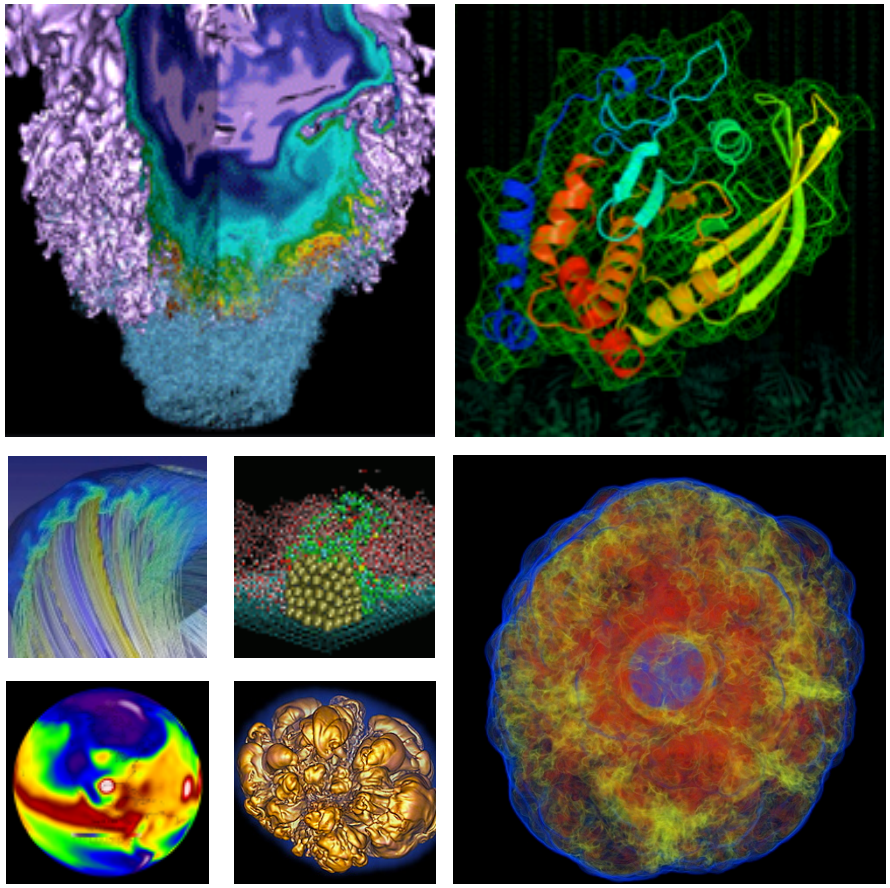
Getting Started at NERSC: Running Jobs



Lisa Gerhardt

NERSC User Services

September 10, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Jobs at NERSC

- **Most jobs are parallel, using 10s to 100,000+ cores**
- **Production runs execute in batch mode**
- **Interactive and debug jobs are supported for up to 30 minutes**
- **Typically run times are a few to 10s of hours.**
  - Each machine has different limits.
  - Limits are necessary because of MTBF and the need to accommodate 5,500 users' jobs
- **Also a number of 'serial' jobs**
  - Typically some kind of pleasantly parallel simulation

What system is best for your job?

Large-Scale Computing Systems

Hopper (NERSC-6): Cray XE6

- 6,384 compute nodes, 153,216 cores
- 144 Tflop/s on applications; 1.3 Pflop/s peak

Edison (NERSC-7): Cray XC30 Phase II

- 5,200 compute nodes, 124,800 cores
- 236 Tflop/s on applications, 2.39 Pflop/s peak

Midrange

Thousands of cores

Carver

- IBM iDataplex cluster, 1,202 nodes
- 9984 cores; 106 TF

PDSF (HEP/NP)

- ~2K core cluster

GenePool (JGI)

- ~5K core cluster
- 2.1 PB Isilon File System

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

- 12.7 PB capacity
- 150 GB/s of bandwidth

HPSS Archival Storage

- 240 PB capacity
- 5 Tape libraries
- 200 TB disk cache

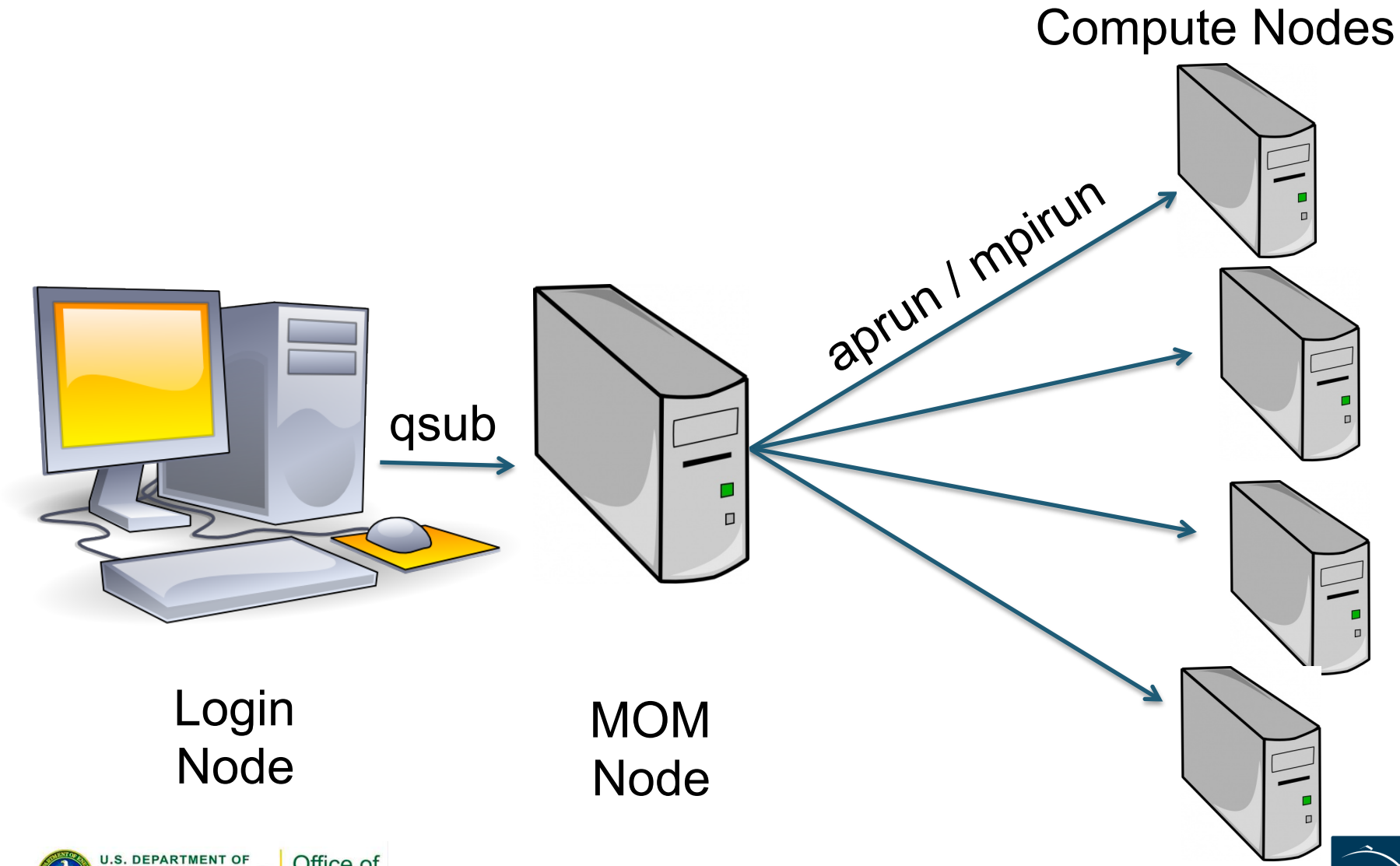
Analytics & Testbeds

Dirac 48 Fermi GPU nodes

Login Nodes and Compute Nodes

- Each supercomputer has 3 types of nodes that you will use
  - Login nodes
  - Compute nodes
  - Job launcher or "MOM" nodes
- Login nodes
  - Edit files, compile codes, run UNIX commands
  - Submit batch jobs
  - Run short, small utilities and applications
- Compute nodes
  - Execute your application; dedicated to your job
  - No direct login access
- Job launcher or "MOM" nodes
  - Execute your batch script commands
  - Carver: "head" compute node; Hopper / Edison: shared "service" node; not a compute node

Launching Parallel Jobs



Launching Parallel Jobs

- A "job launcher" executes your code
  - Distributes your executables to all your nodes
  - Starts concurrent execution of N instances of your program
  - Manages execution of your application
  - On Hopper / Edison: the job launcher is called "aprun"
  - On Carver: "mpirun"
- Only the job launcher can start your job on compute nodes
- You can't run the job launcher from login nodes

Submitting Batch Jobs

- To run a job on the compute nodes you must write a "batch script" that contains
  - Batch directives to allow the system to schedule your job
  - An `aprun` or `mpirun` command that launches your parallel executable
- Submit the job to the queuing system with the `qsub` command
  - `%qsub my_batch_script`

Hopper - Cray XE6

- 153,216 cores, 6,384 nodes
- "Gemini" interconnect
- 2 12-core AMD 'MagnyCours' 2.1 GHz processors per node
- 24 processor cores per node
- 32 GB of memory per node (384 "fat" nodes with 64 GB)
- 216 TB of aggregate memory
- 1.3 GB memory / core (2.7 GB / core on "fat" nodes) for applications
- /scratch disk quota of 5 TB
- 2 PB of /scratch disk
- Choice of full Linux operating system or optimized Linux OS (Cray Linux)
- PGI, Cray, Pathscale, GNU compilers

Hopper's Nodes

- **Login Nodes**
  - Hopper's login nodes run a full Linux operating system and provide support services for the system. When you connect to Hopper with SSH, you land on the login nodes. These nodes are shared by many users; please do not run applications on the login nodes.
- **Compute Nodes**
  - The 6,384 compute nodes are dedicated to running scientific applications. A job is given exclusive access to each node it requests for the entirety of the job's run time. Since Hopper has 24 cores on each node, the minimum number of cores per job is 24.
- **Job Host (MOM) Nodes**
  - MOM nodes are servers that execute batch job commands. These nodes are shared by many users and thus are not intended for compute- or memory-intensive applications.

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 ./my_executable
```

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 ./my_executable
```

Job directives: instructions for the batch system

- Submission queue
- How many nodes to reserve for your job
- How long to reserve those nodes
- Optional: what to name STDOUT files, what account to charge, whether to notify you by email when your job finishes, etc.

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 ./my_executable
```

Change from home directory to job submission directory

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 ./my_executable
```

Launches parallel executable on the compute nodes.
Carries over (partial) login environment and controls how your executable accesses the memory on each processor.

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 ./my_executable
```

mppwidth is number of compute cores requested (/24 for number of nodes). Must match with number of tasks requested (-n)

Sample Hopper Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=192
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 –N 12 ./my_executable
```

-N = number of tasks per node
Might do this to get more memory / task

Hybrid OpenMP/MPI

```
#PBS –q regular
#PBS -l mppwidth=96
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
export OMP_NUM_THREADS=6
aprun -n 16 -d 6 -N 4 -S 1 ./hybrid.x
```

16 tasks (-n), 4 on each node (-N), 6 OpenMP threads / task (-d), separate tasks onto individual NUMA nodes (-S)
Many more examples on www.nersc.gov

Edison - Cray XC30

- Newest NERSC peta-flop machine
- Phase II just started August 23
- 124,800 cores, 5,200 nodes
- "Aries" interconnect
- 4 12-core Intel 'IvyBridge' 2.4 GHz processors per node
- 24 processor cores per node
- 64 GB of memory per node
- 333 TB of aggregate memory
- 2.7 GB memory / core for applications
- 6.4 PB of /scratch disk
- Choice of full Linux operating system or optimized Linux OS (Cray Linux)
- PGI, Cray, Pathscale, GNU compilers
- Still in testing

Sample Edison Batch Script - MPI

```
#PBS -q debug
#PBS -l mppwidth=48
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
aprun -n 96 -j 2 ./my_executable
```

Turn on hyperthreading (-j)

Carver - IBM iDataPlex

- 9,984 compute cores
- 1,202 compute nodes
- 2 quad-core Intel Nehalem 2.67 GHz processors per node
- 8 processor cores per node
- 24 GB of memory per node (48 GB on 80 "fat" nodes)
- 2.5 GB / core for applications (5.5 GB / core on "fat" nodes)
- InfiniBand 4X QDR

- NERSC global /scratch directory quota of 20 TB
- Full Linux operating system
- PGI, GNU, Intel compilers

Sample Carver Batch Script - MPI

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

cd $PBS_O_WORKDIR
mpirun –n 128 ./myexecutable
```

MPI tasks

Sample Carver Batch Script - OpenMP

```
#PBS -q debug
#PBS –l nodes=1:ppn=8
#PBS -l walltime=00:10:00
#PBS -N my_job

export OMP_NUM_THREADS=8
cd $PBS_O_WORKDIR
./myexecutable
```

Number of threads

Interactive Parallel Jobs

- You can run small parallel jobs interactively for up to 30 minutes (ex. is for Hopper)

```
% qsub –I –q interactive –l mppwidth=48
[wait for job to start]
% cd $PBS_O_WORKDIR
% aprun –n 48 ./mycode.x
```

Serial Jobs

- Carver has a special queue for running serial jobs
  - A single process running on a single node
  - Each serial node can run up to 12 jobs from different users

```
#PBS -q serial
#PBS -l walltime=00:10:00
#PBS -N my_job

cd $PBS_O_WORKDIR
./myexecutable
```

Monitoring Your Job

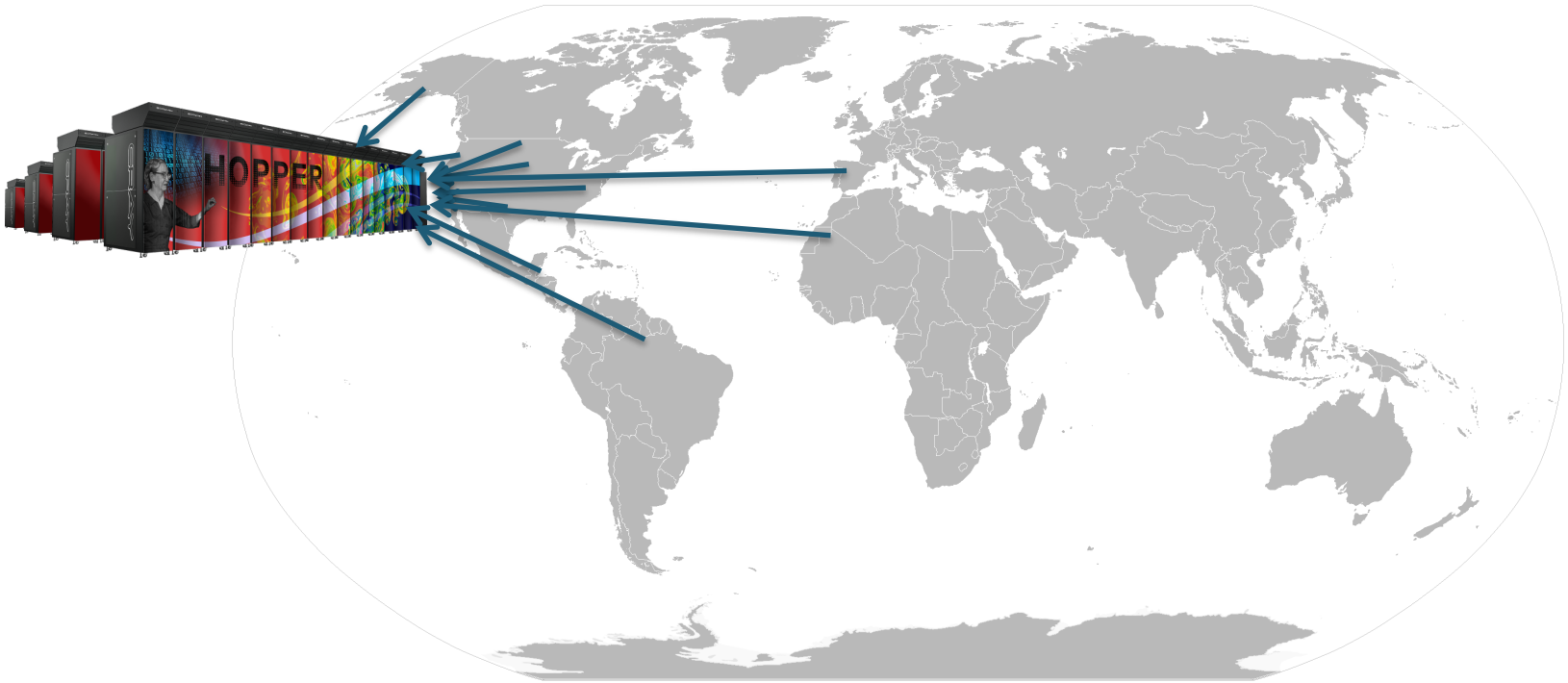
- Once your job is submitted, it will start when resources are available
- Monitor it with
  - qstat –a
  - qstat –u username
  - showq
  - qs
  - NERSC web site “Queue Look”

    `https://www.nersc.gov/users/live-status/global-queue-look/`
  - NERSC web site “Completed Jobs”

    `https://www.nersc.gov/users/job-logs-and-analytics/completed-jobs/`

I qsub'd my job, but it's not running!

- **You are not alone!**



How do we handle the demand?

That's what the batch queue is for!

source: itu.dk

Your jobs will wait until the resources are available for them to run

Your job's place in the queue is a mix of time and priority, so line jumping is allowed, but it may cost more

Job Limits

There are per user, per machine job limits. See the NERSC web site for details. Here are the limits on Hopper as of Sep. 3, 2013.

Specify these queues (with #PBS –q queue_name)

NEVER these!

Submit Queue	Execution Queue[1]	Nodes	Processors	Max Wallclock	Relative Priority	Run Limit[2]	Queued Limit[3]	Queue Charge Factor
interactive	interactive	1-256	1-6,144	30 mins	2	1	1	1
debug	debug	1-512	1-12,288	30 mins	3	1	1	1
regular	reg_1hour	1-256	1-6,144	1 hr	5	8	8	1
	reg_short	1-682	1-16,368	6 hrs	5	16	16	1
	reg_long	1-64	1-1,536	96 hrs	5	2	2	1
	reg_small	1-682	1-16,368	48 hrs	5	16	16	1
	reg_med	683-2,048	16,369-49,152	36 hrs	4	4	4	0.6
	reg_big	2,049-4,096	49,153-98,304	36 hrs	4	2	2	0.6
reg_xbig[4]	4,097-6,100	98,305-146,400	12 hrs	1	2	2	0.6	
--	bigmem[5]	1-384	1-9,216	24 hrs	5	1	1	1
low	low	1-683	1-16,392	24 hrs	7	6	6	0.5
thruput[6]	thruput	1-2	1-48	168 hrs	5	250	500	1.0
scavenger[7]	scavenger	1-683	1-16,392	6 hrs	8	2	2	0
premium	premium	1-2,048	1-49,152	12 hrs	3	1	1	2
ccm_queue[8]	ccm_queue	1-682	1-16,368	96 hrs	4	16	16	1
ccm_int[9]	ccm_int	1-512	1-12,288	30 mins	3	1	1	1
iotask[9]	iotask	1-682	1-16,368	12 hrs	6	1	-	1
xfer[10]	xfer	--	--	12 hrs	--	4	3	0

Tips for jobs

- Submit shorter jobs.
  - Checkpoint if possible
  - Take advantage of 'backfill' opportunities
  - Run jobs just before maintenance
- Make sure the wall clock time you request is accurate. As noted above, shorter jobs are easier to schedule. Many users unnecessarily enter the largest wall clock time possible as a default.

Average Queue Wait Time

	Hours Requested																																																
Nodes	<1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	48+
1	1.6	5.3	4.1	5.6	5.6	8.5	13	38	58	24	9.8	17	18	110	67	15	31	16	42	6.2	17	10	26	26	29	11	19	20	57	26	42	13	12	27	34	21	46		18	26	26	17	20		28	20	21	24	43	29
2	1.7	7.5	8.0	9.6	22	15	25	8.6	21	14	15	13	23	18	22	23	15	14	30	34	34	31	25	30	41	21	35	33	20	19	25		41		36	29	47	40	0.1	33	45		35	17	58	31	33	33	58	76
3	1.4	17	4.8	5.1	4.4	8.0	48	22	11	-2,590.3	15	5.0	19	28	29	19	26	14	23	23	51		35	37	45	66	89		32	40	51		42			50	112	44	31		35				8.3	22	39	37	93	24
4	3.6	11	6.3	14	10.0	12	14	99	16	27	27	29	38	16	24	26	39	20	71	28	48		14	29	52	60	44		35		112		54			69	54		25	65	148							93	65	44
5	2.2	1.1	7.2	6.0	5.9	9.7	20	13	16	6.5	49	8.6	19	13	38	25	12	28	35	33	58	43	25	50	76	61				23	43		47			43	49		44		44		45						25	74	32
6	1.1	6.0	3.5	10	6.4	13	11	13	41	11	40	21	19	24	21	23	34	445	32		43		35	84	80		27		78		58	34	56	49	28	53	59		42		53						56	93	102	34
7	0.4	5.4	2.5	7.5	9.9	6.5	14	3.2	15	7.8	38	31	24								47		83	16	57						37		47			55	19								104					64	35
8	1.8	16	8.4	36	9.8	16	19	17	16	16	59	41	22	4.7	27	36	37		33	57	64			73	68	18	18		56		25		75	44	28		56								86	17		77	67	35
9	0.4	2.1	2.5	5.8	8.7	15	11	12	14		32	18	18			36	23				130				45						236						63				51				90	49	72	59	104	16
10	1.8	7.2	8.6	4.8	13	11	12	17	21	18	28	26	23	50	81	72	40		13	25	80		72	50	110	60					62					36	47				55				113	91		57	87	76
11	10	26	7.0	6.1	16	13	30	15	15	11	22	27	31	38	27	31	25	39	45	61	75			125	56		41				59					47	77	41			52					48	58	81	102	45
12	2.4	9.6	10	5.3	23	47	96	16	36	31	254	18	32	9.5		45	40		22	53	60			113	57	61			34		36						138				38	63			68	37	63	47	59	42
13	0.4	1.4	7.7	5.0	10	19	13	6.4	8.8	10	80		18								48				57						50						15				64						70		66	
14	1.6	1.1	18	12	8.7	22	17	14	27	27	27		29						105		97			37	54						38						37						53		151				109	120
15	12	6.7	15	16	7.7	9.9	62	81	14	47	53	37	25	47	86	38	17	38			72			65	204	54		48			44					11	93									51			77	87
16	1.6	1.9	10	27	11	22	39	18	18	30	27	28	22		124	42	40	43	29	66	58	77		65	65	79	33	45		96	68	140		53		120	53				105							80	55	32
17-19	1.3	26	19	5.3	10	8.0	20	13	46	21	30	11	43		104	27	30		75		76			144	112	132			152	63	36		116			64	98		49		69				67				77	53
20-23	8.9	8.3	9.3	19	20	19	23	24	29	46	32	25	29	68	53	57	154			67	57	76	41	61	66	59				39	75		202			59	61				65		27				44	41	75	36
24-31	2.6	2.6	4.8	11	15	24	48	15	46	30	19	11	25	166	47	43	47	69	57	67	53		28	66	125	78			86	39	66		132	41		102	75				94					100	40	40	87	104
32-47	1.5	9.2	9.5	26	18	27	58	18	36	51	45	28	37	82	51	54	49	65	86	44	60	128	47	66	101	46				21	89		134	43		67	67		84	49	88	47				62	65	79	88	55
48-63	6.0	12	9.3	15	18	18	32	54	83	43	31	65	40		47	64	32		74	74	52			71	114	62			86		64		69			108	72		34		58					106	95	77	80	98
64-127	1.6	23	20	21	31	29	36	45	53	33	39	43	52	77	67	59	73	60	86	61	72	45	86	94	78	63	61		105	119	59		64			127	64			42	65					59	58		70	101
128-255	4.8	-178.1	21	38	40	50	54	45	77	51	33	149	48	66	130	58	73	73	78		63		55	79	101					41	73		34			175	64				45	50			98		109	102	91	63
256-511	7.8	41	47	38	83	47	131	30	67	51	56	32	44	43	71	51	79	2.5	105	58	95			98	86	349				77	95						54				68								112	105	
512-1023	12	53	100	91	48	48	65	30	50	94	44	60	56	62	28	66	54	44	36	41	53		82	32	42	16		32	42		87		144		36	82	76												128	
1024-1535	21	31	32	26	301	29	77	35	56	55	90	39	89	39	41		33	65	93		45				45						72					139	208													
1536-2047	24	43	21	11	35	32	66		48		47		35				39		31		142		19		64												110													
2048-3071	30	32	59	41	32	45	50	49	45		106		21	46		70	203	78	144	101	33			67	53	31			56							50														
3072-4095	25	57	102							130			193						72						31																									
4096-6143	133	75	75	50	50	0.0	129	60		14																																								
6144-9531	0.0	55		29																																														

How Your Jobs Are Charged

- **Your repository account is charged for each node your job was allocated for the entire duration of your job.**
  - The minimum allocatable unit is a node. Hopper and Edison have 24 cores/node, so your minimum charge is 24*`walltime`.

MPP hours = (# nodes) * (# cores / node) * (walltime) * (queue charge factors)

  - mppwidth = 96 for 1 hour. MPP hours = (4) * (24) * (1 hour) *(1) MPP hours charged = 96 hours
  - Serial jobs on Carver are charged differently.
- **If you have access to multiple repos, pick which one to charge in your batch script**
  - `#PBS -A` *`repo_name`*

Charge Factors & Discounts

- **Each machine has a "machine charge factor" (mcf) that multiplies the "raw hours" used**
  - Hopper has mcf=1.0
  - Carver has mcf=1.5
- **Queues have "priority charge factors" (pcf) and corresponding relative scheduling priorities**
  - Premium pcf=2.0
  - Low pcf=0.5
  - Everything else pcf=1.0
- **On Hopper only:**
  - reg_med, reg_big, reg_xbig jobs get a 40% discount

More Information

NERSC Web pages

Hopper:

http://www.nersc.gov/users/computational-systems/hopper/running-jobs/

Edison:

http://www.nersc.gov/users/computational-systems/edison/running-jobs/

Carver:

http://www.nersc.gov/users/computational-systems/carver/running-jobs/

Contact NERSC Consulting:

- Toll-free 800-666-3772
- 510-486-8611, #3
- Email *consult@nersc.gov*.

NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB