Computational Challenges

SOC: Rory Smith and Sarah Caudill
Panelists: Ben Farr, Surabhi Sachdev, Sylvia Biscoveanu, Peter Couvares

Meet Our Panel

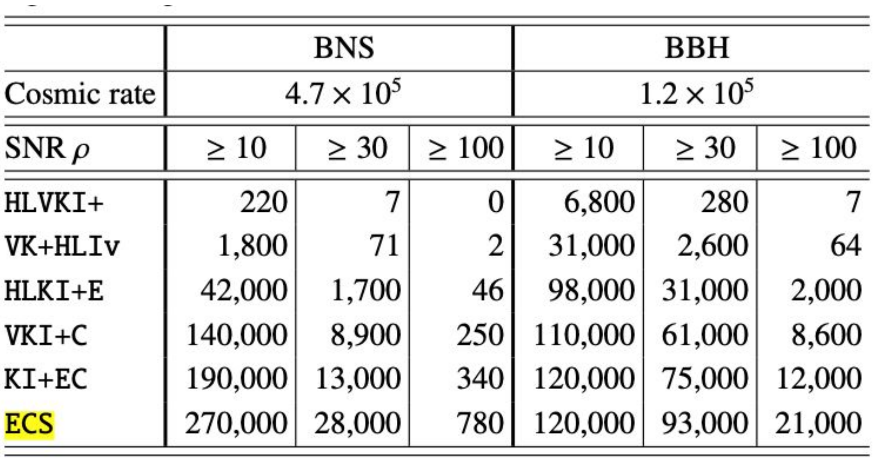
Sylvia Biscoveanu
MIT

Ben Farr
UOregon

Surabhi Sachdev
UW Milwaukee

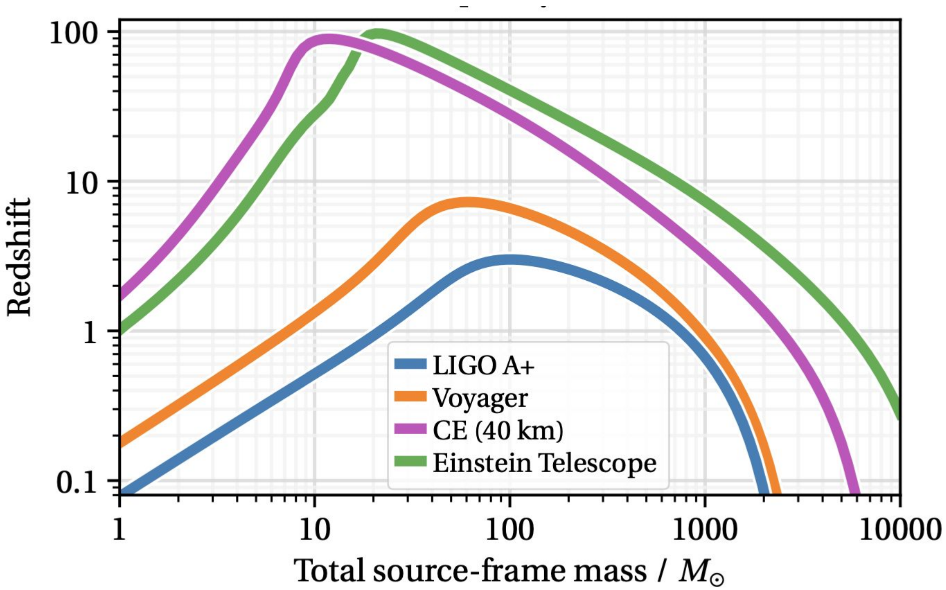
Peter Couvares
LIGO Lab - Caltech

3G Rates and Detection Expectations



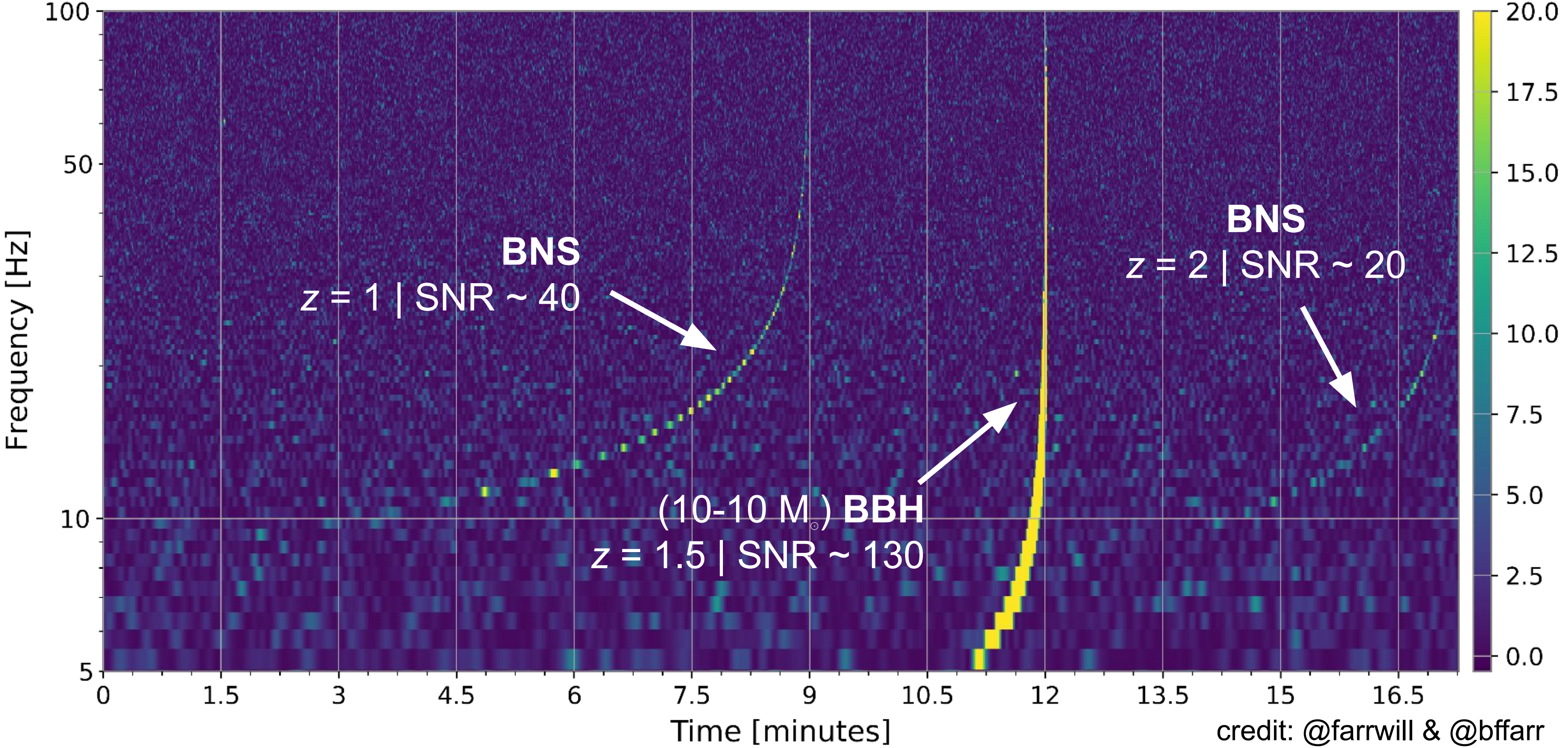
	BNS			BBH		
Cosmic rate	4.7×10^5			1.2×10^5		
SNR ρ	≥ 10	≥ 30	≥ 100	≥ 10	≥ 30	≥ 100
`HLVKI+`	220	7	0	6,800	280	7
`VK+HLIv`	1,800	71	2	31,000	2,600	64
`HLKI+E`	42,000	1,700	46	98,000	31,000	2,000
`VKI+C`	140,000	8,900	250	110,000	61,000	8,600
`KI+EC`	190,000	13,000	340	120,000	75,000	12,000
`ECS`	270,000	28,000	780	120,000	93,000	21,000

Borhanian+ 2022



CEHS, Evans+ 2021

3G Rates and Detection Expectations



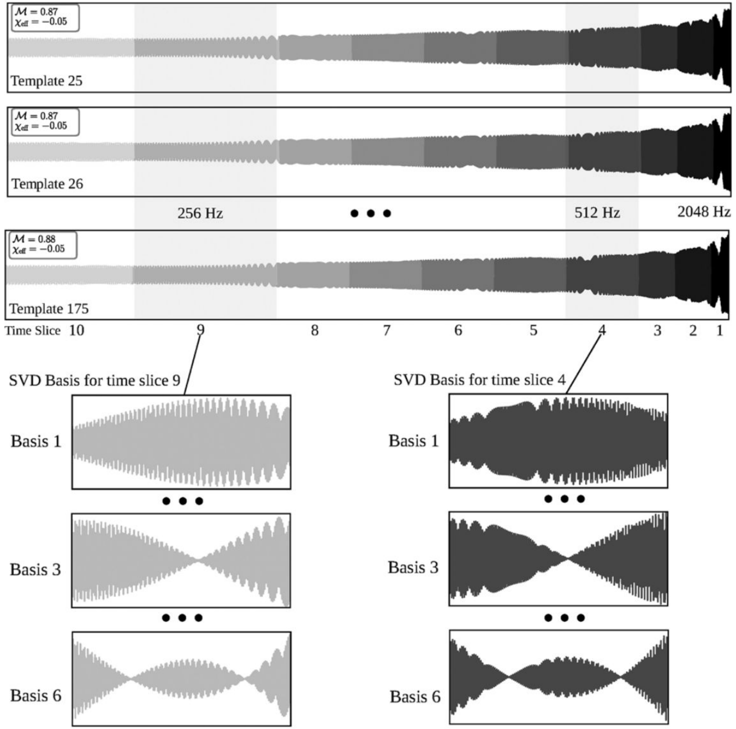
credit: @farrwill & @bffarr

Searches: Overview of challenges

- Challenges:
    - Larger template banks - long duration of signals in band
    - Large number of signals - overlapping signals
- Some techniques used by current searches will scale well - the LLOID method - (Kipp Cannon *et al* 2012) first used by the GstLAL search
    - Overlap in signals is not a problem for matched-filtering as long as the signals are separated in time-frequency
    - Multi-banding will help with analyzing data starting at a lower frequency containing long waveforms
    - SVD decomposition will help with the increased number of templates needed for analysis
    - Meacher *et al.* 2016 have shown that it is possible to search for BNSs starting from 5Hz in ET mock data with current methods

Searches: SVD and multi-banding

- Multi-banding: Divide the waveform into several frequency-bands and sample each according to Nyquist sampling frequency
- Construct an orthogonal basis and use a reduced set that recovers SNRs up-to an accuracy of 99.99%
- Together achieve a reduction in FLOPs by a factor of ~10000



Cody Messick *et al.* 2017

Searches

- Current methods of background estimation used by pipelines will not work in signal dominated 3G data
- We will need to think of innovative background estimation methods - perhaps using deep learning techniques
- Searches will continue to play an important role in the 3G era - we will have a chance to provide hours of early warning for binary neutron star mergers
- We will also be able to provide early warnings for binary black holes

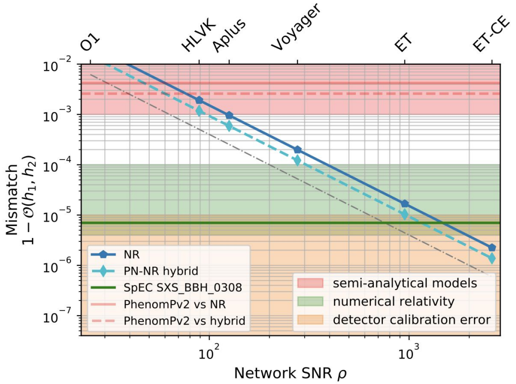
Overlapping signals

- Thousands of individual seconds in a year with multiple mergers (hundreds for BBH and thousands for BNS)
- Parameter estimation biases can be significant using standard methods when the merger times are within ~0.1 seconds
- When two signals with similar source parameters overlap, only the parameters of the louder one are recovered
- The beating in the composite waveform leads to the recovery of highly precessing signals with unequal mass ratios
- Refs: Samajdar et al 2021, Pizzati et al 2021, Antonelli et al 2021, Relton et al 2022

Methods for accelerated parameter estimation

- Current methods like **reduced order models** (Canizares+ 2014, Smith+ 2016, 2021, Morisaki+ 2021a) **relative binning** (aka heterodyning; Zackay+ 2018, Cornish+ 2021, Leslie+ 2021), and **multibanding** (Vinciguerra+ 2017, Morisaki 2021b), all rely on the premise of reducing the number of frequencies where the waveform is evaluated
- Surrogate models for higher-fidelity waveforms use interpolation to speed up waveform evaluation (Pürrer 2014, 2016, Field+ 2014, Blackman+ 2015, 2017, Varma+ 2019, Thomas+ 2022)
- Machine learning techniques for speeding up inference rather than just waveform evaluation (Gabbard+ 2019, Chua+ 2020, Green+ 2020, 2021, Delaunoy+ 2020, Krastev+ 2021, Shen+ 2021, Dax+ 2021)

Sources of systematic uncertainty



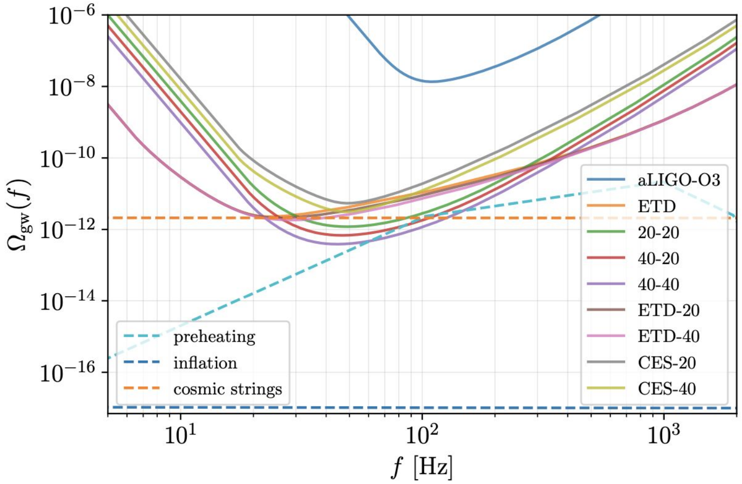
- Waveforms (see session this afternoon)
  - Need to reduce waveform approximant errors by three orders of magnitude for golden binaries (Pürrer+ 2019)
- Calibration (Payne+ 2020, Vitale+ 2020)
  - Sub-dominant effect even at very high SNRs, sky localization most affected for individual sources
- Noise modeling (Rover+2008, 2011, Littenberg+ 2013, 2014, Veitch+ 2014, Edwards+ 2015, Talbot+ 2020, Biscoveanu+ 2020, Chatziioannou+ 2020, Plunkett+ in prep.)
  - Comparable to effect of calibration error for individual-event posteriors
  - Needs to be accounted for when conducting fully bayesian searches

Population inference at scale

- Current methods for population inference that involve “reweighting” posterior samples from individual events will not scale when the population includes thousands of events
- Look to machine learning and density estimation techniques for sensitivity estimation and population inference
  - Wong+ 2020b, Gerosa+ 2020, Wysocki+ 2020, Golomb+2021, Talbot+ 2020
- Interpolation of population synthesis simulations to map observed population to theoretical predictions (see session tomorrow morning)
  - Wong+ 2019, 2020a/c, 2022, Zevin+ 2021

Astrophysical and Cosmological Backgrounds

- The primordial background from early-universe sources will be obscured by a foreground of merging compact binaries
- Even with traditional foreground subtraction methods (Cutler+ 2006, Harms+ 2008, Sharma+ 2008), residual contamination particularly from sub-threshold BNS mergers will contaminate the background (Sachdev+ 2020)
- Fully bayesian method to infer the presence of a compact binary (and marginalize over uncertainty in its parameters) in each segment of data on top of a Gaussian background (Biscoveanu+ 2020)



CEHS, Evans+ 2021

The future is fully bayesian

- Create a complete likelihood that accounts for glitch, CBC, and Gaussian background hypotheses, analyze every segment of data that may contain a merger (~0.1s segments to avoid overlap issues)
    - Measure the merger rate, population properties, glitch probability, etc. all at once (Smith+ 2017)
    - Extension of proposed bayesian detection strategies (Veitch+ 2009, Isi+ 2018, Vajpeyi+ 2021, Ashton+ 2019, 2020, Pratten+ 2020)
- Akin to the LISA “global fit” strategy - fit the parameters of every binary merger in band all at once (Cornish+ 2005, Crowder+ 2007, Littenberg(+) 2011, 2020)

3G Computing (Peter)

- Increasing complexity of computing platforms + increasing focus on computational cost-efficiency from funding agencies == big changes.
- **Code will be harder to write and more expensive to maintain because computing hardware is getting *weirder*.**
- Since ~1990, we've gotten accustomed to targeting essentially one platform (x86 unix), re-running old code on newer hardware, and riding the wave of single-threaded CPU price/performance gains.
- Single-core performance improved 52%/year in 1986–2003 and 23%/year in 2003–2011, but slowed to 7%/year in 2011–2018.
- We will see **very limited price/performance improvements for simple code** in the future**.** CERN projects ~10%/year CPU price/performance gains over the next decade.
- Due to the opportunities for cost savings, we may need to support multiple x86/ARM CPU instruction sets, GPUs, and other coprocessors and treat them each as distinct platforms — **lowest common denominator x86 code no longer good enough**.

3G Computing (Peter)

- Domain-specific code that takes advantage of hardware parallelism will enjoy much larger price/performance improvements. **This requires a large increase in software development and maintenance costs over time.**
- We are moving from a single, *stable-over-time* platform (x86 unix) to many *unstable-over-time* heterogenous platforms: >1 x86 CPU instruction sets, >1 GPU architectures, FPGAs, higher-level data processing engines (e.g., Spark, Dataflow, Ray,), etc.
- Software development will be harder, and potentially more importantly, **code that successfully exploits today's parallelism may need to be rewritten to enjoy tomorrow's parallelism**. A critical code may need to be rewritten for a new arch every 5 years vs. working for 30+ years.
- "Bang for the buck" — it's not always about the fastest possible code; it's about the best performance for a given amount of hardware + scientist/developer FTE effort, measured over many years.

We Will Need Help (Peter)

- We need software engineering effort focused on domain-specific computing for 3G; maintainability and efficiency will be key requirements. We need to plan for how this effort will be organized and funded (c.f., “The Astropy Problem").
- **A major foreseeable problem is how to answer these questions when the ground-based GW data analysis community is fully occupied with development for 2G science runs.** The GW community is still relatively small, and will need to grow and/or divide its efforts between exploiting existing 2G data and preparing for future 3G data.
- We will need expert knowledge and domain experience at the intersection of computer science / computational astrophysics / data processing. Due to demand from industry, necessary scientific computing expertise is scarce & expensive.
- “Centers of excellence” needed to support GW scientists’ computing — “roll our own” isn’t going to work. **In the same way that the US-LHC community recognized the need for external computing research support with its IRIS-HEP Institute, the GW/Astro community may need to develop a similar support organization**.
- Some existing building blocks and partners: PATh / OSG / HTCondor (NSF) for workflow management and distributed high-throughput computing support, CILogon (NSF) for distributed identity & access management, etc.

Back-up slides

Removal of astrophysical background