Slurm Workload Manager Project Report

Slurm BOF
SC12
November 15, 2012

Morris Jette and Danny Auble
[jette,da]@schedmd.com

Go Team!
Over 100 Individual Contributors

Ramiro Alba (Centre Tecnologic de Tranferencia de Calor, Spain)
Amjad Majid Ali (Colorado State University)
Par Andersson (National Supercomputer Centre, Sweden)
Don Albert (Bull)
Ernest Artiaga (Barcelona Supercomputing Center, Spain)
Danny Auble (SchedMD, LLNL)
Susanne Balle (HP)
Ralph Bean (Rochester Institute of Technology)
Alexander Bersenev (Institute of Mathematics and Mechanics, Russia)
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)
Hongjia Cao (National University of Defense Technology, China)
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 (LLNL/Bringham Young University)
Josh England (TGS Management Corporation)
Kent Engstrom (National Supercomputer Centre, Sweden)
Carles Fenoy (Barcelona Supercomputing Center, Spain)
Damien Francois (Universite catholique de Louvain, Belgium)
Jim Garlick (Lawrence Livermore National Laboratory)
Didier Gazen (Laboratoire d'Aerologie, France)
Raphael Geissert (Debian)
Yiannis Georgiou (Bull)
Mark Grondona (Lawrence Livermore National Laboratory)
Andriy Grytsenko (Massive Solutions Limited, Ukraine)
Takao Hatazaki (HP)
Matthieu Hautreux (CEA, France)
Chris Holmes (HP)
David Hoppner
Nathan Huff (North Dakota State University)
David Jackson (Adaptive Computing)
Alec Jensen (SchedMD)
Moe Jette (SchedMD, LLNL)
Klaus Joas (University Karlsruhe, Germany)
Greg Johnson (Los Alamos National Laboratory)
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)
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)
Bjorn-Helge Mevik (University of Oslo, Norway)
Chris Morrone (Lawrence Livermore National Laboratory)
Pere Munt (Barcelona Supercomputing Center, Spain)
Mark Nelson (IBM)
Michal Novotny (Masaryk University, Czech Republic)
Bryan O'Sullivan (Pathscale)
Gennaro Oliva (Institute of High Performance Computing and Networking, Italy)
Remi Palancher
Alejandro Lucero Palau (Barcelona Supercomputing Center, Spain)
Daniel Palermo (HP)
Martin Perry (Bull)
Dan Phung (LLNL/Columbia University)
Ashley Pittman (Quadrics, UK)
Vijay Ramasubramanian (University of Maryland)
Krishnakumar Ravi[KK] (HP)
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)
Aleksej Saushev
Rod Schultz (Bull)
Jason Sollom (Cray)
Tyler Strickland (University of Florida)
Jeff Squyres (LAM MPI)
Prashanth Tamraparni (HP, India)
Jimmy Tang (Trinity College, Ireland)
Kevin Tew (LLNL/Bringham Young University)
John Thiltges (University of Nebraska-Lincoln)
Adam Todorski (Rensselaer Polytechnic Institute)
Stephen Trofinoff (Swiss National Supercomputing Centre)
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)
Anne-Marie Wunderlin (Bull)
Nathan Yee (SchedMD)

Go Team!
Over 100 Individual Contributors

Ramiro Alba (Centro Tecnológico de Tranferencia de Calor, Spain)
Amjad Majid Ali (Colorado State University)
Par Andersson (National Supercomputer Centre, Sweden)
Don Albert (Bull)
Ernest Artiaga (Barcelona Supercomputing Center, Spain)
Danny Auble (SchedMD, LLNL)
Susanne Balle (HP)
Ralph Bean (Rochester Institute of Technology)
Alexander Bersenev (Institute of Mathematics and Mechanics, Russia)
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)
Hongjia Cao (National University of Defense Technology, China)
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 (LLNL/Bringham Young University)
Josh England (TGS Management Corporation)
Kent Engström (National Supercomputer Centre, Sweden)
Carles Fenoy (Barcelona Supercomputing Center, Spain)
Damien Francois (Université catholique de Louvain, Belgium)
Jim Garlick (Lawrence Livermore National Laboratory)
Didier Gazen (Laboratoire d'Aerologie, France)
Raphael Geissert (Debian)
Yiannis Georgiou (Bull)
Mark Grondona (Lawrence Livermore National Laboratory)
Andriy Grytsenko (Massive Solutions Limited, Ukraine)
Takao Hatazaki (HP)
Matthieu Hautreux (CEA, France)
Chris Holmes (HP)
David Hoppner
Nathan Huff (North Dakota State University)
David Jackson (Adaptive Computing)
Jacob Jensen (SchedMD)
Moe Jette (SchedMD, LLNL)
Klaus Joas (University Karlsruhe, Germany)
Greg Johnson (Los Alamos National Laboratory)
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)
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)
Bjorn-Helge Mevik (University of Oslo, Norway)
Chris Morrone (Lawrence Livermore National Laboratory)
Pere Munt (Barcelona Supercomputing Center, Spain)
Mark Nelson (IBM)
Michal Novotny (Masaryk University, Czech Republic)
Ryan O Sullivan (Nathscale)
Gennaro Oliva (Institute of High Performance Computing and Networking, Italy)
Jose Pancorbo
Alejandro Lucero Palau (Barcelona Supercomputing Center, Spain)
Daniel Palermo (HP)
Martin Perry (Bull)
Dan Phung (LLNL/Columbia University)
Ashley Pittman (Quadrics, UK)
Vijay Ramasubramanian (University of Maryland)
Krishnakumar Ravi[KK] (HP)
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)
Aleksej Saushev
Rod Schultz (Bull)
Jason Sollom (Cray)
Tyler Strickland (University of Florida)
Jeff Squyres (LAM MPI)
Prashanth Tamraparni (HP, India)
Jimmy Tang (Trinity College, Ireland)
Kevin Tew (LLNL/Bringham Young University)
John Thiltges (University of Nebraska-Lincoln)
Adam Todorski (Rensselaer Polytechnic Institute)
Stephen Trofinoff (Swiss National Supercomputing Centre)
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)
Anne-Marie Wunderlin (Bull)
Nathan Yee (SchedMD)

Thank You
Merci
Gracias
Kiitos

SchedMD LLC
http://www.schedmd.com

Contributing Organizations
Funding and/or Work

- Barcelona Supercomputing Center
- Bright Computing
- Bull
- CEA
- Fred Hutchinson Cancer Research Center
- Greenplum/EMC
- HP
- Intel
- Lawrence Livermore National Laboratory
- National University of Defense Technology (China)
- NVIDIA
- Oak Ridge National Laboratory
- SchedMD
- Swiss National Supercomputing Centre (CSCS)

Version 2.5 Enhancements

- Record of power consumption by job
  - New *acct_gather_energy* plugin infrastructure

- User control over CPU frequency
  - New *srun --cpu-freq* option

- Added ability to reserve all nodes in a partition
  - Reservation updated when nodes added to or removed from a partition

- Boards added to node topology information
  - In addition to sockets, cores, thread, and Gres

Version 2.5 Enhancements

- Support for Intel MIC (Xeon Phi) Processor
  - Offloading only, not running as stand alone node

- Substantial performance improvements
  - Throughput up to 630 jobs per second

- Integration with IBM Parallel Environment
  - New launch plugin infrastructure
    - New *launch/slurm, launch/poe,* and *launch/runjob* plugins
    - New *switch/nrt plugin*

Version 2.5 Enhancements

- Advanced reservation of cores, not nodes
  - Not currently available for BlueGene

- Ability to exclude accounts and users from reservation
  - Example: account=science users-=adam

- Node's CPU_Load information available

- Streamlined installation on Cray with RPMs
  - *Launch/aprun* plug for better srun support

Release Status

- Version 2.5.0-rc1 (release candidate one) available now and undergoing testing
- Release of version 2.5.0 planned late in November
- Release version 2.6 planned 2nd quarter 2013
  - Continue with major release about every 6 months

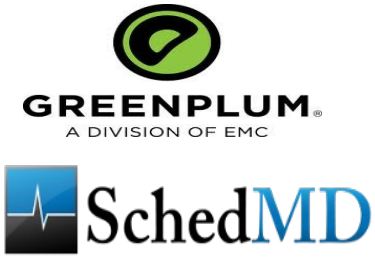
Version 2.6 Enhancements
(Preliminary)

- Scheduling optimized for energy efficiency
  - Based upon infrastructure/hardware power limits and job energy needs
  - Temperature aware

- License Management integration with FlexLM/Flexnet Publisher

Version 2.6 Enhancements (Preliminary)

- Integration with MapReduce
  - Orders of magnitude performance improvement



- Improved support for Intel MIC (Xeon Phi) Processor
  - Use as stand-along Slurm compute node

- Finer-grained BlueGene resource management
  - Partitions/queues and advanced reservations managed by c-node rather than midplane

Version 2.6 Enhancements (Preliminary)

- Partition parameter *MaxCPUsPerNode*
  - Use to reserve some CPUs for use with GPUs


```
# Excerpt from slurm.conf
#
JobSubmitPlugins=limit_gpu_use_by_partition  # Site-specific script
#
NodeName=tux[1-128] CPUs=12 Gres=gpu:1
#
PartitionName=cpu Default=yes Nodes=tux[1-128] MaxCPUsPerNode=10
PartitionName=gpu Default=no  Nodes=tux[1-128] MaxCPUsPerNode=12
```