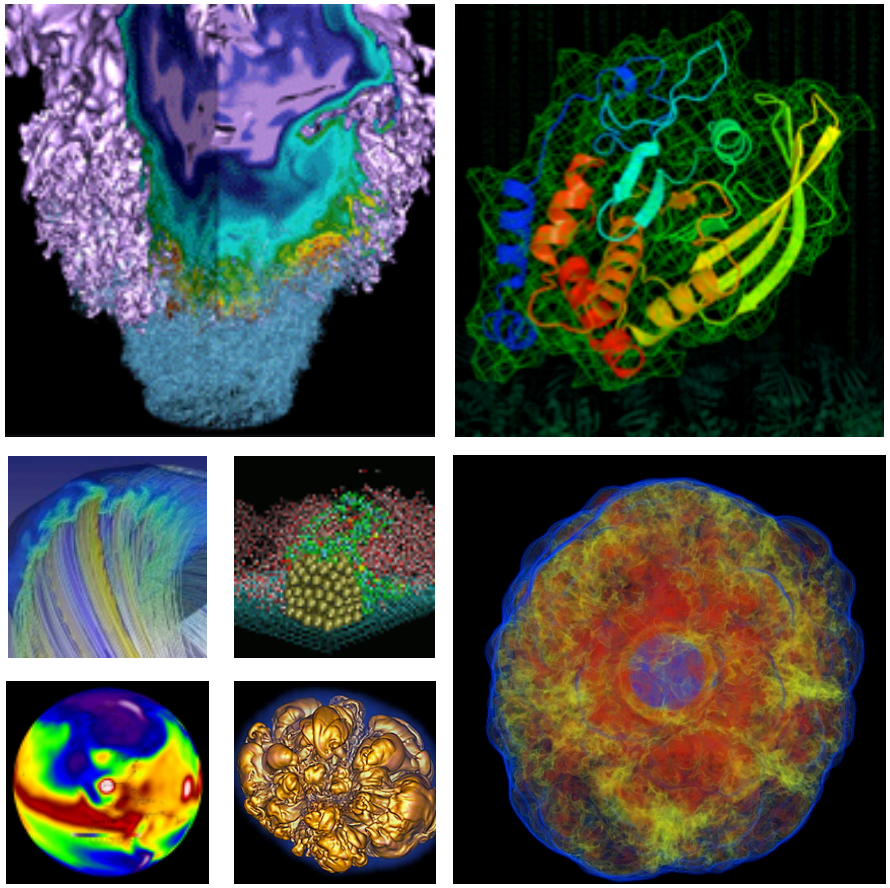
September 2016 NERSC Update



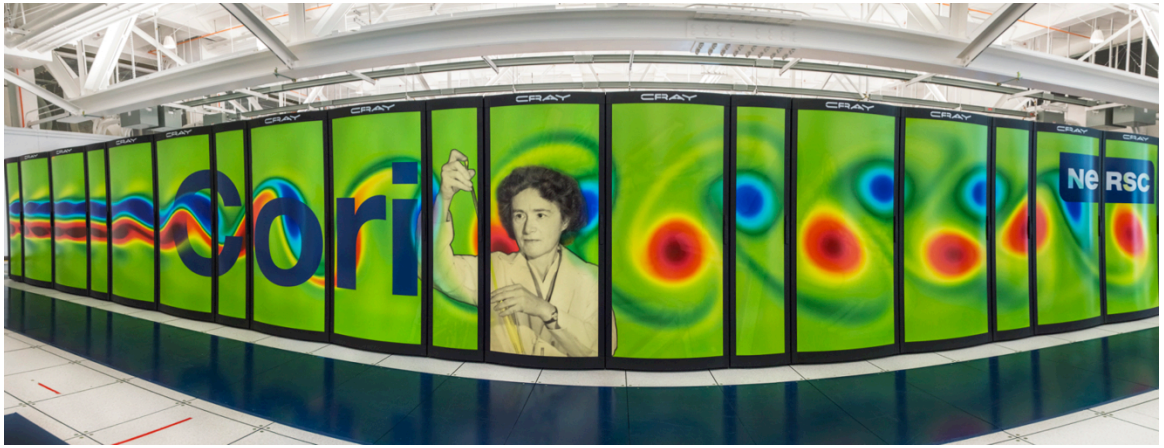
Richard Gerber

High Performance Computing Department Head
Senior Science Advisor

September 7, 2016

Compute Systems

Edison
Cray XC 30
Intel Xeon (Ivy Bridge)
~2 B NERSC Hours



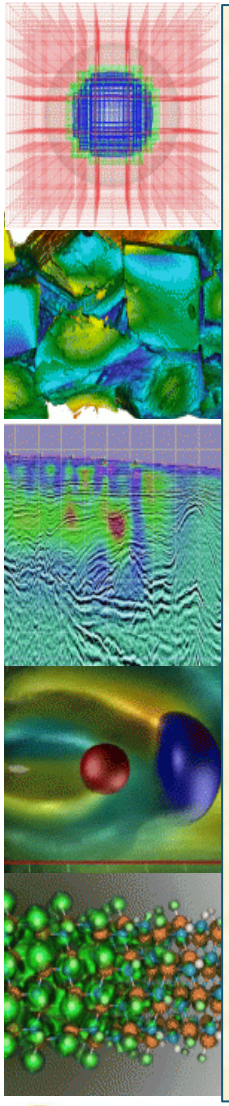
Cori Phase 1
Cray XC 40
Intel Xeon (Haswell)
~1 B NERSC Hours

Cori Phase 2
Cray XC 40
Intel Xeon Phi (KNL)
~6 B NERSC Hours

The NERSC-8 System: Cori

- **Phase 1 "Haswell" system has been running productively throughout 2016**
  - Two week outage in June to upgrade OS to Cray "Rhine/Redwood" needed for Cori Phase 2 operation
- **Phase 2 partition with 9,300 Intel "Knights Landing" compute nodes has arrived and is powered up in Berkeley**
  - Cray has successfully run various checkout and stress checks
  - Staff is evaluating readiness for integration with Phase 1
- **We are planning to integrate Phase 1 and 2 will into a single system**
  - Go/No Go decision Thursday
  - Priorities: (1) Keep Phase 1 up enough to deliver on 2.4 hours committed to DOE production and (2) give users access to Phase 2 for early science in October
  - If all is OK, downtime for Phase 1/2 integration may start as early as Friday 9/9/2016
  - Planning for up to 6 weeks of downtime on Cori Phase 1

Application Readiness: NESAP



ASCR

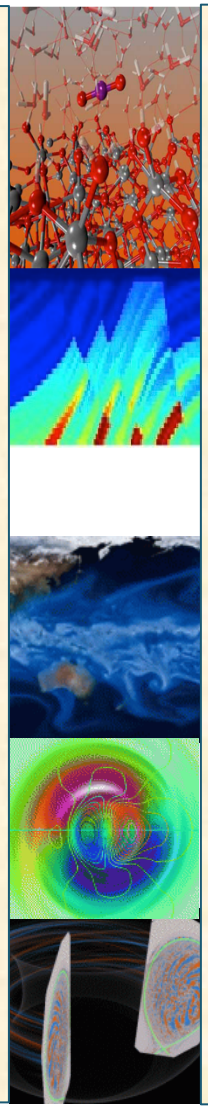
Almgren (LBNL)	**BoxLib**
Trebotich (LBNL)	**Chombo-crunch**

HEP

Vay (LBNL)	**WARP & IMPACT**
Toussaint(Arizona)	**MILC**
Habib (ANL)	**HACC**

NP

Maris (Iowa St.)	**MFDn**
Joo (JLAB)	**Chroma**
Christ/Karsch (Columbia/BNL)	**DWF/HISQ**



BES

Kent (ORNL)	**Quantum Espresso**
Deslippe (LBNL)	**BerkeleyGW**
Chelikowsky (UT)	**PARSEC**
Bylaska (PNNL)	**NWChem**
Newman (LBNL)	**EMGeo**

BER

Smith (ORNL)	**Gromacs**
Yelick (LBNL)	**Meraculous**
Ringler (LANL)	**MPAS-O**
Johansen (LBNL)	**ACME**
Dennis (NCAR)	**CESM**

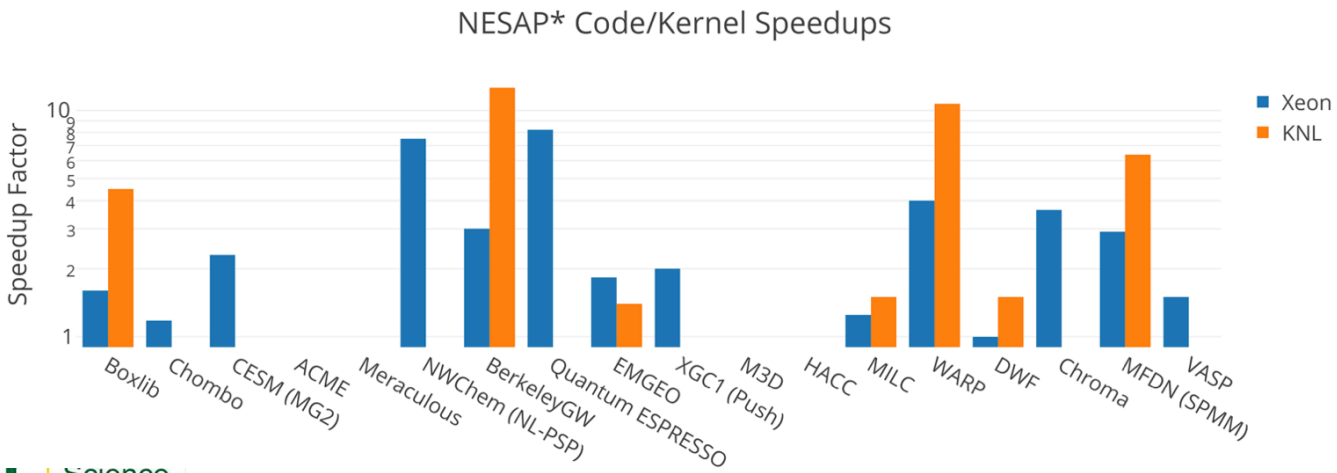
FES

Jardin (PPPL)	**M3D**
Chang (PPPL)	**XGC1**

NESAP Code Status (Work in Progress)

	GFLOP/s KNL	Speedup HBM / DDR	Speedup KNL / Haswell
Chroma (QPhiX)	388 (SP)	4	2.71
MILC	117.4	3.8	2.68
CESM (HOMME)			1.8
MFDN (SPMM)	109.1	3.6	1.62
BGW Sigma	279	1.8	1.61
HACC	1200		1.41
EMGEO (SPMV)	181.0	4.2	1.16

	GFLOP/s KNL	Speedup HBM / DDR	Speedup KNL / Haswell
DWF	600 (SP)		0.95
WARP	60.4	1.2	1.0
Meraculous			0.75
Boxlib		1.13	1.1
Quantum ESPRESSO			1
XGC1 (Push-E)	8.2	0.82	0.2-0.5
Chombo			0.5-1.5

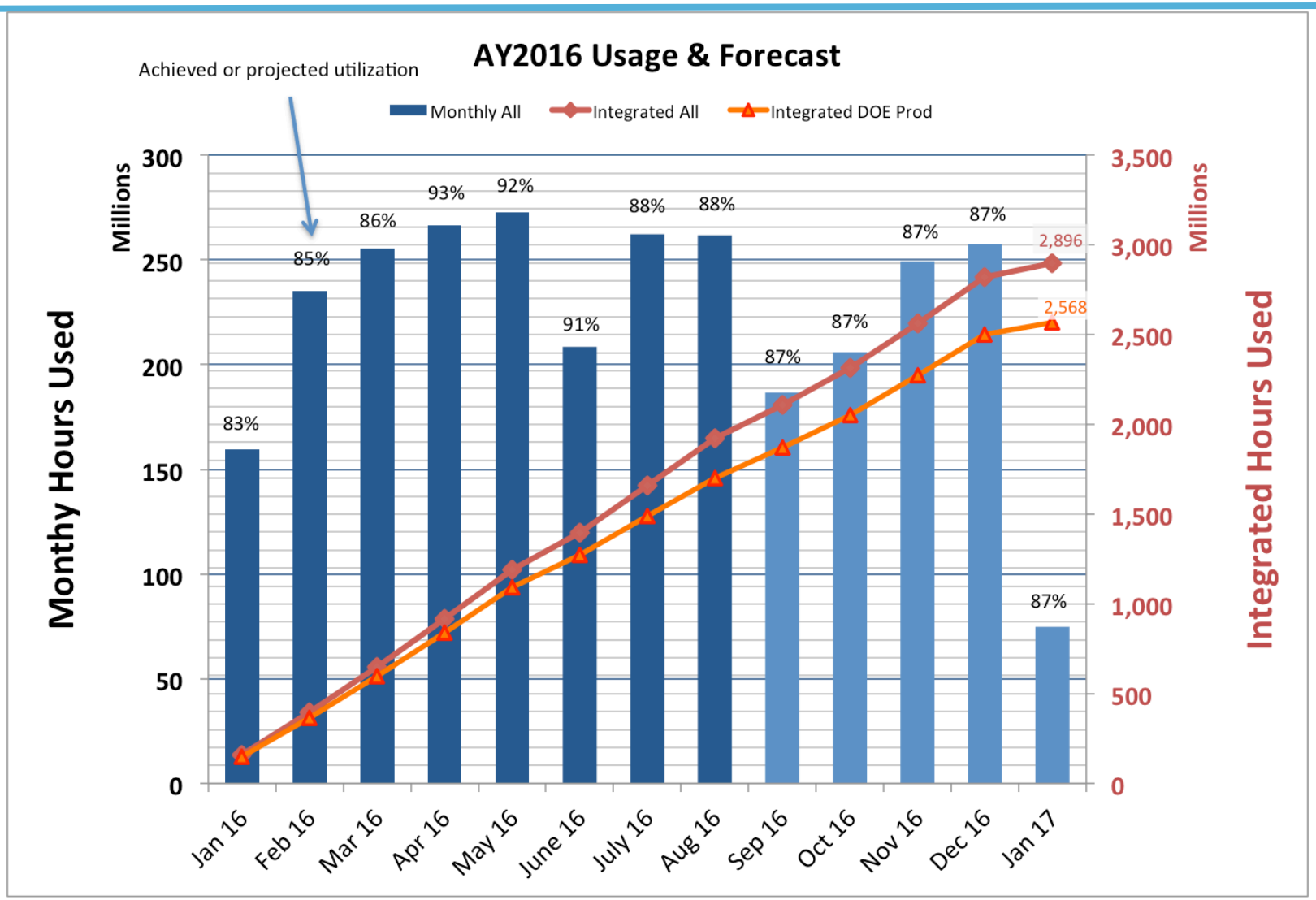


*Speedups from direct/indirect NESAP efforts as well as coordinated activity in NESAP timeframe

Additional Hours from Cori Phase 2 Early User Access

- **When Cori Phase 2 becomes usable, NESAP teams will get excusive early access for 4-6 weeks**
- **Then all users will be able to use a small number of nodes to test and optimize their codes for Xeon Phi**
- **When teams can demonstrate readiness for the Xeon Phi architecture, they will get full access**
  - We do not want unprepared users to have a bad experience on Cori Phase 2 or use inefficiently
  - Questionnaire / worksheet is being prepared
- **As of today, we anticipate giving all users access to the full Cori system (Phase 1 + 2) when production computing begins in July 2017**
  - We are not planning to allocate "Xeon Hours" and "Xeon Phi Hours"
  - We are hoping users will run where it makes sense for them and PMs will be given enough data to make informed allocation decisions

2016 Usage (including planned outages)



Usage and Forecast Overview 2016

Allocation Pool	Allocated (M Hrs)	Used 8/31/16	Remaining Commitment to DOE
DOE Production	2,477*	1,702	775
ALCC	223*	131	92
DDR	142 (158 unallocated)	54	88
TOTAL	2,814	1,228	955

Estimated for all of AY2016, considering planned outages and 87% overall availability at other times:

NERSC is estimated to deliver 974 M more hours in AY2016
NERSC has a remaining commitment to DOE of 955 M Hours

2,896 (2,814 June est.) M Hours Total Will Be Used in AY2016

* ALCC gave 77 M hours to DOE Production

Used vs. Charged & Hours Remaining

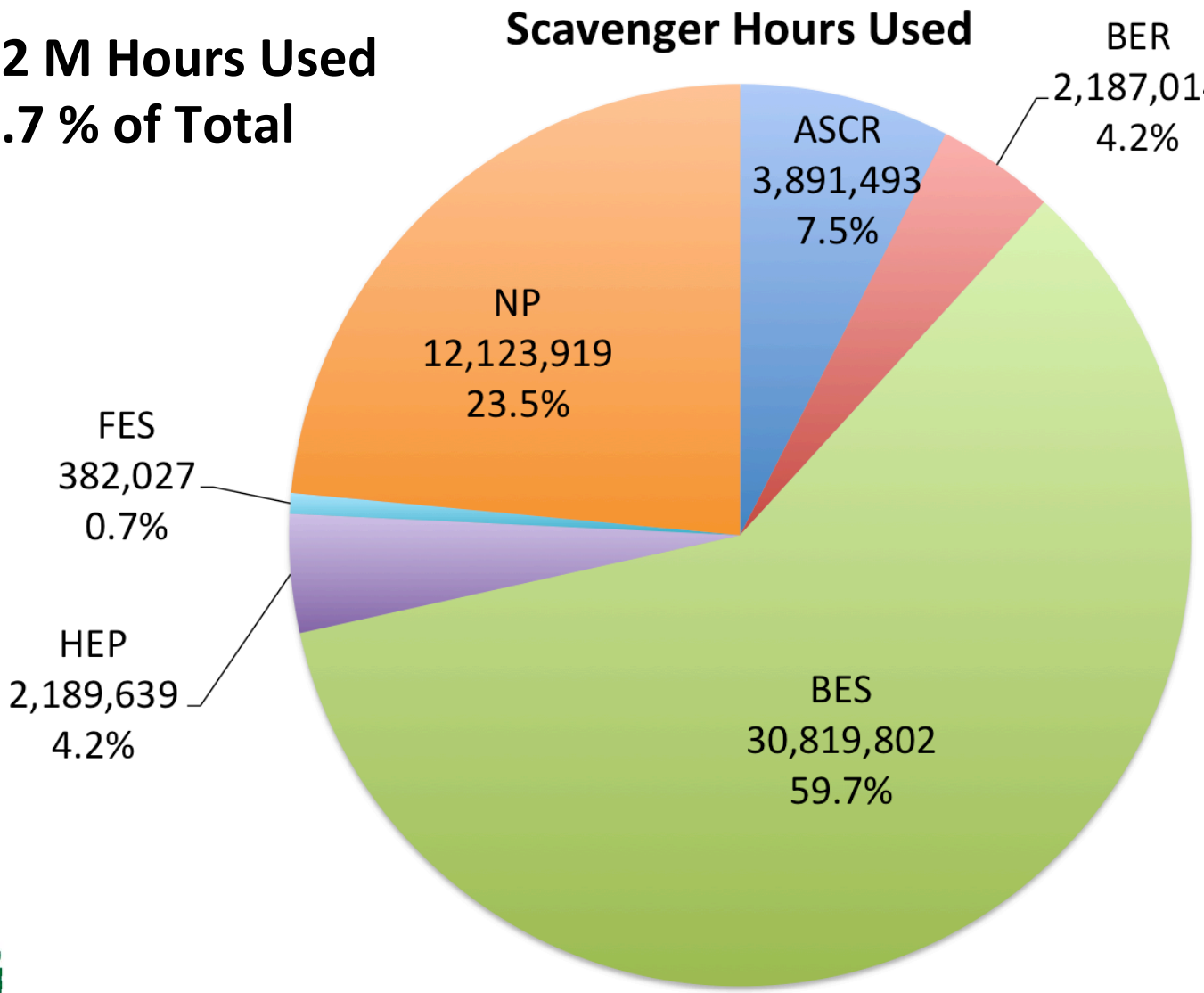
- **Hours used is greater than hours charged**
  - Mostly because of large job discount on Edison
  - Low job priority charging
  - Scavenger backfill computing
  - Refunds & free time
- **Result is that repo & reserves have more time than available the rest of the year**
  - DOE Production Balance: 1,012 M
  - ALCC Balance: 188 M
  - DDR Balance: 88 M
  - Total Balance: 1,288 M
    - vs. 955 M available (34% oversubscription)

Additional Hours: Scavenger Computing

- **Beginning this year when a repo runs out of time, it can run jobs in the scavenger queue**
  - Early in the year, throughput in scavenger is terrible
  - This will improve, but only if we remain resolute and don't create additional time ("print money")
- **NERSC will not "rescue" repos that are out of time**
  - They will have to run in scavenger or get time from DOE
  - Advantage: repos that still have allocation remaining do not have to compete in the regular queues with "rescued" repos
- **DOE program managers do not need to rescue either**
  - Additional time is not needed to enable access to NERSC computing
  - Adding time to a repo will have the effect of giving it much greater priority in the queues

2016 Scavenger Hours

52 M Hours Used
2.7 % of Total



Once we have some SLURM / NIM development done, may want to limit repos to some % of allocation in scavenger

100% over?

Cori Phase 2 Supplemental Allocation and Application Readiness

- **While Cori Phase 2 will greatly increase NERSC capability and capacity, not all codes will be able to run efficiently on the Xeon Phi partition**
- **NERSC is identifying codes and repos that will be ready to run well in production mode on Cori Phase 2 by the time it goes into production in July 2017**
- **NERSC proposes**
  - Allocating 2.4 billion NERSC hours for DOE Production computing for 2017 during the normal ERCAP cycle
  - Making an additional ~2.4 billion allocation in about May 2017, once the program managers have info about what projects can run on the Xeon Phi Cori Phase 2 partition

NERSC AY 2017 Allocations Forecast

System	"NERSC Hour" Charge per Node Hour	Nodes in System	~Hours in a Year	Overall System Availability Estimate	~Total NERSC Hours for AY2017 (M)	DOE Prod NERSC Hours (M) (80%)	ALCC NERSC Hours (M) (10%)	Directors Reserve NERSC Hours (M) (10%)
Edison	48	5576	8760	.85	2,000	1,600	200	200
Cori P1	80	1630	8760	.85	1,000	800	100	100
Cori P2 (6 months)	96*	9300	8760	.40 (6 months)	3,000†	2,400†	300‡	300†
2017					**6,000**	**4,800**	**600**	**600**
2016					3,000	2,400	300	300

* - Estimate, may adjust once we measure application performance on system
† - Supplemental allocation in Spring 2017
‡ - Applies to 2017-18 ALCC allocation cycle
Assumes Cori Phase 2 goes into production in mid 2017
Multiply the shaded columns to get the Total NERSC Hours Available for AY2017
Numbers are approximate (but pretty close to actual values!)

Take Away Summary

- **NERSC is on pace to deliver committed hours to DOE Production and ALCC for 2016**
- **There is no "NERSC reserve" time due to Cori Phase 2 integration and required OS upgrades.**
  - Most Director's Reserve will not be allocated until it is clear that NERSC can meet its DOE commitments
- **NERSC will not "rescue" repos that are out of time and has no time to give to needy or new projects**
- **Free early user time on Cori Phase 2 will help, as will returning from planned outages early and good system availability**
- **Allocations in 2017 will double, but codes need to be ready to use the Xeon Phi and program managers need to consider readiness in allocation decisions**
- **2.4 B DOE Production hours will be allocated to start 2017 with another 2.4 B supplemental allocation in ~May 2017 for Cori Phase 2 production (expected July 2017)**