Improving the prospects for multimessenger astronomy with early-warning detection of compact binary coalescence.

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

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

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

PACS numbers: 95.55.Ym, 84.30.Vn, 95.75.Wx

1. Introduction

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

Would the introduction be more effective without this paragraph?

As a compact binary system loses energy to gravitational waves, its orbital separation decreases. This causes a run-away inspiral with the gravitational-wave amplitude and frequency increasing until the system eventually merges near the innermost stable circular orbit (ISCO). If a neutron star is involved it may

become tidally disrupted near the merger. This disrupted matter can fuel a bright electromagnetic counterpart in the system's final moments as a binary [6].

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

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

I think we should say also something about how we can also get the full SNR *within moments after the coalescence, it is obvious maybe, but we should be explicit*

The gravitational-wave community initiated a project to send alerts when potential gravitational-wave transients are observed. In October 2010, LIGO completed its sixth science run (S6) and Virgo completed its third science run (VSR3). While both LIGO detectors and Virgo were operating, several all-sky detection pipelines operated in a low-latency configuration, namely MBTA, ihope, Coherent WaveBurst, and Omega [20, 21]. The S6 analyses achieved latencies of 30–60 minutes, which were dominated by a human vetting process. Candidates were sent for electromagnetic followup to several telescopes; Swift, ROTSE, TAROT, and Zadko [22, 20] took images of likely sky locations. MBTA achieved the best gravitational-wave trigger generation latencies of 2–5 minutes. We assume that in the advanced detector era the vetting process will be automated, so current trigger generation and telescope actuation would then dominate latency.

Citation needed for LOOC-UP

Get references for these low-latency pipelines.

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

Fortunately, the morphology of inspiral signals can be exploited to offset some of the computational complexity of low latency algorithms. First, the signals evolve slowly in frequency, so that they can be broken into contiguous bandlimited time

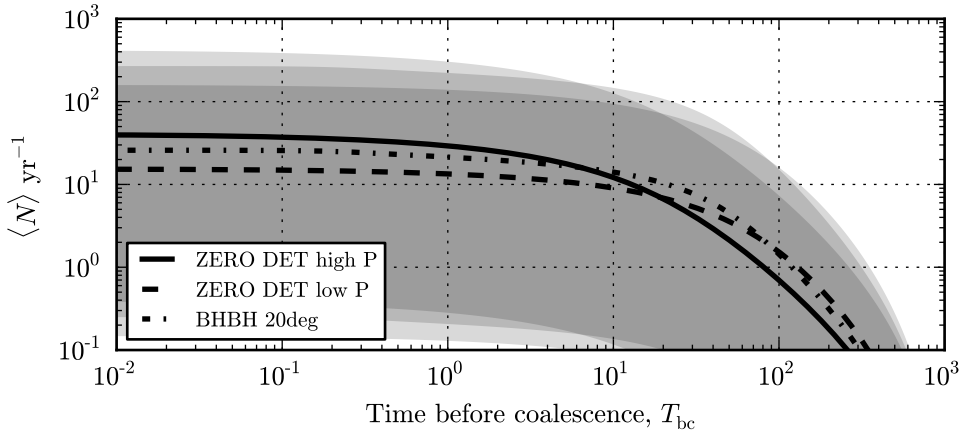


Figure 1: Expected number of NS–NS sources that will be detectable $T_{\rm bc}$ seconds before coalescence. The solid line is the most likely yearly rate estimate $N_{\rm re}$ for Advanced LIGO and the shaded region is the interval $N_{\rm low}$ to $N_{\rm high}$ from [5]. Note that assuming SNR 8 is sufficient for detection and that we observe $N_{\rm re} = 40$ events per year with a detector having the ZERO_DET_high_P noise model described in [17], ~ 10 sources detectable within 10 seconds of merger and ~ 1 sources detectable within 100 seconds of merger. Other noise models, ZERO_DET_low_P [18] and BHBH_20deg [19] provide better results for early detection but at the cost of fewer total events observed above SNR 8.

intervals and processed at possibly lower sample rates. Second, inspiral filter banks consist of highly similar templates, admitting principal component analysis to reduce the number of templates. We described a rank-reduction scheme based on singular value decomposition in [23]. We will use both aspects to demonstrate that a very low latency analysis with predictive detection of compact binary sources is possible with current computing resources. Assuming other aspects of gravitational-wave observation latency can be reduced significantly, this should allow the possibility for prompt alerts to be sent to the astronomical community.

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

2. Early warning searches for compact binary coalescence

In this section we describe a decomposition of the compact binary parameter space that reduces low-latency filtering cost sufficiently to allow for the possibility of early-warning detection with modest computing requirements. We expand on the ideas of [24, 25] that describe a multiband decomposition of the compact binary parameter space that resulted in search with ∼minutes latency in LIGO's S6 and Virgo's VSR2 science runs. We combine this with the orthogonal decomposition described in [23] that exploits the redundancy of the template banks.

2.1. Conventional CBC *matched filter searches*

Inspiral signals are parameterized by a set of intrinsic parameters $\bar{\theta}$ that determine the amplitude and phase evolution of the gravitational wave strain. For systems where the effects of spin can be ignored, the intrinsic parameters are the component masses of the binary, $\bar{\theta} = (m_1, m_2)$. Searches for gravitational waves from compact binary coalescence typically employ matched filter banks that discretely sample the possible intrinsic parameters [26]. The filters for the waveforms $h(t, \bar{\theta})$, known as templates, are the waveforms weighted by the inverse detector noise amplitude spectral density. To construct a template bank, templates are chosen with discrete signal parameters $\theta_1, \theta_2, \cdots, \theta_N$ to assure a bounded loss of SNR [27, 28]. That is, any possible signal within the search space will have a cross-correlation of ≥ 0.97 with at least one template. Such a template bank is said to have a *minimum match* of 0.97. Data are filtered against each template to produce an SNR time-series. Local peak-finding across time and templates determines detection candidates. ‡

Change this to a harpoon or vector symbol.

We will denote the i^{th} filter with parameters $\bar{\theta}_i$ as a function of time $x_i(\tau)$. In this work we will discuss transformations to a set of filters $\{x_i(\tau)\}$. Some of these transformations are not useful or practical over the entire parameter space. For that reason we assume from here onward that the set of filters $\{x_i(\tau)\}$ refer to a set of near-neighbor filters that can be chosen as a subset of the full parameter space. Several such subsets can be chosen until all of the filters are a member of one local set.

Filtering the detector data $h(t)$ involves a convolution of the data with the filter. For a unit-normalized filter, and whitened detector data, the result can be interpreted as the signal-to-noise ratio, $\rho_i(t)$ and is defined as

$$\rho_i(t) = \int x_i(\tau)\, h(t-\tau)\, d\tau \tag{1}$$

$$= \int \tilde{x}_i(f)\, \tilde{h}(f)\, \mathrm{e}^{-2\pi \imath f t} df, \tag{2}$$

nvf: I don't think that the splitting of the bank is a crucial point here. Perhaps later if we discuss the up-front cost of the SVD.

where the second line is a result of the convolution theorem and $\tilde{x}_i(f)$ is the Fourier transform of $x_i(\tau)$ as is $\tilde{h}(f)$ the Fourier transform of $h(t)$.

The evaluation of the integrals in (1) and (2) are implemented as sums over sample points for the digitized gravitational-wave detector output. Discrete Fourier transforms can be computed efficiently numerically. For that reason (2) is typically far faster, computationally. To evaluate (1) requires $\mathcal{O}[N_{x_i} N]$ floating point operations per filter x_i, where N_{x_i} is the number of sample points in the filter x_i and N is the number of sample points in the data h. Assuming N_t filters are required, the total cost is $\mathcal{O}[N_t N_{x_i} N]$ However, (2) requires only $\mathcal{O}[N \log N]$ operations per filter assuming transform lengths that are longer than the filter (i.e. $N > N_{x_i}$), resulting in a total cost of $\mathcal{O}[N_t N \log N]$. In most cases $N_{x_i} \gg \log N$ and the computational savings by choosing the frequency-domain integral form (2) is clear. However, to take full advantage of the computational efficiency of (2) requires an acausal knowledge of the detector data $h(t)$, which implies an inherent latency. In contrast, (1) can be updated every time a new sample point of detector data is taken.

‡ There are two gravitational-wave polarizations, $+$ and $\times$. A given detector will observe a combination of these polarizations that will largely be degenerate with an overall unknown constant phase. This can be maximized over by filtering for quadrature phases and taking the magnitude of the result. For simplicity we will ignore that aspect in this work, as it is straightforward to generalize but not necessary for understanding any of the points that will be made.

2.2. Proposed method

In order to minimize latency we propose using the time-domain convolution presented in (1). However, because the brute-force evaluation of (1) is far too costly to be useful, we will consider an approximation to (1) that can reduce substantially the cost of real-time filtering. This approximation has the form

$$\rho_i(t) \approx \sum_{k}^{N_{\mathrm{ts}}} \sum_{j}^{N_{\mathrm{u_j},k}} \int_{\tau_k}^{\tau_{k+1}} v_{ijk}\sigma_{jk}u_{jk}\left(\tau\right) h(t-\tau)\, d\tau \tag{3}$$

where $u_{jk}(\tau)$ is an orthogonal basis set of filters spanning the space of $\{x_i(\tau)\}$ and $\sigma_{jk}v_{ijk}$ is a tensor relating the filters $u_{jk}(\tau)$ to the original filter set $\{x_i(\tau)\}$. We claim that with a suitable choice of filters $u_{jk}(\tau)$ one can reduce the computational cost of (1) sufficiently to feasibly search for gravitational waves from compact binary coalescence in real-time. This requires 1) exploiting the time-frequency characteristics of the binary waveforms and 2) exploiting the redundancy of the template bank. We describe our procedure for producing the decompostion in (3) in the remainder of this section.

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

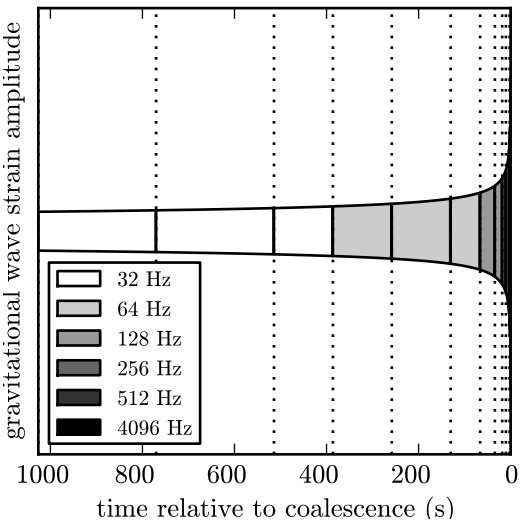
The first step of the orthogonal decomposition described in (3) is to divide the templates into *time slices*. This is a time-domain analogue to the frequency-domain decomposition described in [24, 25, 29, 30]. A matched filter is constructed for each time slice. The outputs form an ensemble of partial 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 singular value docomposition, is sufficient to enable a computationally efficient time-domain search and furthermore is an essential part of an early-warning detection scheme.

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

$$f = \frac{1}{\pi\mathcal{M}}\left[\frac{5}{256}\frac{\mathcal{M}}{-t}\right]^{3/8}. \tag{4}$$

Here, $\mathcal{M}$ is the chirp mass of the binary in units of time (where $GM_\odot/c^3 \approx 5\mu\mathrm{s}$) and t is the time relative to the coalescence of the binary [26, 31]. Usually the template is truncated at some prescribed time t_0, or equivalently frequency $f_{\max}$. This is often chosen to correspond to the ISCO defined previously. An inspiral signal will enter the detection band at a low frequency, $f = f_{\mathrm{low}}$ corresponding to a time t_{low}. The template is assumed to be zero outside the interval $[t_{\mathrm{low}}, t_0)$ and is said to have a duration of $t_0 - t_{\mathrm{low}}$. It is critically sampled at a rate of $2f_{\max}$.

The monotonic time-frequency relationship of (4) 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 filter integral at a lower sample rate. Specifically we consider here time slice boundaries with the next highest power-of-two sample rate that critically samples the time sliced filter. The time slices for this template consist of the k intervals $(t_k, t_{k-1}], \ldots, (t_2, t_1], (t_1, t_0]$ sampled at frequencies $f_{\mathrm{k}-1}, \ldots, f_1, f_0$ where $f_0 \geq 2f_{\mathrm{ISCO}}$, $t_0 = t_{\mathrm{ISCO}}$, $f_{k-1} \geqslant 2f_{\mathrm{low}}$, and $t_k \leqslant t_{\mathrm{low}}$. An example time-slice design satisfying these constraints for a $1.4 - 1.4\,M_\odot$ binary is shown in table 1.



f_k (Hz)	$(t_{k+1},t_k]$ (s)	$N_{\mathrm{ts},k}$
4096	$(-0.25,0]$	1024
512	$(-2.25,-0.25]$	1024
256	$(-10.25,-2.25]$	2048
128	$(-18.25,-10.25]$	1024
128	$(-34.25,-18.25]$	2048
128	$(-66.25,-34.25]$	4096
64	$(-130.25,-66.25]$	4096
64	$(-258.25,-130.25]$	8192
64	$(-386.25,-258.25]$	8192
32	$(-514.25,-386.25]$	4096
32	$(-770.25,-514.25]$	8192
32	$(-1026.25,-770.25]$	8192

Table 1: Example of nearly critically sampled, power-of-two time slices for a $1.4 - 1.4\,M_\odot$ template extending from $f_{\mathrm{low}} = 10\,\mathrm{Hz}$ to $f_{\mathrm{ISCO}} = 1571\,\mathrm{Hz}$ with a time frequency structure given by (4). f_k is the sample rate of the time slice, $(t_{k+1}, t_k]$ are the boundaries in seconds preceeding coalescence and $N_{\mathrm{ts},k}$ are the number of sample points in the k^{th} filter.

Rather than applying a unique time-slice decomposition to each template waveform, we find a decomposition that is adequate for the entire set $\{x_i\}$. This is facilated by choosing templates in the set $\{x_i\}$ with similar chirp masses, $\mathcal{M}$. The time-slice decompositions of the filters $x_i(\tau)$ lead to new filters $y_{ik}(\tau)$ satisfying

$$x_i(\tau) \approx \sum_k^{N_{\mathrm{ts}}} y_{ik}(\tau) \tag{5}$$

where N_{ts} is the number of time slices required. Note that we write the relationship as approximate since the resampling implementation may have some inherent loss in quality. We note that by construction these filters are orthogonal over the index k since they are disjoint in time. In the next section we examine how to reduce the number of filters $y_{ik}(\tau)$ via singular value decomposition to construct a set of filters that is also orthogonal within a given time slice.

2.2.2. Reducing the number of filters with the singular value decomposition As described previously, the template banks are, by design, highly correlated. It is possible to greatly reduce the number of filters required to achieve a particular minimum match by designing an appropriate set of orthonormal *basis templates*. A purely numerical technique based on the application of the singular value decomposition (SVD) to inspiral waveforms is demonstrated in [23]. Using the results of [23] we establish that the filters of the previous section can be approximated to

Symbol	Definition
N	The number of sample points in the data
N_{x_i}	The number of sample points in template x_i
N_t	The number of templates in the set $\{x_i\}$
N_{ts}	The number of time slices
$N_{\mathrm{uj},k}$	The number of orthogonal filters in time slice k
$N_{\mathrm{ts},k}$	The number of samples in time slice k
$f_{\max}$	The maximum frequency of a filter
f_k	The sample frequency of the k^{th} filter
N_{r}	The number of sample points in the resample filter

Table 2: Notation used to describe filtering. This table provides a quick reference for symbols used.

high accuracy by the expansion in the singular value basis

$$y_{ik}(\tau) \approx \sum_{j}^{N_{\mathrm{uj},k}} \sigma_{jk} v_{ijk} u_j(\tau) \tag{6}$$

where σ_{jk} is the j^{th} singular value for the k^{th} time slice, v_{ijk} is an orthogonal matrix for the k^{th} time slice and u_{jk} is a new orthogonal basis filter set for the k^{th} time slice. The authors of [23] showed that to high reconstruction accuracy far fewer filters are needed than were in the original template bank. We find that when combined with the time-slice decomposition, the number of SVD filters, $N_{\mathrm{uj},k}$ is much smaller than the original number of filters N_t. We combine (6) with (5) to arrive at (3). In the next section we compute the expected computational cost scaling of this decomposition and compare it with the brute-force implementation in (1) and higher latency FFT methods.

2.3. Comparison of computational costs

We now examine the computational cost scaling of the approximate implementation of (1) as (3). An actual implementation of this decomposition in a working analysis pipeline is discussed in the next section along with measured computational requirements. For convenience, table 2 is a recap of the meaning of various symbols used in this calculation.

In table 3 we present the computational cost scaling in floating point operations per sample for common tasks in the pipeline.

The filter bank can be implemented using finite impulse response (FIR) filters, which are just sliding window dot products. If there are N_t templates of length N_{x_i} and the data stream contains N samples, then applying the filter bank requires $2N_t N_{x_i} N$ operations.

More commonly, the matched filters are implemented using the FFT convolution. This entails applying FFTs to blocks of D samples, with $N_{x_i} \leq D$, each block overlapping the previous one by $D - N_{x_i}$ samples. There are $N/(D - N_{x_i})$ such blocks required to filter N samples of data. Modern implementations of the Cooley-Tukey FFT, such as the ubiquitous `fftw`, require about $4N \lg N$ operations to evaluate a DFT

Process	**ops/sample**
FIR matched filter, N_t templates of length N_{x_i}	$2N_tN_{x_i}$
FFT matched filter, N_t templates of length N_{x_i} blocks of length D	$\frac{4(N_t+1)\lg D+2N_t}{1-N_{x_i}/D}$
FIR resampling filter, length N_r for each of N_t templates and sample rates $f_1 < f_2$	$2N_tN_\mathrm{r}\ f_1/f_2$
multiply $M \times L$ real matrix by $L \times 1$ real vector	$2ML$

Table 3: Number of floating point operations per sample (multiplications and divisions) required for a selection of signal processing operations used in LLOID.

of size N [32]. A D sample cross-correlation consists of a forward FFT, a Dsample dot product, and an inverse FFT totaling $8D\lg D + 2D$ operations per block. Per sample, this is $(8\lg D + 2)/(1 - N_{x_i}/D)$ operations. As this expression indicates the number of operations increases as the block size D approaches the filter length N_{x_i}.

The FIR filter implementation has the advantage that it has no intrinsic latency, whereas the FFT convolution has latency of $D - N_{x_i}$. However, the FIR filter implementation has the disadvantage of much greater overhead per sample than the FFT convolution. For a 1 ks template sampled at 4096 Hz, the FIR implementation requires about about $N_{x_i}/8\lg 2N_{x_i} = 2.2 \times 10^4$ times more operations per sample than the FFT implementation. We will now consider the computational cost of the FIR filter implementation described in the previous sections.

This is more commonly known as "overlap-save". We should find someone else's operation count and cite it. Drew: Why don't we change this to an overlap of m samples so we can see what happens as we increase the overlap to reduce latency.

It is convenient to express the computational cost of the entire filtering procedure in floating point operations per second flop/s. The cost will be the sum of the cost of the FIR filtering for the orthogonal filter in each time slice plus the cost of reconstructing the original waveforms with matrix operations and resampling. Using the formulas in table 3 we arrive at

$$\text{C.C.} \propto 2\sum_{k}^{N_\mathrm{ts}}(N_{\mathrm{ts},k} + N_t)N_{\mathrm{u_j},k}f_k r \sum_{k,f_k\neq f_{f_\mathrm{max}}}^{N_\mathrm{ts}} N_tN_\mathrm{r}f_k \tag{7}$$

where the first sum is the total cost for filtering and reconstructing the orthogonal filters u_{jk}. The second sum is the cost of resampling the reconstructed time-slice outputs to the original sample rate f_max. It assumes a FIR filter with N_r sample points. Note that the cost for resampling only occurs for the downsampled time slices. The resampling cost largely depends on N_r. The computational cost of (7) is dominated by the highest frequency terms in the sum. Comparing just the highest frequency term of (7) with (1) shows $\mathcal{O}[(N_{\mathrm{ts,max}} + N_t)N_{\mathrm{u_j,max}}f_\mathrm{max}]$ flop/s versus $\mathcal{O}[N_tN_{x_i}f_\mathrm{max}]$ flop/s. The full calculation for a particular patch of parameter space is shown in table **??**. FIXME we need to actually say what went into this

From here down introduce an example, but don't go through the FFT methods in lloid focus just on the comparison of the TD method without lloid tricks, the FD method without lloid tricks, and the TD lloid method (no comp detection statistic though), remember that the goal is near realtime detection not computational cost

3. Implementation

In this section we describe an implementation of the method described in section **??** suitable for rapid gravitational-wave searches for compact binary coalescence. The

operations/sample	**latency (s.)**	**method**
3,714,580,480	2.4×10^{-4}	conventional FIR method (1)
49,937	1.8×10^{3}	conventional FFT method (2)
120,000	9.0×10^{2}	conventional FFT method (2)
120,673	2.4×10^{-4}	FIR method with time slices and SVD (3)

Table 4: FIXME: Operation counts per sample for four different detection methods.

previous method requires several computations that can be done before the analysis is actually underway. Thus we divide the procedure into two stages 1) an offline planning stage and 2) an online, low latency filtering stage. The offline stage can be done before the analysis is started an updated asychronously, whereas the online stage must keep up with the detector output and produce search results as rapidly as possible. In the next two subsections we describe what these stages entail.

3.1. Planning stage

The choice of filter waveforms and singular value decomposition can be done in advance and will be valid as long as the detector noise spectrum remains roughly constant. New filter waveforms can be computed asynchronously using updated spectrum estimates as they are available.

The planning stage begins with choosing templates that cover the space of mass parameters with a hexagonal grid [33] in order to satisfy the minimum match criterion, which assures a user specified maximum loss in SNR for signals that fall in-between the chosen templates. 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. Dividing the mass space into smaller sub-banks also aids in the computational cost of the singular value decomposition. Using our understanding of the time-frequency evolution of the templates, we choose time slice boundaries as in (**??**) 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 downsampled to the appropriate sample rate. Finally, the SVD is applied to each time slice in order to produce a set of orthogonal FIR filters and a reconstruction matrix that maps them back to the original templates. The downsampled orthogonal FIR filter coefficients, the reconstruction matrix, and the time slice boundaries are all saved to disk.

3.2. Filtering stage

We have implemented a prototype of the low latency filtering stage using an open source signal processing environment called GStreamer [34]. 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. It turns out that it is also uniquely suited to gravitational wave data analysis. In our application, GStreamer excels at queueing, synchronizing, adding, and bookkeeping many different signals at different sample rates. It also provides some useful signal processing primitives such as decimators, FIR filters, and interpolators.

Most importantly, it permits us to recruit all of the host system's CPUs without having to write complicated and error-prone multithreaded code.

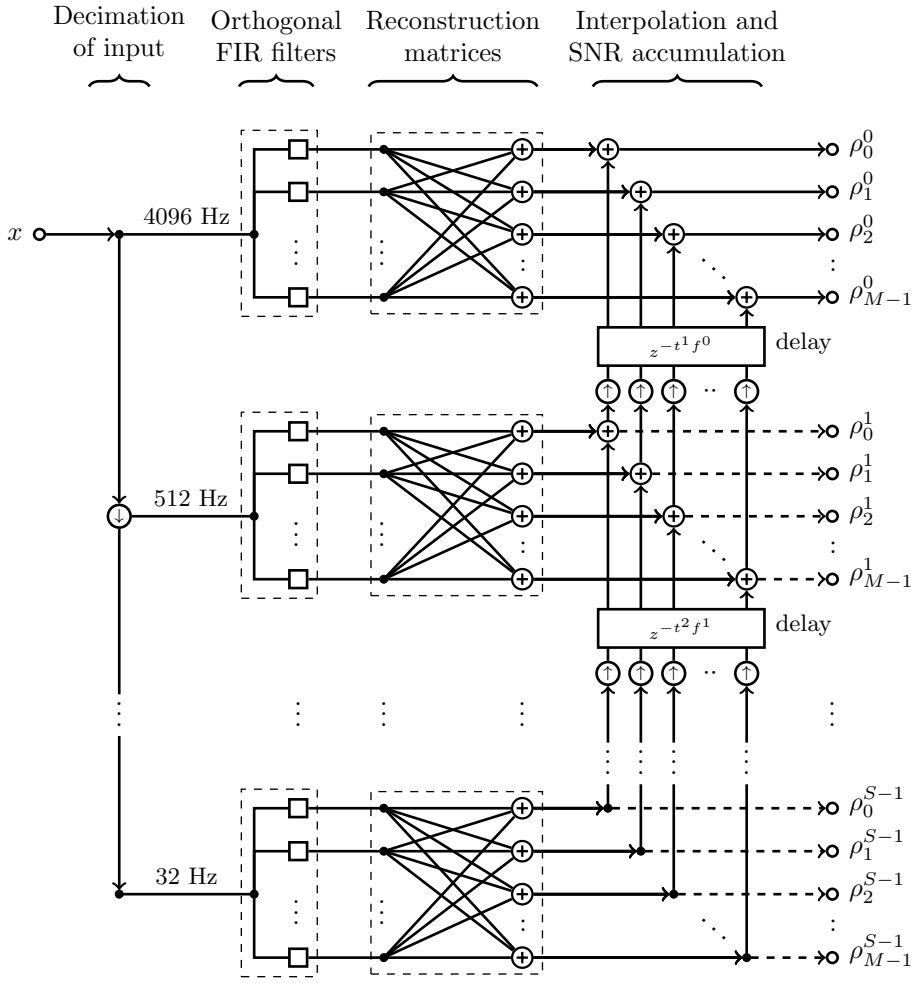


Figure 2: Schematic of LLOID pipeline illustrating signal flow. Circles with arrows represent upsampling ⊙↑ or downsampling ⊙↓. Circles with plus signs represent summing junctions ⊕. Squares □ stand for FIR filters. Sample rate decreases from the top of the diagram to the bottom.

The filter pipeline consists of six distinct stages.

3.2.1. Decimation First, the sample rate of the whitened detector data is reduced to successively lower sample rates by decimation. Decimation involves applying an antialiasing filter to the data, and then downsampling by deleting samples. We use a 192-tap FIR decimator provided by GStreamer.

3.2.2. Time delays Each decimated detector data stream becomes the input for one time slice, but it must be appropriately delayed.

3.2.3. Orthogonal FIR *filters*

3.2.4. Reconstruction

3.2.5. Interpolation

3.2.6. SNR accumulation

4. Results

[I would like to see this section changed to emphasize the low-latency time domain filtering 1. Use a live simulated white noise source (this ignores the latency of whitening, but that goes beyond the scope of this paper, and we mention this and perhaps suggest some exporation 2. Use TD filtering of N templates 3. Present the performance and latency and provide estimates for number of cores required for realtime ALIGO search based on current infrastructure 4. Compute the impulse responses of the templates and histogram the SNR loss for the SVD and time/slice/resampling 5. Put a tee after some of the low frequency stages and perhaps test the possibility of predictive filtering (also with latency measurments) This will remove the complications of running a full pipeline. We need to sort things out before we do that, and this paper should not be delayed any further. The above tests will make the point we need to make]

4.1. Detector noise characteristics

We tested the new detection method with mock Advanced LIGO data having a power spectrum prescribed by the "zero detuning, high power" noise model in [35].

5. Conclusions

We have demonstrated a computationally feasible procedure for the rapid detection of gravitational waves emitted during the coalescence of neutron stars and stellar-mass black holes. These sources are expected to produce prompt electromagnetic signals and may be the progenitors of some short hard gamma-ray bursts. Rapid alerts to the broader astronomical community may improve the chances of detecting an electromagnetic counterpart in bands from X-ray down to radio. We antipate requiring no more than 1000 modern computer cores to analyze a four-detector network of gravitational-wave data for such systems assuming spin effects can be ignored in parameterizing the expected signals [**?**]. This is within the current computing capabilities of the LSC Data Grid [36].

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 of 125 ms. Data aggregation from the observatories will travel over various networks, each capable of high bandwith but perhaps only modest latency. This could amount to a similar latency of $\sim$ 100 ms. Lastly, unless a realtime infrastructure is adopted post data acquisition, it is likely that there will be an inherent latency introduced by such infrastructure. We have shown a prototype implementation using `gstlal` that is capable of $\sim$1 s latency. In

our opinion, significant work would have to be done in order to improve upon this number. However, it should be considered for third generation detector design. For example, a tighter integration of analysis and data acquisition would be beneficial.

We have omitted discussion of source localization. Localization is known to be poor for signals of low SNR [37]. However, one should not immediately dismiss the practical usefulness of a poorly localized source. Even with poor localization, it should be possible to begin downselecting what observatories could view a potential signal and for such observatories to begin any necessary prerequisite activities. In future works we wil explore more rigorously the pointing prospects with realistic simulations using the infrastructure and techniques described in this work.

Latency budget, including 'before' and 'after' quotes for:

- Data acquisition
- Calibration
- Data aggregation
- Analysis
- Localization
- Alert
- Telescope actuation
- Total

Future work:

- Sub-solar mass search
- Hierarchical detection

Acknowledgments

LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0107417. CH would like to thank Ilya Mandel for many discussions about rate estimates and the prospects of early detection. NF would like to thank Alessandra Corsi for illuminating discussions on astronomical motivations.

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

References

[1] Advanced LIGO `www.advancedligo.mit.edu`
[2] Adanced Virgo `https://wwwcascina.virgo.infn.it/advirgo`
[3] GEO 600 `http://www.geo600.org`
[4] Large-scale Cryogenic Gravitational wave Telescope (LCGT) `http://gw.icrr.u-tokyo.ac.jp/lcgt/`
[5] Abadie J *et al.* 2010 *Class. Quant. Grav.* URL `http://iopscience.iop.org/0264-9381/27/17/173001`
[6] Shibata M and Taniguchi K 2008 *Phys. Rev.* **D77** 084015–+ (*Preprint* `arXiv:0711.1410`)
[7] Lee W H, Ramirez-Ruiz E and Granot J 2005 *Astrophys. J.* **630** L165–L168
[8] Nakar E 2007 *Phys. Rept.* **442** 166–236 (*Preprint* `astro-ph/0701748`)
[9] Sari R and Piran T 1999 *The Astrophysical Journal* **520** 641 URL `http://stacks.iop.org/0004-637X/520/i=2/a=641`

[10] Akerlof C, Balsano R, Barthelmy S, Bloch J, Butterworth P, Casperson D, Cline T, Fletcher S, Frontera F, Gisler G, Heise J, Hills J, Kehoe R, Lee B, Marshall S, McKay T, Miller R, Piro L, Priedhorsky W, Szymanski J and Wren J 1999 *Nature* **398** 400–402 (*Preprint* arXiv:astro-ph/9903271)
[11] Fox D W, Yost S, Kulkarni S R, Torii K, Kato T, Yamaoka H, Sako M, Harrison F A, Sari R, Price P A, Berger E, Soderberg A M, Djorgovski S G, Barth A J, Pravdo S H, Frail D A, Gal-Yam A, Lipkin Y, Mauch T, Harrison C and Buttery H 2003 *Nature* **422** 284–286
[12] Jelínek M, Prouza M, Kubánek P, Hudec R, Nekola M, Řídký J, Grygar J, Boháčová M, Castro-Tirado A J, Gorosabel J, Hrabovský M, Mandát D, Nosek D, Nožka L, Palatka M, Pandey S B, Pech M, Schovánek P, Šmída R, Trávníček P, de Ugarte Postigo A and Vítek S 2006 *Astronomy & Astrophysics* **454** L119–L122 (*Preprint* arXiv:astro-ph/0606004)
[13] Mundell C G, Melandri A, Guidorzi C, Kobayashi S, Steele I A, Malesani D, Amati L, DAvanzo P, Bersier D F, Gomboc A, Rol E, Bode M F, Carter D, Mottram C J, Monfardini A, Smith R J, Malhotra S, Wang J, Bannister N, OBrien P T and Tanvir N R 2007 *The Astrophysical Journal* **660** 489 URL http://stacks.iop.org/0004-637X/660/i=1/a=489
[14] Racusin J L *et al.* 2008 *Nature* **455** 183–188 (*Preprint* 0805.1557)
[15] Gruber D, Krühler T, Foley S, Nardini M, Burlon D, Rau A, Bissaldi E, von Kienlin A, McBreen S, Greiner J, Bhat P N, Briggs M S, Burgess J M, Chaplin V L, Connaughton V, Diehl R, Fishman G J, Gibby M H, Giles M M, Goldstein A, Guiriec S, van der Horst A J, Kippen R M, Kouveliotou C, Lin L, Meegan C A, Paciesas W S, Preece R D, Tierney D and Wilson-Hodge C 2011 *Astronomy & Astrophysics* **528** A15+ (*Preprint* 1101.1099)
[16] Nissanke S, Holz D E, Hughes S A, Dalal N and Sievers J L 2010 *Astrophys. J.* **725** 496–514 (*Preprint* 0904.1017)
[17] Anticipated zero detuning, high power, advanced LIGO spectrum. LIGO t0900288-v3. URL https://dcc.ligo.org/public/0002/T0900288/003/ZERO_DET_high_P.txt
[18] Anticipated zero detuning, low power, advanced LIGO spectrum. LIGO t0900288-v3. URL https://dcc.ligo.org/public/0002/T0900288/003/ZERO_DET_low_P.txt
[19] Anticipated bhbh 20 deg detuned advanced LIGO spectrum. LIGO t0900288-v3. URL https://dcc.ligo.org/public/0002/T0900288/003/BHBH_20deg.txt
[20] Hughey B 2011 Electromagnetic follow-ups of candidate gravitational wave triggers in the recent ligo and virgo science runs (Presented at the Gravitational wave physics and astronomy workshop (GWPAW))
[21] TBD 2011 *in preparation*
[22] Kanner J, Huard T L, Marka S, Murphy D C, Piscionere J, Reed M and Shawhan P 2008 URL arXiv:0803.0312v4[astro-ph]
[23] Cannon K, Chapman A, Hanna C, Keppel D, Searle A C and Weinstein A J 2010 *Physical Review D* **82** 44025 (c) : URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010PhRvD..82d4025C&link_type=ABSTRACT
[24] Marion F and the Virgo Collaboration 2004 *Proc. Rencontres de Moriond on Gravitational Waves and Experimental Gravity 2003*
[25] Buskulic D, the LIGO Scientific Collaboration and the Virgo Collaboration 2010 *Classical and Quantum Gravity* **27** 194013 URL http://stacks.iop.org/0264-9381/27/i=19/a=194013
[26] Allen B A, Anderson W G, Brady P R, Brown D A and Creighton J D E 2005 (*Preprint* gr-qc/0509116)
[27] Owen B J 1996 *Phys. Rev. D* **53** 6749–6761
[28] Owen B J and Sathyaprakash B S 1999 *Phys. Rev. D* **60** 022002
[29] FBeauville, M-ABizouard, LBlackburn, LBosi, PBrady, LBrocco, DBrown, DBuskulic, FCavalier, SChatterji, NChristensen, A-CClapson, SFairhurst, DGrosjean, GGuidi, PHello, EKatsavounidis, MKnight, ALazzarini, NLeroy, FMarion, BMours, FRicci, AVicere and MZanolin 2006 *J. Phys. Conf. Ser.* **32** 212 URL arXiv:gr-qc/0509041v1
[30] Beauville F, Bizouard M A, Blackburn L, Bosi L, Brocco L, Brown D, Buskulic D, Cavalier F, Chatterji S, Christensen N, Clapson A C, Fairhurst S, Grosjean D, Guidi G, Hello P, Heng S, Hewitson M, Katsavounidis E, Klimenko S, Knight M, Lazzarini A, Leroy N, Marion F, Markowitz J, Melachrinos C, Mours B, Ricci F, Vicer A, Yakushin I and Zanolin M 2008 *Class. Quant. Grav.* **25** 045001 (*Preprint* http://www.iop.org/EJ/abstract/0264-9381/25/4/045001/) URL arXiv:gr-qc/0701027v1
[31] E K L, M W C and G W A 1992 *Class. Quantum Grav.* **9** L125–31
[32] Johnson S and Frigo M 2007 *Signal Processing, IEEE Transactions on* **55** 111 – 119 URL http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=&arnumber=4034175&queryText%253D%2528%2528Document+Title%253Aa+modified+split-radix+fft+with+fewer%2529%2529%2526openedRefinements%253D*%2526sortType%253Ddesc_Publication+Year%

2526matchBoolean%253Dtrue%2526rowsPerPage%253D50%2526searchField%253DSearch+All
[33] Cokelaer T 2007 *Phys. Rev. D* **76** 102004
[34] GStreamer: open source multimedia framework URL http://gstreamer.freedesktop.org
[35] Shoemaker D 2009 Advanced ligo anticipated sensitivity curves Tech. rep. URL https://dcc.ligo.org/cgi-bin/DocDB/ShowDocument?docid=2974
[36] The LSC Data Grid URL https://www.lsc-group.phys.uwm.edu/lscdatagrid/
[37] Fairhurst S 2009 *New Journal of Physics* **11** 123006