Astronomy and Computing 25 (2018) 139–148

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

The Apertif Monitor for Bursts Encountered in Real-time (AMBER) auto-tuning optimization with genetic algorithms

K. Mikhailov[a,b,*], A. Sclocco[c]

[a] Anton Pannekoek Institute for Astronomy, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[b] ASTRON, The Netherlands Institute for Radio Astronomy, Postbus 2, 7990 AA, Dwingeloo, The Netherlands
[c] NLeSC, Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 17 May 2018
Accepted 19 September 2018
Available online 28 September 2018

Keywords:
Pulsars
Astronomical instrumentation
Methods and techniques
Algorithms
Genetic algorithms

ABSTRACT

Real-time searches for faint radio pulses from unknown radio transients are computationally challenging. Detections become further complicated due to continuously increasing technical capabilities of transient surveys: telescope sensitivity, searched area of the sky, number of antennas or dishes, temporal and frequency resolution. The new Apertif transient survey on the Westerbork telescope happens in real-time on GPUs by means of the single-pulse search pipeline AMBER (Sclocco, 2017). AMBER initially carries out auto-tuning: it finds the most optimal configuration of user-controlled parameters per each of four pipeline kernels so that each kernel performs its task as fast as possible. The pipeline uses a brute-force (BF) exhaustive search which in total takes 5–24 h to run depending on the processing cluster architecture. We apply more heuristic, biologically driven genetic algorithms (GAs) to limit the exploration of the total parameter space, tune all four kernels together and reduce the tuning time to few hours. Our results show that after only few hours of tuning, GAs always find similar or even better configurations for all kernels together than the combination of single kernel configurations tuned by the BF approach. At the same time, by means of their genetic operators, GAs converge into better solutions than those obtained by pure random searches. The explored multi-dimensional parameter space is very complex and has multiple local optima as the evolution of randomly generated configurations does not always guarantee global solution.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Various radio transient surveys constantly search for new pulsars, rotating radio transients (RRATs, McLaughlin et al., 2006), and fast radio bursts (FRBs, Lorimer et al., 2007; Petroff et al., 2016), especially at less explored extragalactic distances in dense environments. More such discoveries can help us better classify transients and study intergalactic medium (IGM). Even though distant radio transients are hard to localize, better localization can more comprehensively explore Galactic and extragalactic source populations in terms of stellar evolution and star formation that should depend on the type of host galaxy.

New discoveries of single bursts with Parkes, UTMOST, and ASKAP (Caleb et al., 2017; Bannister et al., 2017; Bhandari et al., 2018) and one repeating source of bursts with Arecibo (Spitler et al., 2016; Chatterjee et al., 2017) reveal new properties of radio bursts. Searches for much fainter and more distant bursts require more fine-grained searches and lead to new processing challenges (Magro et al., 2011; Barsdell et al., 2012; Sclocco et al., 2016). Just like standard pulsar searches, transient lookups are performed in the two-dimensional, time–frequency space for every unit of dispersion measure (DM, third dimension). Modern searches (see Table 1) are performed in real-time to trigger multi-frequency follow-up. They also require very high time sampling and frequency resolution to better determine the burst structure. Growing data rates and computational costs require larger supercomputers and search pipelines based on graphics processing units (GPUs) rather than central processing units (CPUs).[1]

Transient surveys are also technically limited in their ability to detect new bursts. The number of antennas or dishes in the survey relates to the corresponding amount and size of beams they can produce. This determines how large the observing area of the sky would be. The system equivalent flux density $S_{\rm sys} = T_{\rm sys}/G$, where $T_{\rm sys}$ is a total system temperature and G is a system

---

* Corresponding author at: Anton Pannekoek Institute for Astronomy, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.
E-mail addresses: K.Mikhailov@uva.nl (K. Mikhailov), a.sclocco@esciencecenter.nl (A. Sclocco).

[1] Other options, such as FPGAs and ASICs, are also available. However, FPGAs are very hard to program, and floating point performance is not comparable with GPUs, whereas ASIC are expensive to design and produce.

https://doi.org/10.1016/j.ascom.2018.09.007
2213-1337/© 2018 Elsevier B.V. All rights reserved.

Table 1
Modern radio pulsar and transient surveys and their main characteristics. FoV is a survey field of view in square degrees, N_{beam} is a number of facilitated beams, t_{samp} is a sampling resolution in micro-seconds, n_{pol} is a number of polarizations, ν_{cntr} and $\Delta\nu$ are central frequency and available bandwidth, both in megahertz, T_{sys}/G and $S_{min}^{3\,ms}$ are system noise and minimum detectable flux density for a $10\,\sigma$ single-pulse threshold and 3 ms pulse width, both in Janskys.

Parameter	CHIME[a]	UTMOST[b]	SUPERB[c]	ASKAP[d]	ALERT[e]	SKA-Low[f]	SKA-Mid[f]
Status	Commissioning	Ongoing	Ongoing	Ongoing	Commissioning	Future	Future
FoV (deg^2)	220	9	0.6	30	8.7	27	0.49
N_{beam}	1024	352	13	288	2600	500	1500
t_{samp} (μs)	2.5	655.36	64	–	40.92	50	50
n_{pol}	2	1	4	2	2	4	4
ν_{cntr} (MHz)	600	835.5	1382	1400	1400	250	800
$\Delta\nu$ (MHz)	400	31.25	400	336	300	100	300
N_{chan}	1024	320	1024	336	1536	8192	4096
T_{sys}/G (Jy)	45	28.5	60	1800	70	2×10^{-5}	7×10^{-6}
$S_{min}^{3\,ms}$ (SP, Jy)	0.25	0.9	0.3	13.4	1.6	1.8×10^{-7}	3.7×10^{-8}

[a] Based on the CHIME system overview (The CHIME/FRB Collaboration et al., 2018).

[b] Based on the UTMOST system overview (Bailes et al., 2017; Caleb et al., 2017).

[c] Based on the SUPERB survey overview (Keane et al., 2018; Bhandari et al., 2018).

[d] Based on ASKAP survey description (Bannister et al., 2017).

[e] Based on Apertif Incoherent Search setup (Maan and van Leeuwen, 2017).

[f] Based on the updated SKA review (Dewdney, 2013; Braun, 2015; Levin et al., 2017).

gain. Together with frequency bandwidth $\Delta\nu$, number of polarizations n_{pol}, single-pulse threshold and single-pulse width, this determines down to what extent of radio transient brightness we can possibly search (radiometer equation, Lorimer and Kramer, 2004). Finally, the temporal and frequency resolution of the instrument set the limits to which the intrinsic structure of the pulse can be studied. Within all such limitations, the data should be optimally distributed on CPUs and GPUs for the signal processing: this includes de-dispersion (appropriate shift and integration of frequency channels that removes frequency dispersion), signal smoothing, and signal-to-noise evaluation. One way to find configurations that allow for fast data distribution and processing is to perform auto-tuning. In this case every configuration gets tested in terms of the best possible performance (fastest processing time in case of radio transient surveys). In the end, the most optimal configuration that allows the fastest search gets chosen. Auto-tuning is widely applied in computer science (Williams, 2008), but has also seen applications in other domains such as computational finance (Grauer-Gray et al., 2013) or astronomy (Sclocco et al., 2012). Auto-tuning for radio transient surveys also shows promising results in terms of performance portability (Sclocco et al., 2015).

The new real-time Apertif survey on the Westerbork (WSRT) telescope (ALERT, the Apertif Lofar Exploration of the Radio Transient Sky[2]) is now equipped with a new 160×GPU cluster that achieves 1.3 Pflops of peak performance and a data rate of 4 Tbit/s, and has 2 PBytes of available storage space (Maan and van Leeuwen, 2017). Such computational capacity enables deep searches up to 42 μs time and 0.195 MHz frequency resolution, respectively. Apertif front-ends on 12 WSRT dishes produce more than 400 tied array beams that in total cover 8.7 deg^2 of the sky, searched between 1100 and 1750 MHz with a tunable bandwidth of 300 MHz. Commissioning data from a targeted search toward FRB121102 already suggested a detection (Oostrum et al., 2017).

All hardware and software constraints require an optimized distribution of processing resources on the cluster to allow for the fastest real-time search: GPU threads and items, local memory re-use, and loop transformations. Auto-tuning allows for an automated search of these parameters. Although such tuning is performed only once for a running survey, it should be invoked again in case the survey undergoes hardware changes (e.g. front-end or back-end upgrades) or the search pipeline itself gets extended or improved (e.g. by adding new processing steps). Besides, it should be easily portable to any other survey pipelines.

Section 2 introduces the current search pipeline for ALERT and its current auto-tuning. We introduce a more heuristic approach for auto-tuning with genetic algorithms in Section 3. Section 4 shows achieved performance based on different algorithm input parameters as well as comparison with the pure random search. We discuss auto-tuning parameter space in terms of complexity and degeneracy in Section 5 and draw our conclusions in Section 6.

2. AMBER auto-tuning

The real-time search for new single bursts on WSRT is performed via the single-pulse search pipeline AMBER (The Apertif Monitor for Bursts Encountered in Real-time,[3] Sclocco, 2017). The pipeline can be divided into four main operations or kernels: a two-step de-dispersion,[4] de-dispersed time series downsampling (smoothing) and subsequent signal-to-noise (S/N) computation. Before the search, each kernel of the pipeline gets tuned to find its most optimal processing configuration.[5]

The parallel framework of choice for the accelerators is OpenCL, because it is vendor independent. In this regard GPU threads are referred as work-items, and GPU blocks of related threads are referred as work-groups. In AMBER, OpenCL kernels operate in three dimensional grids, but the pipeline uses only two dimensions, time and DM. These two dimensions limit the amount of available parallelism on both work-groups and work-items. The pipeline configuration is based on survey constraints and processing capabilities (see Table 2) as well as 8 different types of user-controlled parameters[6] (see Table 3) that altogether define a single computational configuration for a specific many-core accelerator.[7]

All user-controlled tuning parameters need to be generated within the corresponding boundary conditions. For de-dispersion kernels, AMBER may or may not utilize local memory (`localMem`)

---

[2] http://alert.eu.

[3] https://github.com/AA-ALERT/AMBER/.

[4] For a single DM, the frequency channels are first united into subbands such that the radio pulse signal first gets de-dispersed along subbands (step one, subband de-dispersion), and then within each subband (step two, intra-subband de-dispersion).

[5] https://github.com/AA-ALERT/AMBER_setup/tree/ARTS_tender.

[6] Previous pipeline version additionally had one more parameter `splitSeconds` responsible for manipulation between different kernels.

[7] Other input files contain downsampling factors, GPU cache line size, and frequency channels that need to be zapped due to terrestrial radio frequency interference (RFI) contamination.

Download English Version:

https://daneshyari.com/en/article/11031578

Download Persian Version:

https://daneshyari.com/article/11031578

Daneshyari.com