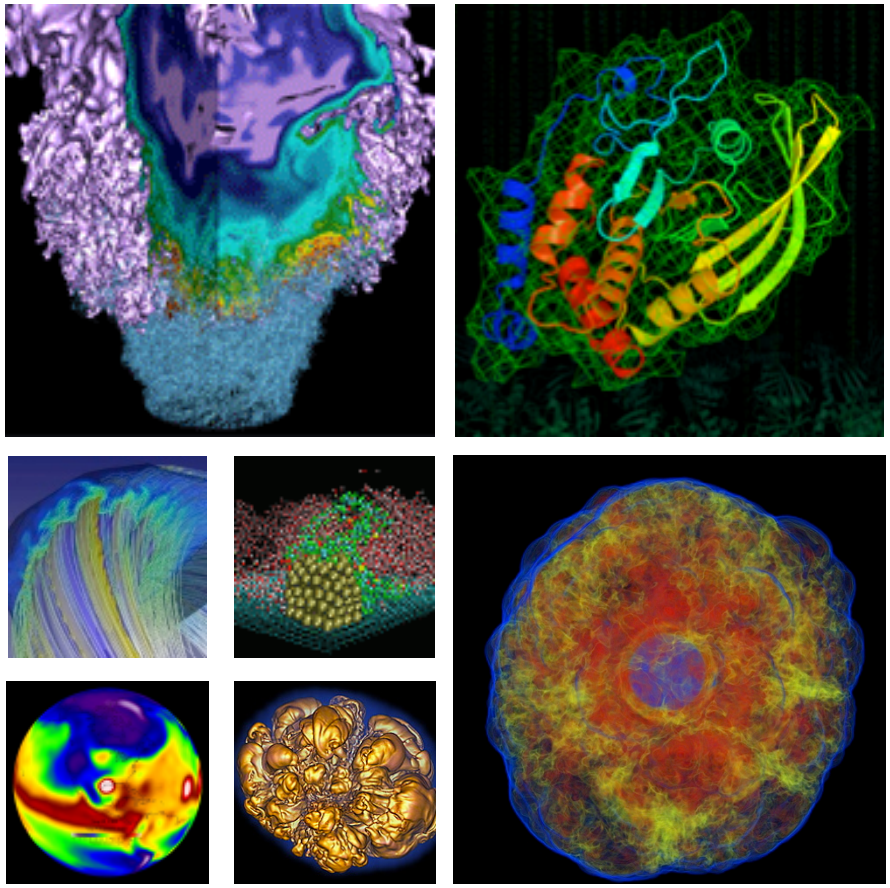
# NERSC Allocations 2016 - 2017



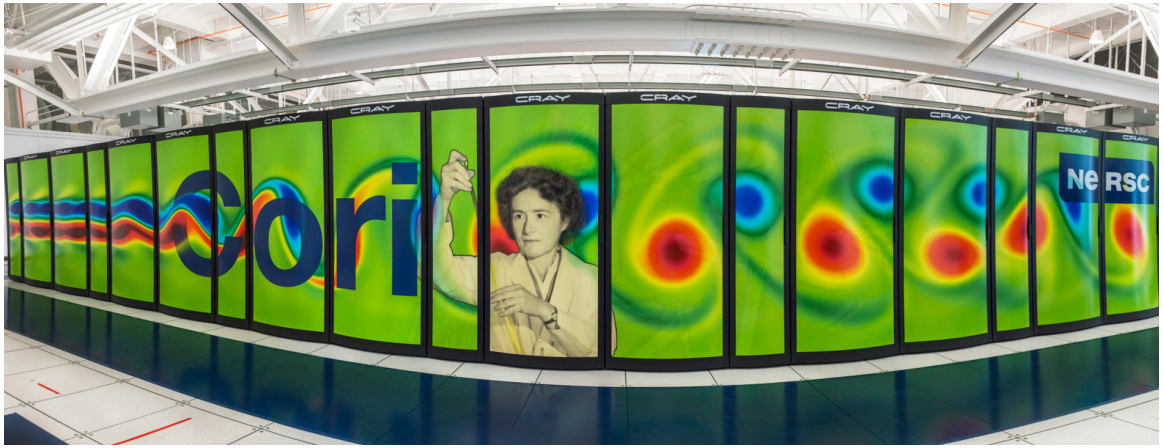

## Richard Gerber

**High Performance Computing Department Head**
**Senior Science Advisor**


June 14, 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

- **Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9,300 Intel Knights Landing compute nodes – mid 2016**
  - Self-hosted, (not an accelerator) manycore processor with 68 cores per node
  - On-package 16 GB high-bandwidth memory; 96 GB DDR memory
- **To run efficiently on Cori, application codes will have to**
  - Increase thread parallelism (OpenMP, …)
  - Exploit data parallelism (vectorization)
  - Improve memory locality (exploit high bandwidth memory)
- **Robust Application Readiness Plan (NESAP)**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams

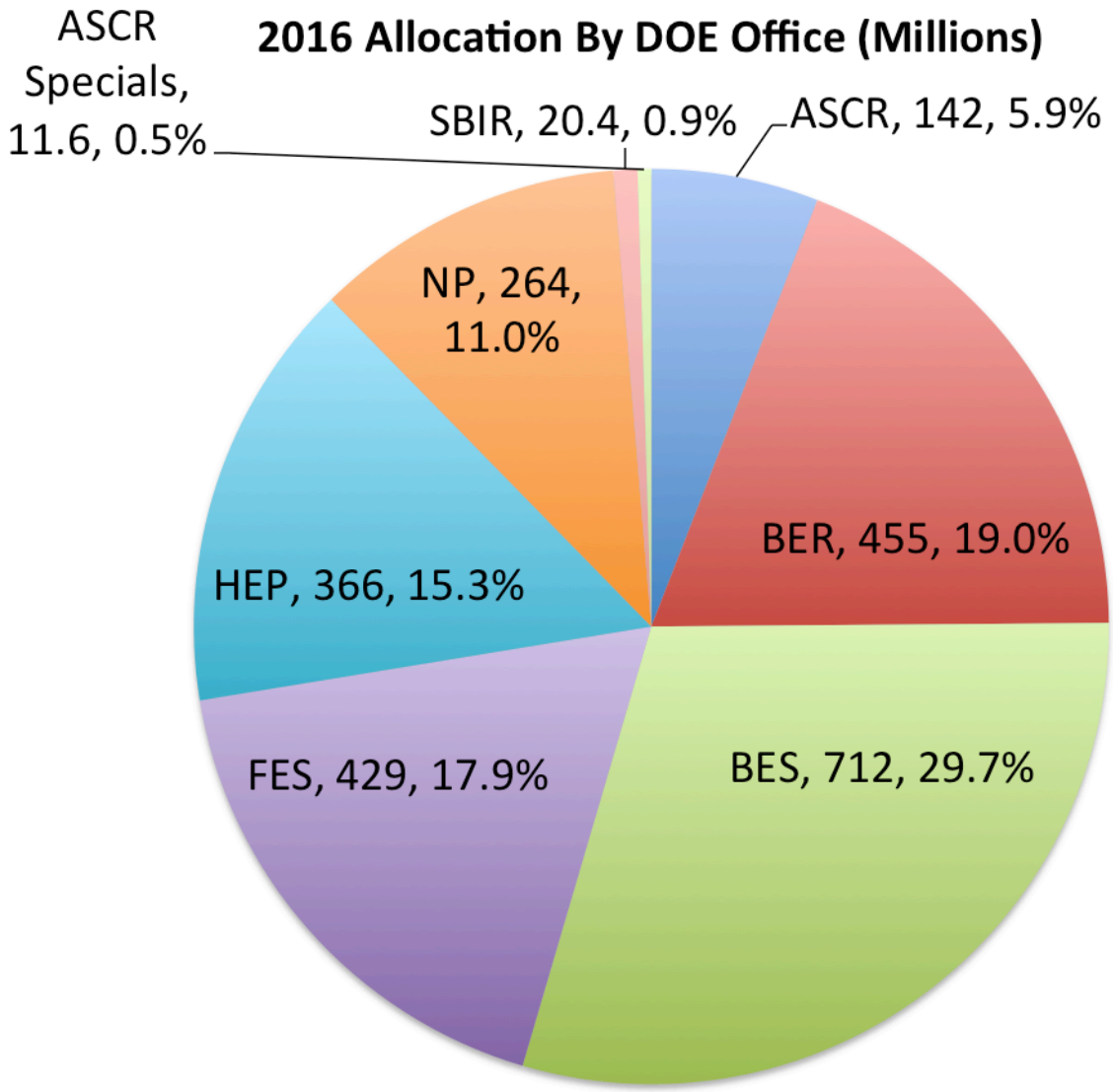
# Allocation Pools and Commitments

- **Commitment to DOE in AY2016: 2.7 Billion NERSC Hours**
  - 2,400 M for DOE Production (mission computing): DOE Managers
  - 300 M for ALCC (ASCR Leadership Computing Challenge): ASCR competition
  - 300 M for Director's Discretionary Reserve: NERSC (brings total to 3B)
- **Additional time set aside for miscellaneous: ~72 M hours**
  - NERSC Overhead – 65 M
  - Startup projects – 5 M
  - Education – 1.5 M
  - Guests – 500 K
- **Additional time is available if system downtimes are less than estimated or new resources become available (e.g. preproduction systems)**

# Charged Hours vs. Used Hours

- **NERSC tracks usage by the "NERSC Hour"**
  - 48 "NERSC Hours" per hour of computing on an Edison node (24 cores)
  - 80 "NERSC Hours" per hour of computing on a Cori Phase 1 node (32 cores)
- **NERSC Allocates and Charges "NERSC Hours"**
  - There are discounts and extra charges
  - Low (half charge rate) and premium (double charge rate) job priorities
  - Edison jobs that use >684 nodes get a 40% discount
  - "Hours Used" ≠ "Hours Charged"
- **We track and report "Hours Used" for meeting our commitment to DOE**
- **User and repo "banking" deals with "Hours Charged"**

# Initial Allocation Distribution Among Offices for 2016

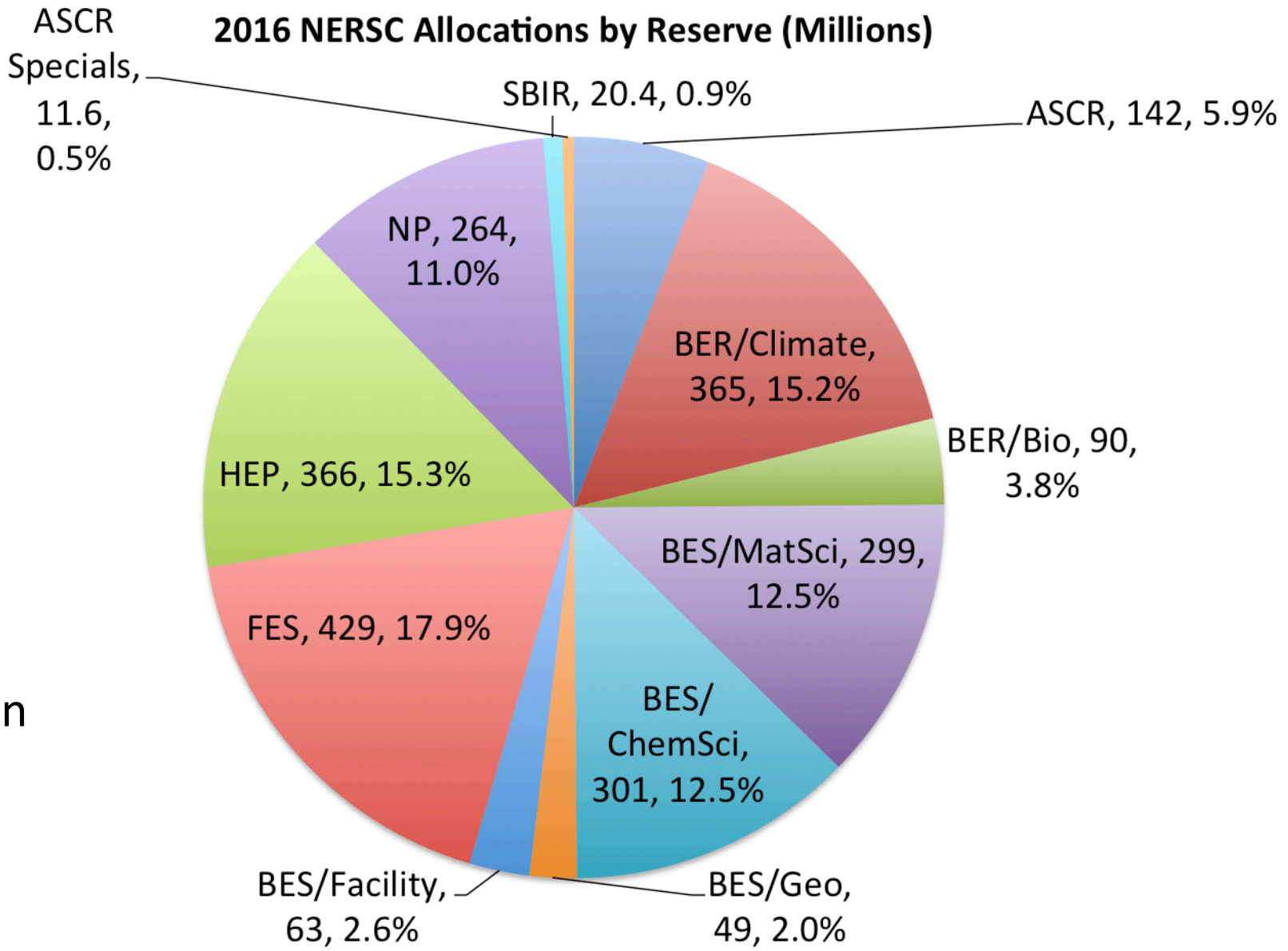



2,400 Million Total

# Allocations Divided Up Into Reserves

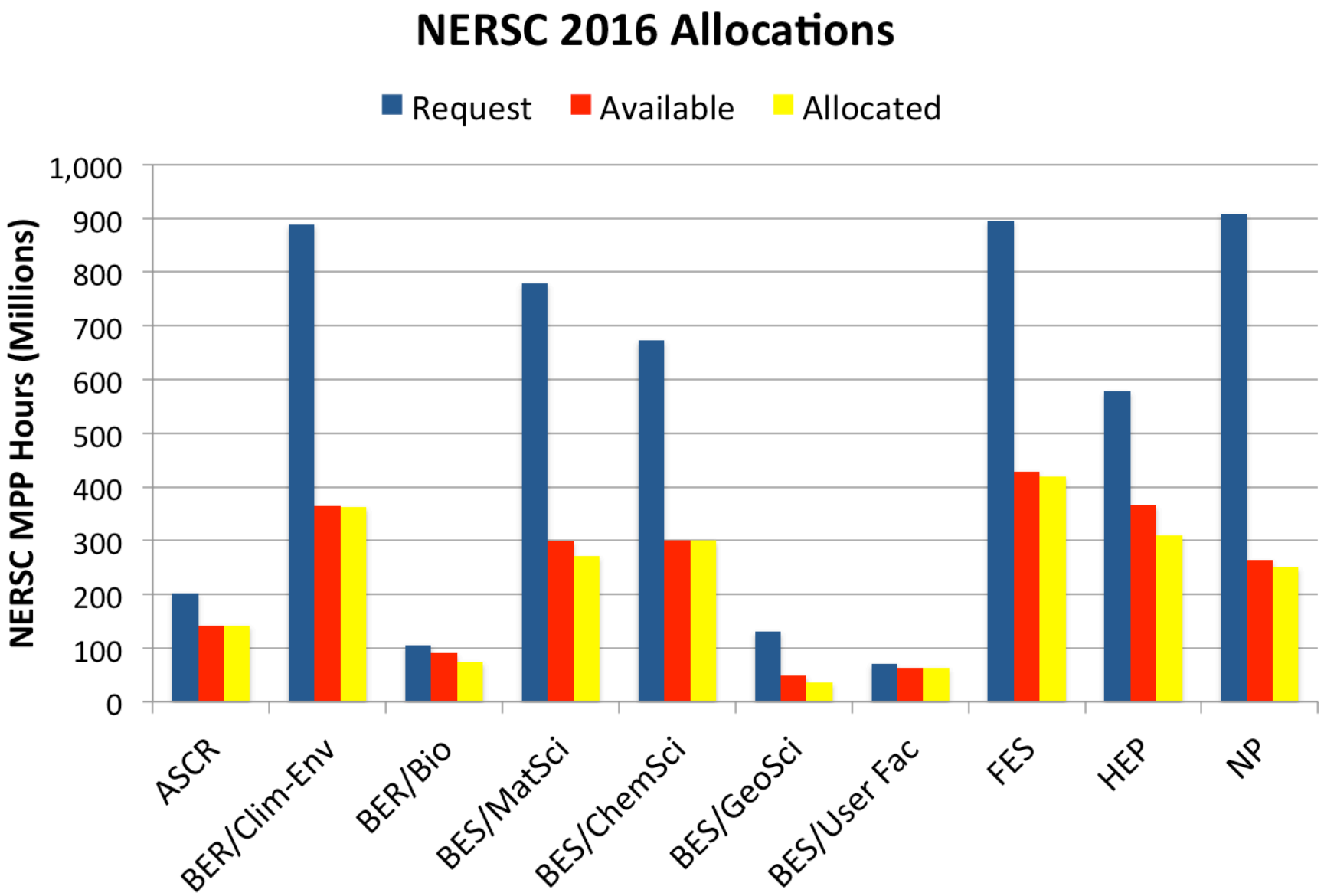
| Office | Reserve | Initial AY2016 Allocation |
|---|---|---|
| **ASCR** | ASCR | **142,000,000** |
| **BER** | | 455,000,000 |
| | Biosciences | **90,000,000** |
| | Climate & Environment | **365,000,000** |
| **BES** | | 712,000,000 |
| | Chemical Sciences | **301,000,000** |
| | Geosciences | **49,000,000** |
| | Materials Science | **299,000,000** |
| | User Facilities | **63,000,000** |
| **FES** | Fusion | **429,000,000** |
| **HEP** | High Energy Physics | **366,000,000** |
| **NP** | Nuclear Physics | **264,000,000** |
| **SBIR/ASCR Specials** | SBIR/ASCR Specials | **20,400,000 / 11,600,000** |

# Initial Allocation Distribution Among Reserves for 2016




2,400 Million Total

# Demand is Greater Than Supply



# Usage is On Track

DOE Production, 41% of Year Passed (June 8), 77 M Hours Given by ALCC

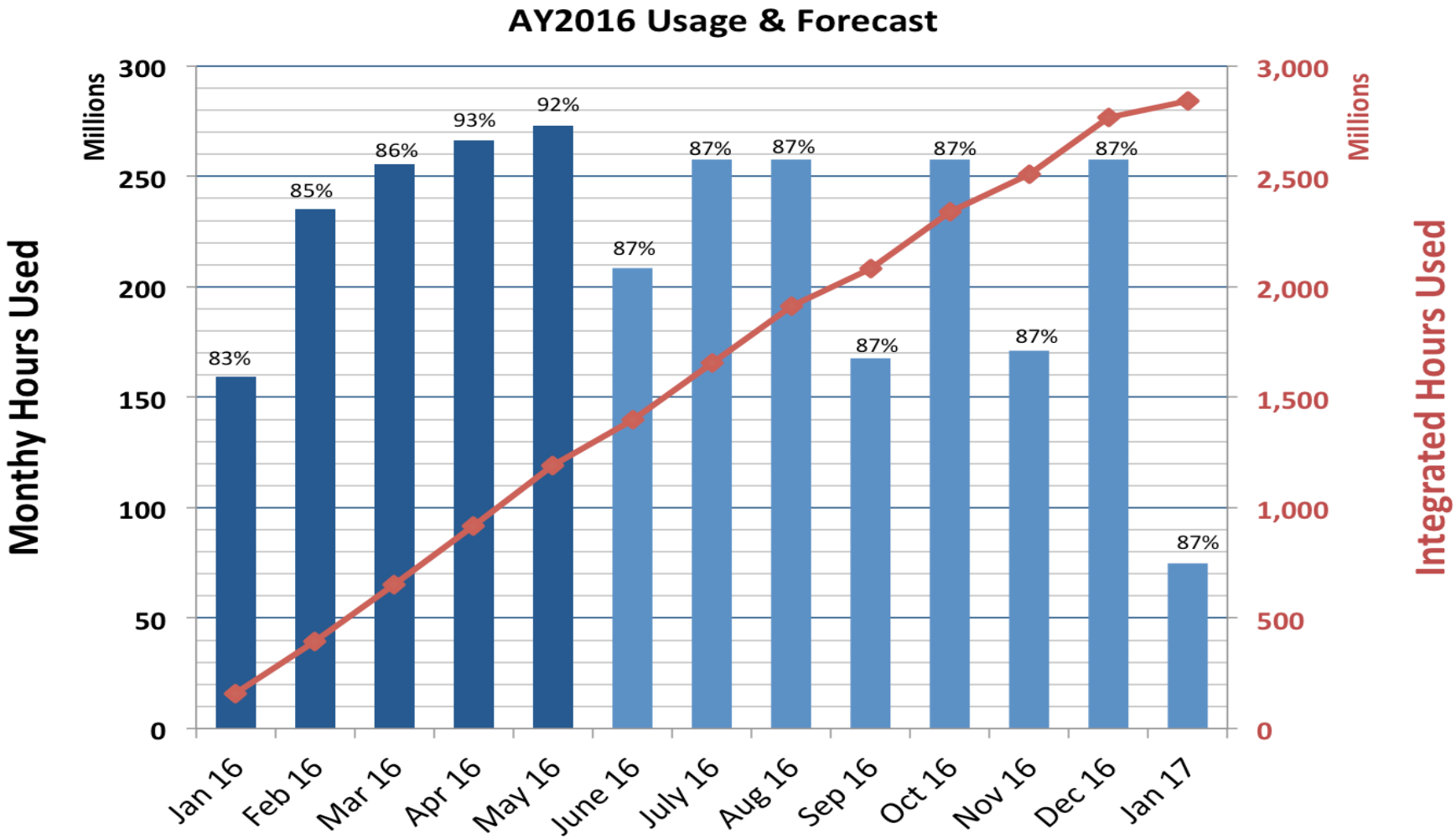
| Office | Initial Allocation | Charged | % of Alloc Charged | Reserve (June 8) |
|---|---|---|---|---|
| ASCR | 153,002,000 | 74,475,509 | 49% | 88,000 |
| BER | 474,268,400 | 220,479,745 | 46% | 1,840,600 (Climate/Env)<br>16,773,080 (BioSci) |
| BES | 750,875,000 | 308,615,642 | 41% | 24,174,000 (ChemSci)<br>14,717,000 (GeoSci)<br>23,479,000 (MatSci)<br>1,898,000 (UserFac) |
| FES | 445,904,000 | 161,492,957 | 36% | 13,010,000 |
| HEP | 373,206,000 | 134,210,376 | 36% | 68,568,000 |
| NP | 276,520,000 | 128,608,700 | 47% | 20,129,000 |

# Issues That Affect Usage the Rest of 2016

- **Cori Phase 2 starts arriving in July 2016**
  - Installation and testing ongoing until full system in place
- **Phase 1 (Xeon) and 2 (Xeon Phi) will be integrated into one system in September 2016**
  - Expect ~4-6 weeks of outage on Cori Phase 1 in September
  - ~2 weeks Cori Phase 1 outage for required OS upgrade starting June 11, 2016
  - ~2 weeks Edison outage in November for OS upgrade (independent of Cori integration; we can change date if needed)
- **We hope these are overestimates of required downtimes, but we want to be realistic and perhaps a little conservative**

# 2016 Usage (including planned outages)

Usage through June 2016
Percentages shown are overall utilization
Dark blue: used, light blue: projected



# Usage and Forecast Overview 2016

| Allocation Pool | Allocated (M Hrs) | Used 6/8/16 | Remaining Commitment to DOE |
|---|---|---|---|
| DOE Production | 2,477* | 1,158 | 1,319 |
| ALCC | 223* | 50 | 173 |
| DDR | 114 (186 unallocated) | 20 | 94 |
| TOTAL | 2,814 | 1,228 | 1,596 |

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

1,600 M Hours Remaining in AY2016

2,814 M Hours Total Will Be Used in AY2016

Overall availability >87% and/or less downtime than anticipated will provide more hours



* ALCC gave 77 M hours to DOE Production

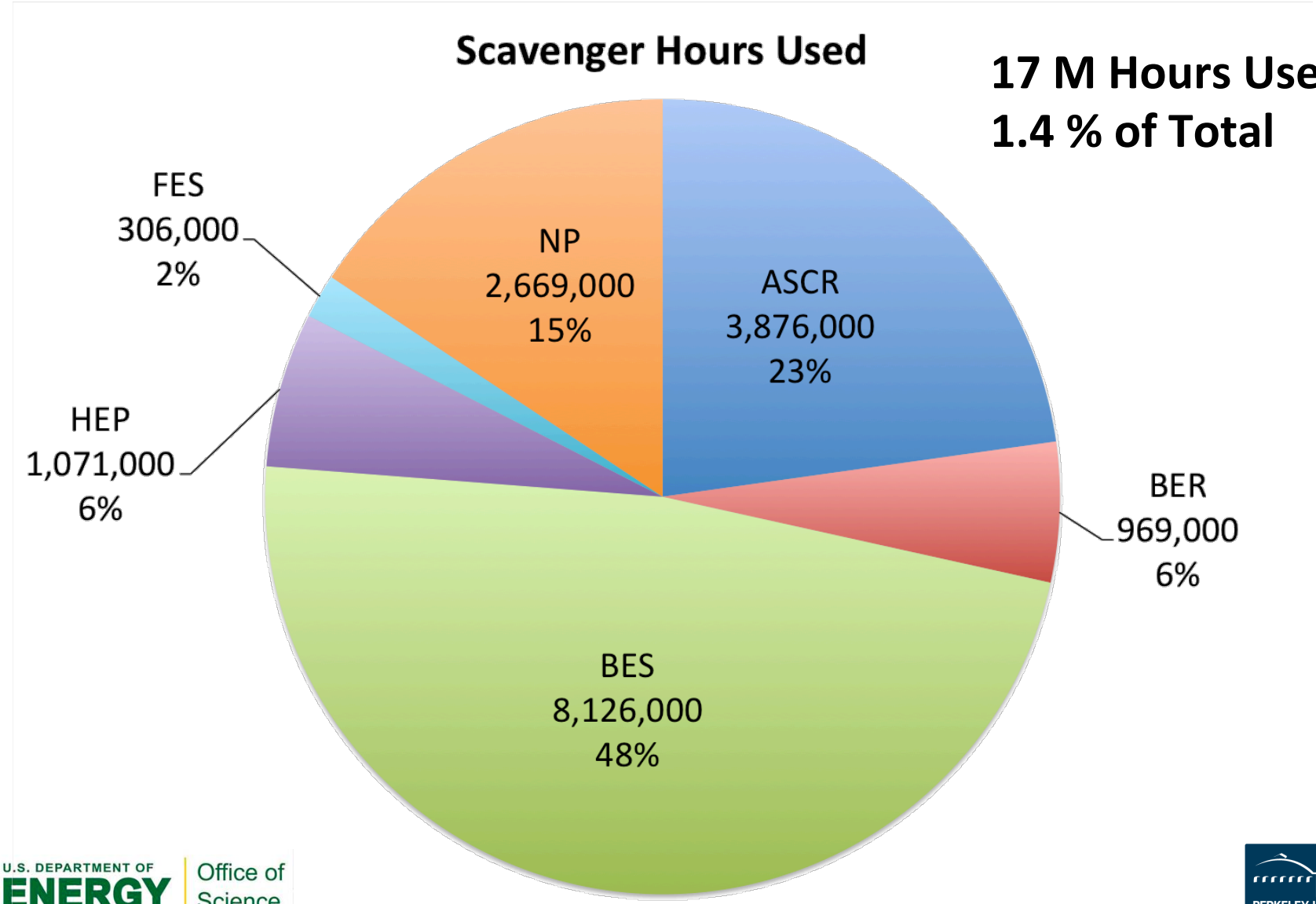
# Additional Hours from Cori Phase 2 Early User Access

- **When Cori Phase 2 becomes usable, NESAP teams will get excusive early access for a few 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
- **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

# 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



# 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 little "NERSC reserve" time due to Cori Phase 2 integration and required OS upgrades
- Free early user time on Cori Phase 2 will help, as will returning from planned outages early and good system availability
- NERSC will not "rescue" repos that are out of time and has little time to give to needy or new projects
- 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)