Masters Dissertation

IMPLEMENTATION OF N-DETECTORS IN A GRAVITATIONAL WAVE DETECTION PIPELINE

Thomas Hill Almeida (21963144)*

Supervisors: Prof. Linqing Wen†, Qi Chu†

2020-10-16

*Software Engineering, University of Western Australia

†Department of Physics, University of Western Australia

Contents

1 Introduction

Gravitational waves have been postulated to exist since Albert Einstein's publication of his general theory of relativity, as massive accelerating objects would cause 'ripples' in the curvature of spacetime [1]. Direct detection of gravitational waves, however, remained beyond the reach of the scientific community until 2015, when the Laser Interferometric Gravitational-Wave Observatory (LIGO [see 2]) reported an observation on the 14th of September [3, 4].

Due to their design, the detectors in use for gravitational wave detection have a significant amount of noise from other sources, whilst the gravitational waves themselves have very weak signals. As such, a large amount of data processing must be done to the outputs produced by the detectors in order to filter and extract any possible gravitational waves. These data processors are known as "pipelines", and have historically been created by research groups that are a part of the LIGO Scientific Collaboration (LSC [see 5]), and are used throughout observation runs for real-time data analysis.

The Summed Parallel Infinite Impulse Response (SPIIR [see 6]) pipeline, based on the SPIIR method originally implemented by Shaun Hooper in 2012, uses a number of IIR (infinite impulse response) filters to approximate possible gravitational wave signals for detection [7]. [8] states that the output of the ith IIR filter can be expressed with the equation:

$$y_k^i = a_1^i y_{k-1}^i + b_0^i x_{k-d_i}, \tag{1}$$

where a_1^i and b_0^i are coefficients, k is time in a discrete form and x_{k-d_i} denotes input with some time delay d_i. After summing the output of the filters, the resulting signal undergoes coherent post-processing (see section 5.1 and [9, chapter 4]) to determine the likelihood of an event having occurred.

The pipeline is currently thought to be the fastest of all existing pipelines, is the only pipeline that implements coherent search, and has participated in every observation run since November 2015, successfully detecting most events that were seen in more than one detector.

The SPIIR pipeline uses GStreamer, a library for composing, filtering and moving around signals, in addition to the GStreamer LIGO Algorithm Library (`gstlal`) [10]. After receiving data from the detectors, the pipeline performs data conditioning and data whitening, followed by the usage of the IIR filters. The data is then combined for post-processing, where events are given sky localization and then inserted into the LIGO event database [8].

The structure of the SPIIR pipeline can be seen in figure 1.

1.1 Research Aims

At the time of the start of this research, the SPIIR pipeline supported the use of two or three detectors for gravitational wave detection — the two American LIGO detectors and the Italian Virgo detector — although additional interferometers are likely to be introduced soon. This presents several issues with the existing pipeline design.

As with many of the other gravitational wave detection pipelines, providing support for additional

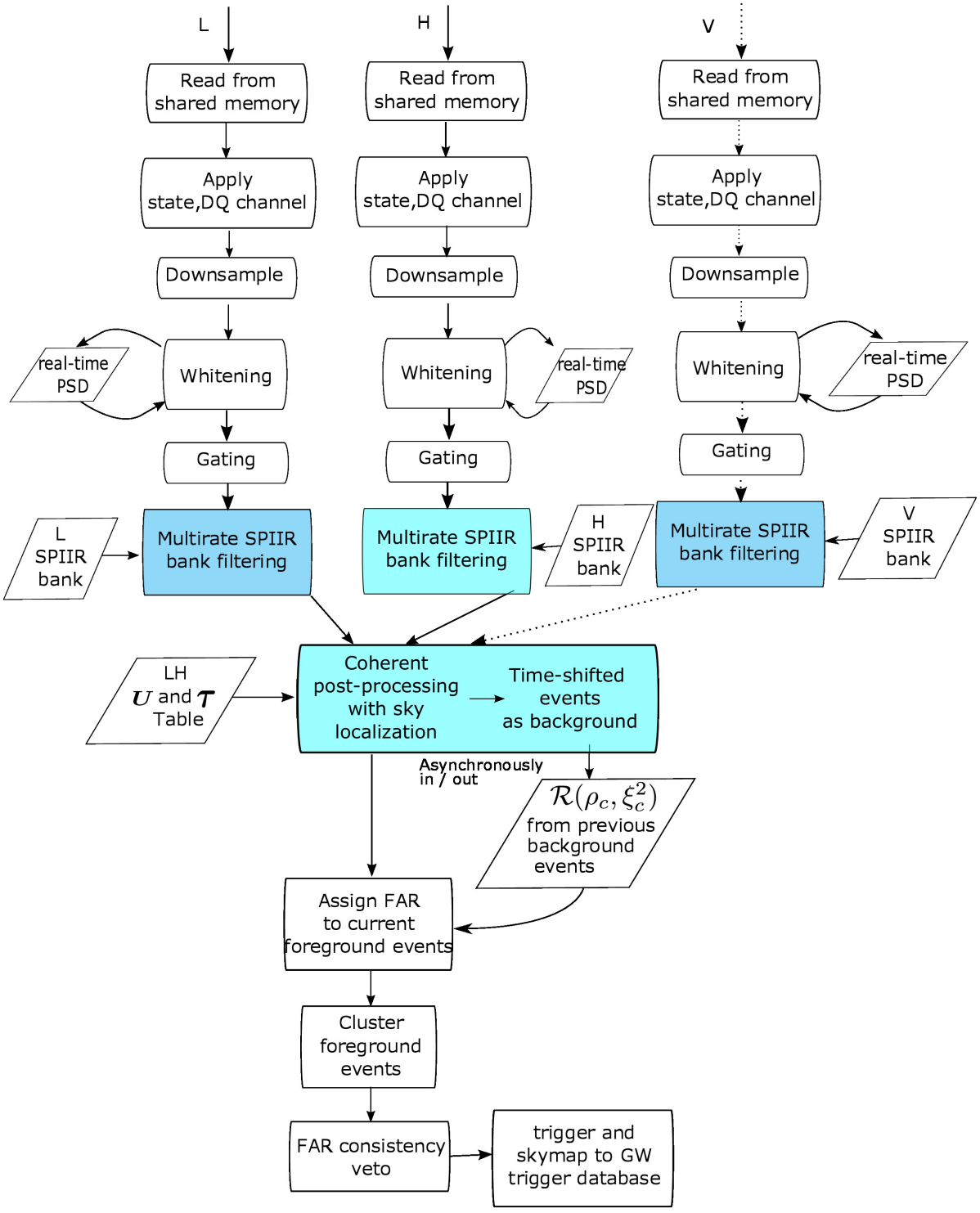

Figure 1: The structure of the SPIIR pipeline

detectors is a significant undertaking for the development team, with many hours of work and testing that need to be completed. As the number of available interferometers continues to grow, development work that could be spent on improving the optimisation, precision, or accuracy of the pipeline would instead have to be spent allowing for those detectors to be used.

Thus, this dissertation aims to layout the design and implementation work done to provide the SPIIR pipeline with the ability to support any number of detectors (N-detectors).

Section 2 shall explore the existing literature on the implementation of N-detectors in other gravitational wave detection pipelines, in addition to exploring CUDA and complexity analysis, two tools that will be used in the analysis of the final design. Section 3 will look at the existing pipeline structure and its deficencies for implementing N-detectors. It will also determine any constraints a design for implementing N-detectors would have, and will provide a framework for evaluating the success of the resulting design. The final design will be presented in section 4, and will explore the implementation details of the design and how it interacts with the existing programming interface that SPIIR provides. A discussion about the implications of this design will be done in section 5, with additional research on those implications being presented there. Finally, section 6 will provide some suggestions for future research that can be based on this research.

2 Literature Review

2.1 N-detector work in other gravitational wave detection pipelines

There are, of course, other gravitational wave detection pipelines that may also have to consider the issue of dealing with a growing number of detectors. As we are in the process of designing a new architecture to allow for any number of detectors to be used, it is well worth examining the methods that the other pipelines may use to shape our own design.

Most detection pipelines use a "coincidence" search to determine whether a gravitational wave event has occurred. Coincidence search, as defined by [9, chapter 3], is a process which includes finding candidates for gravitational waves from individual detectors, identifying temporal coincidences and producing measures to rank candidate events. In more simple terms, a coincidence search considers events which can be seen in a single detector, and then sanity checks that other detectors may have seen them at the same time.

In contrast, the SPIIR pipeline uses a "coherent" search to determine whether a gravitational wave event has occurred, using the maximum likelihood ratio principle to consider specific parameters of a potential signal [9, chapter 4]. As such, it is unlikely that all of the principles for dealing with additional detectors may translate directly to being able to be used for the SPIIR pipeline, as the method of search differs.

PyCBC is one of the most well known toolkits for gravitational wave astronomy, and was one of the pipelines used in the original 2015 gravitational wave detection [11]. From an examination of the codebase of the latest version of PyCBC [12, October 2020], it can be observed that the codebase itself makes no direct mention of detectors — instead it provides a generic `Detector` class as a wrapper around

LALSuite's [13] `LALDetector` structure for validation, which in turns provides utilities for returning information about the detector as well as methods for getting readings from it.

This allows PyCBC to provide an entirely generic gravitational wave searching algorithm library for any input detectors, although the algorithm only supports using two detectors at a time, providing that the detectors are in the LALSuite library. Thus, we can note two things; for PyCBC to allow for additional detectors to be used, they simply update their dependency on the LALSuite library, and; PyCBC doesn't quite support N-detectors in the sense described in section 1.1, instead it allows for any supported detectors to be used in a two-detector search. This means that whilst we cannot use PyCBC when considering how to support support any number of detectors within a gravitational-wave searching algorithm nor its outputs, we can use PyCBC to consider a programming interface with which to support other detectors.

GstLAL is gravitational wave detection library, that exposes components of LALSuite [13] as GStreamer elements for use in other analysis pipelines — including the SPIIR pipeline — as well as providing its own pipeline for processing raw signals from detectors into lists of gravitational wave candidates [14]. The GstLAL's pipeline hard-codes the detector names into both its inputs and its outputs, however the algorithm used for detection itself is actually generic on which detectors are used [15, 16].

This means that the process of adding support for detectors involves changing a number of different files, as well as modifying several internal data structures [17] — which means that whilst GstLAL does not support N-detectors.

Thus any work done to design N-detector support for the SPIIR pipeline will be novel work.

2.2 CUDA

CUDA [18] is an extension of the C++ programming language created by NVIDIA that allows for the development of GPU-accelerated applications. In [8], the SPIIR pipeline had multiple components rewritten in CUDA to take advantage of the high number of simultaneous threads available compared to CPUs. As such, it is worth understanding the computational model of CUDA for the analysis of the SPIIR pipeline.

In CUDA, each individual sequence of instructions being executed is called a *thread*. By its nature, a highly-parallelised environment such as GPUs will run many individual threads, which are partitoned into *warps*, a group of (typically 32) threads. Warps are the smallest unit that GPUs schedule, and all threads in a warp must execute the same instruction – although each thread maintains its own instruction pointer and can branch independently from the warp at a small performance cost. The performance cost of branching within a warp means that a major optimization that does not affect computational complexity in CUDA can be simply reducing the number of branches. Warps are further organised into thread blocks, which contain a small amount of fast memory shared between the threads in the block. Blocks in CUDA are typically executed on the same Simultaneous Multiprocessor (SM). The CUDA Programming Guide ([19]) states that the number of blocks and warps that can reside and be processed together on an SM depends on the number of registers and shared memory available on the SM, as well as on a CUDA defined maximum number of blocks and warps.

For the purpose of actual time-based computation, the maximum number of threads that can run at

any given time is determined by a few factors of the CUDA runtime; the maximum number of resident warps per SM; the maximum number of resident threads per SM; the number of 32-bit registers per thread; the number of 32-bit registers per SM; the number of 32-bit registers per thread block; and the amount of shared memory in each of those divisions. Thus, one major determining factor in any speed-up given by a CUDA operation can be determined by the ability to split the workload across threads and thread blocks such that the number of registers and used memory is well balanced across threads.

2.3 Parallelised Complexity Analysis

As the pipeline changes to accommodate additional detectors, it is important that the impact that this has on the pipeline's runtime is considered. The SPIIR pipeline is designed to be as low latency as possible, and an asymptotic complexity analysis of components that may be impacted by the addition of new detectors allows for the measurement of the potential runtime cost of doing so. The SPIIR pipeline has been parallelised using CUDA [8], and thus determining the asymptotic complexity of components of the pipeline requires different considerations to that of a sequential program.

According to [20], the theoretical efficiency of a multi-threaded or parallelised algorithm can be measured using the metrics of 'span', 'work', 'speed-up' and 'parallelism', all of which should be considered in the context of a directed acyclic graph (DAG) of operations in the algorithm. The ***work*** of a parallelised computation is the total time to execute the entire computation sequentially on a single processor, and can be found by summing the total work of every vertex in the DAG. An example of ***work*** for a merge-sort like algorithm can be seen in figure 2, which has a work of $O(N \log N)$. In comparison, the ***span*** of a parallelised computation is the maximum time taken to complete any path in the DAG. An example of ***span*** for a merge-sort like algorithm can be seen in figure 3, which has a work of $O(N)$. It should be noted that the actual running time of a parallelised computation also depends on the number of processors available for computation and how they are allocated to perform different tasks in the DAG, and thus denoting the running time of parallelised computation on P processors as T_P is also common practice. This leads to work being denoted at T_1 (the time taken to run on a single processor) and span being denoted as T_∞ (the time taken on an infinite number of processors). Another helpful metric is ***speed-up***, which shows how the algorithm scales with additional processors as $S_P = \frac{T_1}{T_P}$. We can also then define ***parallelism*** as the maximum possible speed-up on an infinite number of processors, and thus as $p = \frac{T_1}{T_\infty}$.

Using the above definitions, we can re-derive several laws that provide lower bounds on the running time of T_P.

In one step, a computer with P processors can do P units of work, and thus in T_P time can perform PT_P units of work. As the total work to be done as per above is T_1, the ***work law*** states that [20]:

$$T_P \geq \frac{T_1}{P}. \tag{2}$$

It is also evident that a computer with P processors cannot run any faster than a computer with an infinite number of processors, as the computer with an infinite number of processors can emulate a computer with P processors by using a subset of its processors, leading to the ***span law*** [20]:

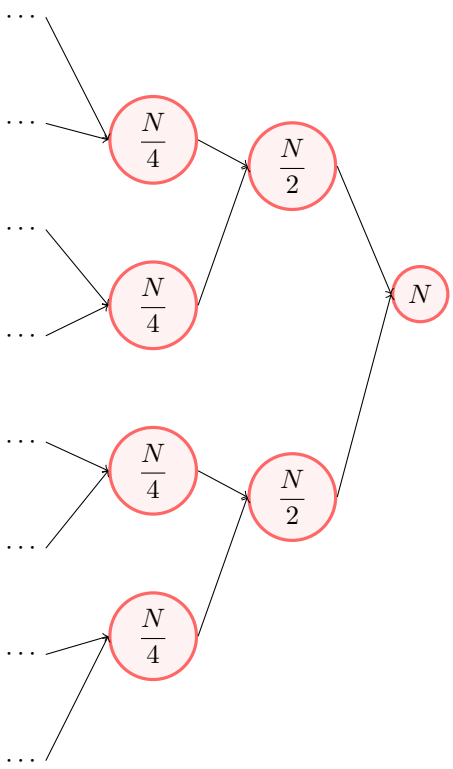

Figure 2: An example of calculating work for a merge-sort like algorithm. Summed nodes are in red.

$$T_P \geq T_\infty. \tag{3}$$

It is also useful to use the metrics of 'cost' and 'efficiency' when analysing parallel algorithms [21]. The ***cost*** of a parallel algorithm is minimised when all of processors are used at every step for useful computation and thus can be defined as $C_P = P \times T_P$. ***Efficiency*** is closely related to cost and describes speed-up per processor and can be defined as:

$$e_P = \frac{S_P}{P} = \frac{T_1}{C_P}. \tag{4}$$

Another helpful theorem for analysis is ***Brent's Theorem***, which states that for an algorithm that can run in parallel on N processors can be executed on $P < N$ processors in a time of approximately [22]

$$T_P \leq T_N + \frac{T_1 - T_N}{P}. \tag{5}$$

This can be approximated with the upper bound of $O(\frac{T_1}{P} + T_N)$ [21].

Determining the span, work, parallelism, efficiency and cost, and examining the application of Brent's theorem to the computations at hand will allow us to analyse the computational complexity of the SPIIR pipeline.

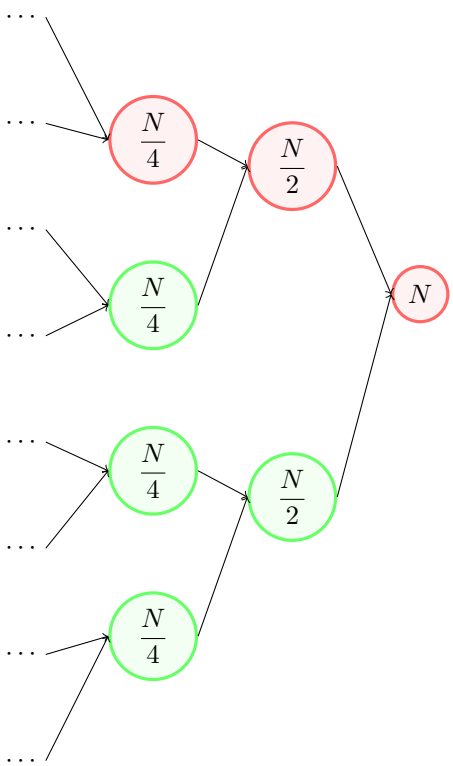

Figure 3: An example of calculating span for a merge-sort like algorithm. Summed nodes are in red.

3 Design process

3.1 Requirements

3.2 Employed tools

3.3 Relevant code

3.4 Evaluation criteria

4 Final Design

4.1 Patches

4.2 Testing

4.3 Evaluation

5 Discussion

5.1 A complexity analysis of the parallel post-processing of the SPIIR pipeline

Coherent post-processing was introduced in [9] by Qi Chu et al as an alternative to the coincidence post-processing used by all other pipelines (see section 2.1). [9] states that the multi-detector maximum log likelihood ratio to be equal to the coherent signal to noise ratio ρ_c^2, which can be expressed as:

$$\rho_c^2 = \underset{\max\{A_{jk},\theta,t_c,\alpha,\delta\}}{\ln \mathcal{L}_{NW}}, \tag{6}$$

where A_{jk} describes the relative inclination of the detector to the source, θ is the mass of the source, α and β are the sky directions of the source and $\mathcal{L}_{NW}$ is the network log likelihood network.

In [9], the computational cost of the coherent search is estimated to be $O(2N_d^3 N_m N_p)$, where N_d is the number of detectors, N_m is the number of sets of IIR filters (called templates), and N_p is the number of potential sky locations. Further optimizations were made to the pipeline in 2018 [8], including moving to using GPU acceleration. Whilst [8] discusses a number of constant time optimizations made to the pipeline, the computational cost of the overall process is not discussed. In addition, [8] parallelised the pipeline, leading to additional potential changes to the potential overall cost.

5.1.1 Motivation

5.1.2 Maximum element reduction

One of the more common operations in the SPIIR pipeline is the concept of a “maximum element reduction”. Reduction is the idea of taking some array of data and producing a single summary output from that array, whether it is the total sum of the array or the maximum value of the array and its index in the array as it is in this case.

[23] discusses the computational complexity of reduction algorithms in a parallelised context, noting that the best complexity according to Brent's Law is $O(\frac{N}{\log N})$ threads each doing $O(\log N)$ sequential work, resulting in a total overall cost of $O(\frac{N}{\log N} \times \log N) = O(N)$.

We can note from our own analysis, that the process of reduction can be parallelised by the use of a binary tree of operations, where each vertex in the binary tree combines the results of the two parent vertices. In the case of determining the maximum of two numbers, each vertex is identical in the amount of work done, and thus we can determine each vertex to be a unit of work. As there are N elements in the original array, we can note that the height of the binary tree is $\log N$, and each level of the binary tree has $N_l/2$ vertices, thus the total number of vertices in the binary tree is $\sum_{i=0}^{\log N} 2 \times i = N$. Using this information, we can determine that the ***work*** of a parallelised reduction is $T_1 = O(N \times 1) = O(N)$, and that the ***span*** of the reduction is $T_\infty = O(\log N \times 1) = O(\log N)$. Thus, the ***parallelism*** of the reduction is:

$$p = \frac{N}{\log N}.$$

Using the span and work laws, we can observe that any algorithm using the above method is bounded by the inequalities $O(\log N) \leq T_P, \frac{O(N)}{P} \leq T_P$. This means that best possible time complexity with P processors is $O(\log N)$ (equation 3). We can determine the minimum number of processors required to achieve this runtime using the formula $T_P = O(\log N) = \frac{O(N)}{P}$, which can be rearranged to

$$P = \frac{N}{\log N},$$

thus the time complexity cannot improve past $P = N/\log N$ processors. We can also observe that using $P = N/\log N$ processors gives a ***cost*** of $C_P = N$, which is identical to the sequential algorithm.

Functions that include maximum element reduction will be denoted for clarity with $M(x)$, where x is the size of the array being reduced.

5.1.3 Determining the number of samples over a signal-to-noise threshold

The coherent post-processing in SPIIR determines the number of samples over a signal-to-noise (SNR) threshold in order to not do more work than is required. The function that is used for determining the number of samples over the threshold (`peaks_over_thresh`) is a sequential algorithm that runs on the CPU, and shall be analysed as such, although there is an alternative GPU-based implementation that is not used.

Initially, the function performs a maximum element reduction to get the maximum SNR from the combined IIR filters (templates) for each sample. Recalling from section 5.1.2 that for maximum element reduction $T_1 = O(N)$, and that this operation is performed S times, where S is the number of samples, we can determine that this initial reduction has a time complexity of $O(ST)$, where T is the number of templates.

The function then determines the maximum SNR across the templates found from the previous step by stepping through every combination of samples and removing SNR samples that are using the same

template and have a lower SNR, resulting in a step with a time complexity of $O(S^2)$.

The function then determines the maximum overall SNR for the input samples ($O(S)$) and cycles through every maximum SNR to cluster maxima that are close together to be a single combined maximum. The number of maxima is bounded by ($O(\min\{S,T\})$) as there cannot be more maxima than there are samples or templates.

This gives the overall function a time complexity of $O(ST + S^2 + S + \min\{S,T\})$, which can be reduced to the dominating terms of:

$$O(ST + S^2).$$

5.1.4 Transposing the input matrices

The full post-processing function requires that the input matrix is transposed for better memory access such that each row is a different template, and each column is a different sample. To transpose the matrix, the GPU function `transpose_matrix` is used, thus this should be analysed as a parallel algorithm.

The algorithm in use works by breaking the original array into tiles of size 32×32, and then inserting the transpose of the tile into an output array. The tiles are further broken down eight processors per row, so each thread does four copies. We can conceptualise this as a DAG by observing that each tile does not depend on any other tile to be completed, and that each tile is composed of 32×8 interdependent processors, each doing 4 units of work.

Using this observation, we can see that the ***span*** of the algorithm is $T_\infty = (32 \times 8) \times 4 = O(1024) = O(1)$, and the ***work*** is $T_1 = O(ST)$, where S is the number of samples and T is the number of templates. Thus, the ***parallelism*** of the transpose is $p = ST$.

Using the span and work laws (equations 3 and 2), we can observe that the above method is bounded by the inequalities $O(1) \leq T_P, \frac{O(ST)}{P} \leq T_P$. Thus it can be determined that the best possible time complexity with P processors is bounded by ratio of available processors to the size of the transposed matrix (the work law). This gives the function an overall time complexity of:

$$O(\frac{ST}{P}).$$

5.1.5 Determining the coherent correlation and statistical value of data points

The scoring metric of different templates and times is determined using coherent correlation and determining their statistical value using a chi squared-based distribution. These scoring metrics are performed using the GPU function `ker_coh_max_and_chisq_versatile`, and thus should be analysed as a parallelised function.

In this function, each block looks at a different SNR maximum (as discussed in section 5.1.3) and splits the threads within the blocks for operations on that peak.

Determining the sky direction of the SNR maximum

Initially, each thread within a block looks at a different sky direction and determines the total signal-to-noise ratio (SNR) by summing the SNR of each of the detectors at that given sky direction with the relevant detector arrival time offsets. The time complexity for the calculation of SNR for a given time

offset is $O(D + D^2)$, where D is the number of detectors. The maximum SNR for all the sky directions is then spread across each warp and placed into shared memory before being shared across every thread in the block, which is an application of the parallelised maximum element reduction function discussed in section 5.1.2.

Thus, the ***span*** of determining the sky direction with the highest signal to noise ratio is $T_\infty = O(D + D^2 + M_{T_\infty}(S))$ and the ***work*** is $T_1 = O(S(D + D^2) + M_{T_1}(S)$, where S is the number of sky directions and $M(x)$ is the complexity of the parallelised maximum element reduction function. We can further state that the ***parallelism*** of this is equivalent to the number of sky directions, $S + S/\log S$.

Calculating signal consistency statistics

After having determined the sky direction with the highest SNR for a given maximum, the function then calculates a signal-morphology based statistic ξ_D^2 for each detector D. The statistic is a reduced χ^2 distribution with $D \times 2 - 4$ degrees of freedom and a mean value of 1, and is given in the discrete form by:

$$\xi_D^2 = \frac{\sum_{j=-m}^{m} |\varrho_D[j] - \varrho_D[0]A_D[j]|^2}{\sum_{j=-m}^{m}(2 - 2|A_D[j]|^2)}, \tag{7}$$

where ϱ is the coherent SNR, A_D is the vector of the correlation of the given template with the output from the detector and $2 \times m$ is the number of samples.

The numerator of the statistic is calculated by splitting the number of samples between the threads of a block, followed by combining the results of the statistic across each warp and then each block. The combination of the statistic across each warp and block is a modification of the parallelised maximum element reduction discussed in section 5.1.2 that uses addition instead of maximum as the combining binary function. Thus the ***span*** of calculating the statistic is $T_\infty = O(D \times M_{T_\infty}(N))$ and its ***work*** is $T_1 = O(D \times M_{T_1}(N))$, where N is the number of samples. We can then state that the ***parallelism*** of calculating the statistic is equivalent to the parallelism of the reduction, $O(N/\log N)$.

Generating time-shifted background noise statistics

The function then performs a number of time shifts on background noise for use with the significance estimation. The generation of a single background statistical variant is equal to the total work of the function so far, save that instead of using blocks for every peak, each warp looks at a different time shift. Thus, whilst the theoretical time complexity does not change, the number of processors available is smaller, so the actual runtime each loop is approximately the warp size slower.

Overall computational cost

Overall, this function has a ***span*** of $T_\infty = 2(D + D^2 + M_{T_\infty}(S) + DM_{T_\infty}(N))$, and has $T_1 = P(S(D + D^2) + M_{T_1}(S) + DM_{T_1}(N) + B(S(D + D^2) + M_{T_1}(S)))$ ***work***, where P is the number of SNR maxima and B is the number of times shifts made to background noise.

5.1.6 Calculating heat skymaps

If the coherent SNR exceeds a threshold, the post-processing produces a skymap of the highest SNR in the GPU function `ker_coh_skymap`.

The function determines the highest maximum SNR by using the maximum element reduction technique discussed in section 5.1.2. Following this, the function re-performs the process discussed in section 5.1.5 with additional sky directions and without the reduction to generate the final skymap.

As such, this function has a ***span*** of $T_\infty = M_{T_\infty}(P) + D + D^2$ and total ***work*** of $T_1 = M_{T_1}(P) + S(D + D^2)$.

5.1.7 Overall complexity

The total span and work of the coherent post-processing step in the SPIIR pipeline is the sum of the total spans and works of the internal functions. Conversely, we cannot determine the overall parallelism as the post-processing step spans a number individual functions that can each be run with a different set of processors. As the step to determine the number of peaks over a threshold (see section 5.1.3) is sequential, we can consider its time complexity as contributing to both the span and work of the total pipeline. Another thing to note is that the steps for determining the coherent correlation, statistic value and skymaps (sections 5.1.5 and 5.1.5) will be run for every detector.

With this in mind, we can determine that the ***span*** of the post-processing is:

$$\begin{aligned} T_\infty &= O(NT + N^2 + 1 + D(2(D + D^2 + \log S + D \log N) + \log P + D + D^2)) \\ &= O(NT + N^2 + D^3 + D^2 \log N + D \log S + D \log P), \end{aligned} \tag{8}$$

where D is the number of detectors, S is the number of sky directions, T is the number of templates, N is the number of samples and $P = \max\{S, T\}$.

The total ***work*** of the post-processing is:

$$\begin{aligned} T_1 &= O(NT + NT + S^2 + D(P + S(D + D^2) + P(S(D + D^2) + S + DN + B(S(D + D^2) + S)))) \\ &= O(NT + N^2 + SPD^3 + SPBD^3 + ND^2), \end{aligned} \tag{9}$$

where D is the number of detectors, S is the number of sky directions, T is the number of templates, N is the number of samples, B is the number of is the number of times shifts made to background noise and $P = \max\{S, T\}$

5.1.8 Implications

5.2 Improvements to maximum element reduction

6 Further work

7 Conclusion

References

[1] *Introduction to LIGO & Gravitational Waves*. URL: https://www.ligo.org/science/GW-GW2.php.

[2] *LIGO Lab: Caltech: MIT*. URL: https://www.ligo.caltech.edu/.

[3] B. P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Phys. Rev. Lett.* 116 (6 2016), p. 061102. DOI: 10.1103/PhysRevLett.116.061102. URL: https://link.aps.org/doi/10.1103/PhysRevLett.116.061102.

[4] *LIGO Detected Gravitational Waves from Black Holes*. URL: https://www.ligo.caltech.edu/detection.

[5] *LIGO Scientific Collaboration*. URL: https://ligo.org/.

[6] *Summed Parallel Infinite Impulse Response Pipeline Codebase*. URL: https://git.ligo.org/lscsoft/spiir/.

[7] Shaun Hooper et al. "Summed parallel infinite impulse response filters for low-latency detection of chirping gravitational waves". eng. In: *Physical Review D - Particles, Fields, Gravitation and Cosmology* 86.2 (2012). ISSN: 1550-7998.

[8] Xiaoyang Guo et al. "GPU-Optimised Low-Latency Online Search for Gravitational Waves from Binary Coalescences". eng. In: vol. 2018-. EURASIP, 2018, pp. 2638–2642. ISBN: 9082797011. URL: https://ieeexplore.ieee.org/document/8553574.

[9] Q. Chu. *Low-latency detection and localization of gravitational waves from compact binary coalescences*. eng. 2017.

[10] *LSC Algorithm Library for GStreamer*. URL: https://git.ligo.org/lscsoft/gstlal/.

[11] *PyCBC - Analyze gravitational-wave data, find signals, and study their parameters*. URL: https://pycbc.org/.

[12] Alex Nitz et al. "gwastro/pycbc: PyCBC release v1.16.11". In: (2020). DOI: 10.5281/zenodo.4075326.

[13] LIGO Scientific Collaboration. *LIGO Algorithm Library - LALSuite*. free software (GPL). 2018. DOI: 10.7935/GT1W-FZ16.

[14] *GstLAL documentation*. URL: https://lscsoft.docs.ligo.org/gstlal/.

[15] Cody Messick et al. "Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data". In: *Physical Review D* 95.4 (2017). ISSN: 2470-0029. DOI: 10.1103/physrevd.95.042001. URL: http://dx.doi.org/10.1103/PhysRevD.95.042001.

[16] *gstlal-inspiral/gst/lal/gstlal_itacac.c*. URL: https://git.ligo.org/lscsoft/gstlal/-/blob/master/gstlal-inspiral/gst/lal/gstlal_itacac.c.

[17] Cody Messick. *Added GEO600 support ("G1") to itacac (d43bcfd6)*. URL: https://git.ligo.org/lscsoft/gstlal/-/commit/d43bcfd6096ac4fab33114848b2d5f9ffaf6ca86.

[18] *CUDA Toolkit*. URL: https://developer.nvidia.com/cuda-toolkit.

[19] *CUDA Programming Guide*. URL: https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#hardware-multithreading.

[20] Thomas H. Cormen et al. *Introduction to algorithms*. eng. 3rd ed. MIT electrical engineering and computer science series. Cambridge, Mass: MIT Press, pp. 779–781. ISBN: 0070131430.

[21] Henri Casanova, Arnaud Legrand, and Yves Robert. *Parallel Algorithms*. eng. CRC Press, 2008, pp. 10–12. DOI: 10.1.1.466.8142.

[22] John L. Gustafson. "Brent's Theorem". In: *Encyclopedia of Parallel Computing*. Ed. by David Padua. Boston, MA: Springer US, 2011, pp. 182–185. ISBN: 978-0-387-09766-4. DOI: 10.1007/978-0-387-09766-4_80. URL: https://doi.org/10.1007/978-0-387-09766-4_80.

[23] Mark Harris. *Optimizing Parallel Reduction in CUDA*. eng. URL: https://developer.download.nvidia.com/assets/cuda/files/reduction.pdf.