1. Project Summary

Iowa State University (ISU) Meteorology and Agricultural Meteorology programs have enjoyed a long and fruitful collaboration with Unidata and the Unidata Community. Our programs extensively utilize Unidata technologies to support our teaching, research and outreach efforts. We are thankful for previous Unidata Equipment Awards granted to ISU that helped establish OpenGIS services for the Iowa Environmental Mesonet (IEM) (2005) [1] and create the “mtarchive” thredds/HTTP service (2010, 2012) [2]. Both of which services continue to thrive till this day. This proposal seeks to enhance and expand various IEM web services by purchasing a Dell R7525 (AMD EPYC) database server to provide two orders of magnitude more capacity.

2. Project Description

The IEM website is a heavily trafficked service providing unique data services supporting local teaching and research for the Unidata’s community and beyond. While the website’s name hints at a limited geographical domain (Iowa), the site actually contains a treasure trove of curated datasets around the globe reflecting over 20 years of active development by the project’s programmer (Daryl Herzmann). Feature additions to the website are heavily guided by feedback received by Unidata Community Members. Some of these internet-unique datasets include:

* 1929-present archive of airport weather station (ASOS) data from around the globe
* National Weather Service (NWS) Watch, Warnings, and Advisories since 2005
* Massive archive of raw text products issued by the NWS since 1983
* RAOB (sounding) archive since 1946
* Billions of hydrologic data obs processed from the NWS SHEF encoded feed since 2010.

These datasets and more are stored within a PostgreSQL relational database allowing for various web applications and web services to materialize data on-demand based on the user’s needs. The IEM server farm consists of a fleet of web, application, caching, and database servers. The farm has been cobbled together over the years with various funding mechanisms, so the five current database servers range in performance. For example, the ASOS and SHEF database servers have software throttles placed on the inbound web service request load to keep from being overrun.

The server proposed here would increase the IEM’s backend database capacity by two orders of magnitude due to the increased input/output operations per second (IOPS). The present fastest IEM database server (Dell R630 Intel 6410 SSDs) is able to sustain ~200k IOPS. The proposed server as configured with NVMe drives, should be able to sustain ~2 million IOPS along with lower latencies.

This **proposal advances remote teaching and classroom interaction for the geo-sciences** by increasing the speed and capacity of the IEM web services to deliver raw and aggregated data on-demand. A common usage is for the inclusion of these services within python notebooks via libraries like Unidata’s siphon to materialize a long time series of observational data for further analysis and plotting. Since all these services are openly and freely available on the Internet, the proposal funded server would have **equitable access** to all. In addition to programmatic web-services, the IEM website offers point-and-click download interfaces for various datasets which also ensures more **equitable access** for those less comfortable with writing code and just want spreadsheets.

The IEM curates an extensive archive of hydrologic data provided from the NWS in SHEF format. Much of this data is outside the typical bailiwick of meteorology and is additionally very difficult to find archives of. IEM implementation of web services for this dataset are limited by database capacity and this proposal would further open this **locally-created dataset and derived products to the wider Unidata community.**

2a. Technical Implementation of the Server

The software setup of this server would closely follow some excellent documentation[3] provided by the “Let’s Encrypt” project, which recently setup MySQL on this very same server hardware. Their Dell R7525 was loaded to the gills with expensive components, but the server with this proposal should perform similarly. OpenZFS on Red Hat Enterprise Linux 8 would be used to provide fault tolerance and scalability for the NVMe drives, since hardware RAID is not an option. OpenZFS would also allow compression to be used at the filesystem level, which would greatly increase (2-3x from local testing) the amount of data it could store. **Project PI Herzmann has extensive server administration and database software configuration experience, so no additional support help should be necessary.** He is permitted to administer the equipment provided by this proposal.

The server would reside within the present IEM computing cluster, which spans two data centers on the ISU Campus providing some limited redundancy. Compute intensive jobs are run within a virtual machine Red Hat Enterprise Virtualization (RHEV) environment that is managed centrally by the college. “Bare-metal” servers, like database servers, are permitted within this environment and are communicated with over isolated VLANs and not necessarily managed by RHEV.

Proposing a single server likely raises concerns about fault tolerance and what happens if this server should die. The server has a five year warranty with four hour parts, so that provides some piece of mind. For other types of failures, the IEM utilizes PostgreSQL replication to have a hot standby ready to take on writes should the primary server fail. The write performance necessary for the IEM real-time processing is trivial, so a warm standby with significantly lesser hardware will work. A lengthy fail-over would require very tight throttles to keep the webfarm load from over-running the backup hardware, but the main goal is to not have data loss nor missed real-time ingest.

2b. Addressing proposal review criteria

1. **Intellectual or technical merit:** For the 20 year technical debt of IEM software code, a beefy database server makes up for lots of poor design choices! For many of the datasets mentioned in this proposal, a relational database server remains the best technical solution to run on limited and commodity hardware.
2. **Contribution to Unidata capabilities:** Unidata has been a pioneer in remote / data proximate software systems with libraries like siphon and servers like Thredds helping to provide data to users as-needed and on-demand. This proposal would enhance the sounding archive service used within siphon and the proposal PI commits to helping to implement the other archives mentioned within siphon. Well, PI would like to do that anyway regardless of proposal funding, but this would guilt more Github pull requests!
3. **Broadening Unidata scope and capabilities:** The IEM archives of primarily point data fill in a gap that is not well facilitated by Thredds, which is more geared toward grid and short time series data. The new services exposed within siphon would additionally help MetPy users wanting such data, but finding it lacking within Unidata Thredds servers.
4. **Contributing to the advancement of technology:** This proposal does not really introduce any new technical approach to the classic problem of how to provide data over the web. This just builds more capacity to support more intensive web services. For example, provide me all CONUS growing degree day data with a provided custom base and ceiling along with the climatology between 15 April and 18 August. The IEM presently can’t provide this as the backend database would take a longer time to compute than practical for web sessions.
5. **Contribution to education:** Python notebooks are heavily used at ISU and other Unidata affiliated Universities. Improvements to speed and breath of web services provided will directly support these notebooks.
6. **Contribution to research:** Experience has shown researchers use whatever technology stack they are most comfortable with. The web services supported by this proposal are generally agnostic, so researchers should benefit as well by having access to realtime and massive archives of tough to find datasets.

3. Budget

The included Dell quote for the single server comes in at $13,130.37. ISU has agreed to waive indirect costs. Server installation, setup, and administration will be covered by proposal PI with datacenter costs of housing the server covered by other funding streams.

The budgeted server was goal-seeked to come well underneath the Unidata $20K limit while still fulfilling the goals outlined within the proposal. Hopefully this allows for more flexibility for the number of proposals that could be funded. If there is more money available after the funded proposals by Unidata are decided, we certainly could add more NVMe + memory to this proposal to fulfill even more goals! :) We’ll need to redo the quoting process anyway as the Dell quote will have expired by the time of funded. Regardless of funding availability, any re-quoted server would not exceed the budget proposed here.

4. Project Time-line

ASAP. There is no way such a fancy new toy would sit idle awaiting anything within project PI control. :) Space is currently available for our allocation within the data center, so it should be up and running within days of receipt and some performance bench marking + testing. An approximate time line is as follows:

* May 2021: Proposal awarding and accounting book work.
* Early June 2021: Purchase and delivery of server by Dell.
* Late June 2021: Server provisioned and relevant databases migrated to it.
* May-Sep 2021: Software development of additional web services this new hardware can support. Submission of pull requests to siphon to expose some current + new services there.
* September 2021: Report submitted to Unidata

[1] https://mesonet.agron.iastate.edu/docs/unidata2006/

[2] https://mtarchive.geol.iastate.edu/thredds/catalog.html

[3] https://letsencrypt.org/2021/01/21/next-gen-database-servers.html