
- Networks to support the infrastructure and the use of appropriate technologies to manage data management both within the data lake and in serving data externally. WLCG has developed an overlay network (LHCONE) that the National Research and Education Networks use to manage the science data traffic. This concept has proven extremely versatile and is now used by many more science communities.

In addition to the above infrastructure developments, CERN has pioneered aspects of open access and open data, including the Zenodo publishing platform and the CERN open data portal, as well as the storage infrastructure underpinning these. These could be interesting components of a future service for the GW community. CERN also provides "CERNBox" - a DropBox-like open source solution that is widely used in the HEP community.

To complete the picture, CERN has the "openlab" concept, which is a vehicle for collaboration with industry, disconnected from any expectation of procurement. This provides a mechanism to fund and staff projects of joint interest between CERN and various industrial partners. In recent years this has expanded to include other research institutes. This is another avenue of potential technology transfer to the 3G GW community, enabling the investigation of new computing technologies in collaboration with the companies, and to provide access to their expertise.

As noted in our discussion of core software, optimal performance of that software will depend on carefully targeting a diversity of hardware, and hardware acquisition should be based onoff of systematic benchmarking. Achieving both of these goals can be accelerated by careful involvement of industry partners. Though the success of CERN's openlab is mixed, it provides a useful starting point for a gravitational-wave model. Ideally, such a collaboration will allow early access to testing platforms to inform purchase decisions, training of scientists through fellowships or internships, and engagement on industry-supported coding platforms and software development tools.

From the CERN viewpoint collaborating with the GW community on some or all of these aspects and the related technologies is of great interest and could aid in developing a broad toolset useful to a broad scientific base. In the recent European Strategy for Particle Physics meeting (May 2019), a broad collaboration of HEP with the Gravitational Wave community was seen as strategic, and very much encouraged from both communities.

10. Collaborations with Industry

Collaboration with industry partners can be beneficial in certain areas. The CERN openlab experience was that, in R&D topics, such a collaboration can be extremely useful to both partners, and it is also a very good means for training and engaging young scientists and computing/software engineers, and for funding summer students. It should be recognised that management of such projects and people also potentially takes effort away from core tasks. Thus it must be clear what the real benefits and outcomes of such collaboration will be. Experience with CERN openlab shows that an institute would be the point of contact with the potential industry partners to manage all of the legal and IP concerns. The GW community would need to channel this type of engagement through one or two large institutes, or potentially collaborate with CERN on it.

Use of industrial partnerships to obtain services is potentially a risk. We must consider long-term sustainability of the computing infrastructure. Negotiating a beneficial cost for a service, which then requires specific adaptation may not be a good choice in terms of sustainability, as when the need from our side changes, or the industry partner policies change, the solution may no longer be cost effective, and requires significant effort to replace. There are many examples of this. Also this type of partnership brings very little benefit to the science community, and risks embedding proprietary solutions.

Recommendation: Identify a forum, in collaboration with other Astronomy and HEP organisations, as a means by which to work effectively with industry partners on well identified key projects.

11. Recommendations

The GW community is facing a number of challenges as it moves towards the third generation of experiments. It is clear that a global collaboration between the various experiments will be essential, and that collaboration must apply to the core software and computing activities as well, or the full scientific potential of the instruments will not be realized.

Resources, both physical and human, will be limited, and efforts must be taken to organise the community to be able to use current experience to find new algorithms and computing solutions, exploit synergies, and find new cost efficient workflows, not just between the various GW collaborations, but also with other sciences (Astronomy, Astro-particle, Particle Physics) that are dealing with similar problems, and relying on very similar infrastructures and resources to address them.

In order to do this effectively there are a number of high level recommendations that need to be addressed in order to prepare for the future. These are the following:

- **Recommendation:** Set up now a computing collaboration for GW, with a structure of a management body and appropriate stakeholder review as discussed above. The collaboration could be bound by an MoU structure.
- **Recommendation:** Provide an initial resource plan (hardware and effort) with an outlook of a few years, to be updated yearly, and maintained as the current best outlook.
- **Recommendation:** Provide a computing technical strategy and roadmap, for discussion and later to become implementation plans.
- **Recommendation:** Organize mock-data challenges for 3G data analysis to systematically test competing techniques and ensure that the best approaches are robust and production-ready.
- **Recommendation:** In concert with HEP and Astronomy communities, we recommend that funding agencies provide more stable funding streams and, along with university and institute employers, define and recognize new roles for hybrid domain science / Computer Science staff dedicated to the computing challenges discussed in this report.
- **Recommendation:** Identify (and if necessary, establish) the right forums for the GW community, HEP, and Astro communities to regularly and systematically discuss the overlap in our computing problems, share information, and plan specific technical collaborations where useful.
- **Recommendation:** Identify a forum, in collaboration with other HEP and Astronomy communities, as an appropriate means by which to work with industry partners on well identified key projects.

Bibliography

[1] Sathyaprakash, B. and Kalogera, V. and the GWIC 3G Science Case team. The Next Generation Global Gravitational Wave Observatory: The Science Book, 2020. 1

[2] Stefano Bagnasco. Virgo and gravitational waves computing in europe. In *Proceedings of the 24th International Conference on Computing in High Energy & Nuclear Physics, Adelaide, South Australia, 4-8 November, 2019*, 2020. to appear. 2

[3] Derek Weitzel, Brian Bockelman, Duncan A. Brown, Peter Couvares, Frank Würthwein, and Edgar Fajardo Hernandez. Data access for LIGO on the OSG. *CoRR*, abs/1705.06202, 2017. 2

[4] Pia Astone, Alberto Colla, Sabrina D'Antonio, Sergio Frasca, and Cristiano Palomba. Method for all-sky searches of continuous gravitational wave signals using the frequency-hough transform. *Phys. Rev. D*, 90:042002, Aug 2014. 5

[5] Helge Meinhard, Bernd Panzer-Steindel, Servesh Muralidharan, Peter Wegner, Christopher Henry Hollowell, Michele Michelotto, Andrea Sciabá, Eric Yen, German Cancio Melia, Martin Gasthuber, Shigeki Misawa, and Edoardo Martelli. Trends in computing technologies and markets. In *Proceedings of the 24th International Conference on Computing in High Energy & Nuclear Physics, Adelaide, South Australia, 4-8 November, 2019*, 2020. to appear. 6

[6] Tito Dal Canton, Alexander H. Nitz, Andrew P. Lundgren, Alex B. Nielsen, Duncan A. Brown, Thomas Dent, Ian W. Harry, Badri Krishnan, Andrew J. Miller, Karl Wette, Karsten Wiesner, and Joshua L. Willis. Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors. Phys. Rev. D, 90(8):082004, October 2014. 7

[7] Samantha A. Usman, Alexander H. Nitz, Ian W. Harry, Christopher M. Biwer, Duncan A. Brown, Miriam Cabero, Collin D. Capano, Tito Dal Canton, Thomas Dent, Stephen Fairhurst, Marcel S. Kehl, Drew Keppel, Badri Krishnan, Amber Lenon, Andrew Lundgren, Alex B. Nielsen, Larne P. Pekowsky, Harald P. Pfeiffer, Peter R. Saulson, Matthew West, and Joshua L. Willis. The PyCBC search for gravitational waves from compact binary coalescence. *Classical and Quantum Gravity*, 33(21):215004, November 2016.

[8] Xiangyu Guo, Qi Chu, Shin Kee Chung, Zhihui Du, and Linqing Wen. Acceleration of low-latency gravitational wave searches using Maxwell-microarchitecture GPUs. *arXiv e-prints*, page arXiv:1702.02256, February 2017.

[9] D. Wysocki, R. O'Shaughnessy, Jacob Lange, and Yao-Lung L. Fang. Accelerating parameter inference with graphics processing units. Phys. Rev. D, 99(8):084026, April 2019. 7

[10] S. Klimenko, I. Yakushin, A. Mercer, and G. Mitselmakher. A coherent method for detection of gravitational wave bursts. *Classical and Quantum Gravity*, 25(11):114029, June 2008. 7

[11] Surabhi Sachdev, Sarah Caudill, Heather Fong, Rico K. L. Lo, Cody Messick, Debnandini Mukherjee, Ryan Magee, Leo Tsukada, Kent Blackburn, Patrick Brady, Patrick Brockill, Kipp Cannon, Sydney J. Chamberlin, Deep Chatterjee, Jolien D. E. Creighton, Patrick Godwin, Anuradha Gupta, Chad Hanna, Shasvath Kapadia, Ryan N. Lang, Tjonnie G. F. Li, Duncan Meacher, Alexander Pace, Stephen Privitera, Laleh Sadeghian, Leslie Wade, Madeline Wade, Alan Weinstein, and Sophia Liting Xiao. The GstLAL Search Analysis Methods for Compact Binary Mergers in Advanced LIGO's Second and Advanced Virgo's First Observing Runs. *arXiv e-prints*, page arXiv:1901.08580, January 2019. 7

[12] Elena Cuoco, Jade Powell, Marco Cavaglià, Kendall Ackley, Michal Bejger, Chayan Chatterjee, Michael Coughlin, Scott Coughlin, Paul Easter, Reed Essick, Hunter Gabbard, Timothy Gebhard, Shaon Ghosh, Leila Haegel, Alberto Iess, David Keitel, Zsuzsa Marka, Szabolcs Marka, Filip Morawski, Tri Nguyen, Rich Ormiston, Michael Puerrer, Massimiliano Razzano, Kai Staats, Gabriele Vajente, and Daniel Williams. Enhancing Gravitational-Wave Science with Machine Learning. *arXiv e-prints*, page arXiv:2005.03745, May 2020. 7

[13] Daniel George and E. A. Huerta. Deep neural networks to enable real-time multimessenger astrophysics. Phys. Rev. D, 97(4):044039, February 2018. 7

[14] Daniel George and E. A. Huerta. Deep Learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data. *Physics Letters B*, 778:64–70, March 2018.

[15] Hunter Gabbard, Michael Williams, Fergus Hayes, and Chris Messenger. Matching Matched Filtering with Deep Networks for Gravitational-Wave Astronomy. Phys. Rev. Lett., 120(14):141103, April 2018.

[16] Timothy D. Gebhard, Niki Kilbertus, Ian Harry, and Bernhard Schölkopf. Convolutional neural networks: A magic bullet for gravitational-wave detection? Phys. Rev. D, 100(6):063015, September 2019.

[17] Plamen G. Krastev. Real-time detection of gravitational waves from binary neutron stars using artificial neural networks. *Physics Letters B*, 803:135330, April 2020.

[18] Marlin B. Schäfer, Frank Ohme, and Alexander H. Nitz. Detection of gravitational-wave signals from binary neutron star mergers using machine learning. *arXiv e-prints*, page arXiv:2006.01509, June 2020. 7

[19] Christoph Dreissigacker and Reinhard Prix. Deep-learning continuous gravitational waves: Multiple detectors and realistic noise. Phys. Rev. D, 102(2):022005, July 2020. 7

[20] Christoph Dreissigacker, Rahul Sharma, Chris Messenger, Ruining Zhao, and Reinhard Prix. Deep-learning continuous gravitational waves. Phys. Rev. D, 100(4):044009, August 2019. 7

[21] Stephen R. Green, Christine Simpson, and Jonathan Gair. Gravitational-wave parameter estimation with autoregressive neural network flows. *arXiv e-prints*, page arXiv:2002.07656, February 2020. 7

[22] Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy. *arXiv e-prints*, page arXiv:1909.06296, September 2019.

[23] Alvin J. K. Chua and Michele Vallisneri. Learning Bayesian Posteriors with Neural Networks for Gravitational-Wave Inference. Phys. Rev. Lett., 124(4):041102, January 2020. 7

[24] Daniel George, Hongyu Shen, and E. A. Huerta. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning. Phys. Rev. D, 97(10):101501, May 2018. 7

[25] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. K. Katsaggelos, S. L. Larson, T. K. Lee, C. Lintott, T. B. Littenberg, A. Lundgren, C. Østerlund, J. R. Smith, L. Trouille, and V. Kalogera. Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6):064003, March 2017.

[26] Massimiliano Razzano and Elena Cuoco. Image-based deep learning for classification of noise transients in gravitational wave detectors. *Classical and Quantum Gravity*, 35(9):095016, May 2018. 7

[27] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 473–479. MIT Press, 1997. 7

[28] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 7

[29] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 7

[30] Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404. Morgan Kaufmann, 1990. 7