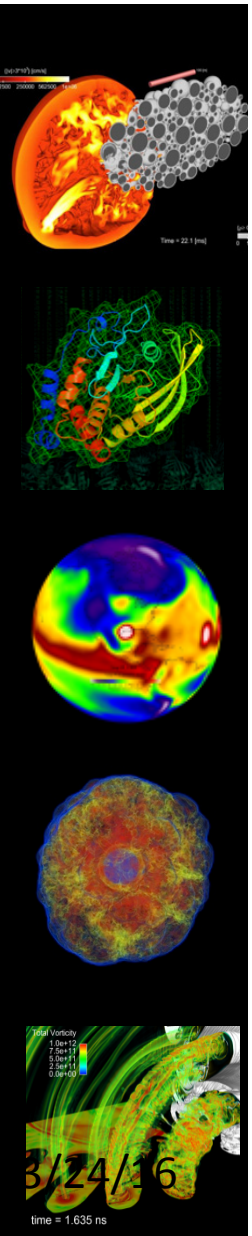



# NERSC-9

Nicholas J. Wright, NERSC-9 Chief Architect

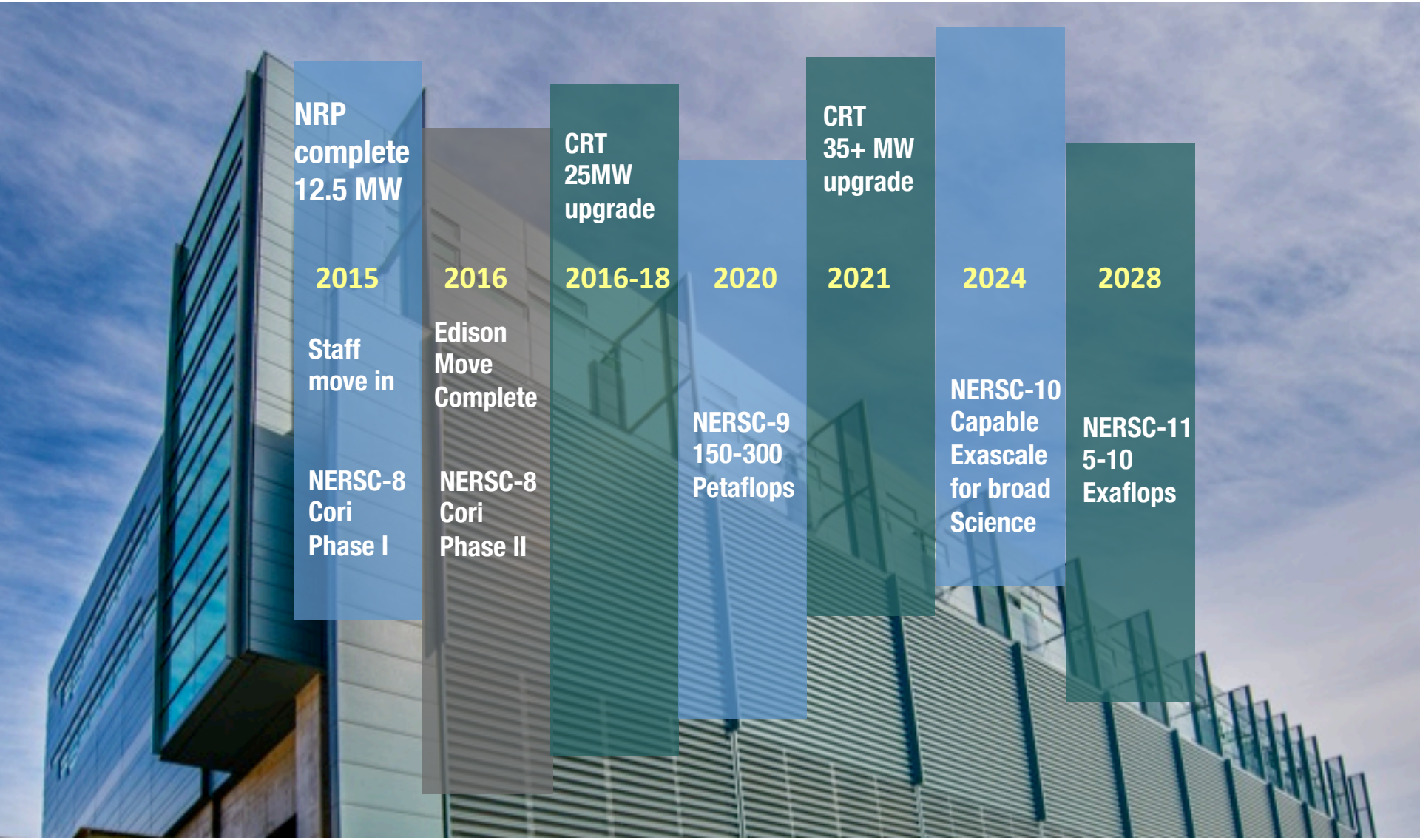
NUG meeting

March 24, LBNL

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

# NERSC Timeline

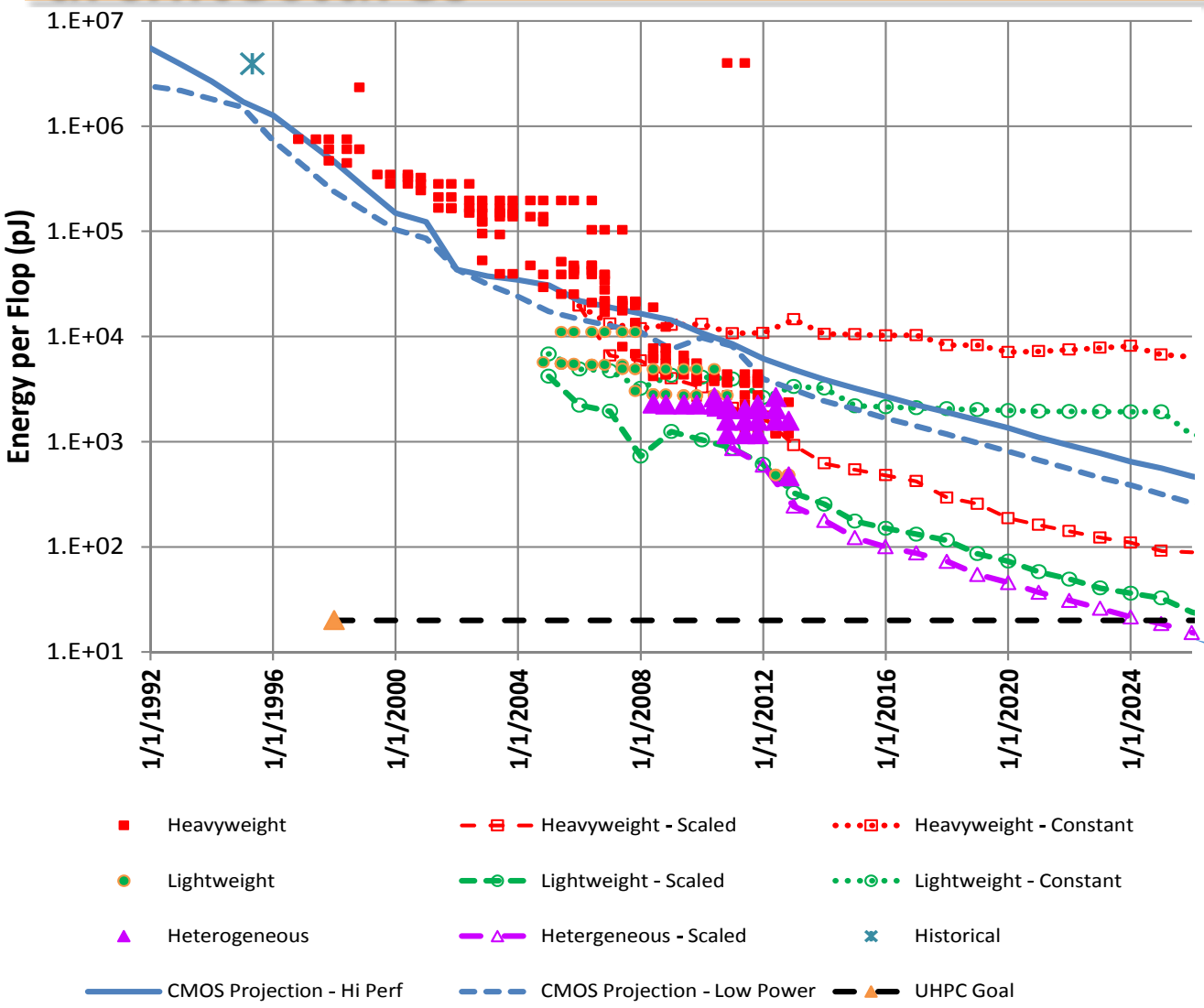


# APEX 2020 Current Status

- 3rd joint SC/NNSA procurement
  - After Trinity/NERSC-8 (2016) & CORAL (2018)
- RFP draft technical specs released Nov. 10, 2015
  - 2nd Draft released March 11th

http://www.lanl.gov/projects/apex/_assets/docs/APEX2020_draft_tech_specs_v2.0.pdf

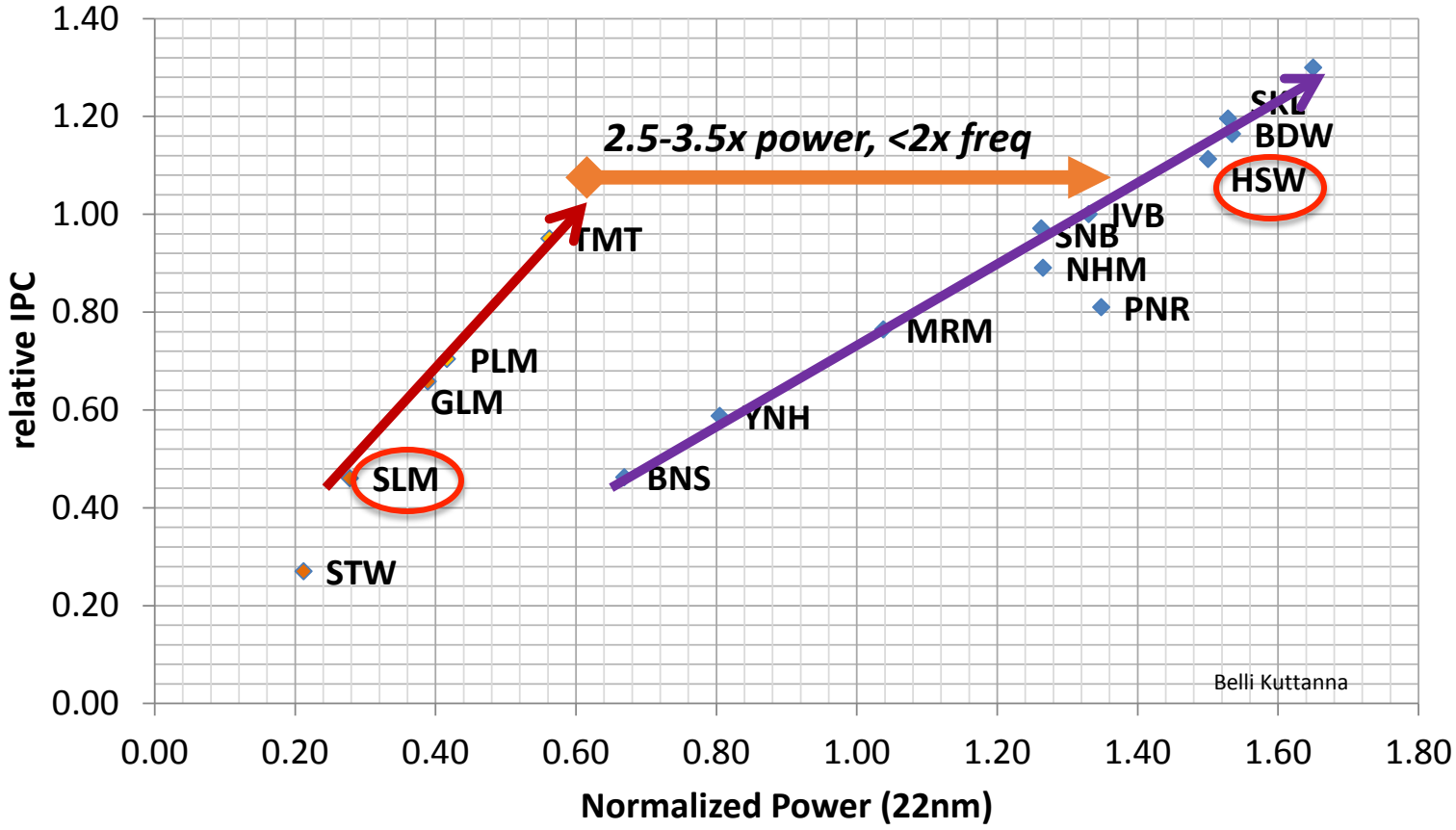
# NERSC needs to transition to energy efficient architectures




Manycore or Hybrid is the only approach that crosses the exascale finish line

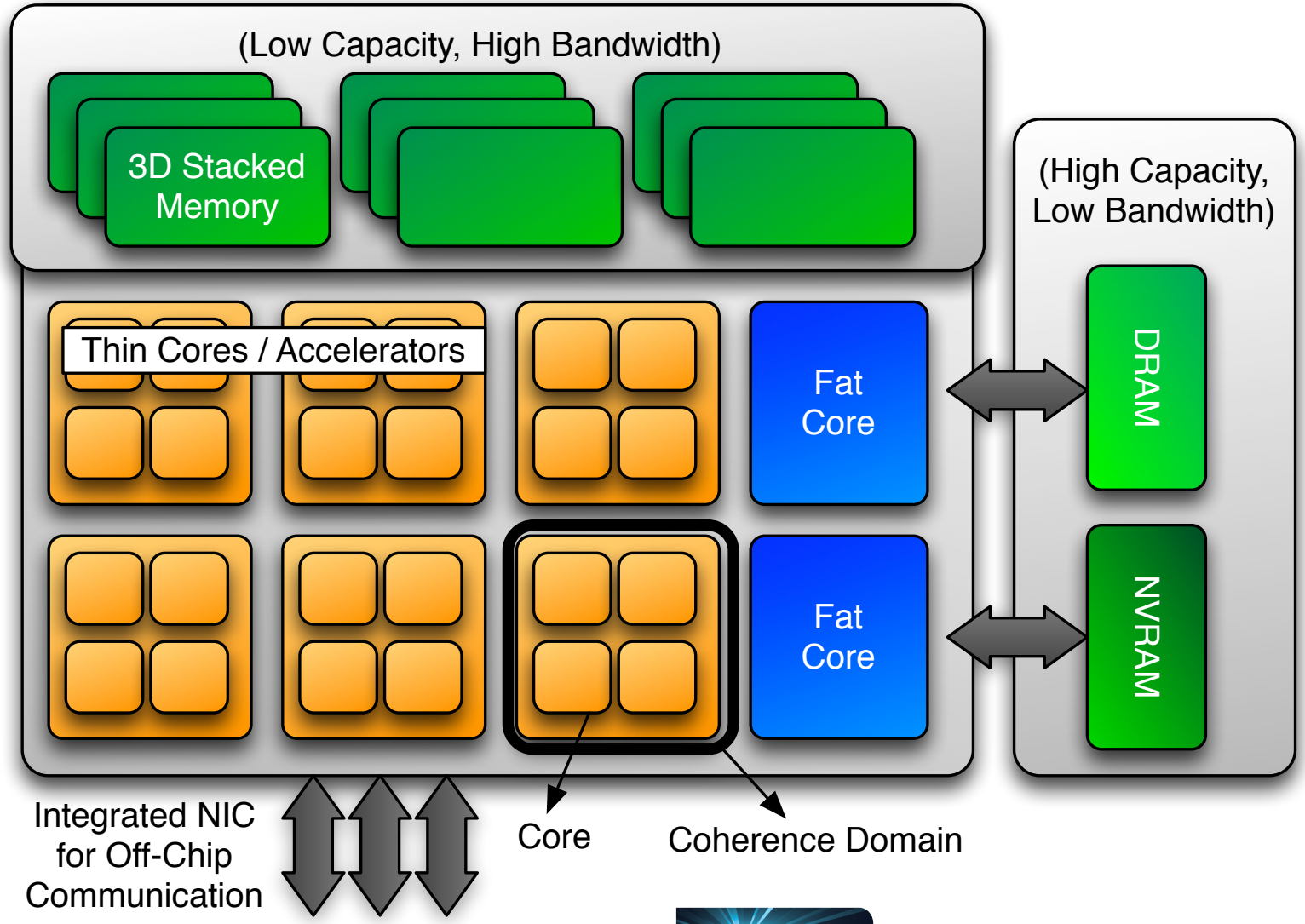
NERSC CROSSROADS

# Throughput vs Single Thread: Perf Trade-off

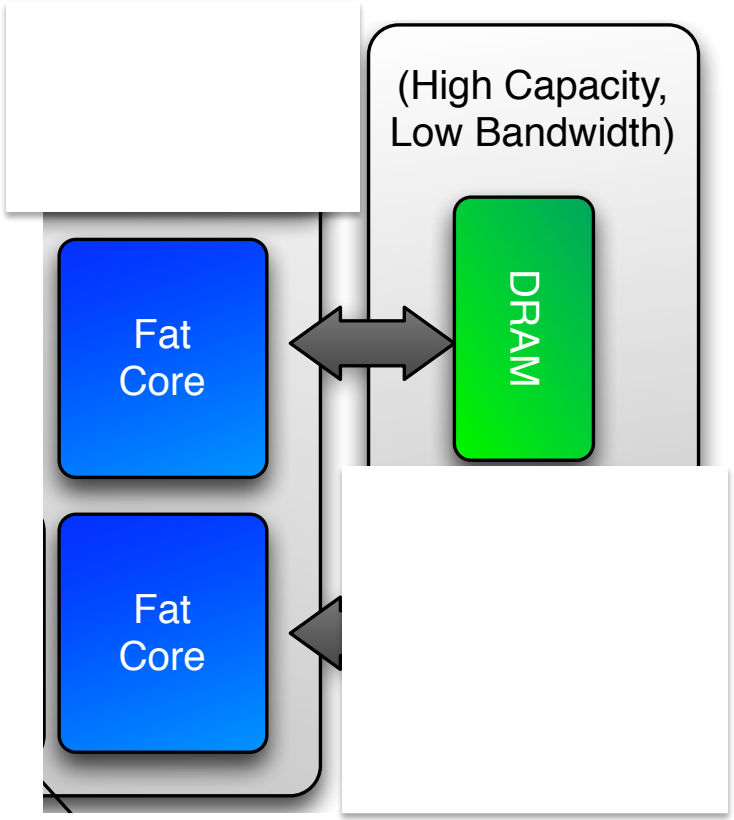



Haswell: Silvermont IPC: ~3x Power: ~5x

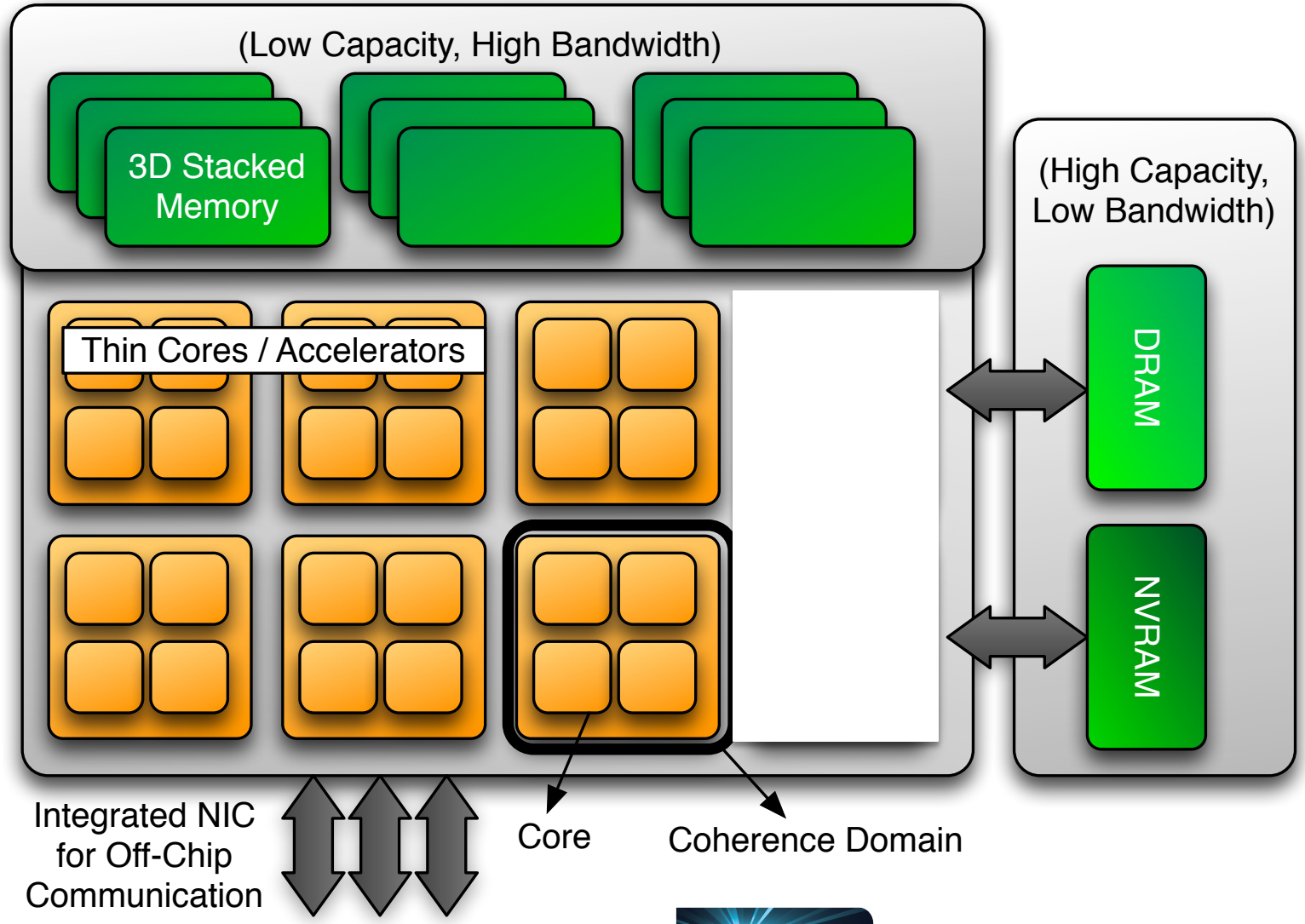
# Abstract Machine Model for Exascale



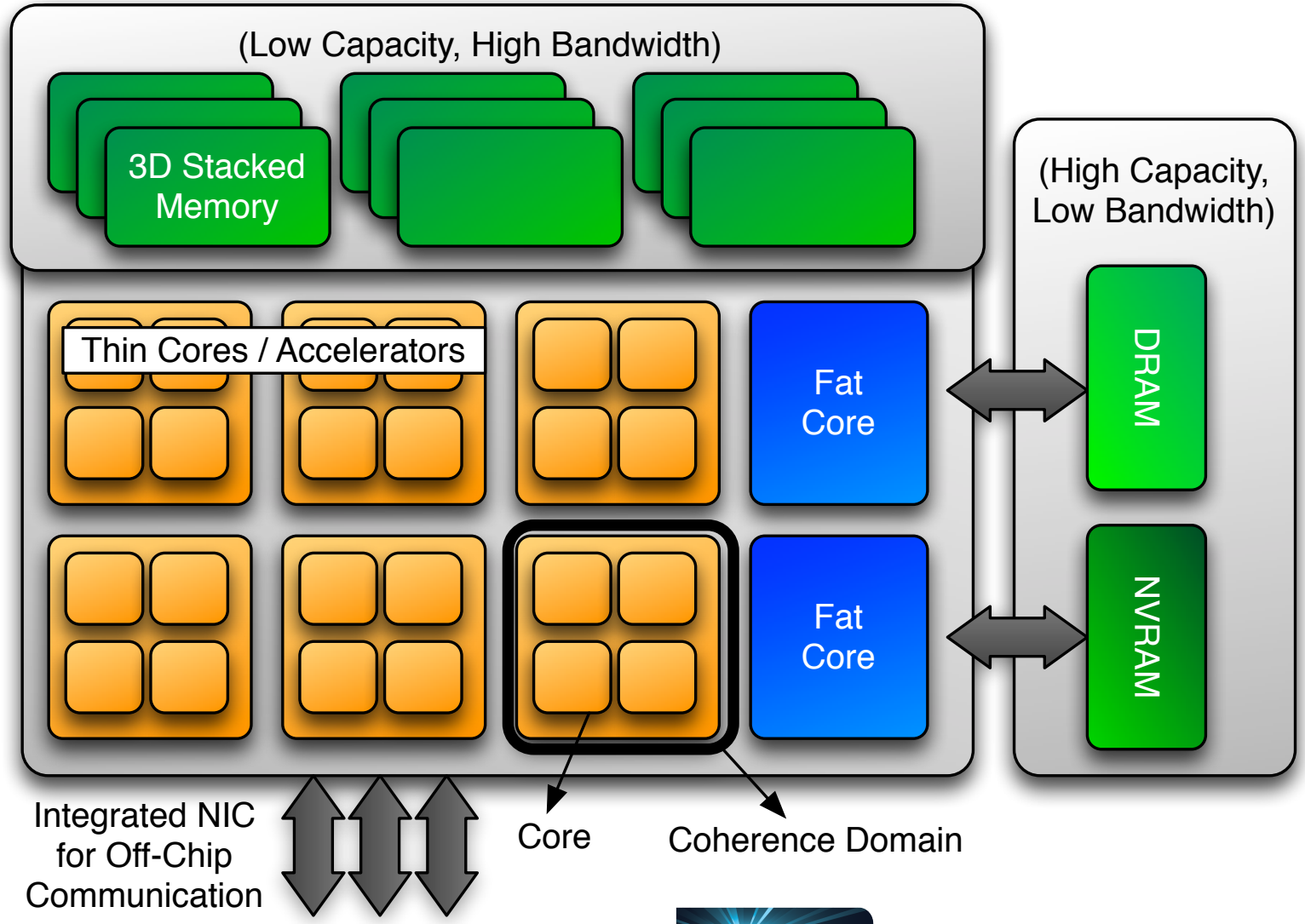
# Edison - 2012



# Cori (NERSC-8)- 2016



# NERSC-9 (2020) ? – An exascale-era architecture



| Layer | NERSC-7 (Edison) 2013 | NERSC-8 (Cori) 2016 | NERSC-9 2020 |
|---|---|---|---|
| **High Bandwidth Memory per node** | None | 16 GB, >400 GB/sec | More ! |
| **DRAM per node** | 64 GB, ~100 GB/sec | 96 GB, 90-100 GB/ sec | Some |
| **NV-DIMM (byte addressable)** | None | None | Maybe |
| **Non-Volatile (Page addressable)** | None | 1.5PB, 1.5 TB/sec | 10s PBs, 10s TB/sec |
| **Spinning Disk – /scratch** | 8PB, 130 GB/sec | 28 PB, 700 GB/sec | Collapsed layer > 50 PBs ~1 TB/sec |
| **Spinning Disk – longer term (/project)** | ~30 PB, ~70 GB/sec | ~50 PB, ~100 GB/sec | |
| **Tape** | ~40 PB, ~10 GB/sec | ~100PB, ~20 GB/sec | ~100s PB, ~10s GB/ sec |

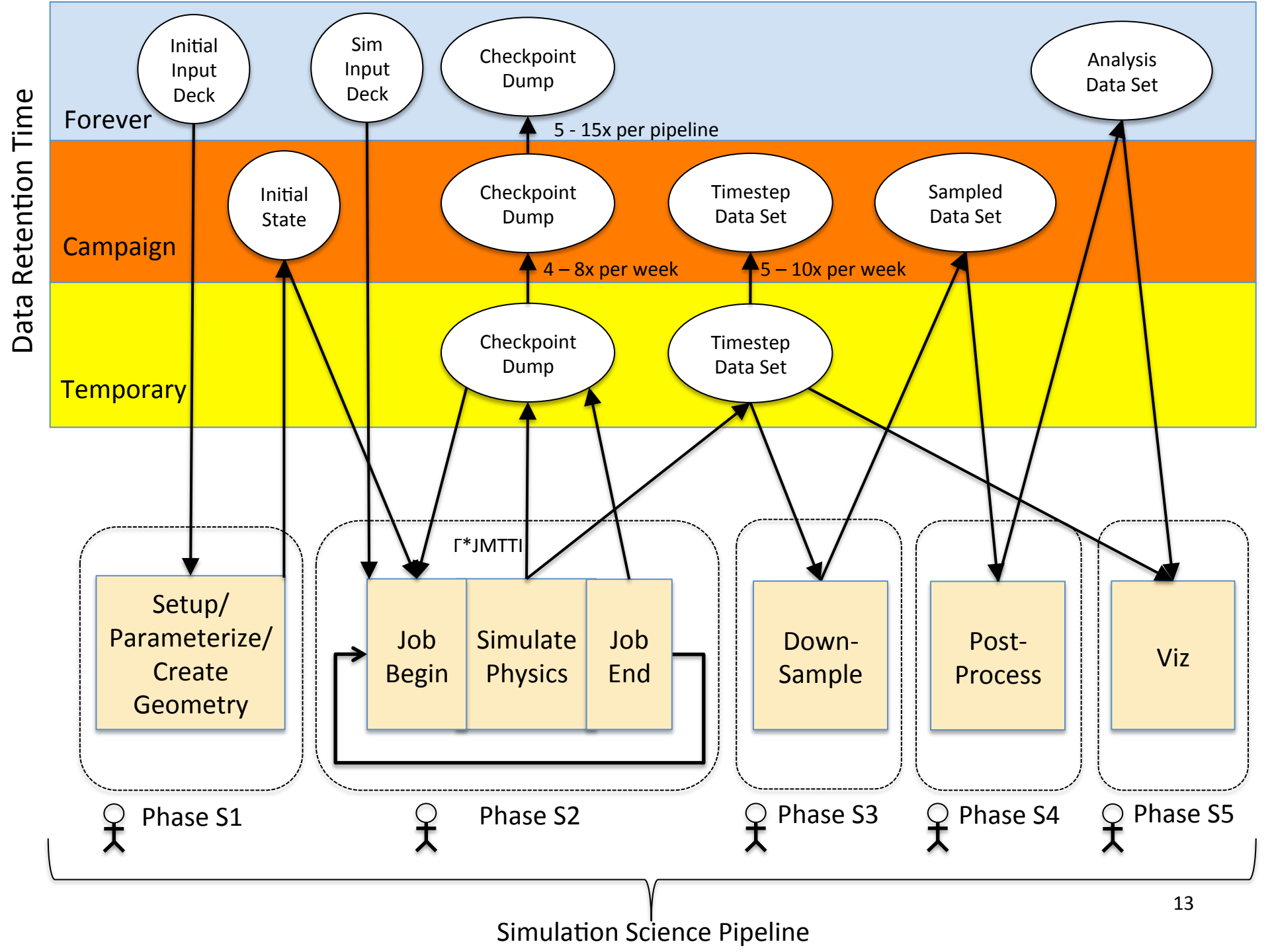
# Market Survey: Storage Technologies are Changing

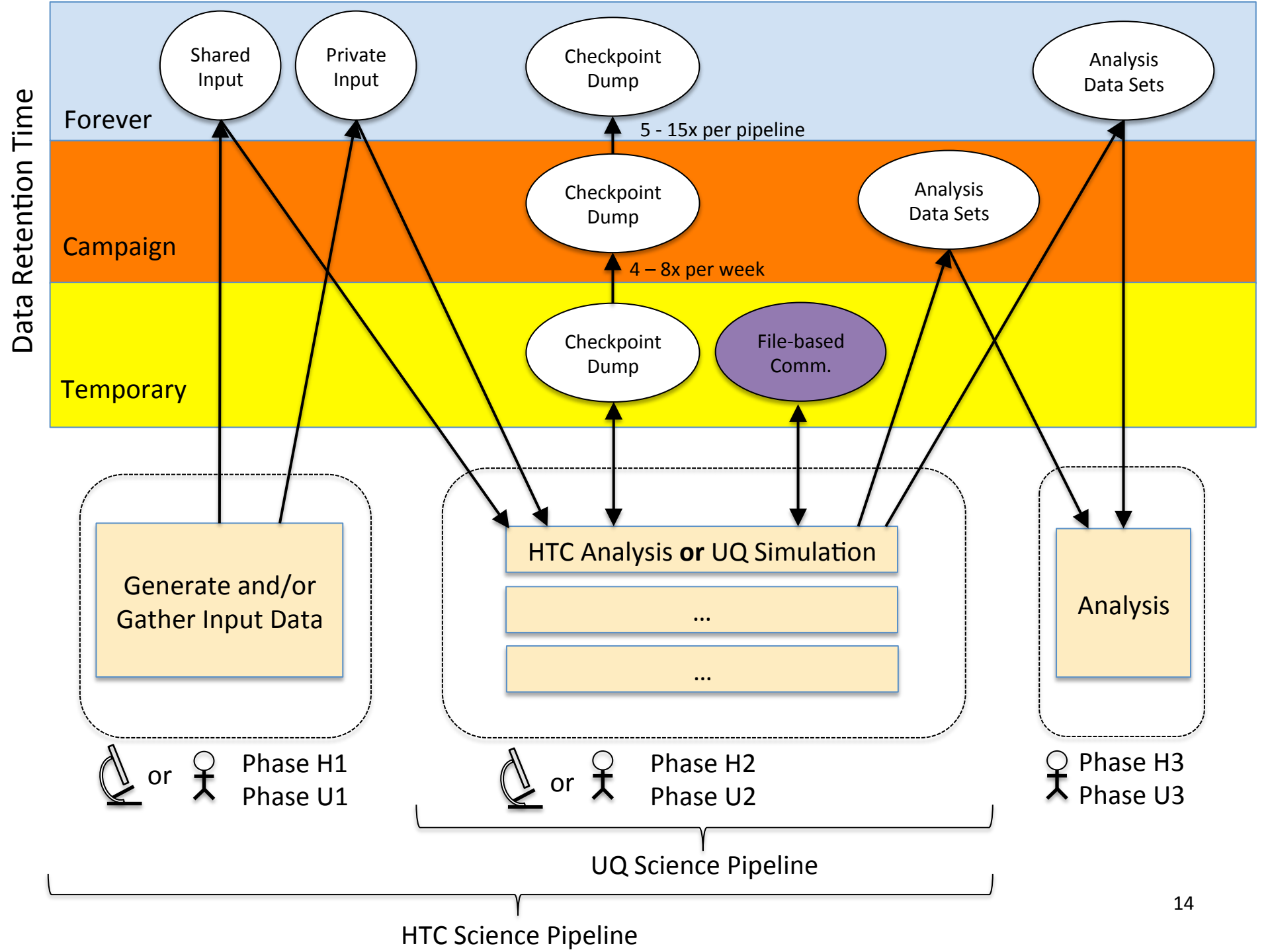
- NVRAM technologies are cost effective for bandwidth today
  - Burst Buffers in Trinity/Cori (2016) & CORAL (2018)
- In 2020
  - Will any spinning disk be needed for capacity? Cost is the limiting factor
  - NVRAM: How much ? What kind(s) ? Where to put it in the machine? What software (runtime/scheduler/ OS) enhancements will be needed?
    - Workflows !
      - Fusion, Climate, QCD, ALS, JGI, Materials, Sky Survey

https://www.nersc.gov/assets/apex-workflows-v2.pdf

# APEX will Define Workflows to Optimize Platform Storage

- A workflow is a description of the steps needed to obtain results in a scientific investigation
- The workflow life cycle typically consists of many computational and data transformation steps
  - Running simulations and/or experiments
  - Analyzing output data
  - Managing data to aid the scientific investigation, including collecting information to benefit future studies and help future validation of results
- Whitepaper released which describes other storage uses cases present in APEX workflows
  - Based upon extensive requirements gathering exercise
  - Includes estimates of data volumes and lifetimes for multiple NERSC, LANL, LLNL and SNL workflows
- Overall goal is to provide a framework to reason about platform storage design decisions
  - Allows vendor to innovate and be flexible


Data Retention Time
Forever
Campaign
Temporary
Initial Input Deck
Sim Input Deck
Checkpoint Dump
5 - 15x per pipeline
Analysis Data Set
Initial State
Checkpoint Dump
Timestep Data Set
Sampled Data Set
4 – 8x per week
5 – 10x per week
Checkpoint Dump
Timestep Data Set
Γ*JMTTI
Setup/ Parameterize/ Create Geometry
Job Begin
Simulate Physics
Job End
Down-Sample
Post-Process
Viz
Phase S1
Phase S2
Phase S3
Phase S4
Phase S5
Simulation Science Pipeline


Data Retention Time
Forever
Campaign
Temporary
Shared Input
Private Input
Checkpoint Dump
5 - 15x per pipeline
Analysis Data Sets
Checkpoint Dump
4 – 8x per week
Analysis Data Sets
Checkpoint Dump
File-based Comm.
Generate and/or Gather Input Data
HTC Analysis or UQ Simulation
...
...
Analysis
or
Phase H1
Phase U1
or
Phase H2
Phase U2
Phase H3
Phase U3
UQ Science Pipeline
HTC Science Pipeline

# Target System Configuration

| | NERSC-8 | NERSC-9 - Target |
|---|---|---|
| SSP | > 5x Edison | > 20x Edison |
| Baseline Memory Capacity | 1.1 PB | > 3 PiB |
| Burst Buffer | 1.5 PB 1.5 TB/s | >90 PB<br>>5 TB/s |
| Disk | 22 PB 744 GB/s | |

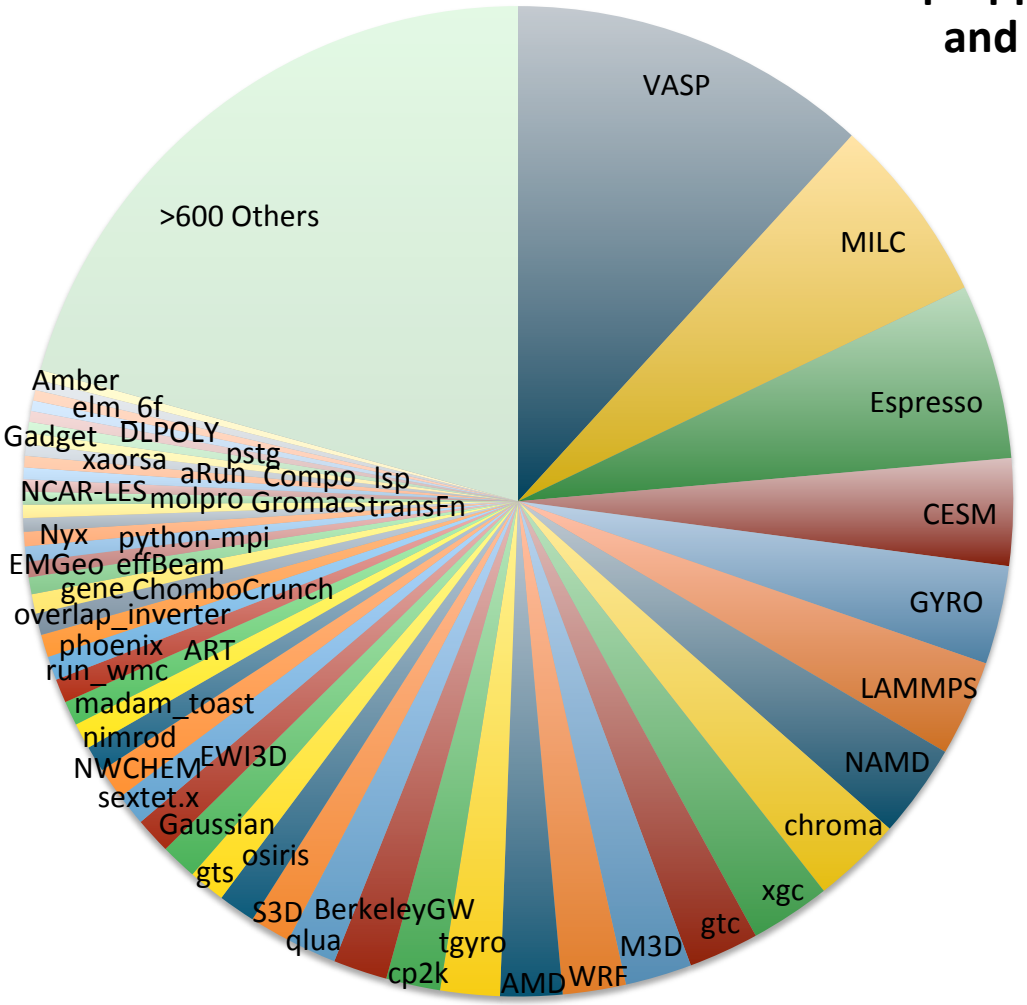
# Market Surveys have Formed the Basis of our Requirements Development

- The Crossroads/NERSC-9 (CN9) teams had many formal (Face-to-Face) and informal (telecon) interactions with vendors over the last 15 months
  - Interactions continue leading up to the RFP release
- Market Surveys and interactions focused on major prime and technology provider candidates:

# Technical Specifications Include Findings From Workload Analysis and Requirements Workshops

- NERSC workload analysis performed as part of the procurement activities
  - http://portal.nersc.gov/project/mpccc/baustin/NERSC_2014_Workload_Analysis_30Oct2015.pdf
- NERSC has held one requirements workshop per office looking at 2017 requirements
  - http://www.nersc.gov/science/hpc-requirements-reviews

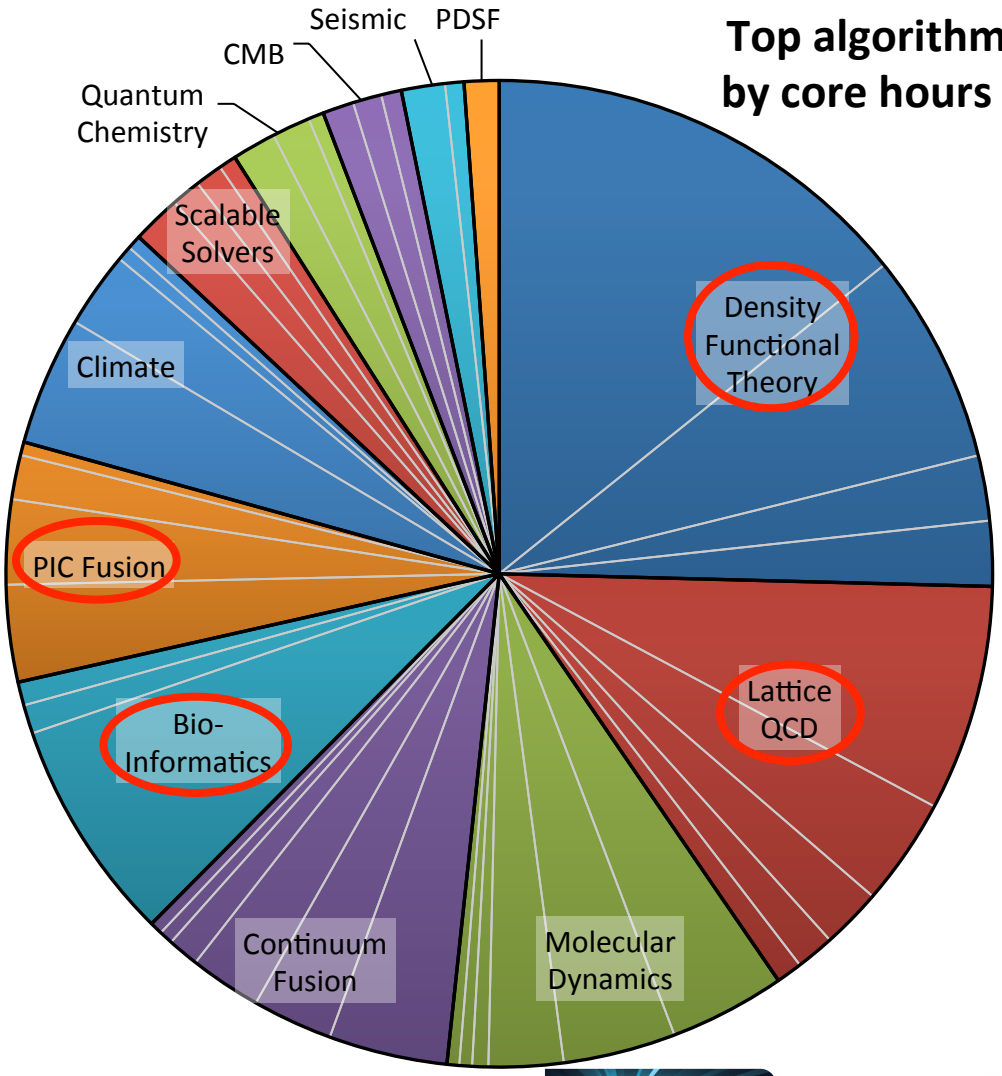
# Over 650 applications run on NERSC resources




**Top Application codes on Hopper and Edison by hours used. Jan – Dec 2014**

- 13 codes make up 50% of workload
- 25 codes make up 66% of workload
- 50 codes make up 80% of workload
- Remaining codes (over 600) make up 20% of workload.

# NERSC Benchmarks Were Chosen to Represent the Workload




**Top algorithms on NERSC systems by core hours used Jan – Dec 2014**

- Regrouped top codes by similar algorithms.
- A small number of benchmarks can represent a large fraction of the workload.
  - miniDFT
  - MILC
  - GTC
  - Meraculous
- Includes Genepool and PDSF systems.

# APEX plans to use "mini-apps", some full apps for system evaluation

| MiniApp | Description | Language |
|---|---|---|
| miniDFT (Quantum Espresso) | Plain-wave Density Functional Theory (DFT) | Fortran |
| MILC | Lattice Quantum Chromodynamics (QCD). Sparse matrix inversion, CG | C |
| GTC-P | Particle-in-cell magnetic fusion | C |
| UMT | Unstructured-Mesh deterministic radiation Transport | C/C++/Fortran |
| SNAP | Neutron particle transport application | Fortran |
| PENNANT | Unstructured finite element | C |
| Meraculous | De novo genome assembly | UPC |
| MiniPIC | Particle in cell for accelerators | C++ |
| HPCG | High Performance Conjugate Gradient | C |

# Goals and Objectives for the NERSC-9 Project

1. Provide a significant increase in computational capabilities over the Edison system, at least 16x on a set of representative DOE benchmarks
2. Platform needs to meet the needs of extreme computing and data users by accelerating workflow performance
3. Platform should provide a vehicle for the demonstration and development of exascale-era technologies
4. Delivery in the 2020 time frame

# NERSC-9 Will Provide Capabilities for DOE Data-Intensive Users in 2020

- NERSC-9 will build upon the successes of the data different components of Cori
- End to end workflow requirements and performance are critical for the design and optimization of the system
- Overall goal is to enable seamless data motion with dynamic allocation and scheduling of resources
  - Enable first steps towards exascale-era storage system
  - Vendor community excited about engagement and collaboration opportunities

# APEX 2020 – NRE on the Path the Exascale

- The APEX 2020 systems NRE topics will target areas that
  - achieve higher application performance,
  - improve support for data-intensive computing, and,
  - enable greater ease of use

  by advancing new technologies on the path to the exascale systems in 2023
- The Crossroads and NERSC-9 platforms NRE topics are
  - Technologies for the exploration of new and novel programming models concepts
  - A platform integrated storage system that supports new models for moving and managing data seamlessly
  - Systems with scalable management capabilities to enhance the reliability, resilience, power and energy usage characteristics

# Summary

- NERSC-9 will be 2020 machine that meets the needs of all NERSC users
- NERSC will continue its NESAP program in support of NERSC-9
- NERSC will partner with vendors on Non-Recurring Engineering projects to maximize the usability and performance of the machine

# Questions?


NERSC

# The Application Transition Program is designed to continue users on the path to exascale

- Technical specifications asks for Center of Excellence
  - Establishment of a collaboration between the Labs, the chosen OEM, and key technology providers, e.g. processor, is essential to meet the goals of the making efficient use of the platform in a timely manner
- Center of Excellence (CoE) based upon previous DOE efforts
  - NERSC Exascale Scientific Applications Program (NESAP)
  - CAAR & ESP programs at ORNL & ANL
- Center of Excellence (CoE) leverages some or all of:
  - SSI metric applications
  - NERSC Exascale Scientific Applications Program (NESAP)
  - Select applications expected to use the machine shortly after operational readiness/acceptance

# The Application Transition Program will provide development resources for users

- Early access to key technologies and programming environments is essential for application transition
  - Programming environment is crucial
- Access to emulation and simulation capabilities as early as possible
  - key contribution of technology providers
- Early Access Development System
  - One or more iterations of increasing scale
    - Eventually 2-10% of final system size
- Development test beds
  - To investigate select advanced technology areas
    - E.g. Network, power management, burst buffer
  - Same or different composition of hardware depending on topic

# APEX Non Recurring Engineering (NRE): Philosophy

- Technical Specifications ask for NRE proposals
- NRE contracts potentially 10-15% of platform budgets
- Other topics that have potential to impact path to exascale will be considered
- Focus on topics that provide added value beyond planned vendor roadmap activities
- NRE collaborations will have impact on follow-on platforms procured by the U.S. Department of Energy's NNSA and Office of Science.