STUDENT CLUSTER COMPETITION REPRODUCIBILITY CHALLENGE

A Brief History

Based on work by: Michela Taufer, Stephen Lien Harrell, Hai Ah Nam, Kris Garrett, Christopher Bross, Scott Michael, and many others

Presented by: Stephen Lien Harrell

REPRODUCIBILITY INITIATIVE AT SC

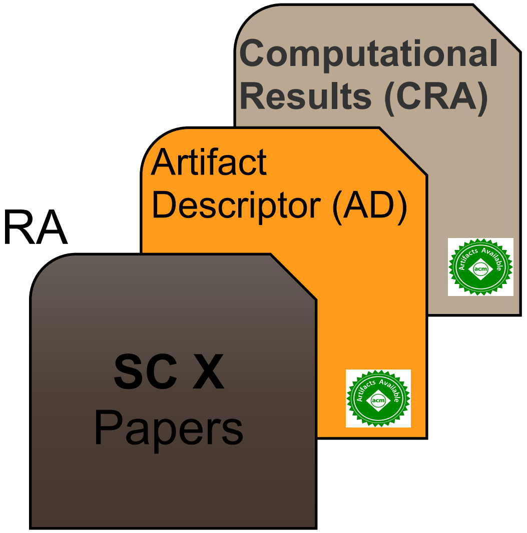
Technical Program @ SC X

Select BP/BSP candidates

Assign badge

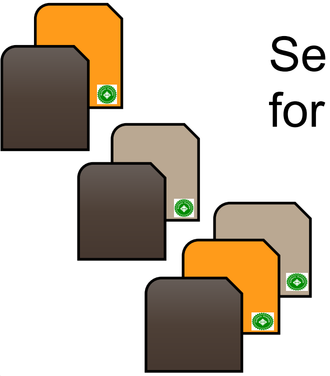
Check AD or CRA

Review papers



Technical Program @ SC X+2

Review ParCo SI paper with SCC reports from SCC @ SC X+1



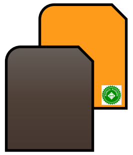
Select one (1) **SC X** paper for **SC X+1** SCC

Technical Program @ SC X+1

Assign badge to SC X paper

Give SIGHPC certificate to SC X paper authors

Present ParCo SI with SCC reports from SCC @ SC X-1



Generate replication benchmark for diverse set of HPC platforms

Student Cluster Competition @ SC X+1

Partner with vendors

Build a cluster

Test performance benchmarks

Replicate **SC X** Paper

Generate replication reports

STUDENT CLUSTER COMPETITION (SCC)

- Teams of 6 undergraduate students build and operate a small HPC cluster on the exhibit floor of SC every year since 2007
- The teams “race” their clusters to run data sets during the competition
- Primary constraint is the machine must run within 3000 watts of power – this means we see many different hardware architectures

REPRODUCIBILITY CHALLENGE DETAILS

- Each team runs a computational experiment from the chosen paper during the competition and attempts to reproduce computational results from the paper
- The teams then write a report on how they implemented the experiment and what their findings were
- These reports were published in PARCO special issues for SC16 and SC17 (volume 70 and 79 respectively)
- Reproducibility Challenge Report Outline (4 Page Max)
    - *Introduction*
        - *State the claims that the paper made and what is trying to be reproduced*
    - *Description of the HPC machine and environment*
    - *Description of the steps taken to reproduce*
    - *Data from the student's experiment*
    - *Compare results to the original paper*
        - *Are they similar, why or why not?*
    - *Conclusion*
        - *Were you able to reproduce the results?*

- Along with the report the students were required to submit the output files from the application

SC16 SCC REPRODUCIBILITY CHALLENGE

- Chosen paper
  - Flick, P., Jain, C., Pan, T., & Aluru, S. (2015, November). A parallel connectivity algorithm for de Bruijn graphs in metagenomic applications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (p. 15). ACM.

- Competition Challenge
  - Application from the paper is ParConnect
  - Used 2 un-released datasets so computation is done at the competition
  - Students were asked to use a profiler to determine MPI timings and reproduce Figure 3 from the paper
  - Students were asked to do a strong scaling study and compare their results to Figure 4

SC17 SCC REPRODUCIBILITY CHALLENGE

- Chosen paper
    - Höhnerbach, M., Ismail, A. E., & Bientinesi, P. (2016, November). The vectorization of the tersoff multi-body potential: an exercise in performance portability. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*(p. 7). IEEE Press.

- Competition Challenge
    - Application from the paper is LAMMPS
    - A new unreleased dataset is used
    - Students were asked to reproduce performance timings in figures based on their architectures (Figure 4 and 5 for CPU and KNL, Figure 6 for GPU)
    - Students were asked to do a strong scaling study on both the original dataset in the paper and the new dataset and compare results to Figure 9.

SC18 SCC REPRODUCIBILITY CHALLENGE – STARTING TODAY!

- Chosen paper
  - Uphoff, C., Rettenberger, S., Bader, M., Madden, E. H., Ulrich, T., Wollherr, S., & Gabriel, A. A. (2017, November). Extreme scale multi-physics simulations of the tsunamigenic 2004 sumatra megathrust earthquake. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (p. 21). ACM.

- Can't speak about the details as they have not been released to the students yet.

SC19 SCC REPRODUCIBILITY CHALLENGE – OPEN QUESTIONS

- We will be reusing a lot of great work from the previous challenges

- Main Question: How do we curate a set of artifacts and release those with the reports
  - What digital artifacts are appropriate to curate?
    - Containers? Metadata? Automation?
  - What information can one gleam by looking at the artifact along with the report to contrast the reports?
    - By architecture or approach for example
  - Is this useful to community at large?
    - It is extra work on top of the base challenge to do this.
    - Will community members ever look at the artifacts or attempt to replicate the work themselves?
      - If they did, would the artifacts of how it was reproduced be useful?

LINKS

- All competition applications and challenges (including reproducibility) from SC15 to present
  - https://scholarworks.iu.edu/dspace/handle/2022/21179

- SC16 Reproducibility Challenge Paper and PARCO Journal
  - Paper: https://doi.org/10.1145/2807591.2807619
  - Journal: https://doi.org/10.1016/j.parco.2017.10.002

- SC17 Reproducibility Challenge Paper and PARCO Journal
  - Paper: https://doi.org/10.1109/SC.2016.6
  - Journal: https://doi.org/10.1016/j.parco.2018.10.001

- SC18 Reproducibility Challenge Paper
  - Paper: https://doi.org/10.1145/3126908.3126948