Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS)

General Co-Chairs: Garth Gibson, Carnegie Mellon University
                        Yong Chen, Texas Tech University
Program Co-Chairs: Shane Canon, Lawrence Berkeley National Laboratory
                        Dean Hildebrand IBM

Abstract: This proposal describes a plan for a Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS) to be hosted at the 2016 ACM/IEEE Supercomputing Conference (SC'16). This workshop proposal is a joint effort and an outcome of intensive discussions from two successful workshop communities with SC in the past a number of years, the Parallel Data Storage Workshop (PDSW) and Data Intensive Scalable Computing Systems (DISCS). Looking at the last three years only, PDSW has averaged 125 attendees per year, publishing about 8 papers of about 24 submitted each year; over its four years DISCS has averaged about 60 attendees per year, publishing about 11 papers of about 21 submitted each year. The objective of this joint workshop effort is to combine two overlapping communities and to better promote and stimulate researchers' interactions to address some of most critical challenges for scientific data storage, management, devices, and processing infrastructure for both traditional compute intensive simulations as well as and data-intensive high performance computing solutions. This proposal describes the proposed joint workshop scope, format, expected outcomes, organization, and timeline.

1. Introduction

Many scientific problem domains have become extremely data intensive. For instance, simulations that model the behavior of plasma in a tokamak fusion device or in the Earth's magnetosphere can generate tens or even hundreds of terabytes of data during a single run [3, 4]. The Linear Coherent Light Source at the Stanford Linear Accelerator Laboratory produced over 2 PB in 2013 alone. Traditional high performance computing (HPC) systems and the programming models for using them such as MPI [2] were designed from a compute-centric perspective with an emphasis on achieving high floating point computation rates. But processing, memory, and storage technologies have advanced at differing rates resulting in a widening performance gap between computation and the data management infrastructure. Hence data management has become the performance bottleneck for a significant number of applications targeting HPC systems.

The explosion of data processing systems using infrastructure like MapReduce [1] has altered the storage and data management landscape feeding different data processing techniques back into traditional HPC data processing workflows for manipulating and exploring large data volumes. The Congressional Office of Management and Budget has informed the Department of Energy that new machines beyond the first exascale machines must address both the traditional simulation workloads as well as data intensive applications. This coming convergence prompts integrating these two workshops into a single entity to address the common coming challenges.

The proposed Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS) is a combined effort of two successful workshops with SC in the past a number of years, Parallel Data Storage Workshop (PDSW) and Data Intensive Scalable Computing Systems (DISCS). The 10 year PDSW workshop series focused on the data storage and management problems and emerging solutions found in peta- and exascale scientific computing environments. It paid special attention to issues in which community collaboration can be crucial for problem identification, workload capture, solution interoperability, standards with community buy in, and shared tools, given the growing and unprecedented demands of storage capacity, performance, concurrency, reliability, availability, and manageability from peta- and exascale computing infrastructures. Since 2013, PDSW has been held in conjunction with the Association of Computing Machinery (ACM) SIGHPC, allowing workshop proceedings to be published via the ACM and IEEE digital libraries. ACM and/or IEEE have also published the proceedings since 2007, the first year with peer-reviewed papers. Attendance over the last 3 years (2015-2013) has been 150, 100, and 120 respectively. The paper acceptance rate over the last 6 years (2015-2010) has been 9/25, 8/29, 8/16, 9/20, 9/21, and 9/18. In addition to the regular papers, a

program of 5-10 short work-in-progress papers or posters have also been a fixture every year.

The DISCS workshop series has been held each of the last four years (since 2012) in conjunction with SC and facilitated dialogue about research aimed at the intersection of data intensive computing and traditional high performance computing. DISCS workshops provided a venue for researchers to discuss recent results and the future challenges of running data intensive applications on both traditional HPC systems and the latest data-centric computer systems. The programs included a diverse collection of rigorously peer-reviewed papers, prestigious invited keynote speakers, and lots of vibrant discussion. Attendance was 40 in 2014 and 100 in 2012. The paper acceptance rate over all 4 years (2015-2012) has been 9/15, 10/15, 10/19, and 14/43. Since 2013, the DISCS workshops have been held in cooperation with the Association for Computing Machinery (ACM) SIGHPC, allowing workshop proceedings to be published via the ACM and IEEE digital libraries. Also, each year we invited authors of accepted workshop papers to submit an extended version of their paper to a journal special issue that is guest edited by the DISCS organizers. The DISCS workshop provided a significant value to the authors who present their research as part of the workshop, and to the SC attendees who participate in the workshop. NUMBERS

2. Scope

The scope of the proposed joint PDSW-DISCS workshop is summarized as:

- Storage architectures, virtualization, emerging storage devices and techniques
- Performance benchmarking, resource management, and workload study from production systems
- Programmability, APIs, and fault tolerance of storage systems
- Parallel file systems, metadata management, complex data management, object and key-value storage, and other emerging data storage/retrieval techniques.
- System architectures interconnection networks, I/O, power efficiency for data intensive computing
- Programming models for data intensive computing (extensions to traditional programming models or to data intensive programming models, non-traditional programming languages/models)

- Runtime systems, inter-node and inter-system communication, data compression and de-duplication, caching and prefetching, and data integrity for data intensive computing
- Productivity tools for data intensive computing, data mining and knowledge discovery tools, mathematical and statistical techniques, tools for performance, debugging, and administration
- Techniques for integrating compute into a complex memory hierarchy facilitating in situ and in transit data processing avoiding I/O bottlenecks.
- Data filtering/compressing/reduction techniques that maintain sufficient scientific validity for large scale compute-intensive workloads.
- (If workflow is in scope from a data management perspective [it was for DISCS]) tools and techniques for managing data movement among compute and data intensive components both solely within the compute area as well as incorporating the memory/storage hierarchy
- Data management support for emerging programming models such as Asynchronous Multi-Task programming models (e.g., Charm++ or Legion)

3. Format

The joint PDSW-DISCS workshop intends to continue and combine the successful format from two previous workshop series. More specifically, the joint PDSW-DISCS workshop is planned to include an invited keynote talk, presentations of peer-reviewed research papers, and a work-in-progress (WIP) or lightning talks session using a full-day, single-track schedule. The workshop will consist of four sessions. After brief welcoming remarks from the workshop organizers, the first session will consist of an hour-long invited keynote talk. Previous keynote speakers include established researchers from both academia and industry settings, as well as visionary program managers of federal agencies. We expect to identify a keynote speaker of similar experience, relevance, and profile for this joint workshop. The remaining three workshop sessions will consist of presentations of peer-reviewed research papers. Each presenter will be given 20-30 minutes for a presentation and to take questions from the audience. Each submitted paper will receive at least three reviews by members of the PDSW-DISCS program committee. The program committee will discuss their reviews during a teleconference moderated by the workshop program co-chairs, and the program co-chairs will select the papers to be presented at the workshop based on the reviews and this discussion. The WIP session intends

to be the last session of the workshop and features lightning talks of ongoing work. We will conclude the workshop with closing remarks from workshop organizers plus a brief "open mic" opportunity for workshop attendees to give feedback about the new joint workshop. The workshop schedule includes mid-morning, lunch, and mid-afternoon breaks that align with the traditional SC Technical Program break schedule.

4. Expected Outcomes

We expect four tangible outcomes from the joint PDSW-DISCS workshop. First, this joint workshop effort intends to combine two communities with an increasingly converged research agenda to better promote and stimulate researchers' interactions to address some of most critical challenges for storage and data-intensive HPC solutions. This outcome and experience can also be valuable to many other workshops at SC and their evolution over years. Second, assuming the joint PDSW-DISCS workshop is granted "In Cooperation with SIGHPC" status, the proceedings of the workshop will be published via the ACM and IEEE digital libraries. Third, the authors of all papers presented at the workshop will be invited to publish an extended version of their article in a journal special issue. The journal has not yet been selected, but past DISCS workshops have produced special issues for the International Journal of Cluster Computing, Parallel Computing, and the International Journal of High Performance Computing Applications. Fourth, if granted "In Cooperation with SIGHPC" status, the workshop organizers will produce a report summarizing the workshop and associated discussion, submit it to SIGHPC, and make it available on the workshop web site.

5. Plan for Attracting Attendees

With the growing overlap of target attendee domains, existing promotion avenues for the previous PDSW and DISCS workshops will be leveraged. Postings on HPC-Announce, tcpp-announce, hipeac-announce, storage-research-list, and the hpc-iodc workshop at ISC email lists will continue. Posting to the LinkedIn distributed computing group and wikicfp.com. Announcements at related venues, such as the HPC-IODC workshop at ISC will seek to expand the audience.

6. Timeline

The proposed joint PDSW-DISCS workshop series are expected to have one full year cycle from the end of the previous year workshop to the end of the current year workshop once it is fully running. A tentative timeline is scheduled as follows:

- By April 2016, the organizers including the steering committee, general chairs, area chairs, and technical program committee members are finalized.
- By May 2016, the advertisement and distribution of call for papers will be started.
- By August 2016, advertising the workshop and soliciting submissions, answering queries
- By September 30th, 2016, paper acceptance finalized
- By October 15th, 2016, provide workshop proceedings to ACM and IEEE digital libraries.
- By October 31st, 2016, re-advertise conference with information about program.
- November 14th, 2016, planned workshop date, held at SC'16
- December 2016 – June 2017, journal special issue follow-up

7. Organization

A well-designed workshop organization is critical for the healthy growth of the workshop and to be sustainable in the long run. The joint PDSW-DISCS workshop combines the communities from two past workshops and is designed to have an organization of five-level bodies, steering committee, general chair, area chairs (including program committee chair, proceedings chair, and publicity chair), and technical program committee.

The role of the steering committee includes: 1) appointing general chair(s) for the workshop each year; 2) overseeing the management of the workshop each year; and 3) advising general chair(s) and area chairs. The initial joint steering committee consists of the combined steering committee from two past workshop series and is shown as below.

The role of the general chair(s) includes: 1) managing the selection of program chair(s), appointing a proceedings chair and publicity chair for the workshop; 2) securing SC and SIGHPC approval for the workshop and managing the relationship with these bodies; 3) pursuing and managing a journal special issue derived from the workshop; and 4) supervising area chairs in managing workshop activities.

The area chairs include program chairs, proceedings chair, and publicity chair. The responsibility of the program chairs includes: 1) inviting technical program committee members and forming the technical program committee; 2) inviting a keynote speaker or a panel of experts; 3) organizing paper review process and answering authors questions; 4) making selections of papers and posters; 5) arranging workshop schedule and chairing the workshop including inviting session chairs; and 6) managing the workshop website together with the publicity chair. The proceedings chair and publicity chair manage the proceeding with the SIGHPC and promotion activities, respectively. The responsibility of the technical program committee includes: 1) reviewing assigned paper submissions; and 2) discussing papers when there are conflicting reviews. Occasionally external reviewers are needed and invited by the technical program committee members. These external reviewers will assist paper reviews, but the inviting program committee member needs to approve the reviews. The external reviewers' names will be collected and acknowledged in the proceedings.

Steering committee members:
- John Bent, EMC
- Ali R. Butt, Virginia Tech
- Yong Chen, Texas Tech University
- Evan J. Felix, Pacific Northwest National Laboratory
- Garth A. Gibson, Carnegie Mellon University

- William D. Gropp, University of Illinois at Urbana-Champaign
- Gary Grider, Los Alamos National Laboratory
- Dean Hildebrand, IBM Research
- Dries Kimpe, KCG, USA
- Jay Lofstead, Sandia National Laboratories
- Darrell Long, University of California, Santa Cruz
- Xiaosong Ma, Qatar Computing Research Institute, Qatar
- Carlos Maltzahn, University of California, Santa Cruz
- Robert Ross, Argonne National Laboratory
- Philip C. Roth, Oak Ridge National Laboratory
- John Shalf, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory
- Xian-He Sun, Illinois Institute of Technology
- Rajeev Thakur, Argonne National Laboratory
- Lee Ward, Sandia National Laboratories

References
The PDSW series programs and papers can be viewed at the following website:
http://www.pdsw.org

The DISCS series programs and papers can be found at the following websites:
http://ft.ornl.gov/discs-2015/
http://discl.cs.ttu.edu/discs-2014/
http://ft.ornl.gov/discs-2013/
http://discl.cs.ttu.edu/discs/