Technique for early-warning detection of compact binary coalescence and its implications for multi-messenger astronomy

Kipp Cannon[1], Romain Cariou[2], Adrian Chapman[3], M Crispin-Ortuzar[4], Nickolas Fotopoulos[3], Melissa Frei [5], Chad Hanna[6], Erin Kara[7], Drew Keppel[8,9], Laura Liao[10], Stephen Privitera [3], Antony C Searle[3], Leo Singer[3] and Alan J Weinstein[3]

[1] Canadian Institute for Theoretical Astrophysics, Toronto, ON, Canada
[2] Département de physique, École Normale Supérieure de Cachan, 61 Avenue du Président Wilson, 94235 Cachan Cedex, France
[3] LIGO Laboratory - California Institute of Technology, Pasadena, CA, USA
[4] Facultat de Física, Universitat de València, E-46100 Burjassot, Spain
[5] The University of Texas at Austin, Austin, TX, USA
[6] Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada
[7] Department of Physics and Astronomy, Barnard College, Columbia University, New York, NY 10027, USA
[8] Albert-Einstein-Institut, Max-Planck-Institut für Gravitationphysik, Hannover, Germany
[9] Leibniz Universität Hannover, Hannover, Germany
[10] Ryerson University, Toronto, ON, Canada

We have only Institution, City, State, Country for some co-authors. Can we get rid of the full addresses, or is this at the discretion of the co-authors? How about author e-mails?

Abstract. The rapid detection of compact binary coalescence with a network of advanced gravitational-wave detectors will offer a unique opportunity for multi-messenger astronomy. Prompt detection alerts to the astronomical community may make it possible to observe the onset of electromagnetic emission from compact binary coalescence. We demonstrate a computationally practical analysis strategy that produces early warning triggers even before gravitational radiation from the final merger has arrived at the detectors. With current rate estimates for the Advanced LIGO design configuration, we should detect ∼10 sources earlier than 10 seconds before merger in 1 year of live time.

PACS numbers: 04.30.-w, 07.05.Kf

1. Introduction

Would the introduction be more effective without this paragraph?, should we find whitepapers?

The coalescence of compact binary systems consisting of neutron stars (NS) and/or black holes (BH) is the most promising source of gravitational radiation for Advanced LIGO, Virgo, GEO and LCGT [1, 2, 3, 4]. Tens of binary coalescence events are expected to be observed in the advanced detector era later this decade [5].

As a compact binary system loses energy to gravitational waves, its orbital separation decreases. This causes a run-away inspiral with the gravitational-wave amplitude and frequency increasing until the system eventually merges near

the innermost stable circular orbit (ISCO). If a neutron star is involved it may become tidally disrupted near the merger. This disrupted matter can fuel a bright electromagnetic counterpart in the system's final moments as a binary [6].

Prompt electromagnetic emission can arise as shells of relativistically outflowing matter collide in the inner shock. Such an inner shock from a compact binary coalescence is believed to be a mechanism for short gamma-ray bursts (short GRBs) [7, 8]. The same inner shocks, or potentially reverse shocks, can produce a bright accompanying optical flash [9]. Prompt emission is a probe into the extreme initial conditions of the outflow, in contrast with afterglows, which are relatively insensitive to initial conditions. Optical flashes have only been observed for a handful of long GRBs [10, 11, 12, 13, 14, 15] by telescopes with extremely rapid response or, in the case of GRB 080319B, by pure serendipity, where several telescopes were observing a previous GRB in the same field [14]. Short GRBs, on the other hand, typically fade too quickly to observe anything before the initial burst of gamma rays and hard x-rays. Rapid gravitational-wave transient alerts could enable the observation of optical flashes from short GRBs. An optical counterpart would vastly boost confidence in the gravitational-wave detection and provide the tight sky localization necessary to allow determination of the source's host galaxy, which leads to a redshift measurement. With both redshift and a coincident gravitational-wave observation, we can produce precision measurements of the Hubble constant [16].

To this end, we have the ambition of reporting gravitational-wave candidates not minutes *after* the merger, but seconds *before*. By looking for threshold crossings before the gravitational-wave signal leaves the detection band, it is possible to trade some signal to noise ratio (SNR) for latency. Figure 1 shows projected early trigger rates for NS–NS binaries in Advanced LIGO assuming the event rate predictions in [5].

Citation needed for LOOC-UP

The gravitational-wave community initiated a project to send alerts when potential gravitational-wave transients are observed. In October 2010, LIGO completed its sixth science run (S6) and Virgo completed its third science run (VSR3). While both LIGO detectors and Virgo were operating, several all-sky detection pipelines operated in a low-latency configuration, namely MBTA, ihope, Coherent WaveBurst, and Omega [18, 19]. The S6 analyses achieved latencies of 30–60 minutes, which were dominated by a human vetting process. Candidates were sent for electromagnetic followup to several telescopes; Swift, ROTSE, TAROT, and Zadko [20, 18] took images of likely sky locations. MBTA achieved the best gravitational-wave trigger generation latencies of 2–5 minutes. We assume that in the advanced detector era the vetting process will be automated, so current gravitational wave search methodology and telescope actuation would dominate latency.

Get references for these low-latency pipelines.

Advance detection of compact binary coalescences (CBCs) will require striking a balance between latency and throughput. CBC searches consist of banks of matched filters, or cross-correlations between the data stream and a bank of nominal "template" signals. There are many different implementations of matched filters, but most have high throughput at the cost of high latency, or low latency at the cost of low throughput. The former are epitomized by the overlap-save algorithm for FFT convolution, currently the preferred method in gravitational wave searches. The most obvious example of the latter is the time domain (TD) convolution, which has no latency at all. However, its computational complexity is quadratic in the length of the templates, so it is prohibitively expensive for long templates.

Fortunately, the morphology of inspiral signals can be exploited to offset some of the computational complexity of low latency algorithms. First, the signals evolve

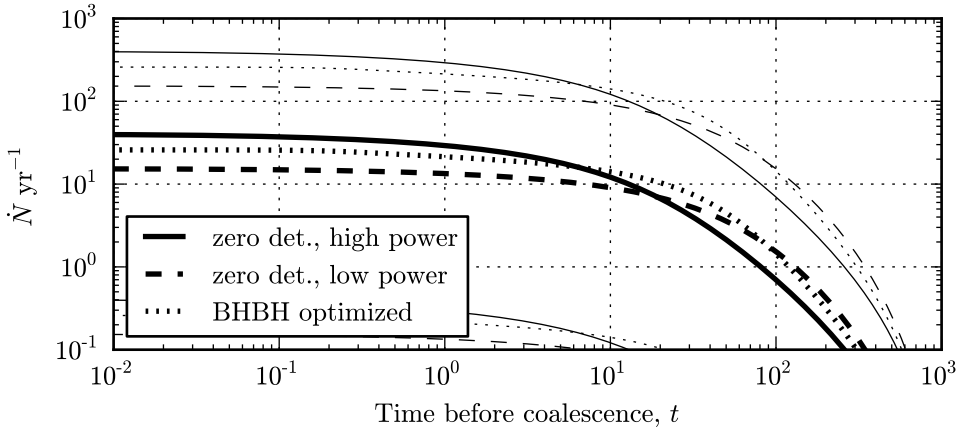


Figure 1: Expected number of NS–NS sources that will be detectable t seconds before coalescence. The heavy solid line is the most likely yearly rate estimate $\dot{N}_{\rm re}$ during Advanced LIGO. The light solid lines represent the confidence interval $[\dot{N}_{\rm low}, \dot{N}_{\rm high}]$ described in [5]. Advanced LIGO noise projections are described in [17]. Assuming that an SNR of 8 is sufficient for detection and that we observe $\dot{N}_{\rm re} = 40$ events yr^{-1} with a detector in the 'zero detuning, high power' configuration, about 10 yr^{-1} will be detectable within 10 seconds before merger and ~ 1 yr^{-1} will be detectable within 100 seconds of merger. The 'zero detuning, lower power' and 'BBH optimized' configurations are still more conducive to advance detection but at the cost of fewer total detectable events above an SNR of 8.

slowly in frequency, so that they can be broken into contiguous band-limited time intervals and processed at possibly lower sample rates. Second, inspiral filter banks consist of highly similar templates, admitting principal component analysis to reduce the number of templates. We described a rank-reduction scheme based on singular value decomposition in [21]. We will use both aspects to demonstrate that a very low latency analysis with advance detection of compact binary sources is possible with current computing resources. Assuming other technical sources of latency can be reduced significantly, this should allow the possibility for prompt alerts to be sent to the astronomical community.

The paper is organized as follows. First we provide an overview of our method for detecting compact binary coalescence signals in an early-warning analysis. We then describe the pipeline we have constructed that implements our method. To validate the approach we present results of simulations and conclude with some remarks on what remains to prepare for the Advanced detector era.

2. Early warning searches for compact binary coalescence

In this section we describe a decomposition of the compact binary parameter space that reduces low-latency filtering cost sufficiently to allow for the possibility of early-warning detection with modest computing requirements. We expand on the ideas of [22, 23] that describe a multiband decomposition of the compact binary parameter space that resulted in a search with minutes latency in LIGO's S6 and Virgo's VSR2

science runs. We combine this with the SVD rank reduction method described in [21] that exploits the redundancy of the template banks.

2.1. Conventional CBC *matched filter searches*

Inspiral signals are continuously parameterized by a set of intrinsic source parameters $\vec{\theta}$ that determine the amplitude and phase evolution of the gravitational wave strain. For systems where the effects of spin can be ignored, the intrinsic source parameters are the component masses of the binary, $\vec{\theta} = (m_1, m_2)$. Searches for inspiral signals typically employ matched filter banks that discretely sample the possible intrinsic parameters [24]. For a given source, the strain observed by the detector is a linear combination of two waveforms corresponding to the '+' and '×' gravitational wave polarizations, so for any value of $\vec{\theta}$ we must implement 2 filters. The coefficients for the M filters are known as templates, and are formed by discretizing and time reversing the waveforms and weighting them by the inverse amplitude spectral density of the detector's noise. To construct a template bank, templates are chosen with the $M/2$ discrete signal parameters $\vec{\theta}_0$, $\vec{\theta}_1$, ..., $\vec{\theta}_{M/2-1}$ to assure a bounded loss of SNR [25, 26]. That is, any possible signal within a given range of intrinsic source parameters will have an inner product that is $\geqslant 0.97$ with at least one template. Such a template bank is said to have a *minimum match* of 0.97. The data from the detector are whitened and convolved with each template to produce $2M$ SNR time series. Local peak-finding across time and templates determines detection candidates.

Can we cut this down to one citation?

Filtering the detector data involves a convolution of the data with the templates. For a unit-normalized template $h_i[k]$ and whitened detector data $x[k]$, both sampled at a rate f^0, the result can be interpreted as the signal-to-noise ratio, $\rho_i[k]$ defined as

$$\rho_i[k] = \sum_{n=0}^{N-1} h_i[n]x[k-n]. \tag{1}$$

Equation (1) can be implemented in the time domain (TD) as an FIR filter, requiring $\mathcal{O}(MN)$ floating point operations per sample. However, it is typically much more computationally efficient to use the convolution theorem and the FFT to implement fast convolution in the frequency domain (FD), requiring only $\mathcal{O}(M \lg N)$ operations per sample.

2.2. The nameless *method*

In the remainder of this section we explore a method for reducing the computational cost of a TD search for compact binary coalescence. We will give a truly zero latency algorithm that competes in terms of floating point operations per second with the conventional FD method, which requires a signficant latency in order to be computationally cheap. Our method, *nameless*, involves two transformations of the template waveforms that produce a set of orthogonal filters with far fewer coefficients than the original templates. The reduction in sample count for all of the filters required to search the entire parameter space is dramatic.

The first transformation is to chop the time-domain templates up into time slices. Since each template slice is disjoint in time, the resulting set for a single template is orthogonal. Given the chirp like structure of the templates, the early time slices have significantly lower bandwidth and can be safely downsampled. The downsampling removes a factor of ~ 100 samples from the early part of the waveform and allows

Where do these numbers come from? CHAD: I have actually changed the statement to 100 for the early time slices.

the filters to be evaluated at about ~ 100 times lower rate. This amounts to more than a factor of 10000 reduction in the floating point operations per second required to filter the early part of the waveform. However, the resulting filters are still not orthogonal across the mass parameter space, and are in fact highly redundant. We use the singular value decomposition to produce an orthogonal filter set from the time sliced templates [21]. We find that this reduces the number of sample points of the filters by another factor of ~ 100. The combined methods reduce the number of floating point operations to the level where they are competitive with the conventional high latency FD matched filter approach. In the remainder of this section we describe the *nameless* algorithm in detail and provide some basic computational cost scaling.

2.3. Selectively reducing the sample rate of the data and template waveforms

The first step of our proposed method is to divide the templates into *time slices* in a time-domain analogue to the frequency-domain decomposition described in [22, 23, 27, 28]. A matched filter is constructed for each time slice. The outputs form an ensemble of partial SNR streams. By linearity, these partial SNR streams can be suitably time delayed and summed to reproduce the SNR of the full template. To wit,

Four citations is way too many! We should pick just one *citation for* MBTA, *and it should be the most recent and complete description of it.*

$$h_i[k] = \sum_{s=0}^{S-1} \begin{cases} h_i^s[k] & \text{for } t^s \leqslant k/f^0 < t^{s+1} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for S integers $\{f^0 t^s\}$ such that $0 = f^0 t^0 < f^0 t^1 < \cdots < f^0 t^S = N$. We will show in the next section that this, combined with the singular value docomposition, is sufficient to enable a computationally efficient time-domain search and furthermore is an essential part of an early-warning detection scheme.

For concreteness and simplicity, we will consider an inspiral waveform in the quadrupole approximation, for which the time-frequency relation is

$$f = \frac{1}{\pi \mathcal{M}} \left[\frac{5}{256} \frac{\mathcal{M}}{t} \right]^{3/8} . \tag{3}$$

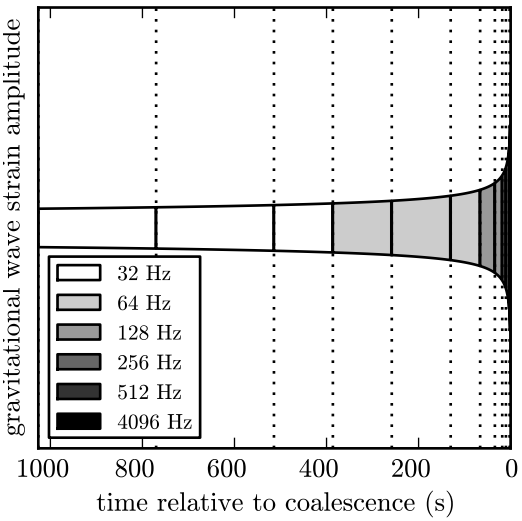
Here, $\mathcal{M}$ is the chirp mass of the binary in units of time (where $GM_\odot/c^3 \approx 5\mu$s) and t is the time relative to the coalescence of the binary [24, 29]. Usually the template is truncated at some prescribed time t^0, or equivalently frequency $f_{\rm hi}$. This is often chosen to correspond to the ISCO. An inspiral signal will enter the detection band at a low frequency, $f = f_{\rm low}$, corresponding to a time $t_{\rm low}$. The template is assumed to be zero outside the interval $[t_{\rm low}, t^0)$ and is said to have a duration of $t^0 - t_{\rm low}$. It is critically sampled at a rate of $2f_{\rm hi}$.

The monotonic time-frequency relationship of equation (3) allows us to choose time-slice boundaries that require substantially less bandwidth at early times in the inspiral. Our goal is to reduce the filtering cost of a large fraction of the waveform by computing part of the convolution at a lower sample rate. Specifically we consider here time slice boundaries with the smallest power-of-two sample rate that sub-critically samples the time sliced template. The time slices consist of the S intervals $\left(t^S, t^{S-1}\right], \ldots, \left(t^2, t^1\right], \left(t^1, t^0\right]$ sampled at frequencies $f^{\rm S-1}, \ldots, f^1, f^0$ where $f^0 \geqslant 2f_{\rm hi}$ and $f^{S-1} \geqslant 2f_{\rm low}$. The time sliced templates may be downsampled

without aliasing, so we define them as

$$h_i^s[k] \equiv \begin{cases} h_i\left[k\dfrac{f}{f^s}\right] & \text{if } t^s \leqslant k/f^s < t^{s+1} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

An example time-slice design satisfying these constraints for a $1.4 - 1.4\,M_\odot$ binary is shown in table 1.



f^s (Hz)	$\left(t^{s+1}, t^s\right]$ (s)	N^s
4096	$(0.25, 0]$	1024
512	$(2.25, 0.25]$	1024
256	$(10.25, 2.25]$	2048
128	$(18.25, 10.25]$	1024
128	$(34.25, 18.25]$	2048
128	$(66.25, 34.25]$	4096
64	$(130.25, 66.25]$	4096
64	$(258.25, 130.25]$	8192
64	$(386.25, 258.25]$	8192
32	$(514.25, 386.25]$	4096
32	$(770.25, 514.25]$	8192
32	$(1026.25, 770.25]$	8192

Table 1: Example of nearly critically sampled, power-of-two time slices for a $1.4 - 1.4\,M_\odot$ template extending from $f_{\text{low}} = 10\,\text{Hz}$ to $f_{\text{hi}} = f_{\text{ISCO}} = 1571\,\text{Hz}$ with a time-frequency structure given by (3). f^s is the sample rate of the time slice, $(t^{s+1}, t^s]$ are the boundaries in seconds preceeding coalescence and N^s is the number of samples in the s^{th} filter.

Since waveforms with neighboring intrinsic source parameters $\vec{\theta}$ have similar time-frequency evolution, it is possible to design computationally efficient time slices for an extended region of parameter space rather than to design different time slices for each template.

We note that the time slice decomposition in equation (2) is manifestly orthogonal since the time slices are disjoint in time. In the next section we examine how to reduce the number of filters within each time slice via singular value decomposition of the time sliced templates.

2.4. Reducing the number of filters with the singular value decomposition

As described previously, the template banks are, by design, highly correlated. It is possible to greatly reduce the number of filters required to achieve a particular minimum match by designing an appropriate set of SVD *basis templates*. A numerical technique based on the singular value decomposition (SVD) of inspiral template banks

is demonstrated in [21]. Similarly, the time sliced templates described above can be approximated to arbitrary accuracy by expansion into a set of SVD *basis templates*, $u_l^s[k]$, related to the original time sliced templates through the *reconstruction matrix*, $v_{il}^s \sigma_l^s$:

$$h_i^s[k] = \sum_{l=0}^{M-1} v_{il}^s \sigma_l^s u_l^s[k] \approx \sum_{l=0}^{L^s-1} v_{il}^s \sigma_l^s u_l^s[k]. \tag{5}$$

The parameter L^s sets the number of basis templates that are kept in the approximation. This determines the SVD tolerance, which affects the SNR loss due to the approximation. The authors of [21] showed that high accuracy could be achieved with far fewer basis templates than templates in the original template bank. We find that when combined with the time slice decomposition, the number of basis templates L^s is much smaller than the original number of templates M. In the next section we describe how we form our novel detection statistic using the time slice decomposition and the SVD.

Say a word or two about SVD tolerance?

2.5. Early warning SNR

In the previous two sections we have described two transformation that greatly reduce the burden of filtering the compact binary parameter space and make TD filtering of the data possible. We are now prepared to define the early warning SNR and to comment on the computational cost of evaluating it. But first, we will introduce some notation refering to the decimation of data, which means reducing the sample rate after low pass filtering,

$$x^{s+1}[k] = \left(H^{\downarrow} x^s\right)[k] \tag{6}$$

and notation refering to the interpolation of data to increase the sample rate,

$$x^s[k] = \left(H^{\uparrow} x^{s+1}\right)[k] \tag{7}$$

From the combination of transformations to the compact binary templates defined in equation (4) and (5) we define the early warning filter output accumulated up to time slice s as,

$$\rho_i^s[k] = \underbrace{\sum_{l=0}^{L^s-1} v_{il}^s \sigma_l^s}_{\text{reconstruction}} \overbrace{\sum_{n=0}^{N^s-1} u_l^s[n] x^s[k-n]}^{\text{orthogonal FIR filters}} + \underbrace{\left(H^{\uparrow} \rho_i^{s+1}\right)[k]}_{\text{SNR from previous time slices}} \tag{8}$$

This quantity is related to the early warning SNR by the normalization of the partial filter defined by the present time slice. This normalization may be computed in the planning stage and can be easily multiplied here. Figure 2 describes the computation of the SNR and early warning SNR schematically. The signal flow diagram drawn in figure 2 contains several distinct stages and parallel branches. The stages are decimation, filtering the data against the orthogonal templates, reconstruction of the original template basis, interpolation and accumulation of the early warning and final SNR.

In the next section we compute the expected computational cost scaling of this decomposition and compare it with the brute-force time-domain implementation of (1) and higher latency frequency-domain methods.

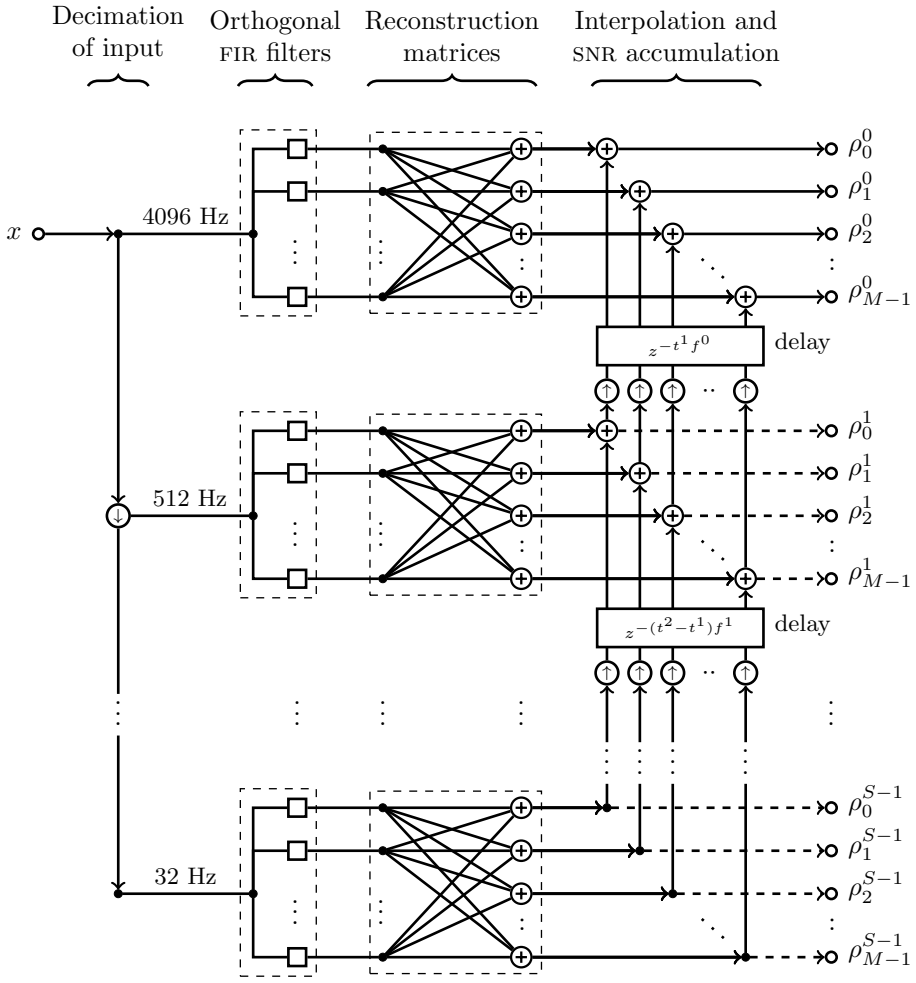


Figure 2: Schematic of LLOID pipeline illustrating signal flow. Circles with arrows represent interpolation ⊕ or decimation ⊕. Circles with plus signs represent summing junctions ⊕. Squares □ stand for FIR filters. Sample rate decreases from the top of the diagram to the bottom. In this diagram each time slice contains three FIR filters that are linearly combined to produce four output channels. In a typical pipeline the number of FIR filters is much less than the number of output channels.

2.6. Comparison of computational costs

We now examine the computational cost scaling of the approximate implementation of a conventional time-domain or frequency-domain matched filter bank as compared with *nameless*. For convenience, table 2 provides a review of the notation that we will need in this section.

This section's two tables are small. Any way to put them side by side or next to another figure?

2.6.1. Conventional time-domain method The conventional time-domain method consists of a bank of FIR filters, or sliding-window dot products. If there are M

Table 2: Notation used to describe filters.

	Definition
M	number of templates
N	number of samples per template
S	number of time slices
L^s	number of orthogonal templates in time slice s
N^s	number of samples in time slice s
f^s	sample rate in time slice s
$N^{\downarrow}$	number of coefficients in decimation filter
$N^{\uparrow}$	number of coefficients in interpolation filter

Table 3: Minimum computational cost of the FIR filter bank, FFT convolution, and *nameless* methods for a bank of 657 templates sampled at 4096 Hz.

Method	flop/s
FIR	2.4×10^{13}
FFT convolution	2.6×10^{8}
nameless	4.7×10^{8}

templates, each N samples in length, then each filter requires MN multiplications and additions per sample, or $2MNf^0$ floating point operations per second (flop/s) at a sample rate f^0.

2.6.2. Conventional frequency-domain method The most common frequency-domain method is known as the *overlap-save* algorithm, described in [30]. It entails splitting the input into blocks of D samples, $D > N$, each block overlapping the previous one by $D - N$ samples. For each block, the algorithm computes the forward FFT of the data and the templates, multiplies them, and then computes the reverse FFT.

Modern implementations of the FFT, such as the ubiquitous `fftw`, require about $2D \lg D$ operations to evaluate a real transform of size D [31]. Including the forward transform of the data and M reverse transforms for all of the templates, the FFT costs $2(M + 1)D \lg D$ operations per block. The multiplication of the transforms adds a further $2MD$ operations per block. Since each block produces $D - N$ usable samples of output, the overlap-save method requires

$$f^0 \cdot \frac{2(M + 1) \lg D + 2M}{1 - N/D} \text{ flop/s.}$$

2.6.3. nameless *method* For time slices s, the *nameless* method requires $2N^s L^s f^s$ flop/s for evaluating the orthogonal filters, $2ML^s f^s$ flop/s for the linear transformation from the L^s basis templates to the M time-sliced templates, and Mf^s flop/s to add the resultant partial SNR stream.

The computational cost of the decimation of the detector data is a little bit more subtle. Decimation is achieved by applying an FIR antialiasing filter and then downsampling, or deleting samples in order to reduce the sample rate from f^{s-1}

to f^s. Naively, an antialiasing filter with $N^{\downarrow}$ coefficients should demand $2N^{\downarrow}f^{s-1}$ flop/s. However, it is necessary to evaluate the antialiasing filter only for the fraction f^s/f^{s-1} of the samples that will not be deleted. Consequently, an efficient decimator that requires only $2N^{\downarrow}f^{s-1} \cdot \left(f^s/f^{s-1}\right) = 2N^{\downarrow}f^s$ flop/s is possible.

The story is similar for the interpolation filters used to change the sample rates of the partial SNR streams. Interpolation of a data stream from a sample rate f^s to f^{s-1} consists of inserting zeros between the samples of the original stream, and then applying a low-pass filter with $N^{\downarrow}$ coefficients. The low-pass filter requires $2MN^{\downarrow}f^{s-1}$ flop/s. However, by taking advantage of the fact that by construction a fraction f^{s-1}/f of the samples are zero, it is possible to build an efficient interpolator that requires only $MN^{\uparrow}f^{s-1} \cdot \left(f^s/f^{s-1}\right) = 2MN^{\uparrow}f^s$ flop/s.

Taking into account the decimation of the detector data, the orthogonal FIR filters, the reconstruction of the time sliced templates, the interpolation of SNR from previous time slices, and the accumulation of SNR, in total *nameless* requires

$$\sum_{s=0}^{S-1} \left(2N^sL^s + 2ML^s + M\right) f^s + 2\sum_{f\in\{f^s\}} \left(N^{\downarrow} + MN^{\uparrow}\right) f \text{ flop/s.}$$

The second sum is carried out over the set of sample rates of all of the time slices rather than over the time slices themselves because some time slices may have the same sample rates.

2.6.4. Extrapolation of the computational cost of the nameless *method to an Advanced LIGO search* Table 3 shows that it requires about 0.5 GFLOPS to conduct the filtering, reconstruction and resampling stages of the *nameless* method for 657 templates chosen around typical NS–NS mass parameters. Given the capability of modern CPUs it should be possible to filter that sub-bank on one CPU. Assuming that the cost holds for the other sub-banks we expect that the full parameter space should require $\sim$ 150 CPUs for realtime filtering with the *nameless* method per detector. Even if real implementations fall short of these estimates by factors of several, there will still be ample computing resources for an early warning search in the advanced detector era.

3. Implementation

In this section we describe an implementation of the *nameless* method described in section 2 suitable for rapid gravitational-wave searches for compact binary coalescence. The *nameless* method requires several computations that can be completed before the analysis is underway. Thus we divide the procedure into two stages 1) an offline planning stage and 2) an online, low-latency filtering stage. The offline stage can be done before the analysis is started and updated asynchronously, whereas the online stage must keep up with the detector output and produce search results as rapidly as possible. In the next two subsections we describe what these stages entail.

3.1. Planning stage

The choice of templates and the SVD can be done in advance and will be valid as long as the detector noise spectrum remains roughly constant. New templates and the

I'd rather not go into this.

Table 4: Filter design for these 657 templates. From left to right, this table shows the sample rate, time interval, number of samples, and number of orthogonal templates for each time slice. The SVD tolerance is varied from $\left(1-10^{-1}\right)$ to $\left(1-10^{-6}\right)$.

			$\log_{10}$ (1 − SVD tolerance)					
f^s (Hz)	$(t^{s+1}, t^s]$ (s)	N^s	−1	−2	−3	−4	−5	−6
4096	$(0.5, 0]$	2048	1	4	6	8	10	14
512	$(4.5, 0.5]$	2048	2	6	8	10	12	16
256	$(12.5, 4.5]$	2048	2	6	8	10	12	15
128	$(76.5, 12.5]$	8192	6	20	25	28	30	32
64	$(140.5, 76.5]$	4096	1	8	15	18	20	22
64	$(268.5, 140.5]$	8192	1	7	21	25	28	30
64	$(396.5, 268.5]$	8192	1	1	15	20	23	25
32	$(460.5, 396.5]$	2048	1	1	3	9	12	14
32	$(588.5, 460.5]$	4096	1	1	7	16	18	21
32	$(844.5, 588.5]$	8192	1	1	8	26	30	33
32	$(1100.5, 844.5]$	8192	1	1	1	12	20	23

SVD can be computed asynchronously using updated spectrum estimates as they are available.

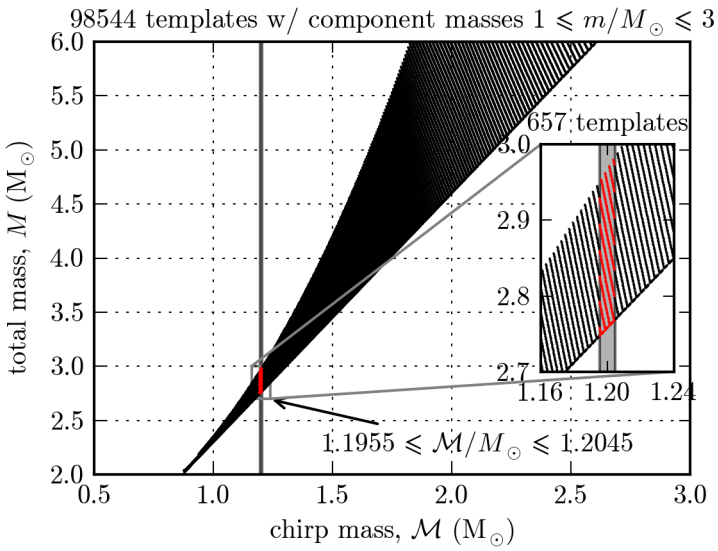


Figure 3: Placement of template parameters used in this paper. The template bank consists of 98544 templates with component masses m_1, m_2, between 1 and 3 $M_\odot$. We shall design a filter bank to search for a small subset of these template parameters with chirp masses $\mathcal{M}$ between 1.1955 and 1.2045 $M_\odot$.

The planning stage begins with choosing templates that cover the space of source parameters with a hexagonal grid [32] in order to satisfy a minimum match criterion. This assures a prescribed maximum loss in SNR for signals that fall between the chosen templates. Typically the minimum match is 0.97 corresponding to a maximum mismatch of 0.03. Next, the templates are subdivided into groups of neighbors called *sub-banks* that are appropriately sized so that each bank can be efficiently handled by a single computer. The neighbors are chosen to have comparable chirp mass, which produces sub-banks with similar time-frequency evolution. Dividing the source parameter space into smaller sub-banks reduces the computational cost of the singular value decomposition and is the approach considered in [21]. We choose time slice boundaries as in equation (4) such that all of the templates within a sub-bank are sub-critically sampled at progressively lower sample rates. Next, the templates within the sub-bank are realized as FIR filter coefficients. For each time slice, the templates are down-sampled to the appropriate sample rate. Finally, the SVD is applied to each time slice in the sub-bank in order to produce a set of orthogonal FIR filters and a reconstruction matrix that maps them back to the original templates as described in equation (5). The down-sampled orthogonal FIR filter coefficients, the reconstruction matrix, and the time slice boundaries are all saved to disk.

3.2. Filtering stage

The *nameless* algorithm could be used in a true sample-in-sample-out real-time system. However, such a system would likely require integration directly into the data acquisition and storage system of the gravitational-wave observatories. A slightly more modest goal is to leverage existing low latency, but not real-time, signal processing infrastructure in order to implement the *nameless* algorithm. For the near-term this should be a viable solution for ultra-low latency searches with $\sim 1s$ intrinsic latency.

We have implemented a prototype of the low-latency filtering stage using an open-source signal processing environment called `GStreamer` [33]. `GStreamer` is a vital component of many Linux systems, providing media playback, authoring, and streaming on devices from cell phones to desktop computers to streaming media servers. Given the similarities of gravitational- wave detector data to audio data it is not surprising that `GStreamer` is useful for our purpose. `GStreamer` also provides some useful stock signal processing elements such as resamplers and filters. We have extended the `GStreamer` framework by developing a library called `gstlal` [34] that provides elements for gravitational-wave data-analysis. Figure 2 describes schematically how we implement the *nameless* algorithm using `gstlal` and `GStreamer` components.

3.2.1. Decimation First, the sample rate of the whitened detector data is reduced to successively lower sample rates by decimation. Decimation involves applying an anti-aliasing filter to the data, and then down-sampling by deleting samples. We use a 192-tap FIR decimator provided by `GStreamer`'s `audioresample` element. The detector data is provided at every power-of-two sample rate required by the template time slices described in (2). These parallel decimated data streams are fed into parallel FIR filter banks in the next stage.

3.2.2. FIR *filters* The FIR filter banks are implemented using a `gstlal` element called `lal_firbank`, which produces N channels of filter output from an $N \times M$ matrix of

FIR filter coefficients. This element is used in parallel branches in the pipeline to implement the SVD basis filters in each time slice. Rather than implement the time-sliced templates as zero-padded FIR filters as described in (2) we instead implement them as shorter filters that contain only the nonzero samples. Adding the appropriate time offset to the filter output later in the pipeline makes up for the lack of explicit zero-padding. The orthogonal filter outputs must next be reconstructed into the output of the underlying time-sliced templates.

3.2.3. Reconstruction From the outputs of the SVD basis filters, we form the partial SNRs for each time slice by multiplying by the reconstruction matrices. This is accomplished by connecting the `gstlal` element `lal_matrixmixer` to the output of `lal_firbank`.

3.2.4. Interpolation In order to form the early-warning SNR from each time slice, we have to add the partial SNR to the early-warning SNR from the subsequent time slices. If the next time slice does not have the same sample rate, its output must first be interpolated. This is done using the same `GStreamer` element as was used for decimation, `audioresample`.

3.2.5. SNR *accumulation* The early-warning SNR for each time slice is formed by accumulating interpolated early warning SNR from the subsequent time slice. This process continuous until we have worked our way to the SNR of the original templates at the full sample rate f^0. In this way, the *nameless* algorithm and this implementation leads to a simple early warning pipeline.

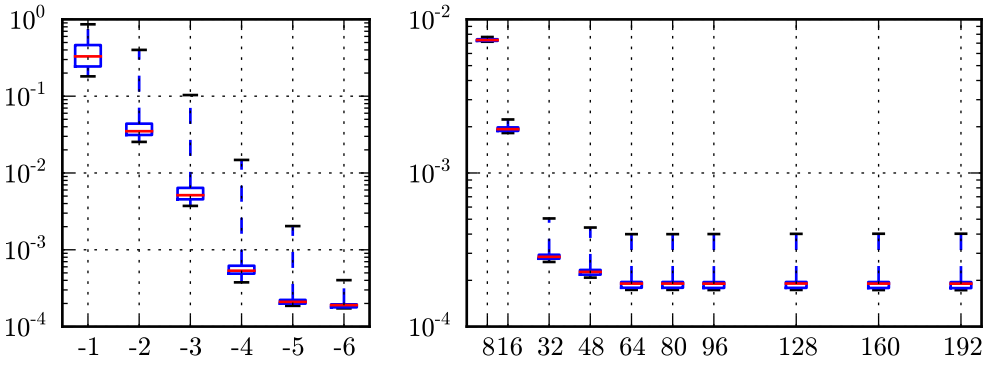
4. Results

In this section we test our implementation of the *nameless* method using the `GStreamer` pipeline described in the previous section. We calculate the measured SNR loss due to the approximations of the *nameless* method and our implementation of it.

4.1. Measured SNR *loss*

We expect two known contributions to the SNR loss to arise in our implementation of the *nameless* algorithm. The first is the SNR loss due to the truncation of the SVD basis and estimates for it exist [21]. The second comes from non ideal implementations of resampling that cause signal loss. The SNR loss is to be compared with the mismatch of 0.03 that arises from the discreteness of the template bank which is typically. We will consider an acceptable SNR loss to be a factor of 10 smaller than this, i.e., no more than ~ 0.003.

It is also within reason that some SNR loss could arise from other suboptimal implementations of the *nameless* algorithm in our test pipeline. To gauge accurately how well the pipeline is performing we tested the response to a unit impulse. By taking the inner product of the impulse response for each channel with the template, we can gauge very accurately the loss in SNR due to the approximations we have made and any inadequacies in our implementation.



(a) Mismatch versus SVD tolerance

(b) Mismatch versus $N^{\uparrow}$

Figure 4: Box-and-whisker plot of mismatch between nominal template bank and *nameless* measured impulse respones. The upper and lower boundaries of the boxes show the upper and lower quartiles; the lines in the center denote the medians. The whiskers represent the minimum and maximum mismatch over all templates. In (a) the interpolation filter length is held fixed at $N^{\uparrow} = 192$, while the SVD tolerance is varied from $\left(1 - 10^{-1}\right)$ to $\left(1 - 10^{-6}\right)$. In (b), the SVD tolerance is fixed at $\left(1 - 10^{-6}\right)$ while $N^{\uparrow}$ is varied from 8 to 192 coefficients.

4.1.1. Checking the effect of the SVD *tolerance* In this section we check the effect of the SVD tolerance parameter defined in [21]. By changing the tolerance of the reconstruction of the physical waveforms one effects directly the mismatch that results from the truncation of the orthogonal filter matrix. The conditions presented here are more complicated than in the original work [21] due to the inclusion of the time sliced templates and resampling. [*However, we find that the results are quite as expected.*] Figure 4a demonstrates that it is possible to acheive typical SNR losses of $\ll 1\%$. However, we do notice that the mismatch saturates with respect to the tolerance. This could be the effect of the resampling, or another SNR loss that we did not model or expect. However the saturation is still an order of magnitude below our target mismatch of ~ 0.003. We find that an SVD tolerance of $\left(1 - 10^{-4}\right)$ is adequate.

4.1.2. Checking the effect of the interpolation filter length Next, keeping the SVD tolerance fixed at $\left(1 - 10^{-6}\right)$, we studied the impact of the length $N^{\uparrow}$ of the interpolation filter. The results are in figure 4b. We find that a filter length of 16 is sufficient to meet our target mismatch of 0.003.

5. Conclusions

We have demonstrated a computationally feasible procedure for the rapid and even advance detection of gravitational waves emitted during the coalescence of neutron stars and stellar-mass black holes. These sources are expected to produce prompt electromagnetic signals and may be the progenitors of some short hard gamma-ray bursts. Rapid alerts to the broader astronomical community may improve the chances

We should drive home the point that our method is as fast as the fft convolution but without any latency at all.

of detecting an electromagnetic counterpart in bands from X-ray down to radio. We anticipate requiring ∼600 modern computer cores to analyze a four-detector network of gravitational-wave data for binary neutron stars and stellar mass black holes. This is within the current computing capabilities of the LSC Data Grid [35].

The algorithm we described has no intrinsic latency. However, there are fundamental and practical latencies associated with the analysis and detection procedure. For example, the LIGO detectors, data acquisition is synchronized to a 1/(16 Hz) cadence introducing an up-front latency. Data aggregation from the observatories will travel over various networks, each capable of high bandwith but perhaps only modest latency. This could amount to a similar latency of ∼100 ms. Lastly, unless a realtime infrastructure is adopted post data acquisition, it is likely that there will be an inherent latency introduced by such infrastructure. We have shown a prototype implementation using `gstlal` that is capable of ∼1 s latency. In our opinion, significant work would have to be done in order to improve upon this number. However, it should be considered for third-generation detector design. For example, a tighter integration of analysis and data acquisition would be beneficial.

We have omitted discussion of source localization though point the reader to some theoretical estimates SNR [36]. In future works we will explore more rigorously the pointing prospects with realistic simulations using the infrastructure and techniques described here.

We need to do this calculation using the flop/s counts and the number of templates. CHAD: Agreed, but if you do it by the books it seems misleading. I have adjusted it to 600. This number is taken from assuming that 1 657 template sub-bank can be filtered on one core and dividing the total number of templates (10^5) by 657 which gives you 150 cores per detector. I put it in section 2.

The way this is worded, it sounds like a big omission.

nvf: It's a bit pie-in-the-sky, but we could also mention reconfiguring the signal recycling mirror to optimize SNR at merger. There's a lot of science there.

Acknowledgments

LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0107417. CH would like to thank Ilya Mandel for many discussions about rate estimates and the prospects of early detection. CH would also like to thank Patrick Brady for countless fruitful conversations about low latency analysis methods. NF would like to thank Alessandra Corsi for illuminating discussions on astronomical motivations. This research is supported by the National Science Foundation through a Graduate Research Fellowship to LS.

This paper has LIGO Document Number LIGO-P0900004-v3.

References

[1] Advanced LIGO `www.advancedligo.mit.edu`
[2] Adanced Virgo `https://wwwcascina.virgo.infn.it/advirgo`
[3] GEO 600 `http://www.geo600.org`
[4] Large-scale Cryogenic Gravitational wave Telescope (LCGT) `http://gw.icrr.u-tokyo.ac.jp/lcgt/`
[5] Abadie J *et al.* 2010 *Class. Quant. Grav.* URL `http://iopscience.iop.org/0264-9381/27/17/173001`
[6] Shibata M and Taniguchi K 2008 *Phys. Rev.* **D77** 084015–+ (*Preprint* `arXiv:0711.1410`)
[7] Lee W H, Ramirez-Ruiz E and Granot J 2005 *Astrophys. J.* **630** L165–L168
[8] Nakar E 2007 *Phys. Rept.* **442** 166–236 (*Preprint* `astro-ph/0701748`)
[9] Sari R and Piran T 1999 *The Astrophysical Journal* **520** 641 URL `http://stacks.iop.org/0004-637X/520/i=2/a=641`
[10] Akerlof C, Balsano R, Barthelmy S, Bloch J, Butterworth P, Casperson D, Cline T, Fletcher S, Frontera F, Gisler G, Heise J, Hills J, Kehoe R, Lee B, Marshall S, McKay T, Miller R, Piro L, Priedhorsky W, Szymanski J and Wren J 1999 *Nature* **398** 400–402 (*Preprint* `arXiv:astro-ph/9903271`)

[11] Fox D W, Yost S, Kulkarni S R, Torii K, Kato T, Yamaoka H, Sako M, Harrison F A, Sari R, Price P A, Berger E, Soderberg A M, Djorgovski S G, Barth A J, Pravdo S H, Frail D A, Gal-Yam A, Lipkin Y, Mauch T, Harrison C and Buttery H 2003 *Nature* **422** 284–286

[12] Jelínek M, Prouza M, Kubánek P, Hudec R, Nekola M, Řídký J, Grygar J, Boháčová M, Castro-Tirado A J, Gorosabel J, Hrabovský M, Mandát D, Nosek D, Nožka L, Palatka M, Pandey S B, Pech M, Schovánek P, Šmída R, Trávníček P, de Ugarte Postigo A and Vítek S 2006 *Astronomy & Astrophysics* **454** L119–L122 (*Preprint* `arXiv:astro-ph/0606004`)

[13] Mundell C G, Melandri A, Guidorzi C, Kobayashi S, Steele I A, Malesani D, Amati L, DAvanzo P, Bersier D F, Gomboc A, Rol E, Bode M F, Carter D, Mottram C J, Monfardini A, Smith R J, Malhotra S, Wang J, Bannister N, OBrien P T and Tanvir N R 2007 *The Astrophysical Journal* **660** 489 URL `http://stacks.iop.org/0004-637X/660/i=1/a=489`

[14] Racusin J L *et al.* 2008 *Nature* **455** 183–188 (*Preprint* `0805.1557`)

[15] Gruber D, Krühler T, Foley S, Nardini M, Burlon D, Rau A, Bissaldi E, von Kienlin A, McBreen S, Greiner J, Bhat P N, Briggs M S, Burgess J M, Chaplin V L, Connaughton V, Diehl R, Fishman G J, Gibby M H, Giles M M, Goldstein A, Guiriec S, van der Horst A J, Kippen R M, Kouveliotou C, Lin L, Meegan C A, Paciesas W S, Preece R D, Tierney D and Wilson-Hodge C 2011 *Astronomy & Astrophysics* **528** A15+ (*Preprint* `1101.1099`)

[16] Nissanke S, Holz D E, Hughes S A, Dalal N and Sievers J L 2010 *Astrophys. J.* **725** 496–514 (*Preprint* `0904.1017`)

[17] 2010 Advanced LIGO anticipated sensitivity curves LIGO-T0900288-v3 URL `https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=2974`

[18] Hughey B 2011 Electromagnetic follow-ups of candidate gravitational wave triggers in the recent ligo and virgo science runs (Presented at the Gravitational wave physics and astronomy workshop (GWPAW))

[19] TBD 2011 *in preparation*

[20] Kanner J, Huard T L, Marka S, Murphy D C, Piscionere J, Reed M and Shawhan P 2008 URL `arXiv:0803.0312v4[astro-ph]`

[21] Cannon K, Chapman A, Hanna C, Keppel D, Searle A C and Weinstein A J 2010 *Physical Review D* **82** 44025 (c) : URL `http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010PhRvD..82d4025C&link_type=ABSTRACT`

[22] Marion F and the Virgo Collaboration 2004 *Proc. Rencontres de Moriond on Gravitational Waves and Experimental Gravity 2003*

[23] Buskulic D, the LIGO Scientific Collaboration and the Virgo Collaboration 2010 *Classical and Quantum Gravity* **27** 194013 URL `http://stacks.iop.org/0264-9381/27/i=19/a=194013`

[24] Allen B A, Anderson W G, Brady P R, Brown D A and Creighton J D E 2005 (*Preprint* `gr-qc/0509116`)

[25] Owen B J 1996 *Phys. Rev. D* **53** 6749–6761

[26] Owen B J and Sathyaprakash B S 1999 *Phys. Rev. D* **60** 022002

[27] FBeauville, M-ABizouard, LBlackburn, LBosi, PBrady, LBrocco, DBrown, DBuskulic, FCavalier, SChatterji, NChristensen, A-CClapson, SFairhurst, DGrosjean, GGuidi, PHello, EKatsavounidis, MKnight, ALazzarini, NLeroy, FMarion, BMours, FRicci, AVicere and MZanolin 2006 *J. Phys. Conf. Ser.* **32** 212 URL `arXiv:gr-qc/0509041v1`

[28] Beauville F, Bizouard M A, Blackburn L, Bosi L, Brocco L, Brown D, Buskulic D, Cavalier F, Chatterji S, Christensen N, Clapson A C, Fairhurst S, Grosjean D, Guidi G, Hello P, Heng S, Hewitson M, Katsavounidis E, Klimenko S, Knight M, Lazzarini A, Leroy N, Marion F, Markowitz J, Melachrinos C, Mours B, Ricci F, Vicer A, Yakushin I and Zanolin M 2008 *Class. Quant. Grav.* **25** 045001 (*Preprint* `http://www.iop.org/EJ/abstract/0264-9381/25/4/045001/`) URL `arXiv:gr-qc/0701027v1`

[29] E K L, M W C and G W A 1992 *Class. Quantum Grav.* **9** L125–31

[30] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2007 *Numerical Recipes* chap 13.1 3rd ed

[31] Johnson S and Frigo M 2007 *Signal Processing, IEEE Transactions on* **55** 111 – 119 URL `http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=&arnumber=4034175&queryText%253D%2528%2528Document+Title%253Aa+modified+split-radix+fft+with+fewer%2529%2529%2526openedRefinements%253D*%2526sortType%253Ddesc_Publication+Year%2526matchBoolean%253Dtrue%2526rowsPerPage%253D50%2526searchField%253DSearch+All`

[32] Cokelaer T 2007 *Phys. Rev. D* **76** 102004

[33] GStreamer: open source multimedia framework URL `http://gstreamer.freedesktop.org`

[34] `gstlal`: GStreamer based gravitational wave analysis software URL `https://www.lsc-group.phys.uwm.edu/daswg/projects/gstlal.html`

[35] The LSC Data Grid URL `https://www.lsc-group.phys.uwm.edu/lscdatagrid/`

[36] Fairhurst S 2009 *New Journal of Physics* **11** 123006