ANTARES: Development Plans

1 Introduction

This document describes the development of the ANTARES project. It lays out the projected path for successful deployment of the prototype. It also includes scenarios that would allow more functionality and flexibility but would also require more resources to build them. The long-term development is less detailed, but we present our vision for future possibilities. A history of the project and the project's engagement with the community are included.

2 Prototype Development

2.1 Complete the Implementation of the Architecture

- Augment the API to add confidence/uncertainty through the computation of the retrieved properties from the Aggregated AstroObject Catalogs, the computed property values, including feature-value confidence/uncertainty from the Touchstones, and of specific calculated rarities, and the final rarity which is combined from that of constituent replicas and combos, each associated with their individual stage code executions whose code is contributed by multiple independent astronomers.

- Have ChaosMonkey shut down and reboot processors to simulate processor failures.

- Continue to test the one-processor failure resiliency code by making ChaosMonkey much more aggressive and by running the system at much higher alert rates for much longer periods of time.

- Ensure that MySQL Cluster is attaining redundancy by examining the coherency of the data store as ChaosMonkey kills MySQL processes as well as rebooting processors running MySQL.

- Test the alert throughput to ensure that all components, in particular MySQL Cluster, can contend with high alert arrival rates.

- Consider how to spread the Aggregated AstroObject Catalog and the Locus-aggregated Alert Database across the cluster to ensure that network traffic does not limit performance. Also test to ensure that the load balancer is achieving adequate spreading of the processing across the cluster.

- Further the implementation of the provenance storage, to ensure that every property value is associated with adequate stored provenance to determine which version of the code computed that value, and which specific values of other properties were utilized in that computation, recursively.

- Start to investigate how to examine, modify, test, and debug stage code that evolves over the ten-year LSST lifespan and that references property values computed by prior versions of that stage code and that of previous stages in the pipeline.

2.2 Nighttime Dashboard, Operations, Live Alert Streams

- Implement a minimal-supervision dashboard. The nighttime dashboard aims to reduce human intervention to a minimum. In normal operations, that should be close to zero.

- Enable comparisons to previous runs. The nighttime dashboard will display information on the health of the live system, especially as it compares to similar historical data.

- Show essential system health information. The basic information displayed during nighttime operation will be memory usage and execution latency and throughput. We call this "performance data."

- Show historical context. In case of abnormal operations, it's important to avoid overcorrection. The nighttime dashboard will be designed to always show the live data in historical context, to enable comparisons to similar models, similar positions in the sky, and similar points in time.

- Allow fast reversal to known-good setup. At the beginning of each night, there will be a designated fallback state consisting of known-good stage code and touchstones. The main operational decision of the nighttime analyst is to revert to this fallback, which will typically happen if the model is either making trivially, obviously-wrong decisions, or if the models are unable to keep with the telescope data throughput. We will use "failed run" to refer to a nighttime problem which requires a reversion to a previously-used setup. The provenance of failed runs is likely more important than that of typical nights, and so will be stored for subsequent daytime analysis. The reversal-to-fallback model is designed to minimize a large loss of current science at the expense of small future improvements.

- Run prototype on smaller-scale live streams of ongoing surveys. The ongoing Catalina Sky Survey, the Palomar Transient Facility (and its planned followup, the Zwicky

Transient Facility) are already generating alerts roughly similar to the ones from LSST. We will use these alert streams to run functional tests of the entirety of the ANTARES broker, well in advance of the first LSST commissioning runs.

- Create operations plan for LSST-scale broker (i.e., beyond the prototype). Development of the system will continue. The Touchstones will grow as more surveys add to our knowledge of transients and variables, the scientific goals of the community will evolve, and filters for the system will be refined. The ANTARES system will have to respond to changing conditions.

2.3 Daytime Dashboard

- Implement the interactive visual exploration tool for ANTARES model runs. The daytime dashboard lets analysts explore the behavior of the models in depth, with the goal of enabling the improvement of the detection stages.

- Add functionality to enable comparison of model performance across different portions of the dataset: split results by apparent magnitude, position in sky, model decisions, etc. This enables analysts to understand data variation, keeping the same model. In addition, add functionality to compare different model decisions on historically-similar data distributions. This enables analysts to understand model variation, conditioned on similar data.

- Add functionality to compare performance data across different runs. Especially in the case of failed runs, analysts will need to understand what caused the run to fail, so that it can be fixed.

- Add functionality to replay nighttime runs. Every nighttime setup is version-controlled and content-addressable. This facilitates the recreation of problems and their correction.

- Implement visualization of provenance information.

- Add functionality to explore the relationship between execution traces and model outcomes. To capture the relationship between data in the touchstones and the decisions made by the nighttime database models, ANTARES records logs of the queries made to the database and the actual stage code executed.

2.4 Expand the Touchstone

Many aspects of Touchstone development are directly related to the choice of filtering algorithms. Changing filtering methods may require reevaluation of some of these tasks.

- Ingest from catalogs of variability, e.g., PTF/iPTF and MACHO surveys. Prefer studies that contain a broad representation of variables that show up in a survey, whether the variables are labeled or not. Requires normalization across catalogs.
- Populate feature sets as needed by filtering algorithms and store in Touchstones.
- Continually expand/modify Touchstones, learning from results of experimentation with filtering stages.
- Develop tool to generate (pseudo) alerts generated from time-domain data-sets already analyzed. Testing ANTARES with such alerts will expose where the Touchstone is lacking, especially inclusion of all variable types.
- For each touchstone instance (matched to a filtering stage), examine the efficacy of adopted features: experiment to find the ones that give most purchase on filtering, and optimize feature set. This is likely to be one of the most challenging tasks in the development process.
- We know that there are not adequate publicly available catalogs for variables in environments that LSST will encounter, notably in or around resolved or semi-resolved external galaxies. New investigations to fill the gap are being taken on by ANTARES team members in collaboration with non-ANTARES personnel.
- Populate touchstone with predicted feature values of examples from theoretical models of “known unknowns.”
- Evaluate the accuracy of alert characterization by nearest-neighbor classification using the available features in Touchstones.

Many of the above are iterative processes: continual curation of the Touchstone source material and the various Touchstone instances is needed to get greater completeness that represents all known classes of variable phenomena. For the proto-type, demonstrating a clear path to such continual improvement may be sufficient.

2.5 Filtering

Implement more filtering stages and evaluate their efficacy, especially to exercise the alert replication and combo functionality with realistic use cases. Test on pseudo alerts generated from time-domain data-sets already analyzed, since these also inform us about how well we are doing.

- VPDF (done)
- Light Curve Anatomy studies. (currently in use - may need improvement)
- Evaluate how the running time of nearest-neighbor-search algorithms scales with the sizes of Touchstones for alert characterization by nearest-neighbor classification.
- Investigate algorithmic enhancements of nearest-neighbor search on a budget and batch nearest-neighbor queries for improved run-time scaling.
- Investigate alternatives to k-nearest-neighbor classification for characterizing alerts, such as kernel-density-based approaches, and other schemes based on non-parametric or parametric modeling of the distributions of known object classes in feature space.
- Generalization of VPDF using time-axis and colors (in development)
- VPDF-like analysis for extra-galactic events
- Extra-galactic transient host association. Host information and correlation with transient probabilities. Research question that may require data collection.
- Incorporate community contributed filtering stages
- Develop specific combo-filtering scenarios

2.6 Community Engagement

- The project does not currently have a well-designed web site that conveys the nature and state of the project. We will complete this in early 2017.
- Continued engagement with the community via meetings and other platforms.
- In conjunction with the Las Cumbres Observatory (LCO), NOAO is hosting a workshop on “Building the Infrastructure for Time-Domain Alert Science in the LSST Era” on 2017 May 22-25. The first 2.5 days of the workshop will bring together astronomers

and others working on a wide variety of technical issues related to time-domain alerts. The second part of the workshop will discuss the science that can be done with the LCO network, bringing focus to specific requirement for the ANTARES broker.

- The software infrastructure is open source and it will need publicly available documentation of the entire system.
- A distributable version of the software, possibly in a container format, that astronomers can use to develop and test filters and other stage code. The portable version of the AstroObject database would have to be limited in size.
- The project needs a pathway for direct engagement with the system via a user interface.
- Curate use cases of interest to community

3 Long-Term Development

3.1 Broker Ecosystem

The ANTARES project is one element in a much larger time-domain ecosystem that will require many software systems to process alerts at varying levels of complexity. The focus of the prototype is the rarest of the rare, but there are many other alerts that will be of interest to astronomers on a wide range of time scales. At longer time scales, a queryable Locus-Aggregated Alert Database is likely to be the main solution (see next section). For other scales beyond the rapid response that the prototype serves, other instances of the ANTARES system could process specific subcategories with new filter sets to provide finer categorization or classification.

As an example, consider periodic variable stars. A significant fraction of the alerts LSST will generate will be continued responses to the same variable stars (Ridgway et al., 2014). These stars will be identified rapidly, and thus diverted at a very early stage in the ANTARES process. These stars can change, though, so real-time monitoring could yield interesting information about possibly dramatic changes in stellar state (e.g., Macri et al., 2001). If the diverted variable stars were channeled to another instance of the broker, a different set of filters could be applied that would be specific to these particular objects. The time limit imposed by keeping up with the LSST observing schedule is not necessary, so more detailed and computationally expensive processing could occur. For periodic variables whose classification is known, there is a prediction of how bright it will be at a particular phase in its light curve. This additional broker could look for deviations from that prediction and issue alerts when that occurs.

Objects may be categorized as periodic variables without having a precise classification. Given the LSST observing cadence, it may take some time before the object is observed at enough different phases in its light curve to make an identification. A separate instance of the ANTARES system could be used to fold in new observations of known variables and then use filters on features specifically designed to provide classifications (e.g., Richards et al., 2011), features that would not be useful for other categories of astronomical alerts.

Replicating the ANTARES software infrastructure would be straightforward. Each instance would require hardware at the scale necessary to process the alerts. The scale of the equivalent of the AstroObject database will depend on the specific goals of each instance. The number of alerts will also cover a wide range. For variable stars, the number would comparable to the scale of the full alert stream, while one that looks only at supernovae only has to handle a few thousand per night. To develop a full set of requirements, one would need to determine the science cases of interest to the community, estimate the number of alerts to process, and evaluate the complexity of filter development. The research and development phase for filter creation could be complex and thus expensive in terms of time and personnel. Simple copies of the system might only require one astronomer and one IT staff person, while more complex systems could require multiple astronomers to develop new AstroObject databases, Touchstones, and filter sets.

3.2 Serving the Locus-Aggregated Alert Database

As described in other documents, the Locus-Aggregated Alert Database provides a mechanism for studying alerts at longer time scales beyond that provided by the core ANTARES system and other brokers in the time-domain ecosystem. The ANTARES project can develop a process to translate the Locus-Aggregated Alert Database out of the system and make it a stand-alone database that serves as a queryable resource to explore LSST alerts along with all the annotations and feature calculations provided. This is, essentially, a Level 3 data product for LSST[1]. The time-scale for this episodic extraction of the Locus-Aggregated Alert Database can set the scale for what is meant by rapid response of the broker ecosystem vs. longer term alert evaluation.

Based on a scale of alerts of hundreds of gigabytes per night, and that most alerts are actually repeats of known variable stars, we can estimate the size of the Locus-Aggregated Alert Database at tens to hundreds of terabytes. This is not an unprecedented scale to serve. The NOAO Data Lab[2] is already planning on serving several databases on the same scale. The addition of this service would require astronomer, developer, and help desk support.

[1] See the LSST Data Products Definition Document, `https://ls.st/dpdd`

[2] http://datalab.noao.edu/

3.3 Build to LSST Scale

The ultimate goal of the ANTARES project is to provide a functional broker that can operate at the scale and rate of alerts that LSST will produce. The prototype has a more limited scope and does not implement all of the capabilities that would be desired for the community-based LSST broker. Nonetheless, the prototype has been developed with the LSST scale in mind, so flexibility and scalability have been considered when building the prototype. Building the full-scale broker will still be a complex task that will require resources well beyond those available for the prototype.

The current plan for alert generation and validation by LSST is that images will be transferred from the telescope to the National Center for Supercomputing Applications (NCSA) at the University of Illinois campus at Urbana-Champaign. The images will be digitally compared to references, sources in the difference image will be evaluated for 'spuriousness,' and the alerts will be broadcast. The budget allows for four alert streams, one of which will accommodate the internal LSST broker. The other three will serve community-based brokers. The ANTARES team has already engaged in informal discussions with NCSA staff about locating the ANTARES infrastructure within the NCSA facility. The bandwidth limitations are less stringent in that case. In addition, the bulk transport of alerts will allow for a more compact format than the VOEvent format of a single XML file for each alert.

The resources required would include hundreds of cores, although thorough testing of stage filters now is still necessary to provide a specific estimate of processing power. The main cost is the people to manage and run the system. The development of a complete operations plan will provide a better guide to the size of the team necessary, but this will require astronomers to vet and test the ongoing addition of filters and algorithms, IT staff to manage the code base and databases, help desk personnel to facilitate public use, astronomers and support staff to manage the allocation of resources, and the cost of bandwidth to broadcast alerts as well as to transfer the Locus-Aggregated Alert Database out to a hosting site that will serve it as a queryable database. With the complexities associated with building such a system, we would also want a project manager.

4 Past and Future Milestones of the ANTARES Project

2009 January The NOAO LSST Science Working Group is formed to develop plans for supporting community-based science in the era of LSST. The broker concept is introduced in the first few months.

2010 June Proposal submitted to NSF with the title "A Framework for Time-Critical Response to Astrophysical Events" with Saha as Principal Investigator. This went to

CISE under the Software Infrastructure for Sustained Innovation program. While given good reviews, it was not funded. Lack of computer science participation was a key criticism.

2011 July A revised version of the above proposal was resubmitted and was again not funded.

2013 January NOAO and the University of Arizona organize a local workshop, "LSST in Tucson," that brings together researchers across a wide range of fields that could be interested in the data LSST will produce, including astronomers, physicists, planetary scientists, chemists, and computer scientists. The ANTARES project grows out of this meeting.

2013 May The ANTARES INSPIRE proposal is submitted to NSF with Rick Snodgrass as Principal Investigator. John Kececioglu, Abi Saha, and Tom Matheson as co-Investigators.

2013 September The ANTARES proposal is funded (INSPIRE: CISE AST-1344204). Given the timing relative to the academic year, the hiring of graduate students and the post-doc is delayed. The four principals continue to refine the architecture design.

2014 September Computer science graduate students and astronomy post-doc begin actual development work.

2015 April Predictive model for stellar variability completed. We demonstrate a functional prototype.

2015 June Carlos Scheidegger of the University of Arizona Computer Science department, a visualization expert, joins the project.

2015 August Second version of functional prototype.

2016 January Third version of prototype presented at American Astronomical Society winter meeting.

2016 May Rob Maier of the University of Arizona Mathematics department joins the project, bringing expertise in statistics.

2016 August Fourth version of prototype presented at LSST 2016 meeting.

2016 September Second astronomy post-doc begins work.

2016 October Proposal titled "In-situ assessment and visualization of data mining in survey astronomy and the ANTARES project" submitted to NSF-CISE with Carlos Scheidegger as Principal Investigator. This proposal aims to fund research and development on the computer science side of the project.

2016 December Review.

2017 May Workshop on time-domain infrastructure.

2017 June Stand-alone version of prototype operating on cluster.

2018 January Operate on ZTF alerts (timing subject to ZTF alert production)

2019 June Public release at contemporary scale of world-wide alerts.

2020 June Public release of LSST scale broker.

5 Community Engagement

Presentations to the community about the ANTARES project.

- Hot Wiring the Transient Universe III (2013 November)
- SPIE Astronomical Telescopes + Instrumentation (2014 June)
- GMT Science Meeting, Washington DC (2014 October)
- American Astronomical Society Meeting 225 (2015 January)
- LSST Project Science Team (2015 January)
- NOAO Big Data Workshop (2015 March)
- Hot Wiring the Transient Universe IV (2015 June)
- Aspen Summer Workshop "The Dynamic Universe: Understanding ExaScale Astronomical Synoptic Surveys" (2015 June)
- International Astronomical Union XXIX (2015 August)
- Air Force Research Lab (2015 October)
- Astro Tweeps (2015 November)

- LSST Project Science Team (2015 December)
- American Astronomical Society Meeting 227 (2016 January)
- SPIE Astronomical Telescopes + Instrumentation (2016 June)
- LSST 2016 (2016 August)
- Hot Wiring the Transient Universe V (2016 October)
- Many local presentations at NOAO, University of Arizona (2013-2016)
- Building the Infrastructure for Time-Domain Alert Science in the LSST Era (2017 May)

6 Engagement with LSST Project and Community

The LSST project is well aware of the development effort for ANTARES. We have been in conversation with members of the project since the inception of ANTARES. This includes the presentations and discussions with the LSST Project Science Team listed above as well as participation in the LSST project-wide All-Hands meetings (Dove Mountain 2014, Bremerton 2015, Tucson 2016). At the Tucson meeting in August 2016, the ANTARES project was specifically called out by the LSST project as a potential community broker. Members of the ANTARES team also contributed to the most recent LSST Joint Technical Meeting in Santa Cruz (2016). We have also begun informal conversations with elements of the LSST Data Management team to discuss alert formats and potential efficiencies to be gained via bulk transport of alerts in different formats.

The LSST Transient and Variable Stars science collaboration is keenly interested in a broker in order to pursue their science interests. We have engaged them in the process of creating 'data challenges' where groups can test filters and techniques to distinguish different types of transients and variables given a minimal amount of information. The collaboration consists of astronomers with a wide range of interests who are also motivated to do the work necessary to prepare for science in the era of LSST.

The potential for an LSST Community Science Center provides a natural venue for interaction between the ANTARES project and the astronomers most interested in utilizing the time-domain alert stream from LSST. Planning and conceptual design for such a center are current NOAO activities.

References

Macri, L. M., Sasselov, D. D., & Stanek, K. Z. 2001. "A Cepheid is No More: Hubble's Variable 19 in M33." ApJL, 550, L159, astro-ph/0102453

Richards, J. W. et al. 2011. "On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data." ApJ, 733, 10, 1101.1959

Ridgway, S. T., Matheson, T., Mighell, K. J., Olsen, K. A., & Howell, S. B. 2014. "The Variable Sky of Deep Synoptic Surveys." ApJ, 796, 53, astro-ph/1409.3265