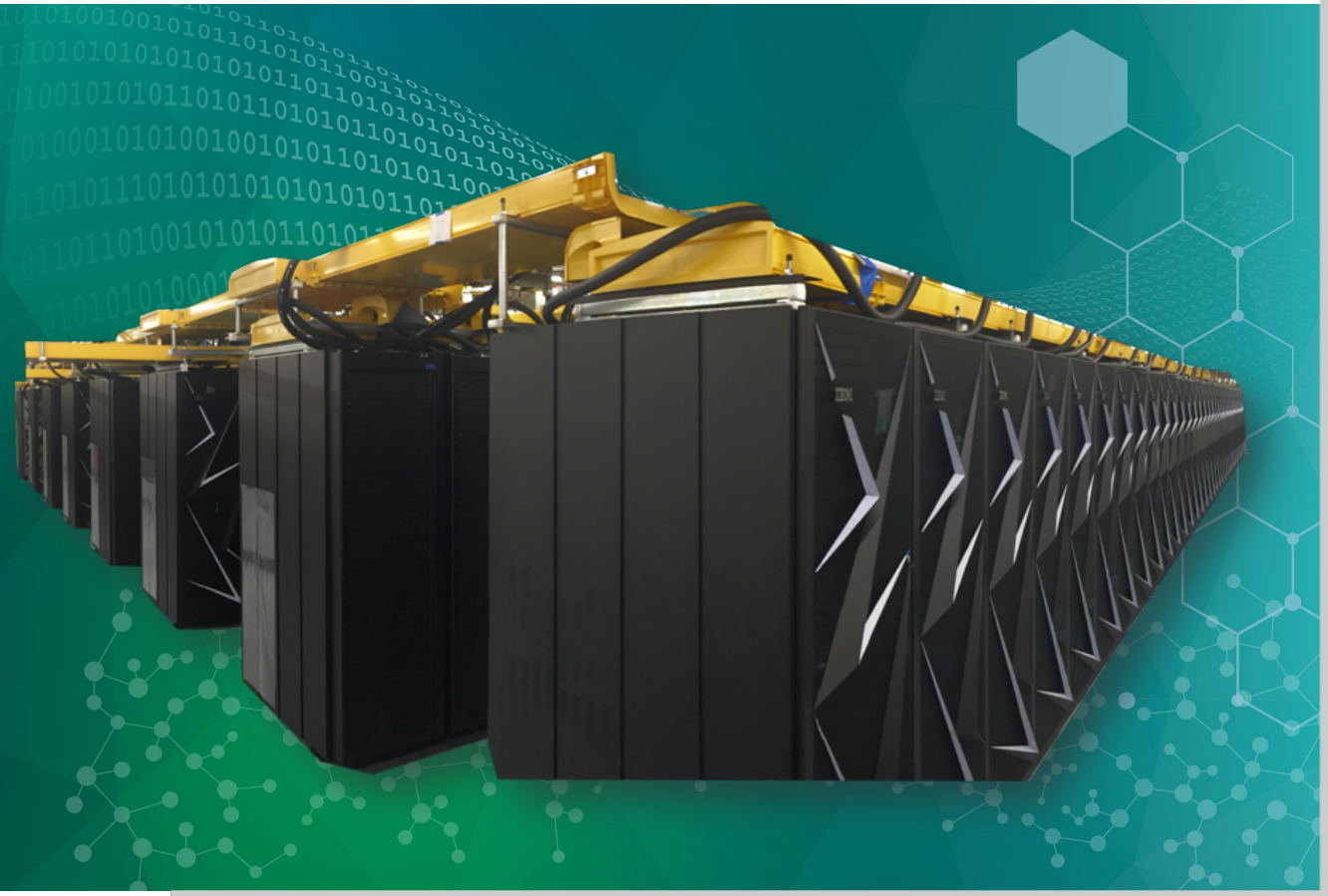
OAK RIDGE National Laboratory

Using the world's most powerful computer

Jack Morrison



SC19 Hands-On with Summit
Denver, CO
November 22, 2019

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

U.S. DEPARTMENT OF ENERGY

- Programming Environment
  - LMOD (modules)
  - Compiler Toolchains
  - Center-provided Software
- Programming Methods for Multi-GPU Nodes
- LSF Resource Scheduler
  - Batch
  - Interactive
- Parallel Job Launch

Programming Environment

- LMOD (modules)
- Compiler Toolchains
- Center-provided Software

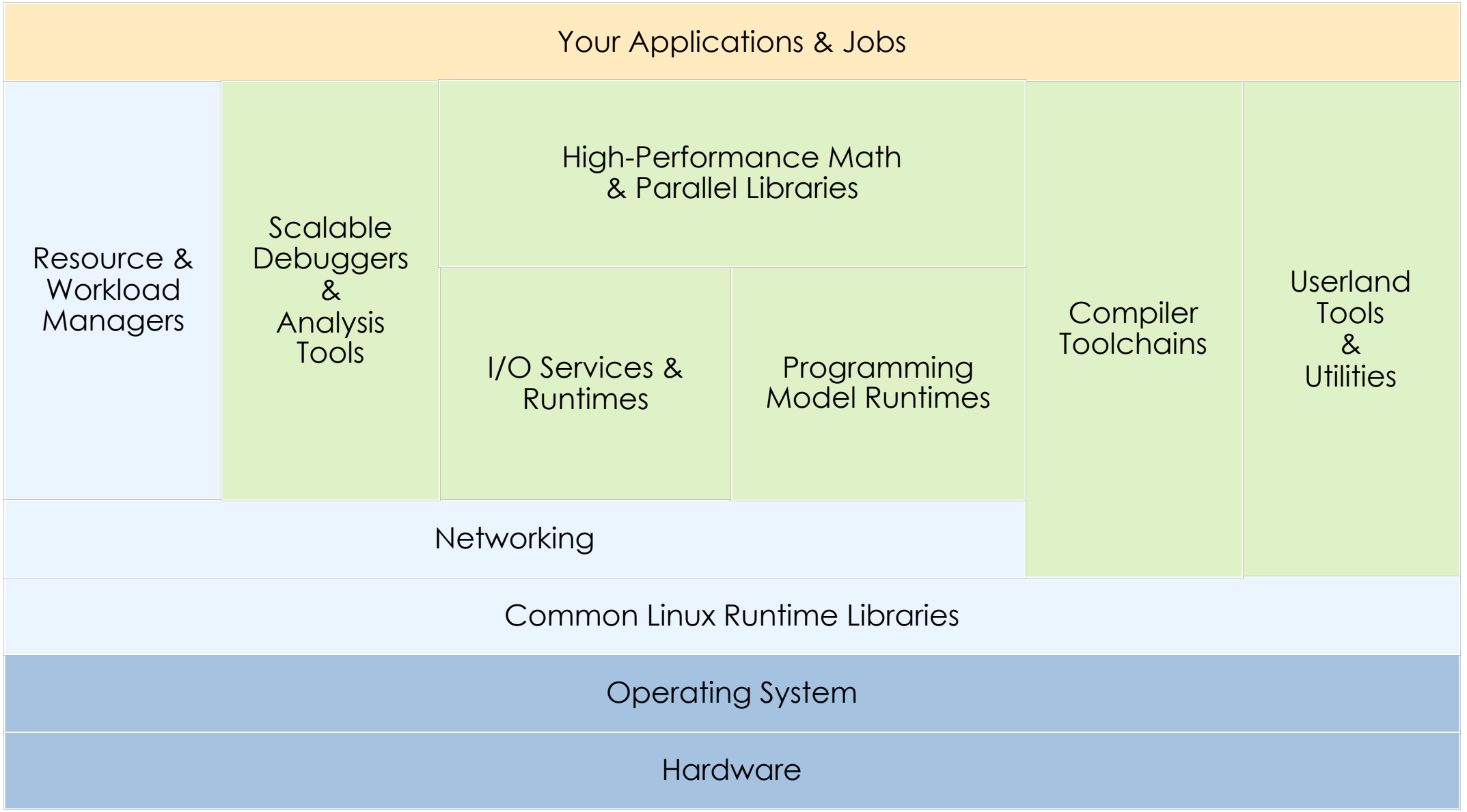
Programming Methods for Multi-GPU Nodes

LSF Resource Scheduler

- Batch
- Interactive

Parallel Job Launch

What is the programming environment?



What is the programming environment?

- At the highest level, the PE is your shell's build- and run-time environment.
  - See output of running `env`
- Software outside of default paths (`/usr/bin, /usr/lib`, etc.)
- Managed via session Environment Variables
  - Search Paths
    - `PATH, LD_LIBRARY_PATH, LIBRARY_PATH, PKG_CONFIG_PATH,...`
  - Program options via Environment Variables
    - `OMPI_*, CC, FC, …`
- Summit runs LMOD to manage environment complexity

LMOD Environment Modules

- Much of the available software cannot coexist simultaneously in your environment.
- Build- and runtime-environment software managed with LMOD (https://lmod.readthedocs.io)
- Usage:

```
$ module -t list                    # list loaded modules
$ module avail                      # Show modules that can be loaded given current env
$ module help <package>             # Help info for package (if provided)
$ module show <package>             # Show contents of module
$ module load <package> <package>…  # Add package(s) to environment
$ module unload <package> <package>… # Remove package(s) from environment
$ module reset                      # Restore system defaults
$ module restore <collection>       # Load a saved collection
$ module spider <package>           # Deep search for modules
$ module purge                      # Clear all modules from env.
```

Module Avail

- The `module avail` command shows **only what can be loaded given currently loaded packages.**
- Full or partial package names limit output to matches.

```
$ module avail
------------ /sw/summit/modulefiles/site/linux-rhel7-ppc64le/Core ----------
   ...
   cuda/9.1.85                      py-nose/1.3.7        (D)
   cuda/9.2.64              (D)     py-pip/9.0.1
   gcc/4.8.5                (L)     python/3.5.2
   gcc/5.4.0                        readline/6.3

  Where:
   L:  Module is loaded
   D:  Default Module

Use "module spider" to find all possible modules.
Use "module keyword key1 key2 ..." to search for all possible modules matching
any of the "keys".
```

Path in `MODULEPATH` where a module exists. Printed in order of priority.

Searching for Modules with Spider

- Use `module spider` (not `avail`) to search for modules
  - Finds packages that cannot be loaded given current environment
  - Shows requirements needed to make package available

```
$ module -t spider hdf5/1.10.0-patch1
----------------------------------------------------------------------------
  hdf5: hdf5/1.10.0-patch1
----------------------------------------------------------------------------
    You will need to load all module(s) on any one of the lines below before
the "hdf5/1.10.0-patch1" module is available to load.
      ...
      gcc/4.8.5
      gcc/4.8.5  spectrum-mpi/10.1.0.4-20170915
      gcc/4.8.5  spectrum-mpi/10.2.0.0-20171117
      ...
```

Spider (cont'd)

- Complete listing of possible modules is only reported when searching for a specific version:
  `` `module spider <package>/<version>` ``
- Can search using limited regular expressions:
  - All modules with 'm' in their name: `module -t spider 'm'`
  - All modules starting with the letter 'm': `module -t -r spider '^m'`

Default Modules

- DefApps meta module
  - XL compiler
  - Spectrum MPI
  - HSI – HPSS interface utilities
  - XAlt – Library usage
  - LSF-Tools – Wrapper utility for LSF
  - darshan-runtime – An IO profiler; unload if using other profilers.

Compiler Environments

IBM XL (default)

- `xl/16.1.1-5 (latest)`
  - Older modules not recommended

New compiler releases added regularly.

GCC

- `Gcc/9.1.0`
- `gcc/8.1.1`
  - OpenACC Capable
- `gcc/7.4.0`
  - Latest w/ CUDA10 NVCC
- `gcc/6.4.0` (default)
- `gcc/5.4.0`
- `gcc/4.8.5`
  - RHEL7 OS compiler in `/usr`
  - "Core" modulefiles

PGI

- `pgi/19.9`
- `pgi/19.7`
- `pgi/19.5`
- `pgi/19.4 (default)`
- `pgi/19.1`
- `pgi/18.10`
- `pgi/18.7`

```
[jackm][login2][~]
▪ module avail

------------------------------------------------ /autofs/nccs-svm1_sw/summit/modulefiles/site/linux-rhel7-ppc64le/spectrum-mpi/10.3.0.1-20190611-aqjt3jo/xl/16.1.1-3 ------------------------------------------------
   adios2/2.2.0                boost/1.59.0 (D)    hdf5/1.8.18       hypre/2.11.1       mpip/3.4.1          netcdf-fortran/4.4.4    netlib-scalapack/2.0.2    parallel-netcdf/1.8.1    petsc/3.7.2-py2
   amgx/2.0.0.130.2-unthreaded fftw/3.3.8          hdf5/1.10.3 (D)   hypre/2.13.0 (D)   netcdf-cxx4/4.3.0   netcdf/4.6.1            parallel-io/2.3.0         parmetis/4.0.3

------------------------------------------------------------------------------ /sw/summit/modulefiles/site/linux-rhel7-ppc64le/xl/16.1.1-3 ------------------------------------------------------------------------------
   autoconf/2.69   (D)    bzip2/1.0.6         gdbm/1.18.1 (D)    libdwarf/20180129       libtool/2.4.2       ncurses/6.1              openssl/1.0.2 (D)    readline/7.0                 (D)    zeromq/4.2.5 (D)
   automake/1.16.1 (D)    cmake/3.14.2        hdf5/1.8.18        libelf/0.8.13           m4/1.4.18     (D)   netlib-lapack/3.8.0      pkgconf/1.5.4 (D)    spectrum-mpi/10.3.0.1-20190611 (L)    zfp/0.5.2    (D)
   boost/1.59.0            diffutils/3.7 (D)   hdf5/1.10.3        libsodium/1.0.15  (D)   metis/5.1.0         numactl/2.0.11     (D)   python/2.7.15        sqlite/3.26.0                (D)    zlib/1.2.11  (D)

------------------------------------------------------------------------------ /sw/summit/modulefiles/site/linux-rhel7-ppc64le/Core ------------------------------------------------------------------------------
   DefApps                          (L)      expat/2.2.5                       hwloc/2.0.2-py3           lsf-tools/1.0             py-certifi/2017.1.23-py3       spectral/20181227
   apr-util/1.6.0                            fontconfig/2.12.3                 hwloc/2.0.2         (D)   lsf-tools/2.0       (L,D) py-docutils/0.13.1-py3         spectral/20190401    (D)
   autoconf/2.69                             freetype/2.9.1                    icu4c/58.2                lz4/1.8.1.2               py-nose/1.3.7-py2              sqlite/3.26.0
   automake/1.16.1                           gcc/4.8.5                         libassuan/2.4.5           m4/1.4.18                 py-nose/1.3.7-py3              subversion/1.9.3-py2
   bison/3.0.5                               gcc/5.4.0                         libbsd/0.9.1              makedepend/1.0.5          py-nose/1.3.7           (D)    subversion/1.9.3     (D)
   c-blosc/1.12.1                            gcc/6.4.0                   (D)   libevent/2.1.8            mercurial/3.9.1           py-pip/9.0.1                   tar/1.31
   cairo/1.16.0-py3                          gcc/7.4.0                         libfabric/1.7.0           mercurial/4.4.1-py3       py-pip/10.0.1-py2              texinfo/6.5
   cmake/3.6.1                               gcc/8.1.0                         libgcrypt/1.8.1           mercurial/4.4.1     (D)   py-pip/10.0.1-py3       (D)    tmux/2.2
   cmake/3.9.2                               gcc/8.1.1                         libgd/2.2.4               nano/2.6.3                py-pygments/2.2.0-py3          valgrind/3.11.0
   cmake/3.11.3                              gcc/9.1.0                         libgpg-error/1.27         npth/1.5                  py-setuptools/25.2.0           valgrind/3.14.0      (D)
   cmake/3.13.4                              gdb/8.0                           libksba/1.3.5             numactl/2.0.11            py-setuptools/40.2.0-py2       vim/7.4.2367
   cmake/3.14.2                              gdb/8.2-py3                 (D)   libnl/3.3.0               openssl/1.0.2             py-setuptools/40.2.0-py3       xalt/0.7.5
   cmake/3.15.2                    (D)       gdbm/1.18.1                       libpciaccess/0.13.5       pango/1.41.0-py3          py-setuptools/40.4.3-py2       xalt/1.1.3
   cuda/9.1.85                               gettext/0.19.8.1                  libpng/1.6.34             papi/5.5.1                py-setuptools/40.4.3-py3 (D)   xalt/1.1.4           (L,D)
   cuda/9.2.148                              git-lfs/2.8.0                     libsodium/1.0.15          papi/5.6.0                py-virtualenv/15.0.1           xl/16.1.1-beta6
   cuda/10.1.105                             git/2.9.3                         libx11/1.6.7              papi/5.7.0          (D)   py-virtualenv/16.0.0-py2       xl/16.1.1-1
   cuda/10.1.168                   (D)       git/2.13.0                        libxcb/1.13               patchelf/0.9              py-virtualenv/16.0.0    (D)    xl/16.1.1-2
   cuda/10.1.243                             git/2.20.1                  (D)   libxdmcp/1.1.2            pcre/8.42                 python/2.7.12                  xl/16.1.1-3          (L,D)
   curl/7.63.0                               glib/2.56.3-py3                   libxext/1.3.3             perl/5.26.2               python/2.7.15                  xl/16.1.1-4
   darshan-runtime/3.1.7-hdf5pre110          gnupg/2.2.3                       libxml2/2.9.8             pgi/18.7                  python/3.5.2                   xl/16.1.1-5
   darshan-runtime/3.1.7-hdf5post110         gnuplot/5.0.1-py3                 libxpm/3.5.12             pgi/18.10                 python/3.7.0                   zeromq/4.2.5
   darshan-runtime/3.1.7           (L,D)     gnuplot/5.0.1               (D)   libxrender/0.9.10         pgi/19.1                  rdma-core/20                   zfp/0.5.0
   darshan-util/3.1.6                        go/1.11.5                         llvm/1.0-20180531         pgi/19.4            (D)   readline/6.3                   zlib/1.2.11
   darshan-util/3.1.7              (D)       gobject-introspection/1.56.1-py3  llvm/1.0-20190225   (D)   pgi/19.5                  renderproto/0.11.1             zstd/1.3.0
   diffutils/3.7                             gsl/2.5                           llvm/9.0.0-1-debug        pgi/19.7                  scons/3.0.1-py2
   emacs/25.1                                harfbuzz/2.1.3-py3                llvm/9.0.0-1              pgi/19.9                  screen/4.3.1
   essl/6.1.0-2                    (D)       hsi/5.0.2.p5                (L)   llvm/9.0.0-2-debug        pixman/0.38.0             serf/1.3.9-py2
   essl/6.2.0-20190419                       htop/2.0.2                        llvm/9.0.0-2              pkgconf/1.5.4             snappy/1.1.7

------------------------------------------------------------------------------ /sw/summit/modulefiles/core ------------------------------------------------------------------------------
   caascade/1.0      (D)    forge/19.0      forge/19.1.3            ibm-wml-ce/1.6.1-3  (D)   perf-reports/19.0.5       python/2.7.15-anaconda2-5.3.0       scorep/5.0_r14272
   caascade/1.1.beta        forge/19.0.2    forge/19.1.4       (D)  ibm-wml/1.6.1       (D)   perf-reports/19.1         python/3.6.6-anaconda3-5.3.0  (D)   scorep/6.0_r14595
   cube/4.4.3               forge/19.0.3    hip/1.5-cuda9           ibm-wml/2019.03-CE        perf-reports/19.1.2       python/3.7.0-anaconda3-5.3.0        tau/2.28.1_patched
   extrae/3.6.1             forge/19.0.4    hip/1.5-cuda10     (D)  perf-reports/18.3         perf-reports/19.1.3       scalasca/2.5                        tau/2.28.1           (D)
   extrae/3.7.0             forge/19.0.5    hpctoolkit/2019.07      perf-reports/19.0.2       perf-reports/19.1.4 (D)   scorep/4.0                    (D)   vampir/9.4.0
   extrae/3.7.1      (D)    forge/19.1      ibm-wml-ce/1.6.1-1      perf-reports/19.0.3       ppt/2.4.0-beta            scorep/4.1                          vampir/9.5.0         (D)
   forge/18.3               forge/19.1.2    ibm-wml-ce/1.6.1-2      perf-reports/19.0.4       ppt/2.4.0-beta2     (D)   scorep/5.0_r13654                   visit/3.0.0

------------------------------------------------------------------------ /sw/summit/lmod/7.7.10/rhel7.3_gnu4.8.5/lmod/lmod/modulefiles/Core ------------------------------------------------------------------------
   lmod/7.7.10    settarg/7.7.10
```

Programming Environment

- LMOD (modules)
- Compiler Toolchains
- Center-provided Software

Programming Methods for Multi-GPU Nodes

LSF Resource Scheduler

- Batch
- Interactive

Parallel Job Launch

Programming Methods

Single Thread, Multiple GPUs	Multiple Threads, Multiple GPUs	Multiple Ranks, Single GPUs	Multiple Ranks, Multiple GPUs
• Conceptually Simple • Requires additional loops • CPU can become a bottleneck • Remaining CPU cores often underutilized	• Conceptually Very Simple • Set and forget the device numbers • Relies on external Threading API • Can see improved utilization • Watch affinity	• Little to no code changes required • Re-uses existing domain decomposition • Probably already using MPI • Watch affinity	• Easily share data between peer devices • Coordinating between GPUs extremely tricky

Programming Methods for Targeting Summit's Multi-GPU Nodes (Jeff Larkin & Steve Abbott)

Programming Environment

- LMOD (modules)
- Compiler Toolchains
- Center-provided Software

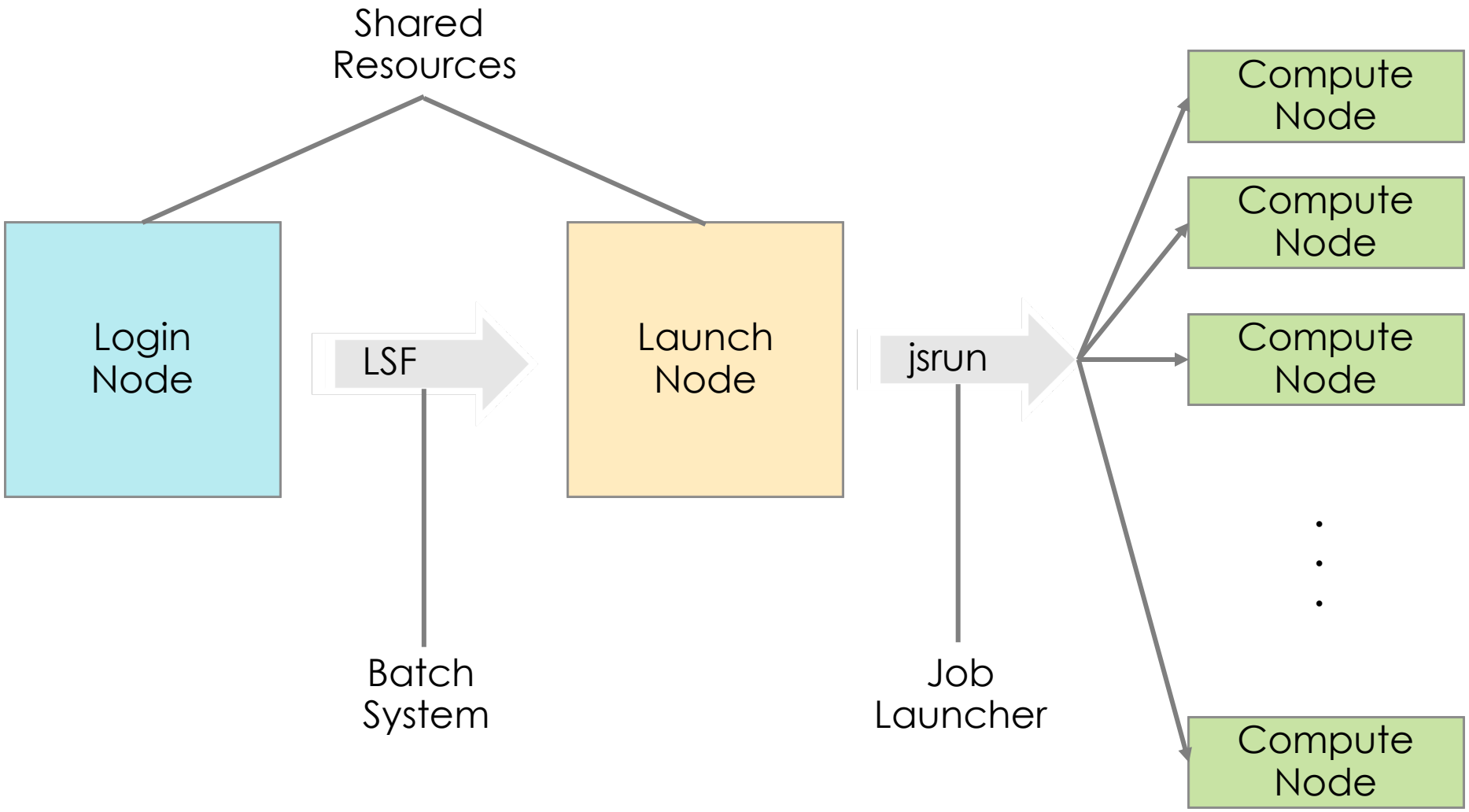
Programming Methods for Multi-GPU Nodes

LSF Resource Scheduler

- Batch
- Interactive

Parallel Job Launch

Login, Launch, & Compute Nodes



Parallel Job Execution

Batch Scheduling System

IBM Load Sharing Facility (LSF)

- Allocates resources
- Batch scheduler
- Similar functionality to PBS/MOAB or Slurm
- Allocates entire nodes (@OLCF)

Parallel Job Launcher

jsrun

- Developed by IBM for the Oak Ridge and Livermore CORAL systems
- Similar functionality to aprun, mpirun, & srun

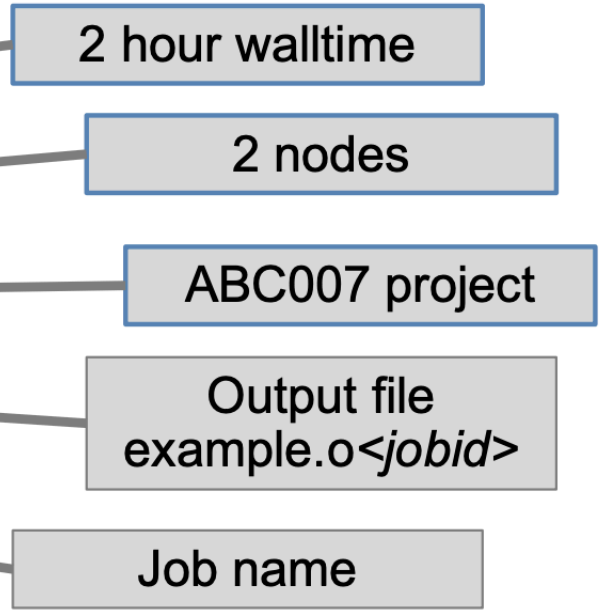
LSF Example Batch Script (non-interactive)

Batch script example

```
#!/bin/bash

#BSUB -W 2:00
#BSUB -nnodes 2
#BSUB -P abc007
#BSUB -o example.o%J
#BSUB -J example

jsrun -n2 -r1 -a1 -c1 hostname
```



Batch submission

```
summit-login1> bsub example.lsf
Job <29209> is submitted to default queue <batch>.
summit-login1>
```

Common bsub Options

Option	Example Usage	Description
-W	#BSUB –W 1:00	Requested Walltime [hours:]minutes
-nnodes	#BSUB –nnodes 1024	Number of nodes (CORAL systems)
-P	#BSUB –P ABC123	Project to which the job should be charged
-J	#BSUB –J MyJobName	Name of the job. If not specified, will be set to 'Not_Specified'
-o	#BSUB –o jobout.%J	File into which job STDOUT should be directed (%J will be replaced with the job ID number) If not specified will be set to 'JobName.%J'
-e	#BSUB –e joberr.%J	File into which job STDERR should be directed
-w	#BSUB –w ended(1234)	Place dependency on previously submitted jobID 1234
-N -B	#BSUB –N #BSUB -B	Send job report via email once job completes (N) or begins (B)
-alloc_flags	#BSUB –alloc_flags gpumps #BSUB –alloc_flags smt1	Used to request GPU Multi-Process Service (MPS) and to set SMT (Simultaneous Multithreading) levels. Setting gpumps enables NVIDIA's Multi-Process Service, which allows multiple MPI ranks to simultaneously access a GPU.

See `man bsub` for full options list

LSF Interactive Jobs

- Allows access to compute resources interactively
- Through batch system similar to batch script submission, but returns prompt on launch node
- Run multiple `jsrun` with only one queue wait, very useful for testing and debugging
- Syntax
  - Use `-Is` and the shell to be started
  - Most other batch flags valid (previous slide)
  - Add batch flags to command line
  - **`-U StudentsSC19`**

```
summit-login1> bsub -Is -P abc007 –nnodes 2 –W 2:00 $SHELL
Job <29507> is submitted to default queue <batch>.
<<Waiting for dispatch ...>>
<<Starting on batch1>>
summit-batch1 307> jsrun -n2 -r1 hostname
a01n01
a01n02
summit-batch1 308>
```

Common LSF Commands

Function	LSF Command
Submit a batch script	bsub
Monitor Queue	bjobs / jobstat
Investigate Job	bhist
Alter Queued Job	bmod
Remove Queued Job	bkill
Hold Queued Job	bstop
Release Held Job	bresume

See manual pages for more info

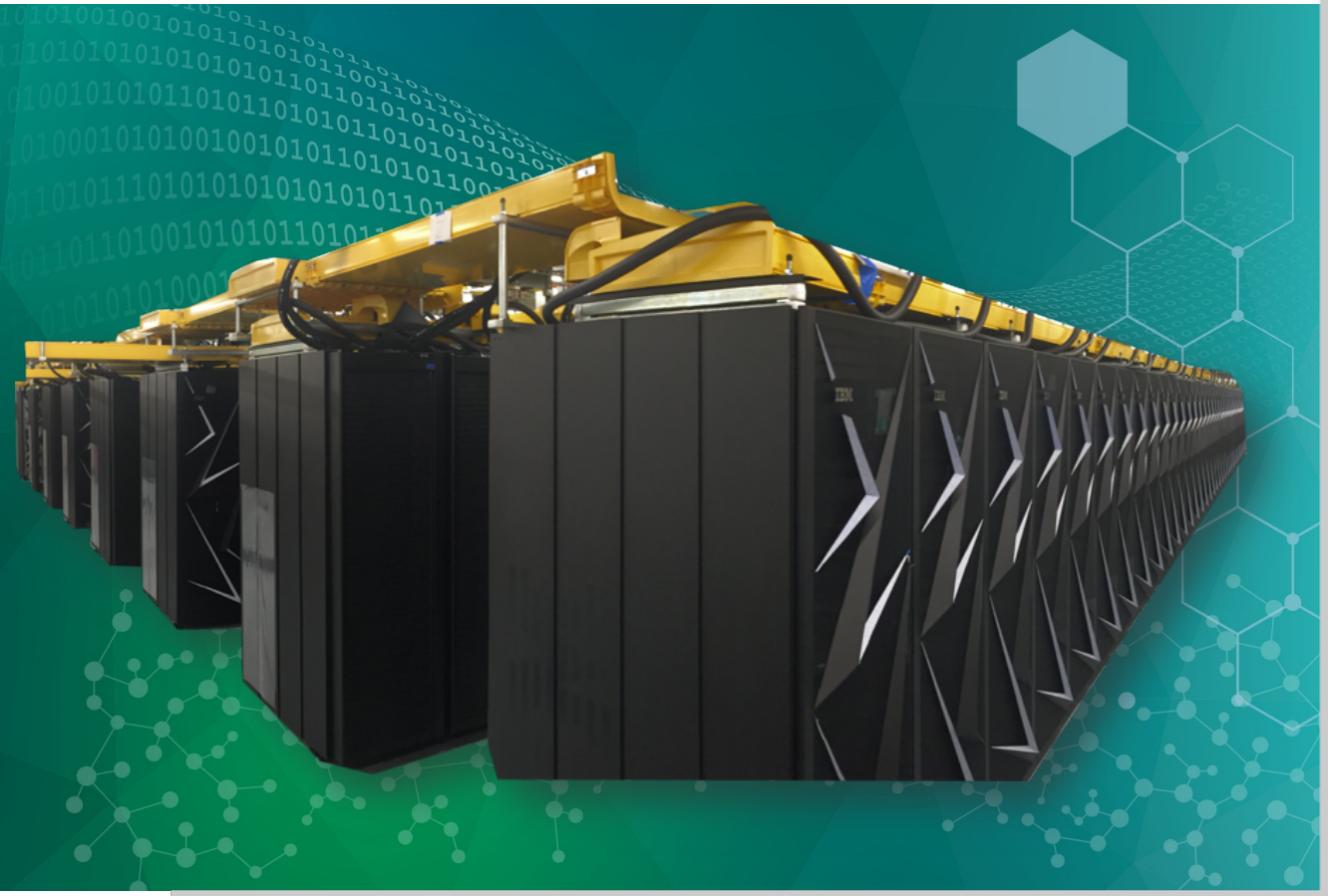
slack

https://bit.ly/2qqNelD

OAK RIDGE National Laboratory

Using the world's most powerful computer

Jack Morrison



SC19 Hands-On with Summit
Denver, CO
November 22, 2019

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

U.S. DEPARTMENT OF ENERGY