

CDI FY17 Request for Proposals

Flocks of a feather dock together: Using Docker and HTCondor to link high-throughput computing across the USGS

Submission Title: Flocks of a feather dock together: Using Docker and HTCondor to link high-throughput computing across the USGS

Lead PI: Richard Erickson

Mission Area: Ecosystems

Region: Midwest Region

Organization: Upper Midwest Environmental Sciences Center

Orcid: 0000-0003-4649-482X

Phone: 6087816353

Email: rerickson@usgs.gov

City: La Crosse

State: WI

Co-PIs and Collaborators:

Type: Collaborator

Name: Jonathan Knudson

Mission Area: Ecosystems

Region: Midwest Region

Organization: Upper Midwest Environmental Sciences Center

Orcid: NA

Phone: 6087816201

Email: jknudson@usgs.gov

City: La Crosse

State: WI

Type: CO-PI

Name: Michael Fienen

Mission Area: Water

Region: Midwest Region

Organization: USGS Wisconsin Water Science Center

Orcid: 0000-0002-7756-4651

Phone: 6088213894

Email: mnfienen@usgs.gov

City: Middleton

State: WI

Type: CO-PI

Name: Sunnie (Grace) McCalla

Mission Area: Ecosystems

Region: Midwest Region

Organization: Upper Midwest Environmental Sciences Center

Orclid: 0000-0003-4292-8694

Phone: 6087816326

Email: smccalla@usgs.gov

City: La Crosse

State: WI

Type: Collaborator

Name: Melvin Bower

Mission Area: Ecosystems

Region: Midwest Region

Organization: Upper Midwest Environmental Sciences Center

Orclid: NA

Phone: 6087816357

Email: mbower@usgs.gov

City: La Crosse

State: WI

Type: CO-PI

Name: Jon Amberg

Mission Area: Ecosystems

Region: Midwest Region

Organization: Upper Midwest Environmental Sciences Center

Orclid: 0000-0002-8351-4861

Phone: 6087836451

Email: jamberg@usgs.gov

City: La Crosse

State: WI

Science Support Framework Element 1: Communities of Practice

Science Support Framework Element 2: Science Data Lifecycle - Analysis

Science Support Framework Element 3: Science Data Lifecycle - Processing

In-Kind Match: \$29,300.00

List of anticipated deliverables from the project: Case studies on flocking HTCondor posted to the USGS CDI BitBucket Repositories; A methods paper on using Docker with HTCondor; Documentation on best practices for deploying and configuring HTCondor and Docker on the USGS CDI BitBucket Repository; and presentations at HTCondor Week on our findings.

Lead Cost Center: Upper Midwest Environmental Sciences Center

Notes, Comments:

Project Description: USGS Science Centers use high-throughput computing to analyze and process data, but are inhibited from sharing these computing resources across centers because of unfamiliarity with these newly-developed tools, especially for USGS use within cybersecurity guidelines. We will develop tutorials and documentation for using HTCondor and Docker within USGS.

Total Budget: \$48,694.00

SECTION 1. PROJECT SUMMARY

Project Title: Flocks of a feather dock together: Using Docker and HTCondor to link high-throughput computing across the USGS

USGS Principal Investigator: Richard A. Erickson

Narrative summary:

USGS Science Centers have begun to use high-throughput computing (HTC) to analyze “Big Data” and run complex models. Specifically, centers have begun using HTCondor to pool their local resources. The Wisconsin, Oregon, Nevada, Texas, and Dakotas Water Science Centers (WSCs) use HTCondor to calibrate hydrologic models and perform uncertainty analysis. The Upper Midwest Environmental Sciences Center (UMESC) uses HTCondor to analyze genomic data from invasive species, long term monitoring data from the Mississippi River, and breeding bird survey data. Fort Collins and the Powell Center use HTCondor for synthesis activities. HTC links many individual computers that each process a small subset of a larger Big Data job. Integrating HTC capabilities in a scientific environment requires networked computers and software to schedule and manage HTC jobs. The specific software used by many USGS Centers is the open source program HTCondor (<http://research.cs.wisc.edu/htcondor/>).

Previously funded CDI work has demonstrated that USGS HTCondor pools can “flock” together—this links multiple pools such that jobs submitted in one pool (e.g. WIWSC) can run locally and also remotely with another pool (e.g. ORWSC). This has been tested as a proof of concept and is ready to be explored on a larger scale. Flocking could also be expanded to other USGS centers. However, some challenges remain before centers can flock. Two specific challenges include **1.)** software must be configured for USGS and DOI security settings and **2.)** software must be “sandboxed” (packed in a small, self-contained fashion) to run on HTCondor. We will provide case studies and example to help address these two problems.

Addressing the first challenge not only requires time to work through technical challenges, but also time to document findings. Often, USGS scientists are funded to complete projects, but not necessarily document their new methods, such as using HTCondor, for other USGS scientists. CDI funding would help us to both develop and document our new techniques. Similarly, the second challenge is unique because of the challenges posed by the Department of Interior’s cybersecurity requirements and the sparse documentation for Docker. CDI funding would enable us to better document case studies for using Docker and setting up local Docker repositories within USGS Centers to reduce the redundancy of effort.

Proposed Solution: We propose to **1.)** Leverage WIWSC expertise to further document best practices for flocking HTCondor between UMESC and the WIWSC and other centers as possible and document robust, best practices for deploying and configuring HTCondor on the USGS CDI BitBucket page; and **2.)** Create tutorials for using HTCondor and Docker that will be posted to the CDI BitBucket Page and collate our tutorials as a methods paper.

SECTION 2. ESTIMATED BUDGET

Budget Category	Federal Funding "Requested"	Matching Funds "Proposed"
1. PERSONNEL (SALARIES including benefits):		
UMESC Federal Personnel Total:	\$ 18,500	\$ 26,000
WWSC Federal Personnel Total:	\$ 18,500	
Contract/Collaborator Personnel Total:		
Total Salaries:	\$ 37,000	\$ 26,000
2. TRAVEL EXPENSES:		
Travel Total (Per Diem, Airfare, Mileage/Shuttle) x # of Trips:	\$ 1,700	\$ 2,700
Other Expenses (e.g. Registration Fees):	\$ 200	\$ 600
Total Travel Expenses:	\$ 1,900	\$ 3,300
3. OTHER DIRECT COSTS: (itemize)		
Equipment (including software, hardware, purchases/rentals):		
Publication Costs:		
Office Supplies, Training, Other Expenses (specify):		
Total Other Direct Costs:		
UMESC Total Direct Costs:	\$ 19,766.67	\$ 29,300
Indirect Costs (Indirect Rate 18.62%):	\$ 3,681	
WWSC Total Direct Costs:	\$ 19,133.33	
Indirect Costs (Indirect Rate 31.953%):	\$ 6,114	
GRAND TOTAL:	\$ 48,694	\$ 29,300