SECTION 1. PROJECT ADMINISTRATIVE INFORMATION

Addressing CDI Framework elements: Science Inputs; Communities of Practice; Computational Tools and Services; Data and Information Assets; Management, Policy, and Standards.

Project Title: Hunting Invasive Species with HTCondor: High Throughput Computing for Big Data and Next Generation Sequencing

Lead USGS cost center requesting funding: Upper Midwest Environmental Sciences Center (UMESC)

USGS Principal Investigator: S. Grace McCalla, USGS-UMESC; ORCID ID 0000-0003-4292-8694; 2630 Fanta Reed Rd., La Crosse, WI, 54603; Ph. 608-781-6326; Fax 608-783-6066; smccalla@usgs.gov

Additional Principal Investigators: Michael Fienen, USGS-WIWSC, mnfienen@usgs.gov; Richard Erickson, USGS-UMESC, rerickson@usgs.gov; Randall Hunt, USGS-WIWSC, rjhunt@usgs.gov; Jon Amberg, USGS-UMESC, jamberg@usgs.gov.

Problem statement and implications: Innovative computing solutions are necessary to effectively process complex data sets in useful timeframes. Integrating high-throughput computing solutions simultaneously across multiple USGS centers leverages already existing hardware to analyze these large data sets. CDI funding could help build data-processing infrastructure between multiple USGS centers. In addition to helping USGS scientists and our partner organizations to analyze data and run large computing jobs, this project would share computing resources to enable better research and work towards meeting computer reduction initiatives. Enabling the use of HTC by UMESC scientists will empower research and build collaborations within USGS and beyond.

Anticipated deliverables: Computing capabilities leveraged to process complex data; Framework for developing high-throughput computing within and between USGS centers for data processing; Webinar on the integration of computing capabilities between USGS centers; USGS blogs entries as the project is developed.

SECTION 2. ESTIMATED BUDGET

Budget Category	Federal Funding "Requested"	Matching Funds "Proposed"
1. PERSONNEL (SALARIES including benefits):		
Federal Personnel Total:	$ 32,289	$ 25,884
Contract/Collaborator Personnel Total:		
Total Salaries:	$ 32,289	$ 25,884
2. TRAVEL EXPENSES:		
Travel Total (Per Diem, Airfare, Mileage/Shuttle) x # of Trips:	$ 2,700	$ 1,500
Other Expenses (e.g. Registration Fees):	$ 200	$ 600
Total Travel Expenses:	$ 2,900	$ 2,100
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:	$ 16,016	$ 27,984
Indirect Costs (Indirect Rate 18.62%):	$ 2,982	
WWSC Total Direct Costs:	$ 19,173	
Indirect Costs (Indirect Rate 29.902%):	$ 5,733	
GRAND TOTAL:	$ 43,904	$ 27,984

SECTION 3. PROJECT SUMMARY

Large and complex data sets pose computational challenges that require innovative solutions to analyze in timeframes useful for today's decision making. The computationally intensive demands of these large data sets, or 'Big Data', surpass the processing capabilities of traditional computing approaches of a single personal computer. The field of genomics, as a Big Data example, is projected to reach 1 zettabyte (1,000,000,000,000 gigabytes) by 2025, and exceed the annual storage needs of YouTube or Twitter (Stephens et al. 2015). Within the USGS, the Upper Midwest Environmental Sciences Center (UMESC) in La Crosse, WI, has begun to generate genomic data using Next Generation Sequencing (NGS) for a wide range of projects that generate large amounts of data that must be analyzed.

The NGS machine at UMESC that processes DNA samples produces 300 GB of data every time it is operated, and multiple sequencing runs can be completed per week. UMESC scientists face increased computational demands to complete their research while the current IT infrastructure within the UMESC Local Area Network cannot handle this size of computing problem. USGS computing approaches must be updated to enable USGS scientists to conduct research and keep pace with the rapidly developing big data scientific fields. Beyond the field of genomics, developing advanced computing capabilities would contribute to data processing for other research that has Big Data needs. As an example, UMESC scientists are evaluating the efficacy of complex sound, CO_2, and seismic water guns as deterrent techniques for Asian carp. To capture the response of each Asian carp to a deterrent, the fish are implanted with radio tags that capture GPS positions for each fish every millisecond. These detections generate billions of correlated data points that must be triangulated to generate a position for every fish during the course of the study. Traditionally processing this data requires 30 days on a personal computer.

One advanced computing area where USGS has begun to invest is "high-throughput" computing (HTC); for example, the Wisconsin Water Science Center (WIWSC) in Middleton, WI, has pioneered implementing and maintaining a widely used open-source HTC code, HTCondor (https://research.cs.wisc.edu/htcondor/), on USGS hardware and within USGS IT security protocols. Note that high-throughput computing is a different computing problem than those addressed by high-performance computing, or those done by "supercomputers" (e.g., see Fienen and Hunt 2015). Supercomputing resources are already supported by the USGS Core Systems Sciences Mission Area.

The WIWSC uses HTC to link 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 that WIWSC implements is the open source program HTCondor (HTCondor Team 2014, http://research.cs.wisc.edu/htcondor/). HTCondor works by passing a large data-processing job to many (100s or even 1000s) different networked computers that each perform a portion of the total workload of the job. HTCondor has the added benefit of leveraging additional use from existing hardware infrastructure because HTCondor can be used with personal computers that are idle. Leveraging HTCondor contributes to satisfying DOI and USGS policies reducing computer purchases and consolidating data centers.

Proposed Solution: We propose to **1.)** Leverage WIWSC expertise to implement HTCondor practices at UMESC within the scope of USGS IT security mandates. **2.)** After UMESC has full HTCondor capabilities, we will link HTCondor capabilities between WIWSC and UMESC to further increase Big Data analysis abilities. Building on this initial framework, this HTC protocol will be standardized from UMESC and WIWSC for possible application throughout the USGS as part of the nascent Advanced Computing Consortium.

References

Fienen, M. N., and R. J. Hunt. 2015. High-Throughput Computing Versus High-Performance Computing for Groundwater Applications. Groundwater 53:180–184.

HTCondor Team. 2014. HTCondor version 8.0 Manual. Technical report, University of Wisconsin-Madison.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. 2015. Big Data: Astronomical or Genomical? PLoS Biol 13:e1002195.