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], Mireia Crispín-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 Searle[3], Leo Singer[3], and Alan Weinstein[3]

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, it should be possible to detect ∼10 such sources earlier than 10 seconds before merger in 1 year of live time using the method we describe.

Subject headings: gamma ray burst: general — gravitational waves — methods: data analysis — methods: numerical

1. Introduction

The coalescence of compact binary systems consisting of neutron stars (NS) and/or black holes (BH) is presently the best understood and most actively pursued target source of gravitational radiation for ground-based gravitational-wave (GW) detectors. As a compact binary system loses energy to GWs, its orbital separation decays, leading to a run-away inspiral with the GW amplitude and frequency increasing until the system eventually merges. It is thought that if a NS is involved, it may become tidally disrupted near the merger and fuel a bright electromagnetic (EM) counterpart (Shibata & Taniguchi 2008). Observation of an EM counterpart to a candidate GW event will vastly boost confidence in the GW detection and will often provide sufficient sky localization to determine the source's host galaxy and therefore measure the source's redshift. The GW observation of this event will also provide a luminosity distance measurement. With both quantities, we can immediately measure the Hubble constant with unprecedented accuracy (Nissanke et al. 2010). The most efficient means of obtaining such joint observations is through the rapid identification and dissemination of candidate GW events.

In the era of first-generation detectors, the GW community initiated a project to send alerts when potential GW transients were observed in order to trigger followup observations by optical telescopes. The best demonstrated latencies were on the order of 1 hour (Hughey 2011), which was an important achievement, but far too late to catch prompt EM

[1] Canadian Institute for Theoretical Astrophysics, Toronto, ON, Canada

[2] Département de physique, École Normale Supérieure de Cachan, Cachan, France

[3] LIGO Laboratory - California Institute of Technology, Pasadena, CA, USA

[4] Facultat de Física, Universitat de València, Burjassot, Spain

[5] 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, USA

[8] Albert-Einstein-Institut, Max-Planck-Institut für Gravitationphysik, Hannover, Germany

[9] Leibniz Universität Hannover, Hannover, Germany

[10] Ryerson University, Toronto, ON, Canada

emission. In order to extract the maximum science from Advanced LIGO CBC events, we have the ambition of reporting GW candidates not minutes *after* the merger, as has already been accomplished, but seconds *before*.

Though compact binary coalescence (CBC) events may be responsible for several classes of EM transients, CBCs are in particular believed to be a mechanism for short gamma-ray bursts (short GRBs) (Lee et al. 2005; Nakar 2007). In this scenario, prompt EM emission arises as shells of relativistically out-flowing matter collide in the inner shock. The same inner shocks, or potentially reverse shocks, can produce a bright accompanying optical flash (Sari & Piran 1999). Prompt emission is a probe into the extreme initial conditions of the outflow, in contrast with afterglows, which arise in the external shock with the local medium and are relatively insensitive to initial conditions.

Optical flashes have only been observed for a handful of long GRBs (Atteia & Boër 2011) 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 (Racusin et al. 2008). The observed optical flashes peaked within tens of seconds and decayed quickly. Short GRBs, on the other hand, typically fade too quickly to even catch the tails of the afterglows in any band. Rapid GW transient alerts could enable the first observation of prompt optical flashes and the rise of afterglows from short GRBs, and, for an exceptional event, perhaps even a glimpse of a precursor (Troja et al. 2010).

We show that by examining the signal to noise ratio (SNR) stream for threshold crossings before the GW signal leaves the detection band, it is possible to trade some SNR and sky localization accuracy for negative latency. Figure 1 shows projected early trigger rates for NS–NS binaries in Advanced LIGO assuming event rates predicted in (Abadie et al. 2010) and anticipated detector sensitivity described in (Shoemaker 2010). The most likely estimates indicate that ∼10 sources will produce an SNR above 8 in the detector 10 seconds or more prior to the merger during one year of observation, though the uncertainty spans orders of magnitude. The gray bands give the 5 to 95% confidence interval as described in (Abadie et al. 2010).

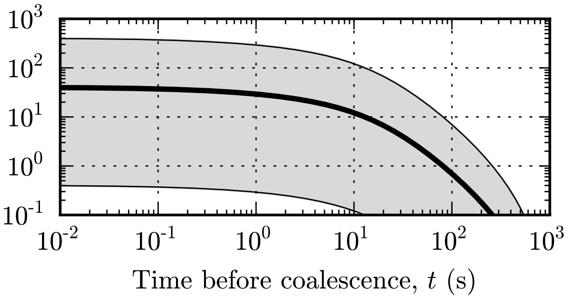


Fig. 1.—: Expected number of NS–NS sources that will be detectable t seconds before coalescence. The heavy solid line is the most likely yearly rate estimate $\dot{N}_{\mathrm{re}}$ during Advanced LIGO. The shaded region represents the confidence interval $[\dot{N}_{\mathrm{low}}, \dot{N}_{\mathrm{high}}]$ described in (Abadie et al. 2010). With an advanced detector in the 'zero detuning, high power' configuration (Shoemaker 2010) and an SNR threshold of 8, we will observe a total of $\dot{N}_{\mathrm{re}} = 40$ events yr^{-1}, but ∼10 yr^{-1} will be detectable within 10 s of merger and ∼1 yr^{-1} within 100 s.

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 to send astronomical alerts, namely MBTA, Coherent WaveBurst, and Omega (Hughey 2011; TBD 2011). The S6 analyses achieved latencies of 30–60 minutes, which were dominated by a human vetting process. Candidates were sent for EM followup to several telescopes; Swift, ROTSE, TAROT, and Zadko (Kanner et al. 2008; Hughey 2011) took images of likely sky locations. MBTA achieved the best GW trigger-generation latencies of 2–5 minutes. We assume that in the advanced detector era the vetting process will be automated, so current GW search methodology and telescope actuation would dominate latency.

Realizing 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 (Press et al. 2007) for frequency domain (FD) convolution, currently the preferred method in GW searches. The most obvious example of the latter is the time domain (TD) convolution, which has no algorithmic latency. 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 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 methods such as singular value decomposition (SVD) to reduce the number of templates (Cannon et al. 2010). 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 CBC waveform space that reduces time-domain (TD) filtering cost sufficiently to allow for the possibility of early-warning detection with modest computing requirements. We expand on the ideas of (Marion & the Virgo Collaboration 2004; Buskulic et al. 2010) that describe a multi-band decomposition of the compact binary parameter space that resulted in a search with minutes latency in LIGO's S6 and Virgo's VSR3 joint science runs (Hughey 2011). We combine this with the SVD rank-reduction method of (Cannon et al. 2010) 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 $\mathbf{\Theta}$ that determine the amplitude and phase evolution of the GW strain. For systems where the effects of spin can be ignored, the intrinsic source parameters are the component masses of the binary, $\mathbf{\Theta} = (m_1, m_2)$. Searches for inspiral signals typically employ matched filter banks that discretely sample the possible intrinsic parameters (Allen et al. 2005). For a given source, the strain observed by the detector is a linear combination of two waveforms corresponding to the '+' and '×' GW polarizations. Thus, for any value of $\mathbf{\Theta}$ 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 $\mathbf{\Theta}_0, \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_{M/2-1}$ to assure a bounded loss of SNR (Owen & Sathyaprakash 1999). 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 M SNR time series. Local peak-finding across time and templates determines detection candidates.

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 SNR, $\rho_i[k]$ 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 as an finite impulse response (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, requiring only $\mathcal{O}(M \lg N)$ operations per sample with increased latency.

2.2. The LLOID method

Here 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 significant latency in order to be computationally cheap. Our method, called LLOID (a backronym for Low Latency Online Inspiral Detection), involves two transformations of the template waveforms that produce a set of orthogonal filters with far fewer coefficients than the original templates.

The first transformation is to chop the TD 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 down-sampled. The down-sampling reduces the number of data samples by a factor of ∼100 from the early part of the waveform and allows the filters to be evaluated at about $\sim 1/100^{\text{th}}$ the sample 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 SVD to produce an orthogonal filter set from the time-sliced templates (Cannon et al. 2010). 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 LLOID 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 TD analogue to the FD decomposition described in (Marion & the Virgo Collaboration 2004; Buskulic et al. 2010). To wit,

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

for S integers $\{f^0 t^s\}$ such that $0 = f^0 t^0 < f^0 t^1 < \cdots < f^0 t^S = N$. 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. We will show in the next section that this, combined with the SVD, is sufficient to enable a computationally efficient TD 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 quadrapole approximation, for which the time-frequency relation is

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

Here, $\mathcal{M}$ is the chirp mass of the binary in units of time (where $GM_\odot/c^3 \approx 5\mu\text{s}$) and t is the time relative to the coalescence of the binary (Allen et al. 2005; E et al. 1992). Usually the template is truncated at some prescribed time t^0, or equivalently frequency f_{hi}. This is often chosen to correspond to the ISCO. An inspiral signal will enter the detection band at a low frequency, $f = f_{\text{low}}$, corresponding to a time t_{low}. The template is assumed to be zero outside the interval $[t^0, t_{\text{low}})$ and is said to have a duration of $t_{\text{low}} - t^0$. It is critically sampled at a rate of $2f_{\text{hi}}$.

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 $\left[t^0, t^1\right)$, $\left[t^1, t^2\right)$, …, $\left[t^{S-1}, t^S\right)$, sampled at frequencies f^0, f^1, …, f^{S-1} where $f^0 \geqslant 2f_{\mathrm{hi}}$ and $f^{S-1} \geqslant 2f_{\mathrm{low}}$. The time sliced templates may be down-sampled without aliasing, so we define them as

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

Since waveforms with neighboring intrinsic source parameters $\mathbf{\Theta}$ 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 SVD of the time-sliced templates.

2.4. Reducing the number of filters with the SVD

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*, previously demonstrated in (Cannon et al. 2010). 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 is 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 (Cannon et al. 2010) 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 and improves on (Cannon et al. 2010) by nearly an order of magnitude. In the next section we describe how we form our early-warning detection statistic using the time-slice decomposition and the SVD.

2.5. Early warning SNR

In the previous two sections we have described two transformations that greatly reduce the burden of filtering waveforms from the 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 referring to the decimation of data, which means reducing the sample rate after low-pass filtering,

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

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

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

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,

$$\rho_i^s[k] = \underbrace{\sum_{l=0}^{L^s-1}}_{\text{reconstruction}} v_{il}^s \sigma_l^s \overbrace{\sum_{n=0}^{N^s-1} u_l^s[n] x^s[k-n]}^{\text{orthogonal FIR filters}} + \underbrace{\left(H^{\uparrow} \rho_i^{s+1}\right)[k]}_{\text{SNR from previous time slices}} \tag{6}$$

This quantity is related to the early-warning SNR by the normalization of the partial filter defined by the present time slice, which may be computed in advance. The signal flow diagram in figure 2 illustrates how this recursion relation can be realized as a multi-rate filter network with outputs for each of the early-warning SNRs.

Chad: are we happy with that statement?

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

2.6. Comparison of computational costs

We now examine the computational cost scaling of the conventional TD or FD matched filter procedure as compared with LLOID. For convenience,

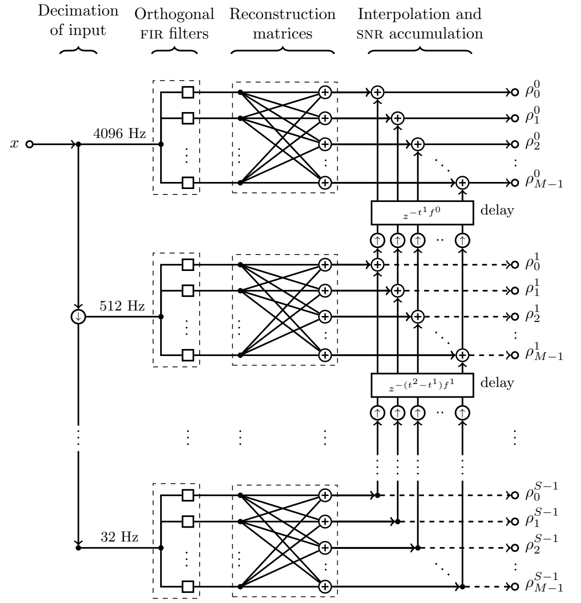


Fig. 2.—: Schematic of LLOID pipeline illustrating signal flow. Circles with arrows represent interpolation ↑⃝ or decimation ↓⃝. Circles with plus signs represent summing junctions ⊕. Squares □ stand for FIR filters. Sample rate decreases from the top of the diagram to the bottom. In this diagram each 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.

table 1 provides a review of the notation that we will need in this section.

2.6.1. *Conventional* TD *method*

The conventional TD method consists of a bank of FIR filters, or sliding-window dot products. If there are M 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* FD *method*

The most common FD method is known as the *overlap-save* algorithm, described in (Press et al. 2007). 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 (Johnson & Frigo 2007). 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. LLOID *method*

For time slices s, the LLOID 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 anti-aliasing filter and then down-sampling, or deleting samples in order to reduce the sample rate from f^{s-1} to f^s. Naively, an anti-aliasing filter with $N^{\downarrow}$ coefficients should demand $2N^{\downarrow} f^{s-1}$ flop/s. However, it is necessary to evaluate the anti-aliasing 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 $2N^{\downarrow} f^{s-1} \cdot \left(f^s/f^{s-1}\right) =$

Table 1:: 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 decimated 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

$2N^\downarrow f^s$ 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 the LLOID algorithm requires

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

The second sum is carried out over the set of distinct sample rates rather than over the time slices themselves.

3. Implementation

In this section we describe an implementation of the LLOID method described in section 2 suitable for rapid GW searches for CBCs. The LLOID method requires several computations that can be completed before the analysis is underway. Thus we divide the procedure into two stages 1) an off-line planning stage and 2) an online, low-latency filtering stage. The off-line 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 planning stage begins with choosing templates that cover the space of source parameters with a hexagonal grid (Cokelaer 2007) 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 SVD and is the approach considered in (Cannon et al. 2010). 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 LLOID 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 GW observatories. A slightly more modest goal is to leverage existing low-latency, but not real-time, signal processing infrastructure in order to implement the LLOID algorithm. For the near-term this should be a viable solution for searches with order seconds of intrinsic latency.

We have implemented a prototype of the low-latency filtering stage using an open-source signal processing environment called `GStreamer`[1]. `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 GW 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

[1] `http://gstreamer.net/`

as re-samplers and filters. We have extended the GStreamer framework by developing a library called gstlal[2] that provides elements for GW data analysis.

3.2.1. Decimation

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). Next, these decimated data streams are fed into parallel FIR filter banks.

3.2.2. FIR filters

The FIR filtering is 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, representing the SVD basis filters in our application. 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.

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 continues until we have worked our way to the SNR of the original templates at the full sample rate f^0. In this way, the LLOID algorithm and this implementation naturally leads to a simple early-warning pipeline.

4. Results

In this section we evaluate the LLOID algorithm using our GStreamer-based implementation described in the previous section. We calculate the measured SNR loss due to the approximations of the LLOID method and our implementation of it. Using settings that give acceptable SNR loss for our chosen parameter space, we then compute the operation counts.

4.1. Setup

We examine the performance of the LLOID algorithm on a small region of compact binary parameter space centered on typical NS–NS masses. We begin by constructing a template bank that spans component masses between 1 and 3 $M_{\odot}$ using a simulated Advanced LIGO noise curve (Shoemaker 2010). Then we create sub-banks by partitioning the parameter space by chirp mass. Figure 3 illustrates this procedure. The result is that we obtain 657 templates with chirp masses between 1.1955 and 1.2045 $M_{\odot}$. With this sub-bank we were able to construct an efficient time-slice decomposition that consisted of 11 time slices with sample rates between 32 and 4096 Hz shown in table 2. We use this template bank and decomposition for the remainder of this section.

4.2. Measured SNR loss

We expect two known contributions to the SNR loss to arise in our implementation of the LLOID algorithm. The first is the SNR loss due to the truncation of the SVD basis and estimates for it exist (Cannon et al. 2010). The second comes from non ideal implementations of re-sampling

[2] https://www.lsc-group.phys.uwm.edu/daswg/projects/gstlal.html

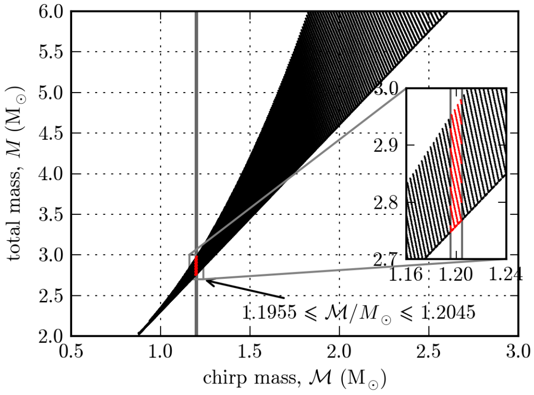


Fig. 3.—: Placement of template parameters used in this paper. The template bank consists of 98544 templates with component masses m_1, m_2, between 1 and 3 $M_\odot$. We design a filter bank to search for a subset of 657 of these templates with chirp masses $\mathcal{M}$ between 1.1955 and 1.2045 $M_\odot$.

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. We will consider a target SNR loss to be a factor of 10 smaller than this, i.e., no more than $\sim$0.003.

It is also within reason that some SNR loss could arise from other suboptimalities of implementation 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.

4.2.1. Effect of the SVD tolerance

In this section we check the effect of the SVD tolerance parameter defined in (Cannon et al. 2010). By changing the tolerance of the reconstruction of the physical waveforms one directly affects the mismatch that results from the truncation of the orthogonal filter matrix. The conditions presented here are more complicated than in the original work (Cannon et al. 2010) due to the inclusion of the time-sliced templates and re-sampling. Figure 4a demonstrates that it is possible to achieve 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 $\sim$0.003. We find that an SVD tolerance of $\left(1 - 10^{-4}\right)$ is adequate to achieve our target SNR loss.

4.2.2. 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.

4.3. Lower bounds on computational cost compared to other methods

We are now prepared to offer the estimated computational cost of filtering this sub-bank of templates compared to other methods. We used the results of the previous subsections to set the SVD tolerance to $1 - 10^{-4}$ and the interpolation filter length to 16. Table 3 lists the theoretical costs for filtering this sub-bank. Both the FD method and LLOID are five orders of magnitude faster than the conventional TD method. However, the FD method has a latency of over half of an hour, whereas LLOID has no latency at all.

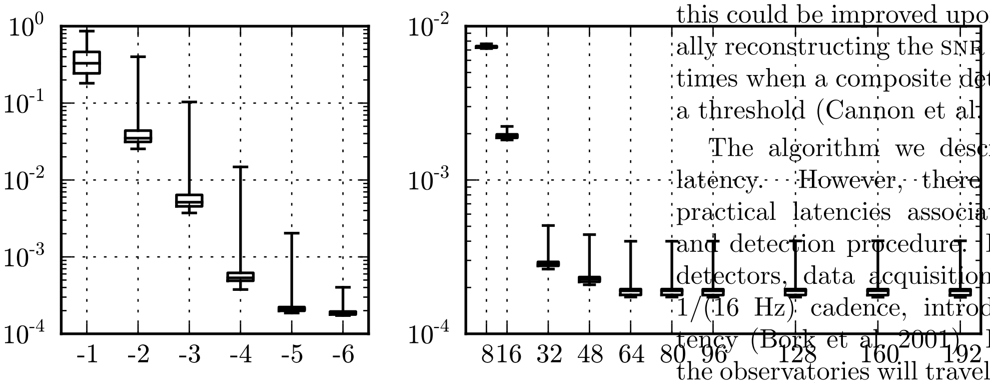
4.4. Extrapolation of the computational cost to an Advanced LIGO search

Table 3 shows that the LLOID method requires $\sim 10^8$ flop/s to cover a sub-bank comprising $\sim 10^2$ out of the total $\sim 10^5$ mass pairs. Given that modern (ca. 2011) workstations can sustain computation rates up to $\sim 10^{10}$ flop/s, and assuming that other regions of the parameter space have similar computational scaling, an entire search could be implemented with $\gtrsim$10 machines. The lengths of templates does vary over the parameter space; lower-mass templates are longer and would require more computation to analyze while higher-mass templates are shorter and would require less. However, we consider the estimates based on this sub-bank to be a reasonable representation.

By comparison, using the TD method to achieve the same latency costs $\sim 10^{13}$ flop/s for this particular sub-bank, and so it would require $\sim 10^6$

Table 2:: 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. We vary SVD tolerance from $\left(1-10^{-1}\right)$ to $\left(1-10^{-6}\right)$.

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



(a) Mismatch versus SVD tolerance

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

Fig. 4.—: Box-and-whisker plot of mismatch between nominal template bank and LLOID measured impulse responses. 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.

present-day machines to search the full parameter space. Presently, the LIGO Data Grid consists of only ~10^4 machines.

5. Conclusions

We have demonstrated a computationally feasible procedure for the rapid and even preemptive detection of GWs emitted during the coalescence of neutron stars and stellar-mass black holes. Our method is as fast as standard FFT convolutions but allows for zero latency, sample-in-sample-out operation. Our search targets 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 will improve the chances of detecting an electromagnetic counterpart in bands from gamma-rays down to radio. We anticipate requiring ~100 modern multi-core computers to analyze a four-detector network of GW data for binary neutron stars and stellar-mass black holes. This is well within the current computing capabilities of the LIGO Data Grid (LDG ????). In the future, this could be improved upon further by conditionally reconstructing the SNR time-series only during times when a composite detection statistic crosses a threshold (Cannon et al. 2011).

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 (Bork et al. 2001). Data aggregation from the observatories will travel over various networks, each capable of high bandwidth but perhaps only modest latency. This could amount to a similar

Table 3:: Computational cost in flop/s of the TD method, the FD method, and LLOID for the bank of 2 × 657 templates sampled at 4096 Hz for chirp masses $1.1955 \leqslant \mathcal{M}/M_{\odot} \leqslant 1.2045$. A block size of $D=2N$ is used for the FFTs in the FD method.

Method	flop/s	Latency (s)
Time domain (TD)	2.4×10^{13}	0
Frequency domain (FD)	2.6×10^{8}	2201
LLOID (this work)	4.7×10^{8}	0

latency of ∼100 ms. Lastly, unless a real-time 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 of LLOID using the open source signal processing software `GStreamer` and `gstlal`. Significant work would have to be done in order to improve upon the latency capability of our implementation, for example, more tightly integrating analysis and data acquisition. This should be considered for third-generation detector design.

Although we have demonstrated a feasible method for advance detection we have not explored the accuracy of sky localization that is possible before merger. Ref. (Fairhurst 2009) discusses some of the theoretical prospects for early sky localization. Our future work will explore the prospects of early-warning detection with realistic simulations of binary mergers using the infrastructure and techniques described here.

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.

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. DK was supported from the Max Planck Gesellschaft. This research is supported by the National Science Foundation through a Graduate Research Fellowship to LS and by a fellowship from The Perimeter Institute for Theoretical Physics to CH.

This paper has LIGO Document Number LIGO-P0900004.

REFERENCES

????, The LSC Data Grid

Abadie, J., et al. 2010, Class. Quant. Grav.

Allen, B. A., Anderson, W. G., Brady, P. R., Brown, D. A., & Creighton, J. D. E. 2005

Atteia, J.-L., & Boër, M. 2011, Comptes Rendus Physique, 12, 255

Bork, R., Abbott, R., Barker, D., & Heefner, J. 2001, eConf - Proceedings of the 8th International Conference on Accelerator and Large Experimental Physics Control Systems

Buskulic, D., the LIGO Scientific Collaboration, & the Virgo Collaboration. 2010, Classical and Quantum Gravity, 27, 194013

Cannon, K., Chapman, A., Hanna, C., Keppel, D., Searle, A. C., & Weinstein, A. J. 2010, Physical Review D, 82, 44025, (c) :

Cannon, K., Hanna, C., Keppel, D., & Searle, A. C. 2011, Phys. Rev. D, 83, 084053

Cokelaer, T. 2007, Phys. Rev. D, 76, 102004

E, K. L., M, W. C., & G, W. A. 1992, Class. Quantum Grav., 9, L125

Fairhurst, S. 2009, New Journal of Physics, 11, 123006

Hughey, B. 2011, in Presented at the Gravitational wave physics and astronomy workshop (GWPAW)

Johnson, S., & Frigo, M. 2007, Signal Processing, IEEE Transactions on, 55, 111

Kanner, J., Huard, T. L., Márka, S., Murphy, D. C., Piscionere, J., Reed, M., & Shawhan, P. 2008, Classical and Quantum Gravity, 25, 184034

Lee, W. H., Ramirez-Ruiz, E., & Granot, J. 2005, Astrophys. J., 630, L165

Marion, F., & the Virgo Collaboration. 2004, Proc. Rencontres de Moriond on Gravitational Waves and Experimental Gravity 2003

Nakar, E. 2007, Phys. Rept., 442, 166

Nissanke, S., Holz, D. E., Hughes, S. A., Dalal, N., & Sievers, J. L. 2010, Astrophys. J., 725, 496

Owen, B. J., & Sathyaprakash, B. S. 1999, Phys. Rev. D, 60, 022002

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007, Numerical Recipes, 3rd edn.

Racusin, J. L., et al. 2008, Nature, 455, 183

Sari, R., & Piran, T. 1999, The Astrophysical Journal, 520, 641

Shibata, M., & Taniguchi, K. 2008, Phys. Rev., D77, 084015

Shoemaker, D. 2010, Advanced LIGO anticipated sensitivity curves, LIGO-T0900288-v3

TBD. 2011, in preparation

Troja, E., Rosswog, S., & Gehrels, N. 2010, The Astrophysical Journal, 723, 1711

This 2-column preprint was prepared with the AAS LaTeX macros v5.2.