SDSC Summer Institute 2017
Supercomputing for the Long Tail of Science

July 31 – August 4, 2017

SDSC Auditorium at UC San Diego

Lesson material repository: https://github.com/sdsc/sdsc-summer-institute-2017

MONDAY, July 31	
8:00 – 8:30AM	Registration, Coffee
8:30 – 8:45	**Welcome** Shawn Strande, Deputy Director, SDSC
8:45 – 9:30	**Orientation** Andrea Zonca, Director of the Summer Institute, SDSC
9:30 – 10:15	**How do I launch and manage jobs on the system?** Mahidhar Tatineni, User Services Manager, SDSC
10:15 – 10:45	Break
10:45 – 12:15	**Launching and Managing Jobs** Mahidhar Tatineni, User Services Manager, SDSC
12:15 – 1:30	Lunch at Café Ventanas
1:30 – 3:00	**How do I manage my data on the file system?** Manu Shantharam, Title, SDSC
3:00 – 3:30	Break
3:30 – 5:00	**How do I know I'm making effective use of the machine?** Bob Sinkovits, Director for Scientific Computing Applications, SDSC
5:30 – 8:30PM	**Reception at Wayne Pfeiffer's home overlooking the Pacific** *Sweater or jacket recommended* Shuttle provided from SDSC driveway

TUESDAY, August 1	
8:00 – 8:30AM	Coffee
8:30 – 10:00	**How do I automate my job pipeline to ensure reproducibility?** *[update with hands-on or Data Science talk]* Ilkay Altintas, SDSC's Chief Data Science Officer, Director, Workflows for Data Science (WorDS) Center of Excellence, SDSC
10:00 – 10:15	Break
10:15 – 12:15	**How do I manage my software?** Andrea Zonca, HPC Applications Support Specialist, SDSC
12:15 – 1:30	Lunch at Café Ventanas
1:30 – 2:30	**What are the benefits of using Science Gateways?** Amit Majumdar, Division Director, Data Enabled Scientific Computing, SDSC
2:30 – 3:30	**When can virtualization help HPC?** Trevor Cooper, High Performance Computing Systems Manager, SDSC
3:30 – 3:45	Break
3:45 – 4:15	SDSC Data Center Tour
4:15 – 5:00PM	Hands-on practice continues with mentors available for questions

WEDNESDAY, August 2 **PARALLEL SESSIONS**		
8:00 – 8:30AM	Coffee	
	Track 1 *Auditorium*	**Track 2** *Synthesis Center E-B143*
Session 1 8:30 – 12:00	**GPU Computing and Programming** Andreas Goetz, Co-Director, CUDA Teaching Center, Co-Principal Investigator, Intel Parallel Computing Center, SDSC This session provides an introduction to massively parallel computing with graphics processing units (GPUs). The use of GPUs is becoming increasingly popular across all scientific domains since GPUs can significantly accelerate time to solution for many problems. Participants will be introduced to essential background of the GPU chip architecture and will learn how to program GPUs via the use of Libraries, OpenACC compiler directives, and CUDA programming. The session will incorporate hands-on exercises for participants to acquire the skills to use and develop GPU aware applications.	**Spark for Scientific Computing** Andrea Zonca, HPC Applications Specialist Mahidhar Tatineni, User Services Manager, SDSC Apache Spark is a cluster computing framework extensively used in Industry to process large amount of data (up to 1PB) distributed across thousands of nodes. It has been designed as a successor of Hadoop focusing on performance and usability. It provides interface in Python, Scala and Java. This session will provide an overview of the capabilities of Spark and how they can be leveraged to solve problems in Scientific Computing. Next it will feature a hands-on introduction to Spark, from batch and interactive usage on Comet to running a sample map/reduce example in Python. The final part will be devoted to two key libraries in the Spark ecosystem: Spark SQL, a general purpose query engine that can interface to SQL databases or JSON files and Spark MLlib, a scalable Machine Learning library.
12:00 – 1:30	Lunch at Café Ventanas	
Session 2 1:30 – 5:00PM	**Performance Optimization** Bob Sinkovits, Director for Scientific Computing Applications, SDSC This session is targeted at attendees who both do their own code development and need their calculations to finish as quickly as possible. We'll cover the effective use of cache, loop-level optimizations, force reductions, optimizing compilers and their limitations, short-circuiting, time-space tradeoffs and more. Exercises will be done mostly in C, but emphasis will be on general techniques that can be applied in any language.	**Visualization** Amit Chourasia, Senior Visualization Scientist, SDSC Visualization is largely understood and used as an excellent communication tool by researchers. This narrow view often keeps scientists from fully using and developing their visualization skillset. This tutorial will provide a “from the ground up" understanding of visualization and its utility in error diagnostic and exploration of data for scientific insight. When used effectively visualization can provide a complementary and effective toolset for data analysis, which is one of the most challenging problems in computational domains. In this tutorial we plan to bridge these gaps by providing end users with fundamental

		visualization concepts, execution tools, customization and usage examples. Finally, a short introduction to SeedMe.org will be provided where users will learn how to share their visualization results ubiquitously.

THURSDAY, August 3 **PARALLEL SESSIONS**		
8:00 – 8:30AM	**Coffee**	
	Track 1 *Auditorium*	**Track 2** *Synthesis Center E-B143*
Session 3 8:30 – 12:00	**Parallel Computing using MPI & Open MP** Pietro Cicotti, Senior Computational Scientist, SDSC This session is targeted at attendees who are looking for a hands-on introduction to parallel computing using MPI and Open MP programming. The session will start with an introduction and basic information for getting started with MPI. An overview of the common MPI routines that are useful for beginner MPI programmers, including MPI environment set up, point-to-point communications, and collective communications routines will be provided. Simple examples illustrating distributed memory computing, with the use of common MPI routines, will be covered. The OpenMP section will provide an overview of constructs and directives for specifying parallel regions, work sharing, synchronization and data scope. Simple examples will be used to illustrate the use of OpenMP shared-memory programming model, and important run time environment variables Hands on exercises for both MPI and OpenMP will be done in C and FORTRAN.	**Machine Learning Overview – AM Session** Mai Nguyen, Data Scientist, SDSC Paul Rodriguez, Research Analyst, SDSC Nicole Wolters, Programmer Analyst III, SDSC Machine learning is an interdisciplinary field focused on the study and construction of computer systems that can learn from data without being explicitly programmed. This track provides an introduction to the machine learning algorithms and techniques used to explore, analyze, and leverage data to construct data-driven solutions applicable to any domain. The morning session will cover the machine learning process, R/RStudio, data exploration, and data preparation. The afternoon session will cover classification, cluster analysis, and tools and procedures to scale up machine learning techniques on Comet. Hands on exercises/demonstrations will be done in R, and Python with Spark.
12:00 – 1:30	Lunch at Café Ventanas	
Session 4 1:30 – 5:00PM	**Python for HPC** Andrea Zonca, HPC Applications Specialist, SDSC Bob Sinkovits, Director for Scientific Computing Applications, SDSC Python is rapidly becoming more widely adopted in the High Performance Computing world. In this session, we will introduce four key technologies in the Python ecosystem that provide significant	**Scalable Machine Learning** Mai Nguyen, Data Scientist, SDSC Paul Rodriguez, Research Analyst, SDSC Nicole Wolters, Programmer Analyst III, SDSC Machine learning is an interdisciplinary field focused on the study and construction of computer systems that can learn from data without being

	benefits for scientific applications run in supercomputing environments. Previous Python experience is not required. (1) IPython Notebook allows users to execute code on a single compute node or cluster and export the Python web interface to the local browser for interactive data exploration and visualization. IPython Notebook supports live Python code, explanatory text, LaTeX equations and plots in the same document. (2) IPython Parallel provides a simple, flexible and scalable way of running thousands of Python serial jobs by spawning IPython kernels (namely engines) on any HPC batch scheduler. It also allows interactive control of the engines from an IPython Notebook session along with the ability to submit more Python tasks to the engines. (3) Numba makes it possible to run pure Python code on GPUs simply by decorating functions with the data types of the input and output arguments. Pure Python prototype code can be gradually optimized by pushing the most computationally intensive functions to the GPU without the need to implement code in CUDA or OpenCL. (4) PyTrilinos is a Python wrapper for the Trilinos, a C++ Distributed Linear Algebra library developed by Sandia National Labs. It provides a high level interface for transparently dealing with complex MPI point-to-point communication strategies for operations involving both dense and sparse matrices and vectors whose data are distributed across an arbitrary number of nodes.	explicitly programmed. This track provides an introduction to the machine learning algorithms and techniques used to explore, analyze, and leverage data to construct data-driven solutions applicable to any domain. The morning session will cover the machine learning process, R/RStudio, data exploration, and data preparation. The afternoon session will cover classification, cluster analysis, and tools and procedures to scale up machine learning techniques on Comet. Hands on exercises/demonstrations will be done in R, and Python with Spark.
5:30 – 9:00PM	**Beach BBQ Dinner at La Jolla Shores Hotel**, *sweater or jacket recommended* 8110 Camino Del Oro, La Jolla, CA 92037 Shuttle provided from SDSC driveway	

FRIDAY, August 4	
8:00 – 8:30	Coffee
8:30 – 9:30	**Emerging Technologies in HPC** Pietro Cicotti, Senior Computational Scientist, SDSC [or Shawn]
9:30 – 11:00	**Lightning Rounds**
11:00 – 11:30	**Wrap up**
11:30AM	**Adjourn** Thank you for attending we hope you enjoyed the week! *(To-go box lunches will be available)*