ICS PhD Portfolio
Anthony J. Christe
achriste@hawaii.edu
February, 2016

Statement of Purpose

Background and Research Interests

There is a large explosion over Washington D.C. Do you send in the troops, the red cross, the firefighters, or the hazmat team? A rolling brownout is traveling across the country, can you predict and mitigate the effects before it hits you? A large tsunami was generated by an earthquake and is heading directly towards Hawaii, do you have enough timely information to make an informed decision on evacuation procedures?

These events are distributed geographically. These events are distributed through time. These events leave clues in waveforms and metadata that can help predict future states of the environment. For the first time ever these events can be studied in great depth using distributed sensor networks. What set of metrics must be present in these types of events that allow us to process the data and maintain users' privacy at the same time?

When we develop the ability to examine the environment at an extremely detailed manner, we must weigh the costs of privacy introduced by the amount of data received vs the information that is necessary for the particular research being performed. Data fusion along with the increasing amount of metadata can make it very easy to uncover information that was intended not to be shared. How do we balance information vs privacy?

The explosion of internet connection devices (the Internet of Things IOT) has made answering these questions possible. By deploying vast distributed sensor networks, it's possible to obtain the data that answers these questions, however there is so much data that finding the signal in the noise can be like finding a needle in a haystack.

My primary research interests lie in the realm of distributed sensor networks and the software architecture that supports distributed sensor networks; namely distributed computing. More specifically, work towards my PhD involves the design and implementation of a framework that can support acquisition, persistence, reporting, and real-time analysis of temporospatial time series data for distributed power quality monitors and distributed infrasound sensors.

The recent and on-going explosion of Big Data has presented a complex challenge for data scientists and system architects. Large sensor networks continuously streaming data will often overwhelm a single server. Traditional database techniques quickly break down when Gigabytes of meta-data need to be continuously written and queried. Trying to process Big Data on a single server quickly becomes an exercise in futility.

My research interests are solving these problems in a unified and transportable way. By leveraging distributed computing, I aim to provide a framework that can meet the demands of this Big Data explosion and to advance the fields of distributed sensor networks and distributed sensor network architectures.

Goals

In the next year I plan to introduce algorithms for dealing with the acquisition of temporospatial data in distributed environments. Services such as TempoDB and OpenTSDB claim to offer a large package of analytics for distributed sensor data, however their acquisition relies on simply metrics such as single temperature values. Current services do not scale when trying to collect data with complicated meta-data and or a large vector of fields per measurement. My algorithms also provide the ability to implement custom aggregation and custom aging strategies dependent on the type of data collected. With current services, you’re stuck with the aggregation that is provided and it is generally very simply.

Within three to five years I hope to implement distributed algorithms for DSP and event detection as part of my PhD work. The current state of the art has distributed computing dominated by JVM based environments which lack solid DSP algorithms. I hope to introduce algorithms that take advantage of the distributed nature of sensor networks to empower data scientists to move from Python on hefty servers to distributed computing.

After receiving my PhD,I hope to continue my work with distributed sensor networks. The amount of sensors is increasing exponentially and will require new techniques to process. My goal is to lay the framework for the upcoming explosion of sensors and data. Advances in these technologies can pave the way to making it easier to deploy and manage sensor networks, make it easier for smaller countries to set up National Data Centers, make it easier for universities to do large scale distributed studies and continue to advance this explosion of IoT connected devices. I would prefer to stay in academia, but am willing to go into industry as well to continue to advance these ideas.

Progress

Over the past three years I've been building a framework to detect transients in power quality data. I picked up a lot of my research foundation by taking masters classes in software engineering for smart grids, advanced algorithms, advanced operating systems, theory of computation, AI, and web design.

Over the past year I've been developing a framework for the collection, analysis, and reporting of temporospatial data. My funded research through the Infrasound Laboratory at the University of Hawaii at Manoa involves detecting, quantifying, and localizing large infrasonic signals by deploying a large number of distributed sensors that continuously stream data.

Through my funding agency I was able to secure a academic cooperation participant (ACP) position with Lawrence Livermore National Labs and have been working with their Big Data scientists to solve issues such as massive distributed data ingestion with type safe persistent queues. My cooperation with LLNL provides a wide-breadth of resources in distributed computing and Big Data.

What information is needed to find the needle in the haystack in a non-np complete way? Which metrics need to be computed as data comes in?