Slurm User Group Meeting
November 19, 2013

SC13

SchedMD LLC
http://www.schedmd.com

Over 130 individual contributors
representing dozens of organizations

Ramiro Alba (Centre Tecnològic de Tranferència de Calor, Spain)
Amjad Majid Ali (Colorado State University)
Pär Andersson (National Supercomputer Centre, Sweden)
Don Albert (Bull)
Ernest Artiaga (Barcelona Supercomputing Center, Spain)
Danny Auble (SchedMD)
Jason W. Bacon
Susanne Balle (HP)
Ralph Bean (Rochester Institute of Technology)
Alexander Bersenev (Institute of Mathematics and Mechanics, Russia)
David Bigagli (SchedMD)
Nicolas Bigaouette
Anton Blanchard (Samba)
Janne Blomqvist (Aalto University, Finland)
David Bremer (Lawrence Livermore National Laboratory)
Jon Bringhurst (Los Alamos National Laboratory)
Bill Brophy (Bull)
Luis Cabellos (Instituto de Fisica de Cantabria, Spain)
Thomas Cadeau (Bull)
Hongjia Cao (National University of Defense Technology, China)
Jimmy Cao (Greenplum/EMC)
Ralph Castain (Intel, Greenplum/EMC, Los Alamos National Laboratory)
François Chevallier (CEA)
Daniel Christians (HP)
Gilles Civario (Bull)
Chuck Clouston (Bull)
Yuri D'Elia (Center for Biomedicine, EURAC Research, Italy)
Francois Diakhate (CEA, France)
Joseph Donaghy (Lawrence Livermore National Laboratory)
Chris Dunlap (Lawrence Livermore National Laboratory)
Phil Eckert (Lawrence Livermore National Laboratory)
Joey Ekstrom (Lawrence Livermore National Laboratory/Bringham Young University)
Josh England (TGS Management Corporation)
Kent Engström (National Supercomputer Centre, Sweden)
Carles Fenoy (Barcelona Supercomputing Center, Spain)
Damien François (Université catholique de Louvain, Belgium)
Jim Garlick (Lawrence Livermore National Laboratory)
Didier Gazen (Laboratoire d'Aerologie, France)
Raphael Geissert (Debian)
Yiannis Georgiou (Bull)
Armin Größlinger (University Passau, Germany)
Mark Grondona (Lawrence Livermore National Laboratory)
Dmitri Gribenko
Andriy Grytsenko (Massive Solutions Limited, Ukraine)
Michael Gutteridge (Fred Hutchinson Cancer Research Center)
Chris Harwell (D. E. Shaw Research)
Takao Hatazaki (HP)
Matthieu Hautreux (CEA, France)
Dave Henseler (Cray)
Chris Holmes (HP)
David Höppner
Nathan Huff (North Dakota State University)

David Jackson (Adaptive Computing)
Alec Jensen (SchedMD)
Morris Jette (SchedMD)
Klaus Joas (University Karlsruhe, Germany)
Greg Johnson (Los Alamos National Laboratory)
Magnus Jonsson (Umeå University, Sweden)
Jason King (Lawrence Livermore National Laboratory)
Yury Kiryanov (Intel)
Aaron Knister (Environmental Protection Agency, UMBC)
Nancy Kritkausky (Bull)
Roman Kurakin (Institute of Natural Science and Ecology, Russia)
Sam Lang
Puenlap Lee (Bull)
Dennis Leepow
Olli-Pekka Lehto (CSC-IT Center for Science Ltd., Finland)
Piotr Lesnicki (Bull)
Bernard Li (Genome Sciences Centre, Canada)
Eric Lin (Bull)
Donald Lipari (Lawrence Livermore National Laboratory)
Komoto Masahiro
Steven McDougall (SiCortex)
Donna Mecozzi (Lawrence Livermore National Laboratory)
Bjørn-Helge Mevik (University of Oslo, Norway)
Chris Morrone (Lawrence Livermore National Laboratory)
Pere Munt (Barcelona Supercomputing Center, Spain)
Denis Nadeau
Mark Nelson (IBM)

Michal Novotny (Masaryk University, Czech Republic)
Bryan O'Sullivan (Pathscale)
Gennaro Oliva (Institute of High Performance Computing and Networking, Italy)
Alan Orth (International Livestock Research Institute, Kenya)
Juan Pancorbo (Leibniz-Rechenzentrum, Germany)
Chrysovalantis Paschoulas (Juelich Supercomputing Centre, Germany)
Rémi Palancher
Alejandro Lucero Palau (Barcelona Supercomputing Center, Spain)
Daniel Palermo (HP)
Martin Perry (Bull)
Dan Phung (Lawrence Livermore National Laboratory/Columbia University)
Ashley Pittman (Quadrics, UK)
Ludovic Prevost (NEC, France)
Vijay Ramasubramanian (University of Maryland)
Krishnakumar Ravi[KK] (HP)
Chris Read
Petter Reinholdtsen (University of Oslo, Norway)
Gerrit Renker (Swiss National Supercomputing Centre)
Andy Riebs (HP)
Asier Roa (Barcelona Supercomputing Center, Spain)
Andy Roosen (University of Deleware)
Miguel Ros (Barcelona Supercomputing Center, Spain)
Beat Rubischon (DALCO AG, Switzerland)
Simon Ruderich
Dan Rusak (Bull)
Eygene Ryabinkin (Kurchatov Institute, Russia)
Federico Sacerdoti (D. E. Shaw Research)
Aleksej Saushev
Rod Schultz (Bull)
Filip Skalski (University of Warsaw, Poland)
Jason Sollom (Cray)
Eric Soyez (Science+Computing)
Marcin Stolarek
Tyler Strickland (University of Florida)
Jeff Squyres (LAM MPI)
Prashanth Tamraparni (HP, India)
Jimmy Tang (Trinity College, Ireland)
Kevin Tew (Lawrence Livermore National Laboratory/Bringham Young University)
John Thiltges (University of Nebraska-Lincoln)
Adam Todorski (Rensselaer Polytechnic Institute)
Stephen Trofinoff (Swiss National Supercomputing Centre)
Garrison Vaughan
Daniel M. Weeks (Rensselaer Polytechnic Institute)
Nathan Weeks (Iowa State University)
Andy Wettstein (University of Chicago)
Tim Wickberg (Rensselaer Polytechnic Institute)
Ramiro Brito Willmersdorf (Universidade Federal de Pemambuco, Brazil)
Jay Windley (Linux NetworX)
Eric Winter
Anne-Marie Wunderlin (Bull)
Yair Yarom (The Hebrew University of Jerusalem, Israel)
Nathan Yee (SchedMD)

SchedMD LLC
http://www.schedmd.com

Five of Top Ten Systems Use Slurm
(Top 500 List, June 2013)

Rank	Site	Manufacturer	Computer	Cores	RMax
1	NUDT	NUDT	Tianhe-2	3,120,000	33.9
3	LLNL	IBM	Sequoia	1,572,864	17.2
6	TACC	Dell	Stampede	462,462	5.17
8	LLNL	IBM	Vulcan	393,216	4.29
10	NUDT	NUDT	Tianhe-1A	186,368	2.57

Slurm use continuing to grow, especially on largest systems

Exascale Focus

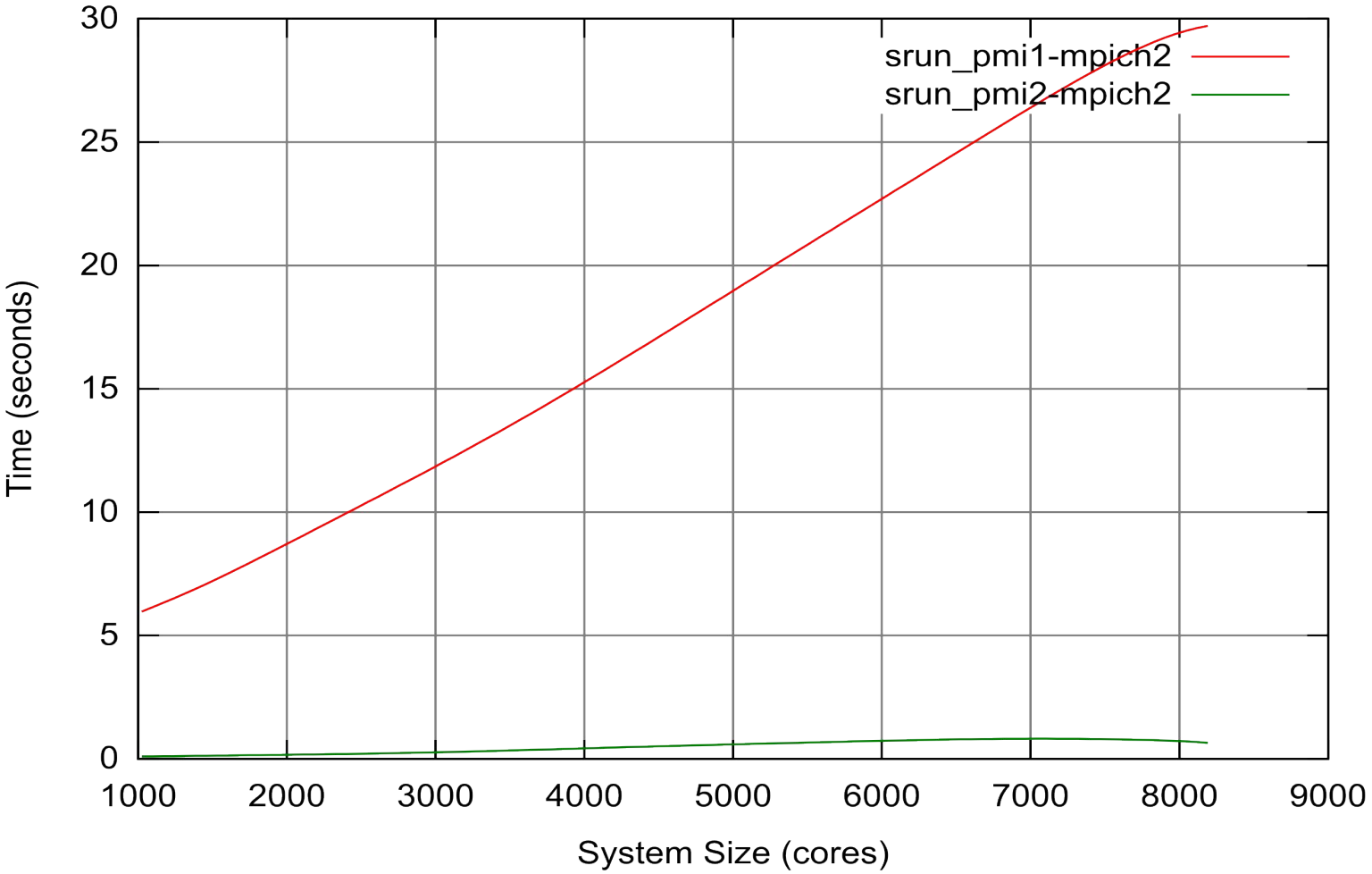
- Heterogeneous Environment
- Scalability
- Reliability
- Energy Efficiency
- New models (Cloud/Virtualization/Hadoop)

Version 2.6 Enhancements

- Released 6 July
- Adds support for job arrays

  $ sbatch –array=1-50000 -N1 -i my_in_%a -o my_out_%a my.bash
- Improved support for PBS/Torque interface
  - More commands and options supported
- Improved Throughput
  - Modified locking for high-throughput computing
  - Added support for pending job steps with call-back
- Adds PMI2 infrastructure (MPICH2 infrastructure)
  - Vast improvement in scalability and performance

PMI1 vs PMI2 Performance



Version 2.6 Enhancements

- Added External Sensors Plugins (IPMI and RAPL)
  - Power consumption of jobs now in accounting records
  - Power consumption data also available by node
- Added job profiling
  - Periodically capture per task resource use (CPU, Memory, Power, InfiniBand and Lustre)
  - Data written to HDF5 file on each node
  - Aggregate data to single HDF5 file on job completion
  - Numerous tools available for analysis

Version 14.03 Enhancements

- Release late March 2014
- Native Cray Cascade support (no ALPS)
- Integration with global resource server (e.g. FlexLM license manager)

Version 14.03 Hadoop

- Hadoop integration
  - Eliminates need for dedicated Hadoop cluster
  - Better scalability
    - Launch: Hadoop/YARN (~N), Slurm (~log N)
    - Wireup: Hadoop/YARN (~N^2), Slurm (~log N)
  - No modifications to Hadoop, transparent to applications

Version 14.03 Failure Management

- Configurable hot-spare nodes
- Running jobs can replace or remove failed or failing nodes
- Jobs can extend time limit based upon failures
- Jobs can drain nodes they perceive to be failing
- Configurable access control lists and limits

Future Work

- Improved scheduling support for job dependencies (e.g. pre-processing, post-processing, co-processing on I/O nodes)
- Optimize system throughput with respect to varying power caps
- Multi-parameter scheduling based on the layout framework
- Fault-tolerance and jobs dynamic adaptation through communication between Slurm , MPI libraries and application
- Network communication scalability optimizations

Audience participation

- Current features most important to you?
- What features are important to add?

- SC13 survey: http://bit.ly/sc13-eval

- Slurm User Group Meeting 2014

  Swiss National Supercomputing Centre

  23-24 September 2014