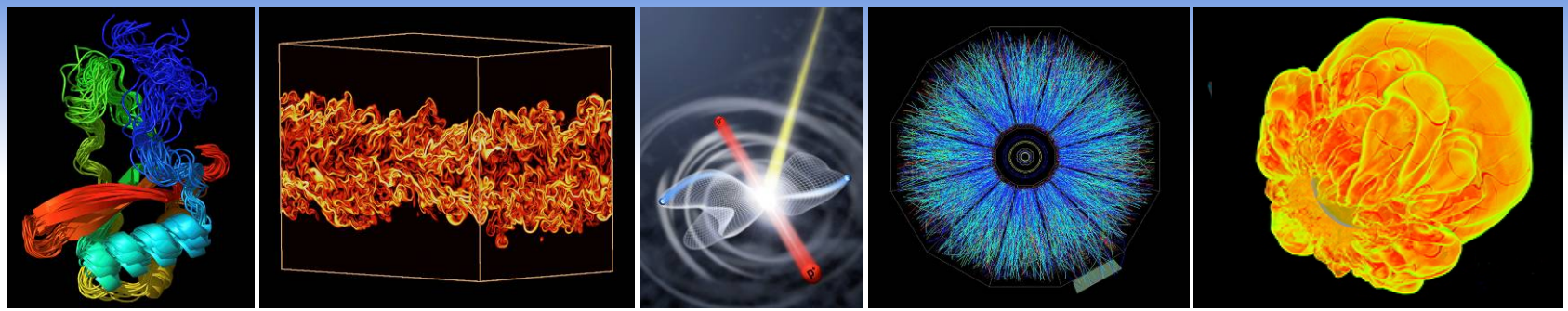

Introduction to Carver

David Turner

NERSC User Services Group

NUG Meeting, October 18, 2010

U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

Tutorial Overview

- Background
- Hardware
- Software
- Programming
- Running Jobs

Background

- Replace Bassi and Jacquard
- Hardware procurement
  - "Scalable Units"
- Two funding sources
  - Carver
    - NERSC program funds
  - Magellan
    - ARRA funds

System Overview

- **IBM iDataPlex System**
  - **14 compute racks**
    - **80 nodes/rack (1120 total compute nodes)**
    - **“Water cooled”**
  - **5 service racks**
    - **Login, I/O, and network nodes**
    - **IB switches**

Nodes

- 2 quad-core Intel Nehalem 2.67GHz processors
  - 8 cores/node
- 24GB DDR3 1333MHz memory
- 48GB DDR3 1066MHz memory
  - 6 login nodes

High Speed Interconnect

- 4X QDR InfiniBand
  - 32 Gbits/sec
- 4 Voltaire switches
- Local fat-trees connected via 2-D mesh

File Systems

- Most nodes “stateless”
- Global homes
  - 40GB quota
- Global scratch
  - 20TB quota
- Project directories
  - 1TB quota

System Split

- Carver
  - 400 nodes
    - 320 with 24GB
    - 80 with 48GB
- Magellan
  - 720 nodes
    - 560 with 24GB
    - 160 with 48GB

Software

- **Scientific Linux**
  - Red Hat distribution
- **PBS-based batch system**
  - Moab scheduler
  - Torque resource manager
- **Open MPI**
  - Evolved from LAM (not MPICH)

Software – cont.

- Compilers
  - PGI, Intel, GCC
- Libraries
  - MKL, NAG, PETSc
  - hdf5, netcdf
  - NCAR graphics

Software – cont.

- Tools
  - totalview, ddt
  - papi, ipm
- Applications
  - gaussian, gamess, molpro, namd, nwchem, qchem, vasp, wien2k
  - IDL, AVS-Express, paraview, VisIt
  - matlab, mathematica, R

Modules

```
% module list
  Currently Loaded Modulefiles:
  1) pgi/10.8          2) openmpi/1.4.2
% module avail
% module load foo
% module unload foo
% module show foo
% module swap foo/1.2.3 foo/1.2.4
```

Modules – cont.

- Intel

```
% module unload pgi openmpi
% module load intel openmpi-intel
```

- GCC

```
% module unload pgi openmpi
% module load gcc openmpi-gcc
```

Programming

- MPI
  - User wrappers
    - mpif90/mpicc/mpiCC

```
% mpif90 mycode.f90
```

- OpenMP
  - PGI

```
% pgf90 -mp mycode.f90
```

  - Intel

```
% ifc -openmp mycode.f90
```

Interactive Jobs

- **Interactive serial**

```
% ./a.out
```

  - **Process limits (ulimits)**
    - **30 minutes, 2GB**
- **Interactive parallel**

```
% qsub -I ...
  qsub: waiting for job
  123456.cvrsvc09-ib to start
  qsub: job 123456.cvrsvc09-ib ready
```

  - **Land on compute node**

Batch Jobs

- **Batch**

```
% qsub myjob.pbs
123456.cvrsvc09-ib
```

Batch Queues

Submit Queue	Execute Queue	Number of Nodes	Number of Cores	Max Wall Time	Relative Priority	User Run Limit
interactive	interactive	1 – 8	1 – 64	30 mins	1	2
debug	debug	1 – 32	1 – 256	30 mins	2	2
regular	reg_short	1 – 16	1 – 128	4 hrs	3	10
	reg_small	1 – 16	1 – 128	48 hrs	3	8
	reg_med	17 – 32	129 – 256	36 hrs	3	7
	reg_big	33 – 64	257 – 512	24 hrs	3	5
	reg_long	1 – 4	1 – 32	168 hrs	3	1
	reg_xlong	1 – 4	1 – 32	504 hrs	3	1
low	low	1 – 32	1 – 256	12 hrs	4	7

Batch Policies

- System-wide user run limit 15
- System-wide user idle limit 30
- Maximum 4 running jobs in reg_long and reg_xlong
- 8 nodes reserved for interactive/debug 05:00 – 18:00

No production in debug!

Batch Scripts

```
#PBS -q debug
#PBS -l nodes=16:ppn=8
#PBS -l walltime=00:10:00
#PBS -N my_job
#PBS -j oe
#PBS -V

cd $PBS_O_WORKDIR
mpirun -np 128 ./my_executable
```

Batch Options

```
#PBS -S shell
#PBS -m [a|b|e|n]
#PBS -A repository
#PBS -l nodes=16:ppn=8:bigmem

qsub -A mpccc -l nodes=4:ppn=8 myjob.pbs
```

Memory Considerations

Process Memory Limits		
Type of Node	Soft Limit	Hard Limit
Login node	2GB	2GB
24GB compute	2.5GB	20GB
48GB compute	5.5GB	44GB

ppn	24GB compute	48GB compute
1	pvmem=20GB	pvmem=44GB
2	pvmem=10GB	pvmem=22GB
4	pvmem=5GB	pvmem=11GB

```
#PBS -l nodes=4:ppn=1:pvmem=20GB
```

mpirun options

```
-v
-hostfile name
-np numprocs
-bycore
-bynode
-output-filename name
-stdin rank
-tag-output
-timestamp-output
```

Additional Information

http://www.nersc.gov/nusers/systems/carver/

http://www.pgroup.com/resources/docs.htm

http://www.open-mpi.org/faq/

http://help.nersc.gov/

consult@nersc.gov

1-800-66-NERSC