TOWARD EARLY-WARNING DETECTION OF GRAVITATIONAL WAVES FROM COMPACT BINARY COALESCENCE

Kipp Cannon[1], Romain Cariou[2], Adrian Chapman[3], Mireia Crispin-Ortuzar[4], Nickolas Fotopoulos[3], Melissa Frei[5,6], Chad Hanna[7], Erin Kara[8], Drew Keppel[9,10], Laura Liao[11], Stephen Privitera[3], Antony Searle[3], Leo Singer[3], and Alan Weinstein[3]

ABSTRACT

Rapid detection of compact binary coalescence (CBC) with a network of advanced gravitational-wave detectors will offer a unique opportunity for multi-messenger astronomy. Prompt detection alerts for the astronomical community might make it possible to observe the onset of electromagnetic emission from CBC. We demonstrate a computationally practical filtering strategy that could produce early-warning triggers before gravitational radiation from the final merger has arrived at the detectors.

Keywords: gamma-ray burst: general — gravitational waves — methods: data analysis — methods: numerical

1. INTRODUCTION

As a compact binary system loses energy to gravitational waves (GWs), its orbital separation decays, leading to a runaway inspiral with the GW amplitude and frequency increasing until the system eventually merges. If a neutron star (NS) is involved, it might become tidally disrupted near the merger and fuel an electromagnetic (EM) counterpart (Shibata & Taniguchi 2008). Effort from both the GW and the broader astronomical communities might make it possible to use GW observations as early warning triggers for EM follow-up. In the first generation of ground-based laser interferometers, the GW community initiated a project to send alerts when potential GW transients were observed in order to trigger follow-up observations by EM telescopes. The typical latencies were 30 minutes (Hughey 2011). This was an important achievement, but too late to catch any prompt optical flash or the onset of an on-axis optical afterglow. Since the GW signal is in principle detectable even before the tidal disruption, one might have the ambition of reporting GW candidates not minutes after the merger, but seconds before. We explore one essential ingredient of this problem, a computationally inexpensive latency free, real-time filtering algorithm for detecting inspiral signals in GW data. We also consider the prospects for advanced GW detectors and discuss other areas of work that would be required for rapid analysis.

Compact binary coalescence (CBC) is a plausible progenitor for most short gamma-ray bursts (short GRBs; Lee et al. 2005; Nakar 2007), but the association is not iron-clad (Virgili et al. 2011). The tidally disrupted material falls onto the newly formed, rapidly spinning compact object and is accelerated in jets along the spin axis with a timescale of 0.1–1 s after the merger (Janka et al. 1999), matching the short GRB duration distribution well. Prompt EM emission including the GRB can arise as fast outflowing matter collides with slower matter ejected earlier in inner shocks. The same inner shocks, or potentially reverse shocks, can produce an accompanying optical flash (Sari & Piran 1999). The 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 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 already observing the afterglow of another GRB in the same field of view (FOV; Racusin et al. 2008). The observed optical flashes peaked within tens of seconds and decayed quickly. For short GRB energy balance and plasma density, however, the reverse shock model predicts a peak flux in radio, approximately 20 minutes after the GRB, but also a relatively faint optical flash (Nakar 2007); for a once-per-year Advanced LIGO event at 130 Mpc, the radio flux will peak around 9 GHz at ∼5 mJy, with emission in the R-band at ∼19 mag. Interestingly, roughly a quarter to half of the observed short GRBs also exhibit extended X-ray emission of 30–100 s in duration beginning ∼10 s after the GRB and carrying comparable fluence to the initial outburst. This can be explained if the merger results in the formation of a proto-magnetar that interacts with ejecta (Bucciantini et al. 2012). Rapid GW alerts would enable joint EM and GW observations to confirm the short GRB-CBC link and allow the early EM observation of exceptionally nearby and thus bright events.

In 2010 October, LIGO[12] completed its sixth science run (S6) and Virgo[13] 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, Coherent Wave-Burst (cWB), Omega, and Multi-Band Template Analysis

[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, MC 100-36, 1200 E. California Blvd., Pasadena, CA, USA; leo.singer@ligo.org

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

[5] Department of Physics, University of Texas at Austin, Austin, TX, USA

[6] Center for Computational Relativity and Gravitation and School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA

[7] Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada

[8] Department of Physics and Astronomy, Barnard College, Columbia University, New York, NY, USA

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

[10] Institut für Gravitationsphysik, Leibniz Universität Hannover, Hannover, Germany

[11] Department of Chemistry and Biology, Ryerson University, Toronto, ON, Canada

[12] http://www.ligo.org/
[13] http://www.ego-gw.it/

(MBTA; Hughey 2011; Abadie et al. 2011a, 2012, 2011c). cWB and Omega are both unmodeled searches for bursts based on time-frequency decomposition of the GW data. MBTA is a novel kind of template-based inspiral search that was purpose-built for low latency operation. MBTA achieved the best GW trigger-generation latencies of 2–5 minutes. Alerts were sent with latencies of 30–60 minutes, dominated by human vetting. Candidates were sent for EM follow-up to several telescopes; *Swift*, LOFAR, ROTSE, TAROT, QUEST, SkyMapper, Liverpool Telescope, Pi of the Sky, Zadko, and Palomar Transient Factory (Kanner et al. 2008; Hughey 2011) imaged some of the most likely sky locations.

There were a number of sources of latency associated with the search for CBC signals in S6/VSR3 (Hughey 2011), listed here.

Data acquisition and aggregation ($\gtrsim$100 ms) — The LIGO data acquisition system collects data from detector subsystems 16 times a second (Bork et al. 2001). Data are also copied from all of the GW observatories to the analysis clusters over the Internet, which is capable of high bandwidth but only modest latency. Together, these introduce a latency of $\gtrsim$100 ms. These technical sources of latency could be reduced with significant engineering and capital investments, but they are minor compared to any of the other sources of latency.

Data conditioning ($\sim$1 min) — Science data must be calibrated using the detector's frequency response to gravitational radiation. Currently, data are calibrated in blocks of 16 s. Within $\sim$1 minute, data quality is assessed in order to create veto flags. These are both technical sources of latency that might be addressed with improved calibration and data quality software for advanced detectors.

Trigger generation (2–5 min) — Low-latency data analysis pipelines deployed in S6/VSR3 achieved an impressive latency of minutes. However, second to the human vetting process, this dominated the latency of the entire EM follow-up process. Even if no other sources of latency existed, this trigger generation latency is too long to catch prompt or even extended emission. Low-latency trigger generation will become more challenging with advanced detectors because inspiral signals will stay in band up to 10 times longer. In this work, we will focus on reducing this source of latency.

Alert generation (2–3 min) — S6/VSR3 saw the introduction of low-latency astronomical alerts, which required gathering event parameters and sky localization from the various online analyses, downselecting the events, and calculating telescope pointings. If other sources of latency improve, the technical latency associated with this infrastructure could dominate, so work should be done to improve it.

Human validation (10–20 min) — Because the new alert system was commissioned during S6/VSR3, all alerts were subjected to quality control checks by human operators before they were disseminated. This was by far the largest source of latency during S6/VSR3. Hopefully, confidence in the system will grow to the point where no human intervention is necessary before alerts are sent, so we give it no further consideration here.

— This work will focus on reducing the latency of trigger production. Data analysis strategies for advance detection of CBCs will have to strike 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 frequency-domain (FD) convolution, currently the preferred method in GW searches. The most obvious example of the latter is direct time domain (TD) convolution, which is latency-free. However, its cost in floating point operations per second is linear in the length of the templates, so it is prohibitively expensive for long templates. The computational challenges of low-latency CBC searches are still more daunting for advanced detectors for which the inspiral signal remains in band for a large fraction of an hour (see the Appendix).

Fortunately, the morphology of inspiral signals can be exploited to offset some of the computational complexity of known 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 the singular value decomposition (SVD) (Cannon et al. 2010) or the Gram-Schmidt process (Field et al. 2011) to reduce the number of templates.

Several efforts that exploit one or both of these properties are under way to develop low-latency CBC search pipelines with tractable computing requirements. One example is MBTA (Marion & the Virgo Collaboration 2003; Buskulic et al. 2010), which was deployed in S6/VSR3. MBTA consists of multiple, usually two, template banks for different frequency bands, one which is matched to the early inspiral and the other which is matched to the late inspiral. An excursion in the output of any filter bank triggers coherent reconstruction of the full matched filtered output. Final triggers are built from the reconstructed matched filter output. Another novel approach using networks of parallel, second-order infinite impulse response (IIR) filters is being explored by Hooper et al. (2010) and Luan et al. (2011).

We will use both properties to demonstrate that a very low latency detection statistic is possible with current computing resources. Assuming the other technical sources of latency can be reduced significantly, this could make it possible to send prompt alerts to the astronomical community.

The paper is organized as follows. First, we discuss prospects for early-warning detection. Then, we provide an overview of our novel method for detecting CBC signals near real-time. We then describe a prototype implementation using open source signal processing software. To validate our approach we present a case study focusing on a particular subset of the NS–NS parameter space. We conclude with some remarks on what remains to prepare for the advanced detector era.

2. PROSPECTS FOR EARLY-WARNING DETECTION AND EM FOLLOW-UP

Before the GW signal leaves the detection band, we can imagine examining the signal-to-noise ratio (S/N) accumulated up to that point and if it is already significant, release an alert immediately, trading S/N and sky localization accuracy for pre-merger detection.

In the quadrupole approximation, the instantaneous frequency of the GW inspiral signal is related to the time t relative to coalescence (section 5.1 of Sathyaprakash & Schutz
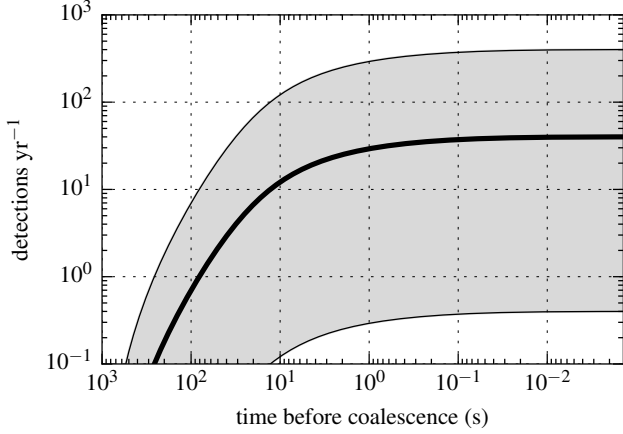
Figure 1. Expected number of NS–NS sources that could be detectable by Advanced LIGO a given number of seconds before coalescence. The heavy solid line corresponds to the most probable yearly rate estimate from Abadie et al. (2010b). The shaded region represents the 5%–95% confidence interval arising from substantial uncertainty in predicted event rates.

2009) through

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

where $\mathcal{M} = M^{2/5}\mu^{3/5}$ is the chirp mass of the binary, $\mathcal{M}_t = G\mathcal{M}/c^3$ is the chirp mass in units of time, M is the total mass, and μ is the reduced mass. The expected value of the single-detector S/N for an optimally oriented (source at detector's zenith or nadir, orbital plane face-on) inspiral source is (Abadie et al. 2010b)

$$\rho = \frac{\mathcal{M}_t^{5/6} c}{\pi^{2/3} D} \sqrt{\frac{5}{6} \int_{f_{low}}^{f_{high}} \frac{f^{-7/3}}{S(f)} df}, \qquad (2)$$

where D is the luminosity distance and $S(f)$ is the one-sided power spectral density of the detector noise. f_{low} and f_{high} are low- and high- frequency limits of integration which may be chosen to extend across the entire bandwidth of the detector. If we want to trigger at a time t before merger, then we must cut off the SNR integration at $f_{high} = f(t)$ with $f(t)$ given by Equation (1) above.

Figure 1 shows projected early detectability rates for NS–NS binaries in Advanced LIGO assuming the anticipated detector sensitivity for the 'zero detuning, high power' configuration described in Shoemaker (2010) and NS–NS merger rates estimated in Abadie et al. (2010b). The merger rates have substantial measurement uncertainty due to the small sample of known double pulsar systems that will merge within a Hubble time; they also have systematic uncertainty due to sensitive dependence on the pulsar luminosity distribution function (Kalogera et al. 2004). The most probable estimates indicate that at a single-detector S/N threshold of 8, we will observe a total of 40 events yr^{-1}; $\sim$10 yr^{-1} will be detectable within 10 s of merger and $\sim$5 yr^{-1} will be detectable within 25 s of merger if analysis can proceed with near zero latency.

We emphasize that any practical GW search will include technical delays due to light travel time between the detectors, detector infrastructure, and the selected data analysis strategy. Figure 1 must be understood in the context of all of the potential sources of latency, some of which are avoidable and some of which are not.
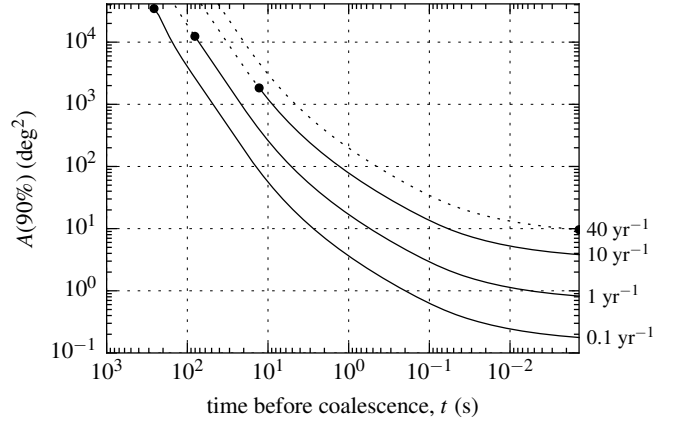

Figure 2. Area of the 90% confidence region as a function of time before coalescence for sources with anticipated detectability rates of 40, 10, 1, and 0.1 yr^{-1}. The heavy dot indicates the time at which the accumulated S/N exceeds a single-detector threshold of 8.

Table 1
Horizon Distance, S/N at Merger, and Area of 90% Confidence at Selected Times Before Merger for Sources with Expected Detectability Rates of 40, 10, 1, and 0.1 yr^{-1}.

Rate yr^{-1}	Horizon (Mpc)	Final S/N	$A(90\%)$ (deg^2)			
			25 s	10 s	1 s	0 s
40	445	8.0	——	——	——	9.6
10	280	12.7	——	1200	78	3.8
1	130	27.4	1300	260	17	0.8
0.1	60	58.9	280	56	3.6	0.2

Note. — A dash (——) signifies that the confidence area is omitted because at the indicated time the SNR would not have crossed the detection threshold of 8.

EM follow-up requires estimating the location of the GW source. The localization uncertainty can be estimated from the uncertainty in the time of arrival of the GWs, which is determined by the signal's effective bandwidth and S/N (Fairhurst 2009). Table 1 and Figure 2 show the estimated 90% confidence area versus time of the loudest coalescence events detectable by Advanced LIGO and Advanced Virgo. This is the *minimum* area; localization is best at high elevation from the plane containing the detectors and worst at zero elevation. Fairhurst also cautions that his Fisher matrix calculation fails to capture disconnected patches of probability, which occur prominently in networks of three detectors where there are generally two local maxima on opposite sides of the plane of the detectors. Aside from the mirror degeneracy, characterizing the uncertainty region by the Fisher matrix alone tends to overestimate, rather than underestimate, the area for low-S/N events, but this effect is generally more than compensated by the source being in an unfavorable sky location. For these reasons, the localization uncertainty estimated from timing is highly optimistic and will only suffice for an order-of-magnitude estimate. Once per year, we expect to observe an event with a final single-detector S/N of $\approx$27 whose location can be constrained to about 1300 deg^2 (3.1% of the sky) within 25 s of merger, 260 deg^2 (0.63% of the sky) within 10 s of merger, and 0.82 deg^2 (0.0020% of the sky) at merger.

It is unfeasible to search hundreds of square degrees for a prompt counterpart. For comparison to some examples

of modern ground-based wide-field survey instruments, the Palomar Transient Factory P48 (Law et al. 2009) has a 3.50×2.31 deg^2 FOV; the Pan-STARRS P1 (Kaiser et al. 2002) has a 7 deg^2 FOV. Even the eagerly awaited LSST will have an FOV of 9.6 deg^2 (Ivezic et al. 2008). However, it is possible to reduce the localization uncertainty by only looking at galaxies from a catalog that lie near the sky location and luminosity distance estimate from the GW signal (Nuttall & Sutton 2010) as was done in S6/VSR3. Within the expected Advanced LIGO NS–NS horizon distance, the number of galaxies that can produce a given signal amplitude is much larger than in Initial LIGO and thus the catalog will not be as useful for downselecting pointings for most events. However, exceptional GW sources will necessarily be extremely nearby. Within this reduced volume there will be fewer galaxies to consider for a given candidate and catalog completeness will be less of a concern. This should reduce the 90% confidence area substantially.

3. NOVEL REAL-TIME ALGORITHM FOR CBC DETECTION

In this section, we describe a decomposition of the CBC signal space that reduces 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 (2003) and Buskulic et al. (2010) that describe a multi-band decomposition of the compact binary signal space that resulted in a search with minutes latency during S6/VSR3 (Hughey 2011). We combine this with the SVD rank-reduction method of Cannon et al. (2010) that exploits the redundancy of the template banks.

3.1. *Conventional CBC searches*

Searches for inspiral signals typically employ matched filter banks that discretely sample the possible intrinsic parameters (Allen et al. 2011). Suppose that the observed data $x[k]$ consists of a known, nominal signal $s[k]$, and additive, zero-mean noise $n[k]$

$$x[k] = s[k] + n[k].$$

A matched filter is a linear filter, defined as

$$y[k] = \sum_{n=0}^{N-1} h[n] x[k-n] = y_s[k] + y_n[k],$$

where y_s is the response of the filter to the signal alone and y_n is the response of the signal to noise alone. The matched filter's coefficients maximize the ratio of the expectation of the filter's instantaneous response to the variance in the filter's output:

$$(\text{signal to noise})^2 = \frac{\mathrm{E}[y[0]]^2}{\mathrm{var}[y[k]]} = \frac{y_s[0]^2}{\mathrm{var}[y_n[k]]}.$$

It is well known (see, for example, Turin 1960) that if $n[k]$ is Gaussian and wide-sense stationary, then the optimum is obtained when

$$\tilde{h}[n] = \tilde{s}^*[n]\, \tilde{S}^{-1}[n],$$

up to an arbitrary multiplicative constant. Here, $\tilde{h}[n]$, $\tilde{s}[n]$, and $\tilde{x}[n]$ are the discrete Fourier transforms (DFTs) of $h[k]$, $s[k]$, and $x[k]$, respectively; $\tilde{S}[n] = \mathrm{E}[\tilde{n}[n]\tilde{n}^*[n]]$ is the folded, two-sided, discrete power spectrum of $n[k]$. It is related to the

continuous, one-sided power spectral density $S(f)$ through

$$\tilde{S}[n] = \begin{cases} S(n) & \text{if } n = 0 \text{ or } n = N/2 \\ S(nf^0/2N)/2 & \text{if } 0 < n < N/2 \\ \tilde{S}[N-n] & \text{otherwise,} \end{cases}$$

where N is the length of the filter and f^0 is the sample rate. (In order to satisfy the Nyquist-Shannon sampling criterion, it is assumed that the detector's continuous noise power spectral density $S(f)$ vanishes for all $f > f^0/2$, or alternatively, that the data are low-pass filtered prior to matched filtering.) The DFT of the output is

$$\tilde{y}[n] = \tilde{s}^*[n]\, \tilde{S}^{-1}[n]\, \tilde{x}[n]$$
$$\equiv \left(\tilde{S}^{-1/2}[n]\, \tilde{s}[n] \right)^* \left(\tilde{S}^{-1/2}[n]\, \tilde{x}[n] \right). \quad (3)$$

The placement of parentheses in Equation (3) emphasizes that the matched filter can be thought of as a cross-correlation of a whitened version of the data with a whitened version of the nominal signal. In this paper, we shall not describe the exact process by which the detector's noise power spectrum is estimated and deconvolved from the data; for the remainder of this paper we shall define $x[k]$ as the *whitened* data stream. Correspondingly, from this point on we shall use $h[k]$ to describe the *whitened* templates, being the inverse DFT of $\left(\tilde{S}^{-1/2}[n]\, \tilde{s}[n] \right)^*$.

Inspiral signals are continuously parameterized by a set of intrinsic source parameters θ that determine the amplitude and phase evolution of the GW strain. For systems in which the effects of spin can be ignored, the intrinsic source parameters are just the component masses of the binary, $\theta = (m_1, m_2)$. For a given source, the strain observed by the detector is a linear combination of two waveforms corresponding to the '+' and '×' GW polarizations. Thus, we must design two filters for each θ.

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 $M/2$ discrete signal parameters $\theta_0, \theta_1, \ldots, \theta_{M/2-1}$. These are chosen such that any possible signal will have an inner product $\geqslant 0.97$ with at least one template. Such a template bank is said to have a *minimal match* of 0.97 (Owen & Sathyaprakash 1999).

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 S/N, $\rho_i[k]$, defined as

$$\rho_i[k] = \sum_{n=0}^{N-1} h_i[n] x[k-n]. \quad (4)$$

This results in M S/N time series. Local peak-finding across time and template indices results in single-detector triggers. Coincidences are sought between triggers in different GW detectors in order to form detection candidates.

Equation (4) can be implemented in the TD as a bank of finite impulse response (FIR) filters, requiring $\mathcal{O}(MN)$ floating point operations per sample. However, it is typically much more computationally efficient to use the convolution theorem and the fast Fourier transform (FFT) to implement fast convolution in the FD, requiring only $\mathcal{O}(M \lg N)$ operations per sample but incurring a latency of $\mathcal{O}(N)$ samples.

3.2. *The LLOID method*

Here we describe a method for reducing the computational cost of a TD search for CBC. We give a zero latency, real-time algorithm that competes in terms of floating point operations per second with the conventional overlap-save FD method, which by contrast requires a significant latency due to the inherent acausality of the Fourier transform. Our method, called LLOID (Low Latency Online Inspiral Detection), involves two transformations of the templates that produce a network of orthogonal filters that is far more computationally efficient than the original bank of matched filters.

The first transformation is to chop the templates into disjointly supported intervals, or *time slices*. Since the time slices of a given template are disjoint in time, they are orthogonal with respect to time. Given the chirp-like structure of the templates, the "early" (lowest frequency) time slices have significantly lower bandwidth and can be safely downsampled. Downsampling reduces the total number of filter coefficients by a factor of ~ 100 by treating the earliest part of the waveform at $\sim 1/100$ of the full sample rate. Together, the factor of 100 reduction in the number of filter coefficients and the factor of 100 reduction in the sample rate during the early inspiral save a factor of $\sim 10^4$ floating point operations per second (flop s^{-1}) over the original (full sample rate) templates.

However, the resulting filters are still not orthogonal across the parameter space and are in fact highly redundant. We use the SVD to approximate the template bank by a set of orthogonal *basis filters* (Cannon et al. 2010). We find that this approximation reduces the number of filters needed by another factor of ~ 100. These two transformations combined reduce the number of floating point operations to a level that is 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.

3.2.1. *Selectively reducing the sample rate of the data and templates*

The first step of our proposed method is to divide the templates into time slices in a TD analog to the FD decomposition employed by MBTA (Marion & the Virgo Collaboration 2003; Buskulic et al. 2010). The application to GW data analysis is foreshadowed by an earlier FD convolution algorithm, proposed by Gardner (1995), based on splitting the impulse response of a filter into smaller blocks. We decompose each template $h_i[k]$ into a sum of S non-overlapping templates

$$h_i[k] = \sum_{s=0}^{S-1} \begin{cases} h_i^s[k] & \text{if } t^s \leqslant k/f^0 < t^{s+1} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for S integers $\{f^0 t^s\}$ such that $0 = f^0 t^0 < f^0 t^1 < \cdots < f^0 t^S = N$. The outputs of these new time-sliced filters form an ensemble of partial S/N streams. By linearity of the filtering process, these partial S/N streams can be summed to reproduce the S/N of the full template.

Since waveforms with neighboring intrinsic source parameters θ 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.

For concreteness and simplicity, consider an inspiral waveform in the quadrupole approximation, for which the time-

frequency relation is given by Equation (1). This monotonic time-frequency relationship allows us to choose time slice boundaries that require substantially less bandwidth at early times in the inspiral.

An inspiral signal will enter the detection band with some low frequency f_{low} at time t_{low} before merger. Usually the template is truncated at some prescribed time t^0, or equivalent frequency f_{high}, often chosen to correspond to the last stable orbit (LSO). The beginning of the template is critically sampled at $2f_{\text{low}}$, but the end of the template is critically sampled at a rate of $2f_{\text{high}}$. In any time interval smaller than the duration of the template, the bandwidth of the filters across the entire template bank can be significantly less than the full sample rate at which data are acquired.

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 rates that subcritically sample the time-sliced templates. The time slices consist of the S intervals $[t^0, t^1), [t^1, t^2), \ldots, [t^{S-1}, t^S)$, sampled at frequencies $f^0, f^1, \ldots, f^{S-1}$, where f^s is at least twice the highest nonzero frequency component of any filter in the bank for the sth time slice.

The time-sliced templates can then be downsampled in each interval without aliasing, so we define them as

$$h_i^s[k] \equiv \begin{cases} h_i\left[k\frac{f}{f^s}\right] & \text{if } t^s \leqslant k/f^s < t^{s+1} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

We note that the time slice decomposition in Equation (5) 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.

3.2.2. *Reducing the number of filters with the SVD*

As noted previously, the template banks used in inspiral searches are by design highly correlated. Cannon et al. (2010) showed that applying the SVD to inspiral template banks greatly reduces the number of filters required to achieve a particular minimal match. A similar technique can be applied to the time-sliced templates as defined in Equation (6) above. The SVD is a matrix factorization that takes the form

$$h_i^s[k] = \sum_{l=0}^{M-1} v_{il}^s \sigma_l^s u_l^s[k] \approx \sum_{l=0}^{L^s-1} v_{il}^s \sigma_l^s u_l^s[k]. \tag{7}$$

where $u_l^s[k]$ are orthonormal *basis templates* related to the original time-sliced templates through the *reconstruction matrix*, $v_{il}^s \sigma_l^s$. The expectation value of the fractional loss in S/N is the SVD tolerance, given by

$$\left[\sum_{l=0}^{L^s-1} \left(\sigma_l^s\right)^2 \right] \left[\sum_{l=0}^{M-1} \left(\sigma_l^s\right)^2 \right]^{-1},$$

determined by the number L^s of basis templates that are kept in the approximation. Cannon et al. (2010) showed that highly accurate approximations of inspiral template banks could be achieved with few basis templates. 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 the rank reduction demonstrated in Cannon et al. (2010) by nearly an order of magnitude.

Because the sets of filters from each time slice form orthogonal subspaces, and the basis filters within a given time slice are mutually orthogonal, the set of all basis filters from all time slices forms an orthogonal basis spanning the original templates.

In the next section, we describe how we form our early-warning detection statistic using the time slice decomposition and the SVD.

3.2.3. *Early-warning output*

In the previous two sections, we described two transformations that greatly reduce the computational burden of TD filtering. We are now prepared to define our detection statistic, the early-warning output, and to comment on the computational cost of evaluating it.

First, the sample rate of the detector data must be decimated to match sample rates with each of the time slices. We will denote the decimated detector data streams using a superscript "s" to indicate the time slices to which they correspond. The operator $H^{\downarrow}$ will represent the appropriate decimation filter that converts between the base sample rate f^0 and the reduced sample rate f^s:

$$x^s[k] = \left(H^{\downarrow}x^0\right)[k].$$

We shall use the symbol $H^{\uparrow}$ to represent an interpolation filter that converts between sample rates f^{s+1} and f^s of adjacent time slices,

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

From the combination of the time slice decomposition in Equation (6) and the SVD defined in Equation (7), we define the early-warning output accumulated up to time slice s using the recurrence relation,

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

Observe that the early-warning output for time slice 0, $\rho_i^0[k]$, approximates the S/N of the original templates. The signal flow diagram in Figure 3 illustrates this recursion relation as a multirate filter network with a number of early-warning outputs.

Ultimately, the latency of the entire LLOID algorithm is set by the decimation and interpolation filters because they are generally time symmetric and slightly acausal. Fortunately, as long as the latency introduced by the decimation and interpolation filters for any time slice s is less than that time slice's delay t^s, the total latency of the LLOID algorithm will be zero. To be concrete, suppose that the first time slice, sampled at a rate $f^0 = 4096$ Hz, spans times $[t^0, t^1) = [0$ s, 0.5 s), and the second time slice, sampled at $f^1 = 512$ Hz, spans $[t^1, t^2) = [0.5$ s, 4.5 s). Then the second time slice's output, $\rho_i^1[k]$, will *lead* the first time slice's output, $\rho_i^0[k]$, by 0.5 s. A decimation filter will be necessary to convert the 4096 Hz input signal $x[k] \equiv x^0[k]$ to the 512 Hz input $x^1[k]$, and an interpolation filter will be necessary to match the sample rates of the two early-warning outputs. In this example, as long as the decimation and interpolation filters are together acausal by less than $t^1 = 0.5$ s, the total S/N $\rho_i^0[k]$ will be available

Table 2
Notation Used to Describe Filters.

	Definition
f^s	Sample rate in time slice s
M	Number of templates
N	Number of samples per template
S	Number of time slices
L^s	Number of basis templates in time slice s
N^s	Number of samples in decimated time slice s
$N^{\downarrow}$	Length of decimation filter
$N^{\uparrow}$	Length of interpolation filter

with a latency of zero samples. When zero latency is important, we may take this as a requirement for the decimation and interpolation filter kernels.

In the next section, we compute the expected computational cost scaling of this decomposition and compare it with the direct TD implementation of Equation (4) and higher latency blockwise FD methods.

3.3. *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, Table 2 provides a review of the notation that we will need in this section.

3.3.1. *Conventional TD method*

The conventional, direct 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, at a sample rate f^0,

$$2MNf^0 \text{ flop s}^{-1}. \quad (9)$$

3.3.2. *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 each of 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 each 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

$$f^0 \cdot \frac{2(M+1)\lg D + 2M}{1 - N/D} \text{ flop s}^{-1}. \quad (10)$$

In the limit of many templates, $M \gg 1$, we can neglect the cost of the forward transform of the data and of the multiplication of the transforms. The computational cost will reach an optimum at some large but finite FFT block size $D \gg N$. In this limit, the FD method costs $\approx 2f^0M\lg D$ flop s^{-1}.

By adjusting the FFT block size, it is possible to achieve low latency with FD convolution, but the computational cost grows rapidly as the latency in samples ($D-N$) decreases. It
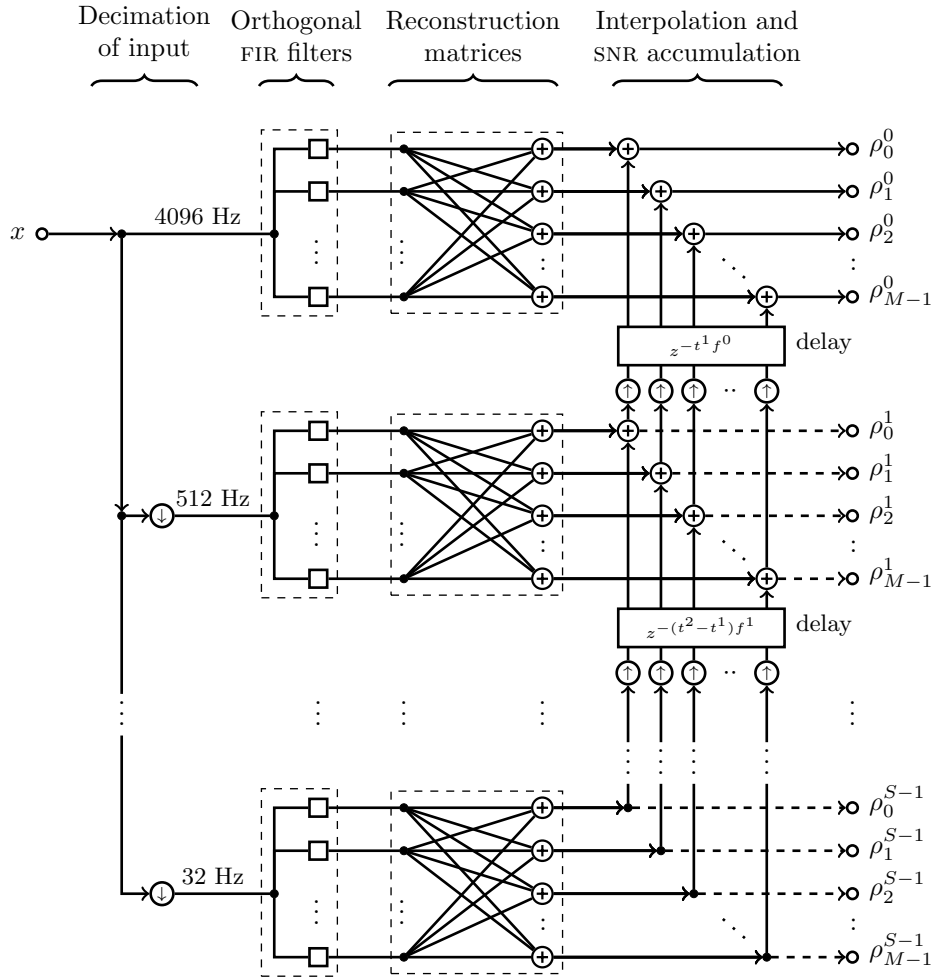
Figure 3. Schematic of LLOID pipeline illustrating signal flow. Circles with arrows represent interpolation ⊕ or decimation ⊖. Circles with plus signs represent summing junctions ⊕. Squares □ stand for FIR filters. Sample rate decreases from the top of the diagram to the bottom. In this diagram, each 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.

is easy to show that in the limit of many templates and long templates, $M, \lg N \gg 1$, the computational cost scales as

$$\left(1 + \frac{\text{template length}}{\text{latency}}\right)\left(2f^0 M \lg N\right).$$

3.3.3. *LLOID method*

For time slice s, the LLOID method requires $2N^s L^s f^s$ flop s^{-1} to evaluate the orthogonal filters, $2ML^s f^s$ flop s^{-1} to apply the linear transformation from the L^s basis templates to the M time-sliced templates, and $M f^s$ flop s^{-1} to add the resultant partial S/N 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 downsampling, or deleting samples in order to reduce the sample rate from f^{s-1} to f^s. Naively, an anti-aliasing filter with $(f^{s-1}/f^s)N^{\downarrow}$ coefficients should demand $2N^{\downarrow}(f^{s-1})^2/f^s$ flop s^{-1}. 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 requires only $2N^{\downarrow} f^{s-1}$ flop s^{-1}. (One common realization is an ingenious structure called a *polyphase decimator*, described in Chapter 1 of Jovanovic-Dolecek (2002).)

The story is similar for the interpolation filters used to match the sample rates of the partial S/N 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 $(f^{s-1}/f^s)N^{\uparrow}$ coefficients. The low-pass filter requires $2MN^{\uparrow}(f^{s-1})^2/f^s$ flop s^{-1}. However, by taking advantage of the fact that by construction a fraction f^s/f^{s-1} of the samples are zero, it is possible to build an efficient interpolator that requires only $2MN^{\uparrow} f^{s-1}$ flop s^{-1}. (Again, see Jovanovic-Dolecek (2002) for a discussion of *polyphase interpolation*.)

Taking into account the decimation of the detector data, the orthogonal FIR filters, the reconstruction of the time-sliced templates, the interpolation of S/N from previous time slices, and the accumulation of S/N, in total the LLOID algorithm requires

$$\sum_{s=0}^{S-1}\left(2N^s L^s + 2ML^s + M\right)f^s + 2\sum_{f^s\in\{f^k:0<k<S\}}\left(N^{\downarrow}f^0 + MN^{\uparrow}f^{s-1}\right) \quad (11)$$

flop s^{-1}. The second sum is carried out over the set of distinct sample rates (except for the base sample rate) rather than over the time slices themselves, as we have found that it is sometimes desirable to place multiple adjacent time slices at the same sample rate in order to keep the size of matrices that enter the SVD manageable. Here we have assumed that the decimation filters are connected in parallel, converting from the base sample rate f^0 to each of the time slice sample rates f^1, f^2, ..., and that the interpolation filters are connected in cascade fashion with each interpolation filter stepping from the sample rate of one time slice to the next.

We can simplify this expression quite a bit by taking some limits that arise from sensible filter design. In the limit of many templates, the cost of the decimation filters is negligible as compared to the cost of the interpolation filters. Typically, we will design the interpolation filters with $N^{\uparrow} \lesssim L^s$ so that the interpolation cost itself is negligible compared

with the reconstruction cost. Finally, if the number of basis templates per time slices L^s is not too small, the reconstruction cost dominates over the cost of accumulating the partial S/N. In these limits, the cost of LLOID is dominated by the basis filters themselves and the reconstruction, totaling $2\sum_{s=0}^{S-1}f^s L^s(N^s+M)$ flop s^{-1}.

3.3.4. *Speedup of LLOID relative to TD method*

If the cost of the basis filters dominates, and the frequency of the templates evolves slowly enough in time, then we can use the time-frequency relationship of Equation (1) to estimate the speedup relative to the conventional, direct TD method. The reduction in flop s^{-1} is approximately

$$\frac{2\sum_{s=0}^{S-1}f^s L^s N^s}{2MNf^0}$$

$$\approx \frac{\alpha}{\left(t_{\text{low}}-t_{\text{high}}\right)\left(f^0\right)^2}\int_{t_{\text{low}}}^{t_{\text{high}}}(2f(t))^2\,\mathrm{d}t$$

$$= \frac{16\alpha\left(t_{\text{low}}f^2(t_{\text{low}})-t_{\text{high}}f^2(t_{\text{high}})\right)}{\left(f^0\right)^2\left(t_{\text{low}}-t_{\text{high}}\right)} \quad (12)$$

where $\alpha \approx L^s/M$ is the rank reduction factor, or ratio between the number of basis templates and the number of templates. This approximation assumes that the frequency of the signal is evolving very slowly so that we can approximate the time slice sample rate as twice the instantaneous GW frequency, $f^s \approx 2f(t)$, and the number of samples in the decimated time slice as the sample rate times an infinitesimally short time interval, $N^s \approx 2f(t)\,\mathrm{d}t$. The integral is evaluated using the power-law form of $f(t)$ from Equation (1). Substituting approximate values for a template bank designed for component masses around (1.4, 1.4) $M_{\odot}$, $\alpha \approx 10^{-2}$, $t_{\text{low}} = 10^3$ s, $f_{\text{low}} = 10^1$ Hz, $f_{\text{high}} = f_{\text{ISCO}} \approx 1570$ Hz, $f^0 = 2f_{\text{ISCO}}$, and $t_{\text{high}} = f_{\text{ISCO}}^{-1}$, we find from Equation (12) that the LLOID method requires only $\sim 10^{-6}$ times as many flop s^{-1} as the conventional TD method.

4. IMPLEMENTATION

In this section we describe an implementation of the LLOID method described in Section 3 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 an offline planning stage and 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.

4.1. *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 minimal match criterion. This assures a prescribed maximum loss in S/N for signals whose parameters do not lie on the hexagonal grid. Next, the grid is partitioned into groups of neighbors called *sub-banks* that are appropriately sized so that each sub-bank can be efficiently handled by a single computer. Each sub-bank contains templates of comparable chirp mass, and therefore similar time-frequency evolution. Dividing the source

parameter space into smaller sub-banks also reduces the off-line cost of the SVD and is the approach considered in Cannon et al. (2010). Next, we choose time slice boundaries as in Equation (6) such that all of the templates within a sub-bank are sub-critically sampled at progressively lower sample rates. For each time slice, the templates are downsampled 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 orthonormal basis templates and a reconstruction matrix that maps them back to the original templates as described in Equation (7). The downsampled basis templates, the reconstruction matrix, and the time slice boundaries are all saved to disk.

4.2. *Filtering stage*

The LLOID algorithm is amenable to latency-free, real-time implementation. However, a real-time search pipeline would require integration directly into the data acquisition and storage systems of the LIGO observatories. A slightly more modest goal is to leverage existing low latency, but not real-time, signal processing software in order to implement the LLOID algorithm.

We have implemented a prototype of the low-latency filtering stage using an open-source signal processing environment called GStreamer[14] (version 0.10.33). 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 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 as resamplers and filters. We have extended the GStreamer framework by developing a library called `gstlal`[15] that provides elements for GW data analysis.

GStreamer pipelines typically operate with very low (in some consumer applications, imperceptibly low) latency rather than in true real time because signals are partitioned into blocks of samples, or *buffers*. This affords a number of advantages, including amortizing the overhead of passing signals between elements and grouping together sequences of similar operations. However, buffering a signal incurs a latency of up to one buffer length. This latency can be made small at the cost of some additional overhead by making the buffers sufficiently small. In any case, buffering is a reasonable strategy for low-latency LIGO data analysis because, as we previously remarked, the LIGO data acquisition system has a granularity of 1/16 s.

5. RESULTS

In this section we evaluate the accuracy of the LLOID algorithm using our GStreamer-based implementation described in the previous section. We calculate the measured S/N loss due to the approximations of the LLOID method and our implementation of it. Using a configuration that gives acceptable S/N loss for our chosen set of source parameters, we then compare the computational cost in flop s^{-1} for the direct TD method, the overlap-save FD method, and LLOID.

5.1. *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 from 1 to 3 $M_\odot$ using a simulated Advanced LIGO noise power spectrum (Shoemaker 2010)[16]. Waveforms are generated in the frequency domain in the stationary phase approximation at (post)$^{3.5}$-Newtonian order in phase and Newtonian order in amplitude (the TaylorF2 waveforms described in Buonanno et al. 2009). Templates are truncated at 10 Hz, where the projected sensitivity of Advanced LIGO is interrupted by the "seismic wall." This results in a grid of 98,544 points, or $2 \times 98,544 = 197,088$ templates. Then we create sub-banks by partitioning the parameter space by chirp mass. Figure 4 illustrates this procedure. We concentrate on a sub-bank with 657 points with chirp masses between 1.1955 and 1.2045 $M_\odot$, or $2 \times 657 = 1314$ templates. With this sub-bank we are able to construct an efficient time
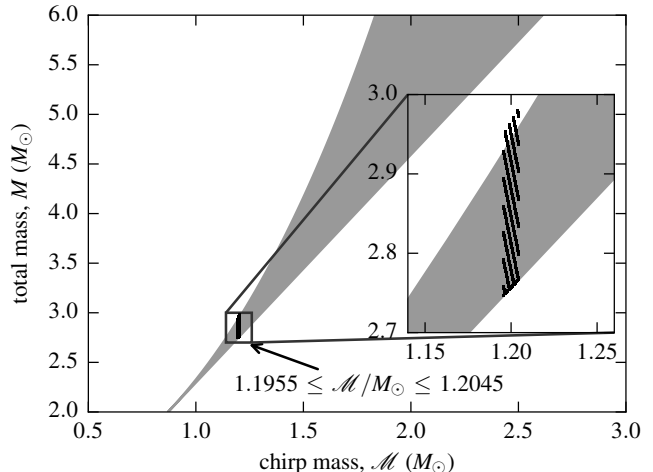

Figure 4. Source parameters selected for sub-bank used in this case study, consisting of component masses m_1 and m_2, between 1 and 3 $M_\odot$, and chirp masses $\mathcal{M}$ between 1.1955 and 1.2045 $M_\odot$.

slice decomposition that consists of 11 time slices with sample rates between 32 and 4096 Hz summarized in Table 3. We use this sub-bank and decomposition for the remainder of this section.

5.2. *Measured S/N loss*

The S/N loss is to be compared with the mismatch of 0.03 that arises from the discreteness of template bank designed for a minimal match of 0.97. We will consider an acceptable target S/N loss to be a factor of 10 smaller than this, that is, no more than 0.003.

We expect two main contributions to the S/N loss to arise in our implementation of the LLOID algorithm. The first is the S/N loss due to the truncation of the SVD at $L^s < M$ basis templates. As remarked upon in Cannon et al. (2010) and Section 3.2.2, this effect is measured by the SVD tolerance. The second comes from the limited bandwidth of the interpolation filters used to match the sample rates of the partial S/N streams. The maximum possible bandwidth is determined by the length of the filter, $N^\uparrow$. S/N loss could also arise if the combination of both the decimation filters and the interpolation filters reduces their bandwidth measurably, if the decimation and interpolation filters do not have perfectly uniform

Table 3
Filter Design Sub-Bank of 1314 Templates.

f^s (Hz)	$[t^s, t^{s+1})$ (s)	N^s	$-\log_{10}(1-\text{SVD tolerance})$ 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

Note. — 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 $(1-10^{-1})$ to $(1-10^{-6})$.

phase response, or if there is an unintended subsample time delay at any stage.

To measure the accuracy of our GStreamer implemention of LLOID including all of the above potential sources of S/N loss, we conducted impulse response tests. The GStreamer pipeline was presented with an input consisting of a unit impulse. By recording the outputs, we can effectively "play back" the templates. These impulse responses will be similar, but not identical, to the original, nominal templates. By taking the inner product between the impulses responses for each output channel with the corresponding nominal template, we can gauge exactly how much S/N is lost due to the approximations in the LLOID algorithm and any of the technical imperfections mentioned above. We call one minus this dot product the *mismatch* relative to the nominal template.

The two adjustable parameters that affect performance and mismatch the most are the SVD tolerance and the length of the interpolation filter. The length of the decimation filter affects mismatch as well, but has very little impact on performance.

Effect of SVD tolerance — We studied how the SVD tolerance affected S/N loss by holding $N^\downarrow = N^\uparrow = 192$ fixed as we varied the SVD tolerance from $(1-10^{-1})$ to $(1-10^{-6})$. The minimum, maximum, and median mismatch are shown as functions of SVD tolerance in Figure 5(a). As the SVD tolerance increases toward 1, the SVD becomes an exact matrix factorization, but the computational cost increases as the number of basis filters increases. 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 interpolation, though we still see that the average mismatch is approximately proportional to the SVD tolerance down to $(1-10^{-4})$. However, as the SVD tolerance becomes even higher, the median mismatch seems to saturate around 2×10^{-4}. This could be the effect of the interpolation, or an unintended technical imperfection that we did not model or

expect. However, this is still an order of magnitude below our target mismatch of 0.003. We find that an SVD tolerance of $(1-10^{-4})$ is adequate to achieve our target S/N loss.

Effect of interpolation filter length — Next, keeping the SVD tolerance fixed at $(1-10^{-6})$ and the length of the decimation filter fixed at $N^\downarrow = 192$, we studied the impact of the length $N^\uparrow$ of the interpolation filter on mismatch. We use GStreamer's stock `audioresample` element, which provides an FIR decimation filter with a Kaiser-windowed sinc function kernel. The mismatch as a function of $N^\uparrow$ is shown in Figure 5(b). The mismatch saturates at $\sim 2 \times 10^{-4}$ with $N^\uparrow = 64$. We find that a filter length of 16 is sufficient to meet our target mismatch of 0.003.

Having selected an SVD tolerance of $(1-10^{-4})$ and $N^\uparrow = 16$, we found that we could reduce $N^\downarrow$ to 48 without exceeding a median mismatch of 0.003.

We found that careful design of the decimation and interpolation stages made a crucial difference in terms of computational overhead. Connecting the interpolation filters in cascade fashion rather than in parallel resulted in a significant speedup. Also, only the shortest interpolation filters that met our maximum mismatch constraint resulted in a sub-dominant contribution to the overall cost. There is possibly further room for optimization beyond minimizing $N^\uparrow$. We could design custom decimation and interpolation filters, or we could tune these filters separately for each time slice.

5.3. *Other potential sources of S/N loss*

One possible source of S/N loss for which we have not accounted is the leakage of sharp spectral features in the detector's noise spectrum due to the short durations of the time slices. In the LLOID algorithm, as with many other GW search methods, whitening is treated as an entirely separate data conditioning stage. In this paper, we assume that the input to the filter bank is already whitened, having been passed
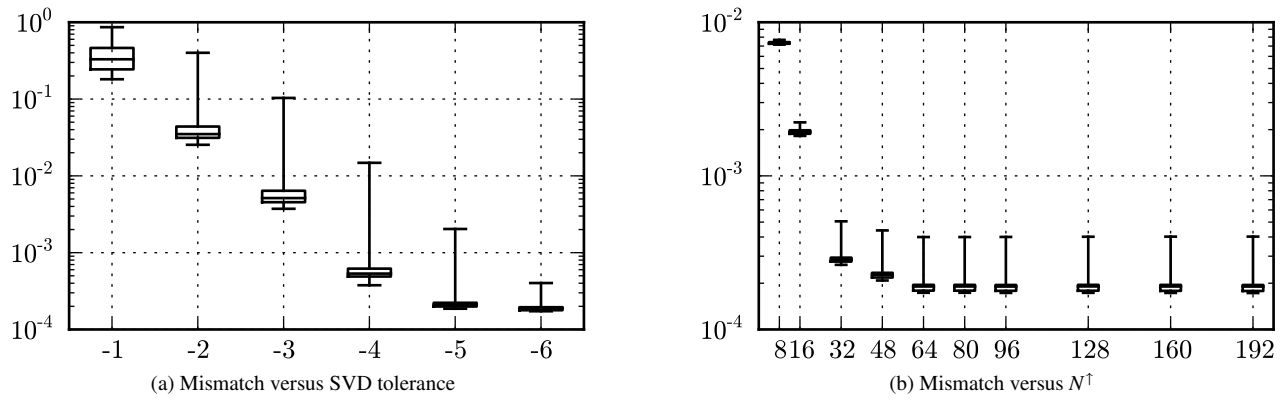
(a) Mismatch versus SVD tolerance

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

Figure 5. 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.

through a filter that flattens and normalizes its spectrum. We elected to omit a detailed description of the whitening procedure since the focus here is on the implementation of a scalable inspiral filter bank.

However, the inspiral templates themselves consist of the GW time series convolved with the impulse response of the whitening filter. As a consequence, the LLOID algorithm must faithfully replicate the effect of the whitening filter. Since in practice the noise spectra of ground-based GW detectors contain both high-Q lines at mechanical, electronic, and control resonances and a very sharp rolloff at the seismic wall, the frequency response of the LLOID filter bank must contain both high-Q notches and a very abrupt high-pass filter. FIR filters with rapidly varying frequency responses tend to have long impulse responses and many coefficients. Since the LLOID basis filters have, by design, short impulse responses and very few coefficients, one might be concerned about spectral leakage contaminating the frequency response of the LLOID filter bank.

The usual statement of the famous Nyquist-Shannon theorem, stated below as Theorem 1, has a natural dual, Theorem 2, that addresses the frequency resolution that can be achieved with an FIR filter of a given length.

Theorem 1. *(After Oppenheim et al. 1997, p. 518) Let $x(t)$ be a band-limited signal with continuous Fourier transform $\tilde{x}(f)$ such that $\tilde{x}(f') = 0 \,\forall\, f' : |f'| > f_M$. Then, $x(t)$ is uniquely determined by its discrete samples $x(n/f^0)$, $n = 0, \pm 1, \pm 2, \ldots$, if $f^0 > 2f_M$.*

Theorem 2. *Let $x(t)$ be a compactly supported signal such that $x(t') = 0 \,\forall\, t' : |t'| > t_M$. Then its continuous Fourier transform $\tilde{x}(f)$ is uniquely determined by the discrete frequency components $\tilde{x}(n\Delta f)$, $n = 0, \pm 1, \pm 2, \ldots$, if $\Delta f < 1/(2t_M)$.*

Another way of stating Theorem 2 is that, provided $x(t)$ is nonzero only for $|t| < 1/(2\Delta f)$, the continuous Fourier transform can be reconstructed at any frequency f from a weighted sum of sinc functions centered at each of the discrete frequency components, namely,

$$\tilde{x}(f) \propto \sum_{n=-\infty}^{\infty} \tilde{x}(n\Delta f)\,\mathrm{sinc}\left[\pi(f - n\Delta f)/\Delta f\right].$$

Failure to meet the conditions of this dual of the sampling theorem results in spectral leakage. For a TD signal to capture spectral features that are the size of the central lobe of the sinc function, the signal must have a duration greater than $1/\Delta f$. If the signal $x(t)$ is truncated by sampling it for a shorter duration, then its Fourier transform becomes smeared out; conceptually, power "leaks" out into the side lobes of the sinc functions and washes away sharp spectral features. In the GW data analysis literature, the synthesis of inspiral matched filters involves a step called *inverse spectrum truncation* (see Allen et al. 2011, Section VII) that fixes the number of coefficients based on the desired frequency resolution.

In order to effectively flatten a line in the detector's noise power spectrum, the timescale of the templates must be at least as long as the damping time τ of the line, $\tau = 2Q/\omega_0$, where Q is the quality factor of the line and w_0 is the central angular frequency. To put this into the context of the sampling theorem, in order to resolve a notch with a particular Q and f_0, an FIR filter must achieve a frequency resolution of $\Delta f \gtrsim \pi f_0/Q$ and therefore its impulse response must last for at least a time $1/\Delta f = Q/\pi f_0$. For example, in the S6 detector configuration known as "Enhanced LIGO," the violin modes (Penn et al. 2007) had $Q \sim 10^5$ and $\omega_0 \sim (2\pi)340$ rad s^{-1}, for a coherence time $\tau \sim 10^2$ s.

In our example template bank, many of the time slices are much shorter than this. However, in summation the time slices have the same duration as the full templates themselves, and the full templates are much longer than many coherence times of the violin mode. For this reason, we speculate that LLOID should be just as robust to sharp line features as traditional FFT-based searches currently employed in the GW field. Future works must verify this reasonable supposition with numerical experiments, including impulse response studies similar to the ones presented here but with detector noise power spectra containing lines with realistically high quality factors.

There could, in principle, be lines with coherence times many times longer than the template duration. For example, the Q of the violin modes may increase by orders of magnitude in Advanced LIGO (Strain & Cagnoli 2006). Also, there are certainly narrow lines that are non-stationary. Both of these cases can be dealt with by preprocessing $h(t)$ with band-stop filters that attenuates the lines themselves but also conservatively large neighborhoods around them. If such bandstops were implemented as an FIR filter, they could be built into the time slices without any difficulty.

Another way to deal with line features with coherence times much longer than the templates would be to entirely 'factor' the whitening out of the LLOID filter bank. Any line features could be notched out in the whitening stage with IIR filters, which can achieve infinitely high Q at just second order. If the detector data were passed through the whitening filter twice, then time-sliced filters need not depend on the detector's noise power spectral density at all. In such a variation on the LLOID method, the basis filters could be calculated from the *weighted* SVD (Gabriel & Zamir 1979; Jackson 2003, Chapter 3.6) of the time-sliced templates, using the covariance of the detector noise as a weight matrix.

5.4. *Lower bounds on computational cost and latency 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 $\left(1 - 10^{-4}\right)$, the interpolation filter length to 16, and the decimation filter length to 48. Table 4 shows the computational cost in flop s^{-1} for the sub-bank we described above. For the overlap-save FD method, an FFT block size of $D = 2N$ is assumed, resulting in a latency of $\left(N/f^0\right)$ seconds. Both the FD method and LLOID are five orders of magnitude faster than the conventional, direct TD method. However, the FD method has a latency of over half of an hour, whereas the LLOID method, with suitable design of the decimation and interpolation filters, has no more latency than the direct TD method.

5.5. *Extrapolation of computational cost to an Advanced LIGO search*

Table 4 shows that the LLOID method requires 6.6×10^8 flop s^{-1} to cover a sub-bank comprising 657 out of the total 98,544 mass pairs. Assuming that other regions of the parameter space have similar computational scaling, an entire single-detector search for NS–NS signals in the 1–3 $M_\odot$ component mass range could be implemented with $(98,544/657) \approx 150$ times the cost, or 9.9×10^{10} flop s^{-1}.

Table 4
Computational Cost in Flop s^{-1} and Latency in Seconds of the Direct TD Method, the Overlap-Save FD Method, and LLOID.

Method	Flop s^{-1} (Sub-bank)	Latency (s)	Flop s^{-1} (NS–NS)	number of Machines
Direct (TD)	4.9×10^{13}	0	3.8×10^{15}	$\sim 3.8 \times 10^5$
Overlap-save (FD)	5.2×10^8	2×10^3	5.9×10^{10}	~ 5.9
LLOID (theory)	6.6×10^8	0	1.1×10^{11}	~ 11
LLOID (prototype)	(0.9 cores)	0.5	————	$\gtrsim 10$

Note. — Cost is given for both the sub-bank described in Section 5.1 and a full 1–3 $M_\odot$ NS–NS search. The last column gives the approximate number of machines per detector required for a full Advanced LIGO NS–NS search.

We computed the computational cost of a full Advanced LIGO NS–NS search a second way by dividing the entire 1–3 $M_\odot$ parameter space into sub-banks of 657 points apiece, performing time slices and SVDs for each sub-bank, and tabulating the number of floating point operations using Expression (11). This should be a much more accurate measure because template length varies over the parameter space. Lower chirp mass templates sweep through frequency more slowly and require more computations while higher chirp mass templates are shorter and require fewer computations. Despite these subtleties, this estimate gave us 1.1×10^{11} flop s^{-1}, agreeing with the simple scaling argument above.

Modern (ca. 2011) workstations can achieve peak computation rates up to $\sim 10^{11}$ flop s^{-1}. In practice, we expect that a software implementation of LLOID will reach average computation rates that are perhaps a factor 10 less than this, $\sim 10^{10}$ flop s^{-1} per machine, due to non-floating point tasks including bookkeeping and thread synchronization. Given these considerations, we estimate that a full Advanced LIGO, single-detector, NS–NS search with LLOID in will require ~ 10 machines.

By comparison, using the conventional TD method to achieve the same latency costs 4.9×10^{13} flop s^{-1} for this particular sub-bank, and so simply scaling up by the factor of 150 suggests that it would require 7.4×10^{15} flop s^{-1} to search the full parameter space. To account for the varying sample rate and template duration across the parameter space, we can also directly calculate the cost for the full TD method search using expression (9), resulting in 3.8×10^{15} flop s^{-1}, agreeing within an order of magnitude. This would require $\gtrsim 10^5$ current-day machines. Presently, the LIGO Data Grid[17] consists of only $\sim 10^4$ machines, so direct TD convolution is clearly impractical.

The overlap-save FD method is slightly more efficient than LLOID for this particular sub-bank, requiring 5.2×10^8 flop s^{-1}. The scaling argument projects that a full FD search would require 7.8×10^{10} flop s^{-1}. The direct calculation from Expression (10) gives 5.9×10^{10} flop s^{-1}, in order-of-magnitude agreement. In this application, the conventional FD search is scarcely a factor of two faster than LLOID while gaining only 0.3% in S/N, but only at the price of thousands of seconds of latency.

5.6. *Measured latency and overhead*

Our GStreamer pipeline for measuring impulse responses contained instrumentation that would not be necessary for

---

[17] https://www.lsc-group.phys.uwm.edu/lscdatagrid/

an actual search, including additional interpolation filters to bring the early-warning outputs back to the full sample rate and additional outputs for recording signals to disk.

We wrote a second, stripped pipeline to evaluate the actual latency and computational overhead. We executed this pipeline on one of the submit machines of the LIGO-Caltech cluster, a Sun Microsystems Sun Fire$^{\text{TM}}$ X4600 M2 server with eight quad-core 2.7 GHz AMD Opteron$^{\text{TM}}$ 8384 processors. This test consumed $\sim 90\%$ of the capacity of just one out of the 32 cores, maintaining a constant latency of ~ 0.5 s.

The measured overhead is consistent to within an order of magnitude with the lower bound from the flop s^{-1} budget. Additional overhead is possibly dominated by thread synchronization. A carefully optimized GStreamer pipeline or a hand-tuned C implementation of the pipeline might reduce overhead further.

The 0.5 s latency is probably due to buffering and synchronization. The latency might be reduced by carefully tuning buffer lengths at every stage in the pipeline. Even without further refinements, our implementation of the LLOID algorithm has achieved latencies comparable to the LIGO data acquisition system itself.

6. CONCLUSIONS

We have demonstrated a computationally feasible filtering algorithm for the rapid and even early-warning detection of GWs emitted during the coalescence of NSs and stellar-mass black holes. It is one part of a complicated analysis and observation strategy that will unfortunately have other sources of latency. However, we hope that it will motivate further work to reduce such technical sources of GW observation latency and encourage the possibility of even more rapid EM follow-up observations to catch prompt emission in the advanced detector era.

CBC events may be the progenitors of some short hard GRBs and are expected to be accompanied by a broad spectrum of EM signals. Rapid alerts to the wider astronomical community will improve the chances of detecting an EM counterpart in bands from gamma-rays down to radio. In the Advanced LIGO era, it appears possible to usefully localize a few rare events prior to the GRB, allowing multi-wavelength observations of prompt emission. More frequently, low-latency alerts will be released after merger but may still yield extended X-ray tails and early on-axis afterglows.

The LLOID method is as fast as conventional FFT-based, FD convolution but allows for latency free, real-time operation. We anticipate requiring $\gtrsim 40$ modern multi-core computers to search for binary NSs using coincident GW data from a four-detector network. In the future, additional computational savings could be achieved by conditionally reconstructing the S/N time series only during times when a composite detection statistic crosses a threshold (Cannon et al. 2011). However, the anticipated required number of computers is well within the current computing capabilities of the LIGO Data Grid.

We have shown a prototype implementation of the LLOID algorithm using GStreamer, an open-source signal processing platform. Although our prototype already achieves latencies of less than one second, further fine tuning may reduce the latency even further. Ultimately the best possible latency would be achieved by tighter integration between data acquisition and analysis with dedicated hardware and software. This could be considered for third-generation detector design. Also possible for third-generation instruments, the LLOID method could provide the input for a dynamic tuning of detector re-

sponse via the signal recycling mirror to match the frequency of maximum sensitivity to the instantaneous frequency of the GW waveform. This is a challenging technique, but it has the potential for substantial gains in S/N and timing accuracy (Meers et al. 1993).

Although we have demonstrated a computationally feasible statistic for advance detection, we have not yet explored data calibration and whitening, triggering, coincidence, and ranking of GW candidates in a framework that supports early EM follow-up. One might explore these and also using the time slice decomposition and the SVD to form low-latency signal-based vetoes (e.g., χ^2 statistics) that have been essential for glitch rejection used in previous GW CBC searches. These additional stages may incur some extra overhead, so computing requirements will likely be somewhat higher than our estimates.

Future work must more deeply address sky localization accuracy in a realistic setting as well as observing strategies. Here, we have followed Fairhurst (2009) in estimating the area of 90% localization confidence in terms of timing uncertainties alone, but it would be advantageous to use a galaxy catalog to inform the telescope tiling (Nuttall & Sutton 2010). Because early detections will arise from nearby sources, the galaxy catalog technique might be an important ingredient in reducing the fraction of sky that must be imaged. Extensive simulation campaigns incorporating realistic binary merger rates and detector networks will be necessary in order to fully understand the prospects for early-warning detection, localization, and EM follow-up using the techniques we have described.

APPENDIX

LOW FREQUENCY CUTOFF FOR INSPIRAL SEARCHES

Ground-based GW detectors are unavoidably affected at low frequencies by seismic and anthropogenic ground motion. The LIGO test masses are suspended from multiple-stage pendula, which attenuate ground motion down to the pole frequency. In the detector configuration in place during S6, seismic noise dominated the instrumental background below about 40 Hz. Considerable effort is being invested in improving seismic attenuation in Advanced LIGO using active

and passive isolation (Harry & the LIGO Scientific Collaboration 2010), so that suspension thermal noise will dominate down to 10–15 Hz. Inspiral waveforms are chirps of infinite duration, but since an interferometric detector's noise diverges at this so-called "seismic wall," templates for matched filter searches are truncated at a low-frequency cutoff f_{low} in order to save computational overhead with negligible loss of SNR.

The expected matched-filter SNR, integrated from f_{low} to f_{high}, is given by Equation (2). The high-frequency cutoff for the inspiral is frequently taken to be the GW frequency at the LSO; for non-spinning systems, $f_{LSO} = 4400(M_\odot/M)$ Hz, where M is the total mass of the binary (section 3.4.1 of Sathyaprakash & Schutz 2009). The choice of f_{low} is based on the fraction of the total SNR that is accumulated at frequencies above f_{low}. To illustrate the relative contributions to the SNR at different frequencies for a (1.4, 1.4) $M_\odot$ binary, we normalized and plotted the integrand of Equation (2), the noise-weighted power spectral density of the inspiral waveform, in Figure 6(b). This is the quantity

$$\frac{1}{\rho^2}\frac{d\rho^2}{df} = \frac{f^{-7/3}}{S(f)}\left(\int_0^{f_{LSO}}\frac{f'^{-7/3}}{S(f')}\,df'\right)^{-1},$$

which is normalized by the total SNR squared in order to put detectors with different absolute sensitivities on the same footing. We used several different noise power spectra: all of the envisioned Advanced LIGO configurations from Shoemaker (2010); the best-achieved sensitivity at LIGO Hanford Observatory (LHO) in LIGO's fifth science run (S5), measured by Abadie et al. (2010a); and the best-achieved sensitivity at LHO during S6, measured by Abadie et al. (2011b). (The noise spectra themselves are shown in Figure 6(a).) It is plain that during S5 and S6 the greatest contribution to the S/N was between 100 and 150 Hz, but for all of the proposed Advanced LIGO configurations the bulk of the S/N is accumulated below 60 Hz. This information is presented in a complementary way in Figure 6(c), as the square root of the cumulative integral from f_{low} to f_{LSO}, interpreted as a fraction of the total "available" S/N,

$$\rho_{frac}(f_{low}) = \sqrt{\left(\int_{f_{low}}^{f_{LSO}}\frac{f^{-7/3}}{S(f)}\,df\right)\left(\int_0^{f_{LSO}}\frac{f^{-7/3}}{S(f)}\,df\right)^{-1}}.$$

Table 5 shows the fractional accumulated S/N for four selected low-frequency cutoffs, 40 Hz, 30 Hz, 20 Hz, and 10 Hz. In S5 and S6, all of the S/N is accumulated above 40 Hz. For the 'high frequency' Advanced LIGO configuration, scarcely *half* of the S/N is accumulated above 40 Hz. For the preferred final configuration, 'zero detuning, high power,' 86.1% of the S/N is above 40 Hz, 93.2% is above 30 Hz, and 98.1% is above 20 Hz. (Since S/N accumulates in quadrature, this means, on the other hand, that under the 'high frequency' configuration a template encompassing *just the early inspiral* from 10 to 40 Hz would accumulate $\sqrt{1-0.533^2} \approx 84.6\%$ of the total S/N! In the 'zero detuning, high power,' configuration, integration from 10 to 40 Hz alone would yield 50.9% of the total S/N, from 10 to 30 Hz, 36.2%, and from 10 to 20 Hz, 19.4%.)

Since the GW amplitude is inversely proportional to the luminosity distance of the source, and the sensitive volume is proportional to distance cubed, the rate of detectable coalescences depends on the choice of low-frequency cutoff. An inspiral search that is designed with a low-frequency cutoff

Table 5
Fractional Accumulated S/N $\rho_{frac}(f_{low})$ for Four Selected Low Frequency
Cutoffs, $f_{low} = 40$ Hz, 30 Hz, 20 Hz, and 10 Hz.

Noise model	40 Hz	30 Hz	20 Hz	10 Hz
LHO (best S5)	100.0	100.0	100.0	100.0
LHO (best S6)	100.0	100.0	100.0	100.0
High frequency	53.3	80.1	97.6	100.0
No SRM	87.8	95.1	98.7	100.0
BHBH 20°	71.1	84.2	96.2	100.0
NSNS optimized	91.5	96.3	99.0	100.0
Zero detuning, low power	67.9	80.0	93.5	100.0
Zero detuning, high power	86.1	93.2	98.1	100.0

at the seismic wall would gain an increase in detection rate of $\rho_{frac}^{-3}(f_{low})$ relative to a search with a low-frequency cutoff of f_{low}. This would represent almost a twofold increase in the rate of detection over a search with a fractional accumulated S/N of 80%, and still a 37% increase over a search with $\rho_{frac} = 90\%$. Existing coalescing binary detection pipelines strive to sacrifice no more than 3% of the available S/N; this forfeits less than a 10% gain in detection rate. In order to satisfy this constraint, the low-frequency cutoff would have to be placed below 30 Hz for all of the conceived Advanced LIGO configurations.

The instantaneous GW frequency, given by Equation (1), is a power law function of time, so the amount of time for the GW frequency to evolve from f_{low} to f_{LSO} depends strongly on f_{low}. The duration of a $(1.4, 1.4)$ $M_\odot$ inspiral is show in Figure 6(d). The inspiral takes only 25 s to evolve from 40 Hz to f_{LSO}, but takes 54 s to evolve from 30 Hz to f_{LSO}, 158 s from 20 Hz, and 1002 s from 10 Hz.

REFERENCES

Abadie, J., et al. 2010a, LIGO-T0900499-v19
—. 2010b, Class. Quantum Grav., 27, 173001
—. 2011a, arXiv:1109.3498v1
—. 2011b, arXiv:1111.7314v3
—. 2011c, ApJ (in press)
—. 2012, arXiv:1112.6005v4
Allen, B. A., Anderson, W. G., Brady, P. R., Brown, D. A., & Creighton, J. D. E. 2011, arXiv:gr-qc/0509116v2
Atteia, J.-L., & Boër, M. 2011, C. R. Phys., 12, 255
Bork, R., Abbott, R., Barker, D., & Heefner, J. 2001, in Proc. 8th Int. Conf. on Accelerator and Large Experimental Physics Control Systems, ed. H. Shoaee (Menlo Park, CA: Stanford Linear Accelerator Center Technical Publications), 19–23
Bucciantini, N., Metzger, B. D., Thompson, T. A., & Quataert, E. 2012, MNRAS, 419, 1537
Buonanno, A., Iyer, B. R., Ochsner, E., Pan, Y., & Sathyaprakash, B. S. 2009, Phys. Rev. D, 80, 084043
Buskulic, D., the LIGO Scientific Collaboration, & the Virgo Collaboration. 2010, Class. Quantum Grav., 27, 194003
Cannon, K., Chapman, A., Hanna, C., Keppel, D., Searle, A. C., & Weinstein, A. J. 2010, Phys. Rev. D, 82, 44025
Cannon, K., Hanna, C., Keppel, D., & Searle, A. C. 2011, Phys. Rev. D, 83, 084053
Cokelaer, T. 2007, Phys. Rev. D, 76, 102004
Fairhurst, S. 2009, New J. Phys., 11, 123006
Field, S. E., Galley, C. R., Herrmann, F., Hesthaven, J. S., Ochsner, E., & Tiglio, M. 2011, Phys. Rev. Lett., 106, 221102
Gabriel, K. R., & Zamir, S. 1979, Technometrics, 21, 489
Gardner, W. G. 1995, J. Audio Eng. Soc, 43, 127
Harry, G. M., & the LIGO Scientific Collaboration. 2010, Class. Quantum Grav., 27, 084006
Hooper, S., Wen, L., Blair, D., Chung, S. K., Chen, Y., & Luan, J. 2010, in AIP Conf. Proc., Vol. 1246, Frontiers of Fundamental and Computational Physics: 10th Intl. Symp., ed. J. G. Hartnett & P. C. Abbott (Melville, NY: AIP), 211–214
Hughey, B. 2011, in Gravitational Wave Physics and Astronomy Workshop (GWPAW)
Ivezic, Z., et al. 2008, arXiv:0805.2366v2
Jackson, J. 2003, A User's Guide to Principal Components, Wiley Series in Probability and Mathematical Statistics (New York: Wiley-Interscience)
Janka, H. T., Eberl, T., Ruffert, M., & Fryer, C. L. 1999, ApJ, 527, L39
Johnson, S., & Frigo, M. 2007, IEEE Trans. Signal Process., 55, 111
Jovanovic-Dolecek, G. 2002, Multirate Systems: Design and Applications (Hershey, PA: Idea Group Publ.)
Kaiser, N., et al. 2002, in Proc. SPIE, ed. J. A. Tyson & S. Wolff, Vol. 4836, 154–164
Kalogera, V., et al. 2004, ApJ, 601, L179, erratum in 614, L137
Kanner, J., Huard, T. L., Márka, S., Murphy, D. C., Piscionere, J., Reed, M., & Shawhan, P. 2008, Class. Quantum Grav., 25, 184034
Law, N. M., et al. 2009, PASP, 121, 1395
Lee, W. H., Ramirez-Ruiz, E., & Granot, J. 2005, ApJ, 630, L165
Luan, J., Hooper, S., Wen, L., & Chen, Y. 2011, arXiv:1108.3174v1
Marion, F., & the Virgo Collaboration. 2003, in Proc. Rencontres de Moriond 2003, ed. J. Dumarchez, Les Arcs
Meers, B. J., Krolak, A., & Lobo, J. A. 1993, Phys. Rev. D, 47, 2184
Nakar, E. 2007, Phys. Rept., 442, 166
Nuttall, L. K., & Sutton, P. J. 2010, Phys. Rev. D, 82, 102002
Oppenheim, A., Willsky, A., & Nawab, S. 1997, Signals and systems, Prentice-Hall signal processing series (Upper Saddle River, NJ: Prentice Hall)
Owen, B. J., & Sathyaprakash, B. S. 1999, Phys. Rev. D, 60, 022002
Penn, S., Kipperman, S., Newman, E., Stephens, P., Harry, G., Saulson, P., & Gretarsson, A. 2007, LIGO-G070553-00-Z
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007, Numerical Recipes, 3rd edn. (Cambridge: Cambridge Univ. Press)
Racusin, J. L., et al. 2008, Nature, 455, 183
Sari, R., & Piran, T. 1999, ApJ, 520, 641
Sathyaprakash, B., & Schutz, B. 2009, Living Rev. Relativity 12, arXiv:0903.0338v1
Shibata, M., & Taniguchi, K. 2008, Phys. Rev. D, 77, 084015
Shoemaker, D. 2010, LIGO-T0900288-v3
Strain, K., & Cagnoli, G. 2006, LIGO-T050267-01-K
Turin, G. 1960, IRE Trans. Inf. Theory, 6, 311
Virgili, F. J., Zhang, B., O'Brien, P., & Troja, E. 2011, ApJ, 727, 109

(a) LIGO noise models

(b) Signal to noise per unit frequency

(c) Accumulated SNR vs. f_{low}

(d) Inspiral duration vs. f_{low}

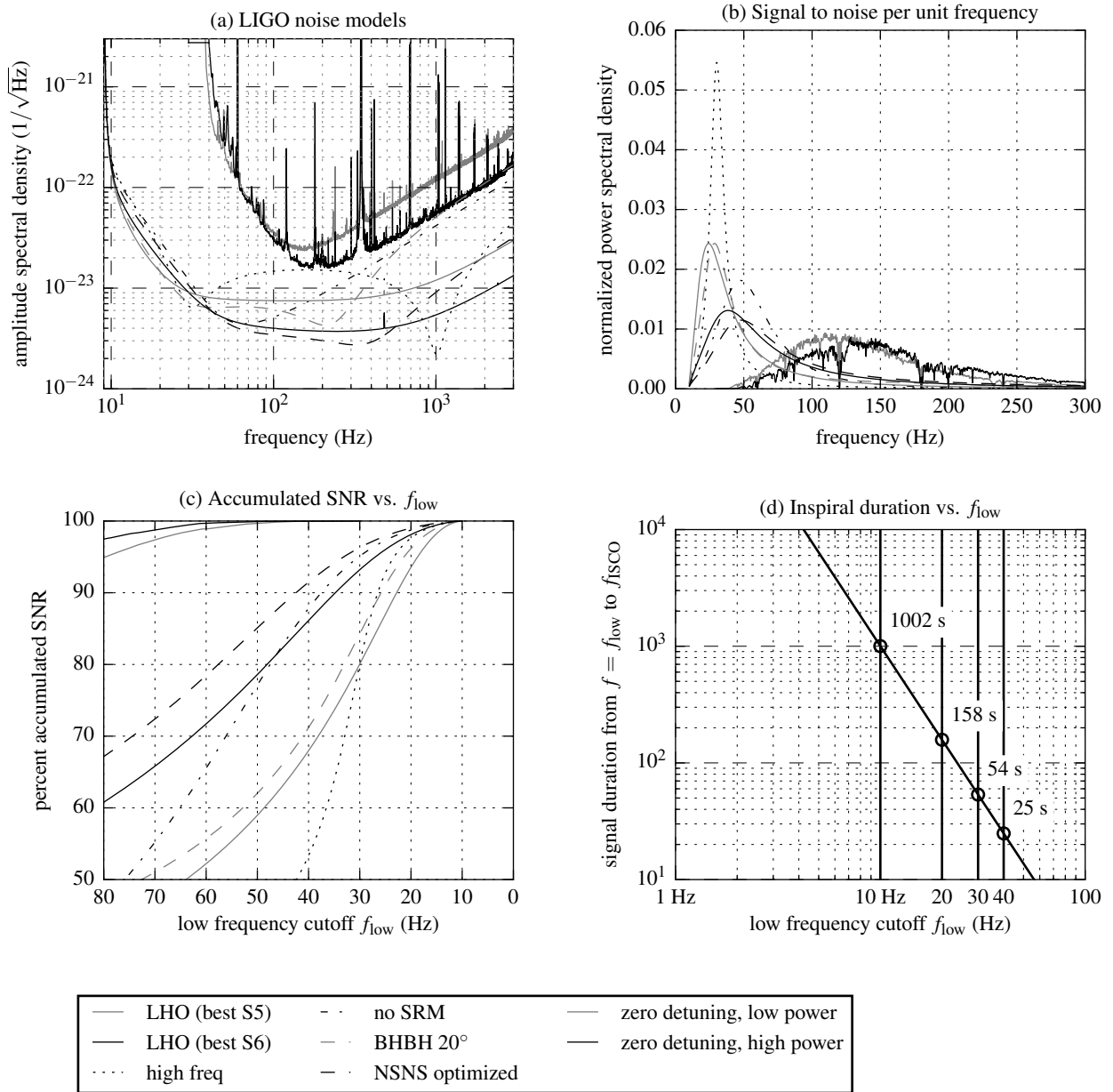|		
—— LHO (best S5)	– · no SRM	—— zero detuning, low power
—— LHO (best S6)	– · BHBH 20°	—— zero detuning, high power
···· high freq	— · NSNS optimized	

Figure 6. From top left: (a) noise amplitude spectral density for a variety of Advanced LIGO noise models, S5, and S6. (b) Normalized signal-to-noise per unit frequency, $(d\rho^2/df)/\rho^2$, for a (1.4, 1.4) $M_\odot$ inspiral. (c) Percentage of S/N that is accumulated from f_{low} to f_{LSO}, relative to S/N accumulated from $f_{low} = 0$ Hz to f_{LSO}. (d) Amount of time for a NS–NS inspiral signal to evolve from frequency f_{low} to f_{LSO}, as a function of f_{low}. For (a)–(c), the line style indicates which noise model was used.