NCAR

Looking towards the future – NCAR's Computing and Storage

Anke Kamrath

Director, Computational and Information Systems Laboratory (CISL)
National Center for Atmospheric Research (NCAR)

anke@ucar.edu

18th Workshop on HPC in Meteorology
September 24, 2018

NCAR

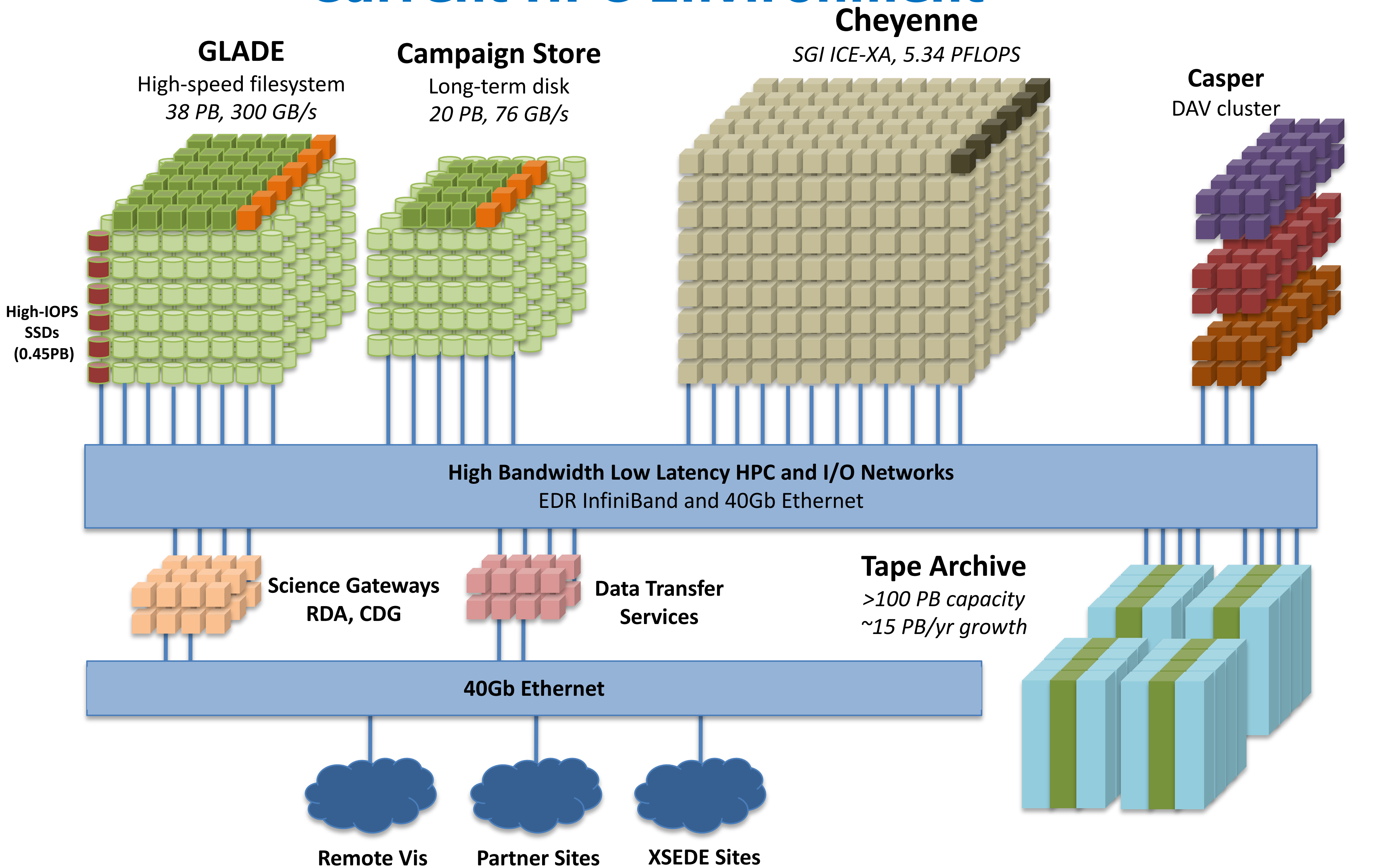
Thanks to Rich Loft, Irfan Elahi, Dave Hart, Rory Kelly of NCAR for their slide contributions

NSF

Overview

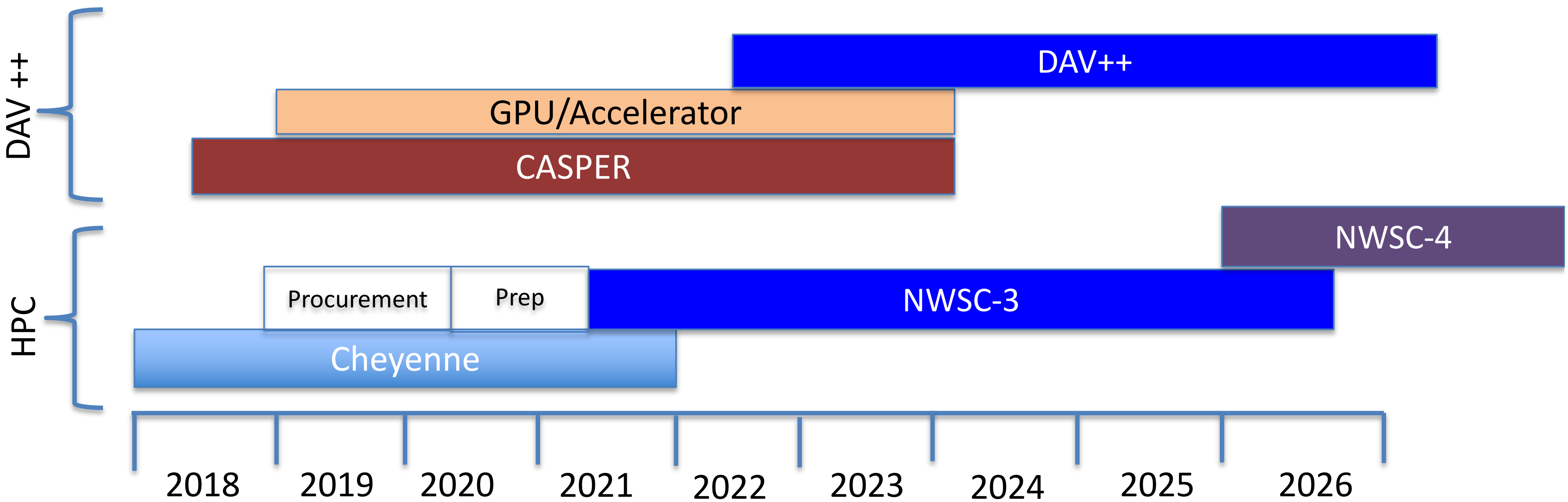
- **Existing NCAR HPC Environment – "Cheyenne"**
- **Procurement Schedule for "NWSC-3"**
- **Next gen of computing and storage**
  - Key Considerations
  - What might it look like?
  - Application preparations
  - Benchmarking Approach
  - Technology Risk Mitigation
  - Procurement Strategy

Current HPC Environment

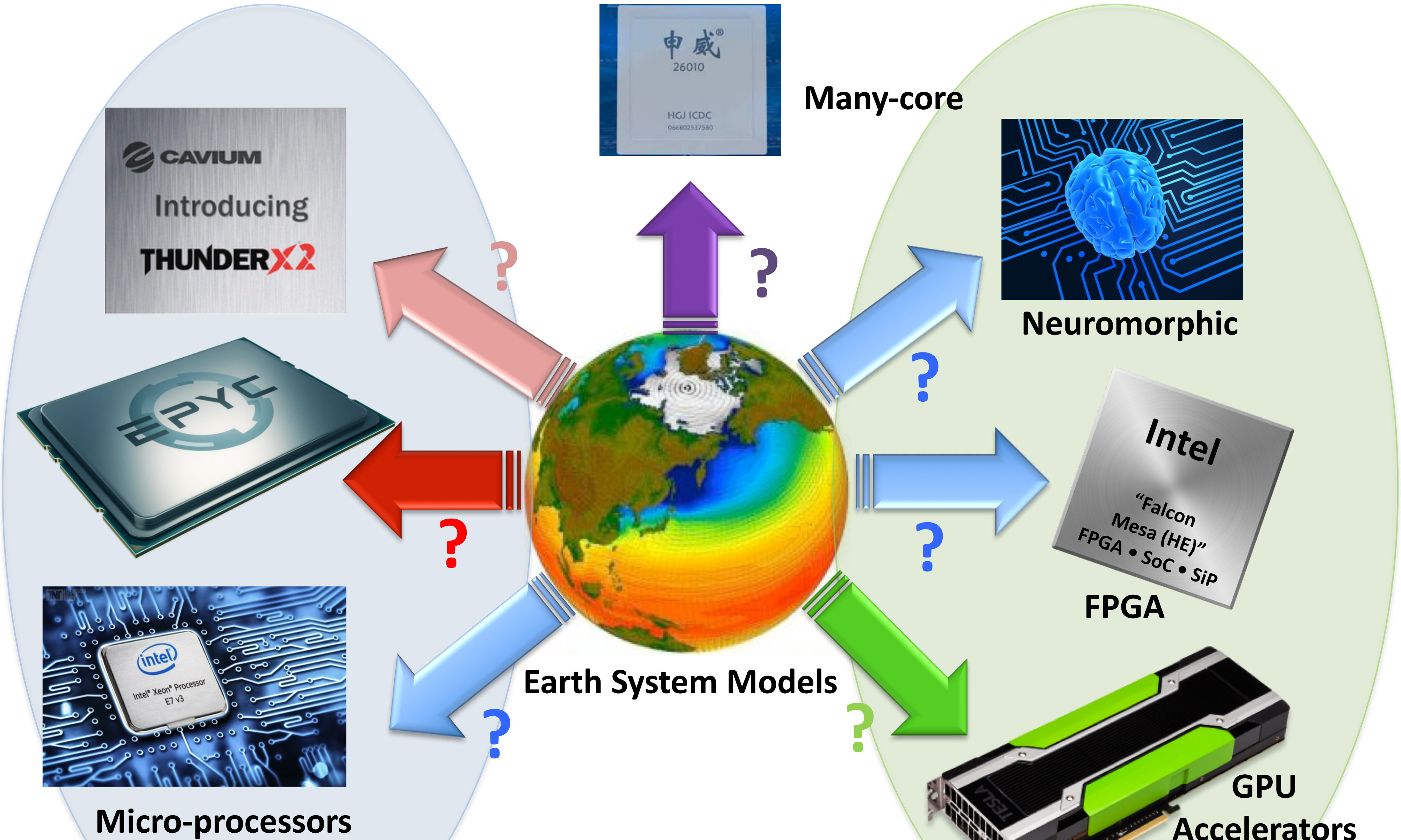


HPC Roadmap

- **Late 2018 - Mid-2019** - Benchmark design
- **Late 2018 - Mid-2019** - Technology Briefings and Architecture Co-Design
- **Feb 2019 - May 2019** - Science Requirements Process
- **Late 2019** - RFP Release.
- **Early to Mid-2020** - Selection and Approval
- **Mid-2020 - Early 2021** - Facility Prep, Vendor System Build and Prep
- **Early 2021 - July 2021** - System Deployment, Install and Acceptance
- **July 2021 -** Production start date, 6-month overlap with Cheyenne

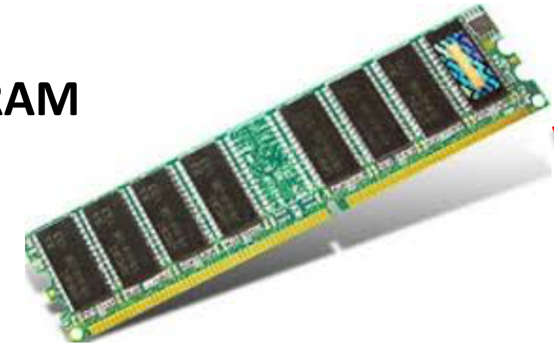


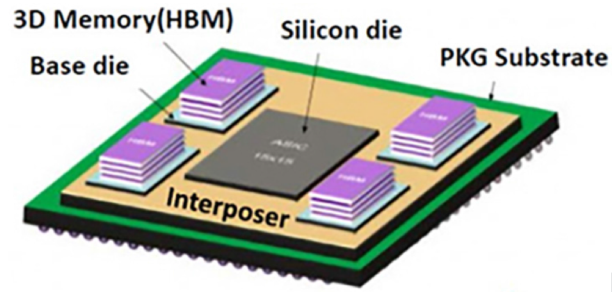
Increasing Computing Complexity



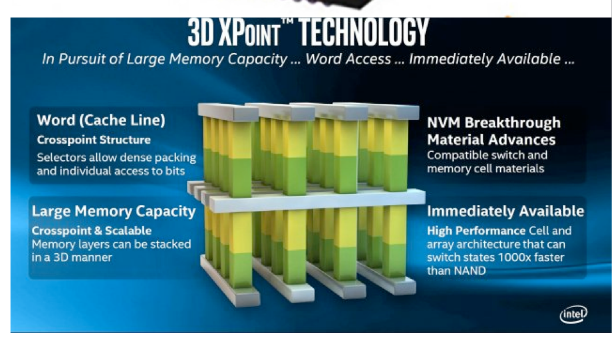
Increasing Storage Complexity

DRAM



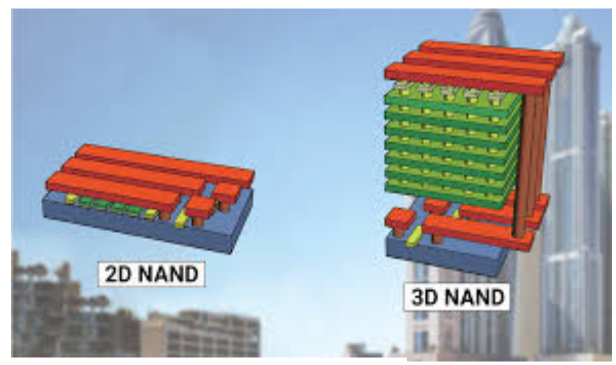


Stacked memory:
Fast, hot & small



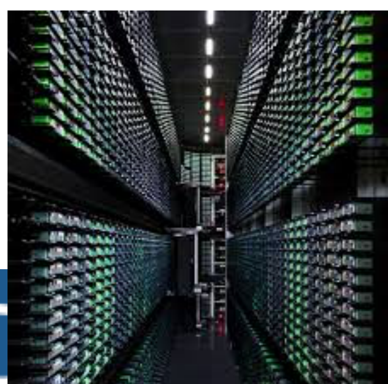
Memory-class storage

DISK



Storage-class memory

TAPE



Cloud-base
object store
public or private

Managing the many choices…

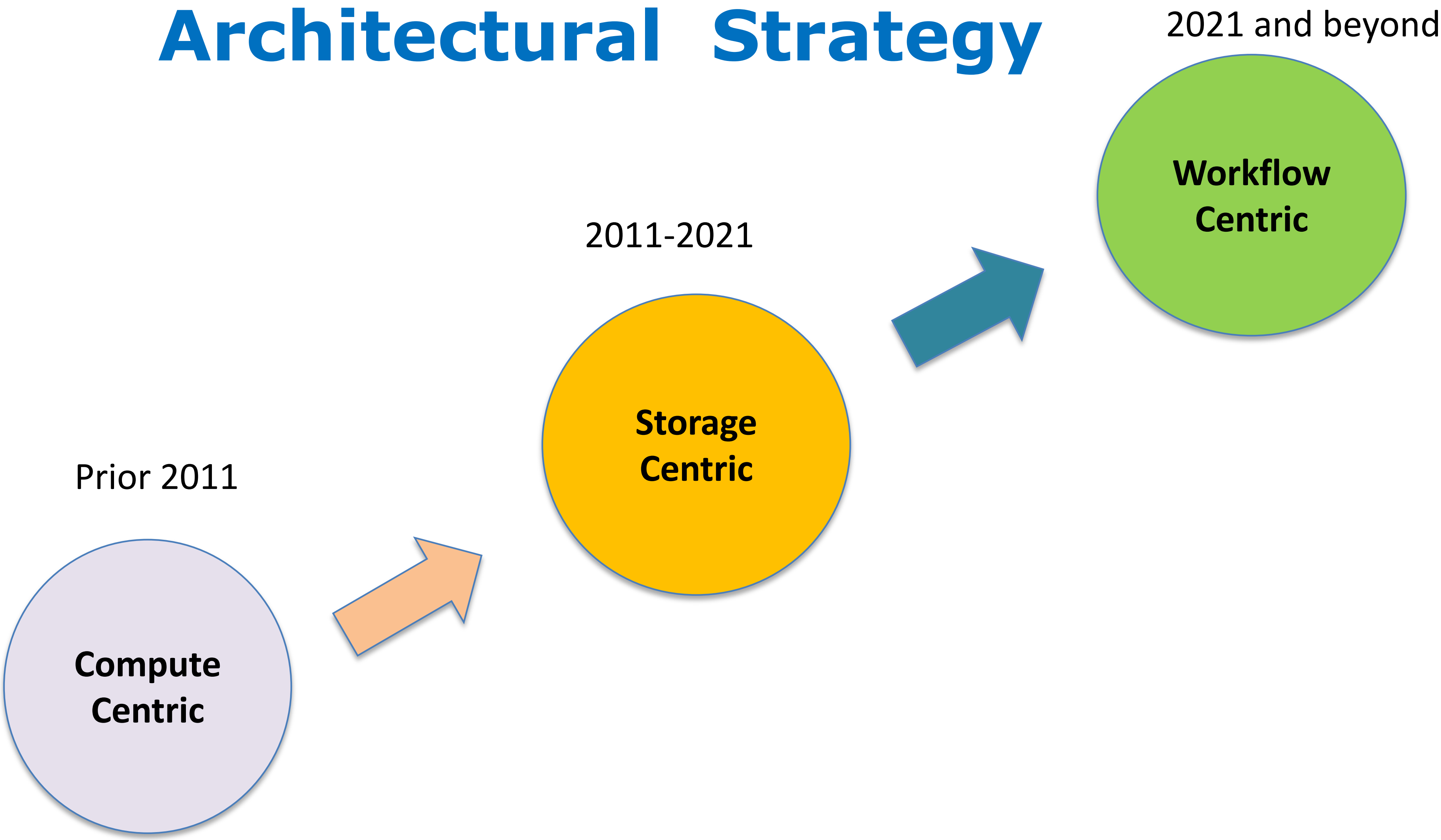
- **Compute Options**
  - **CPUs:** Intel XEON, IBM POWER9, ARM (Cavium ThunderX), AMD (Epyc)
  - **Coprocessors:** GPUs, FPGAs, NEC Aurora
  - **Network:** OPA V2 (Intel), Cray proprietary, Infiniband (HDR), Ethernet?
  - **Machine Learning:** Intel Nervana, NVIDIA tensor cores, Google TPUs
- **Memory & Storage**
  - **High-speed, stacked memory:** HBM2 & MCDATA
  - **High IOPS Non-volatile Storage:** (Optane (3D-xpoint), Flash-based SSDs)
- **Cloud Computing & Storage Considerations**
  - **On-Premise Cloud** (Containers)
  - **Commercial Cloud** (Bursting)
  - **HPC in the Cloud** (Hosting)

NWSC-3 Attributes

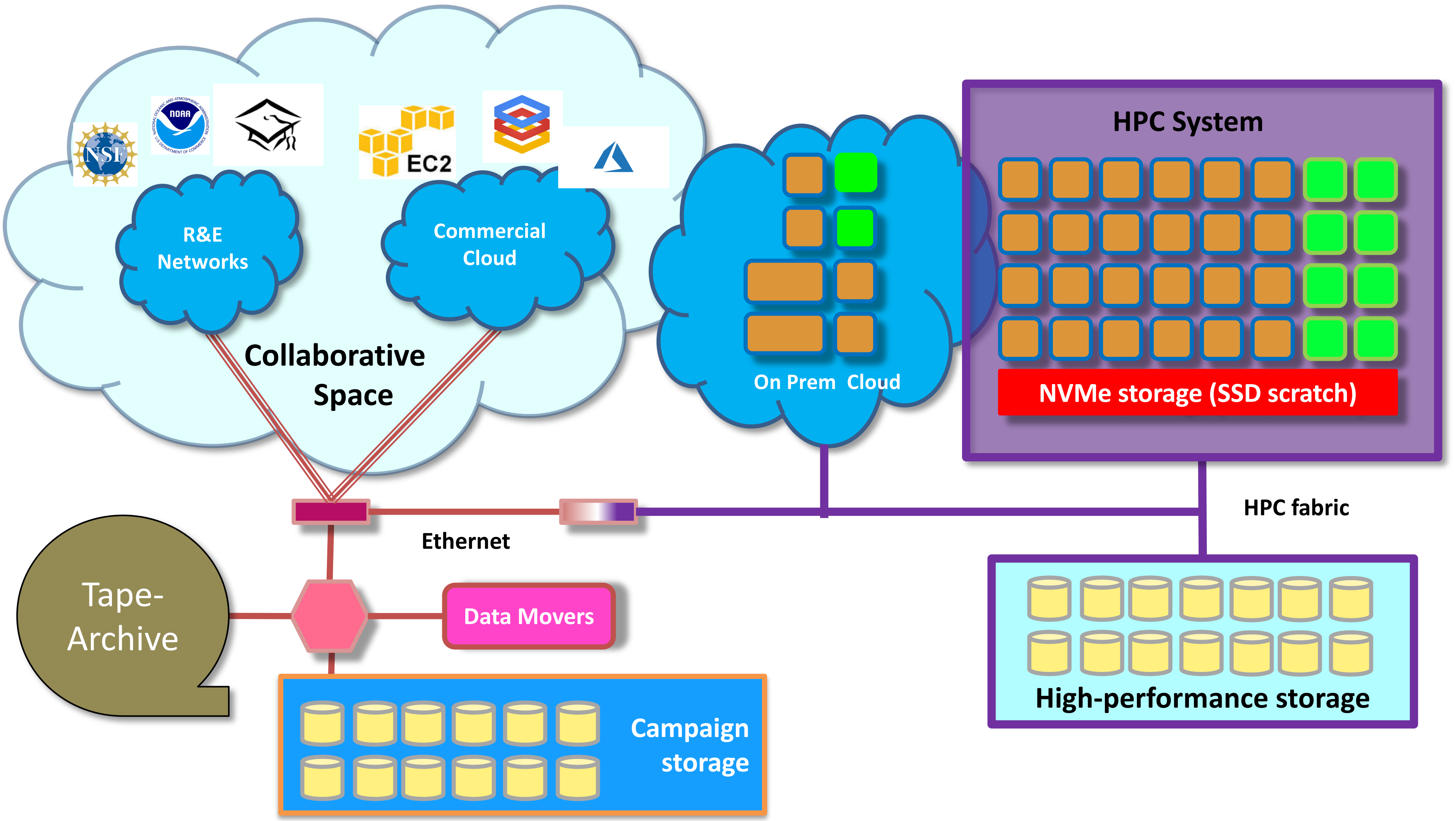
(Focus on a design that enhances the end-to-end rate of science throughput)

- ✓ **Heterogeneous hardware**
- ✓ **High IOPS Storage**
- ✓ **High Bandwidth Memory**
- ✓ **Application containers**
- ✓ **Cloud bursting capability**
- ✓ **Storage tiers**

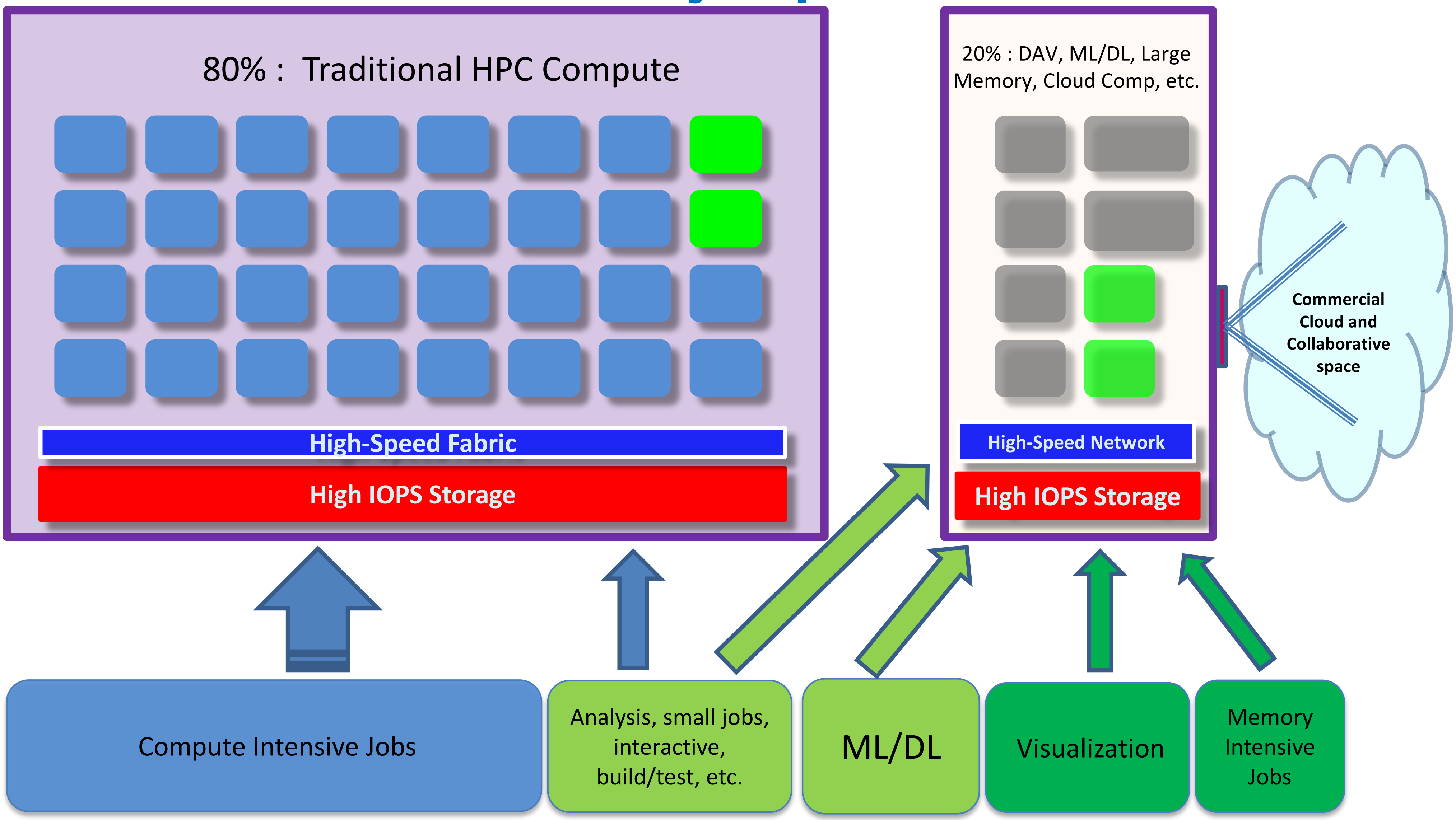
Evolution of HPC Architectural Strategy



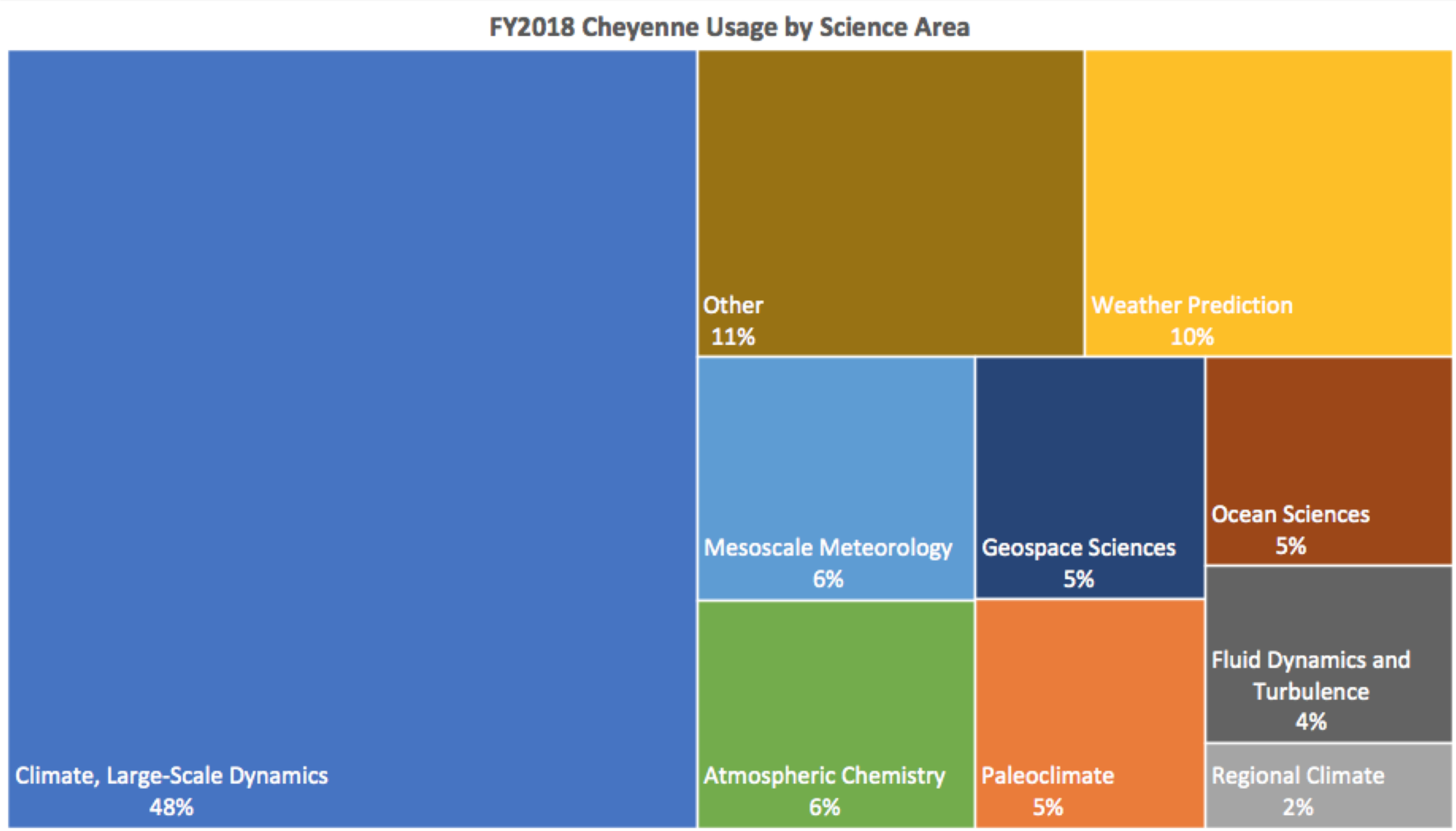
NWSC-3 Design Strawman



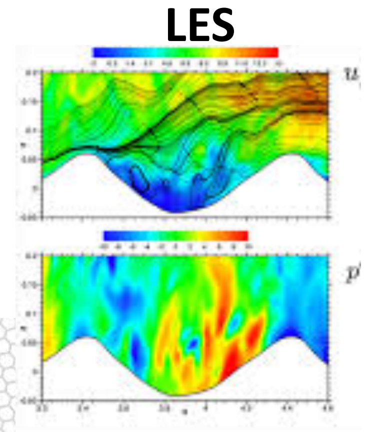
Workflow Centric Architecture
User defined job placement



Simulation Workload



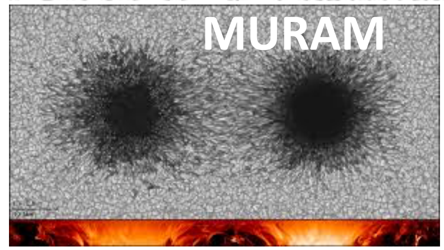
- **Climate** (Climate + Paleoclimate + Regional Climate) = 55% (~CESM)
- **Weather** (Weather + Meso Meteo + Atmos Chem) = 22% (~WRF/MPAS)
- **Geospace** = 5% (~MHD via MURAM)
- **Fluid Dynamics/Turbulence** =5% (~LES)



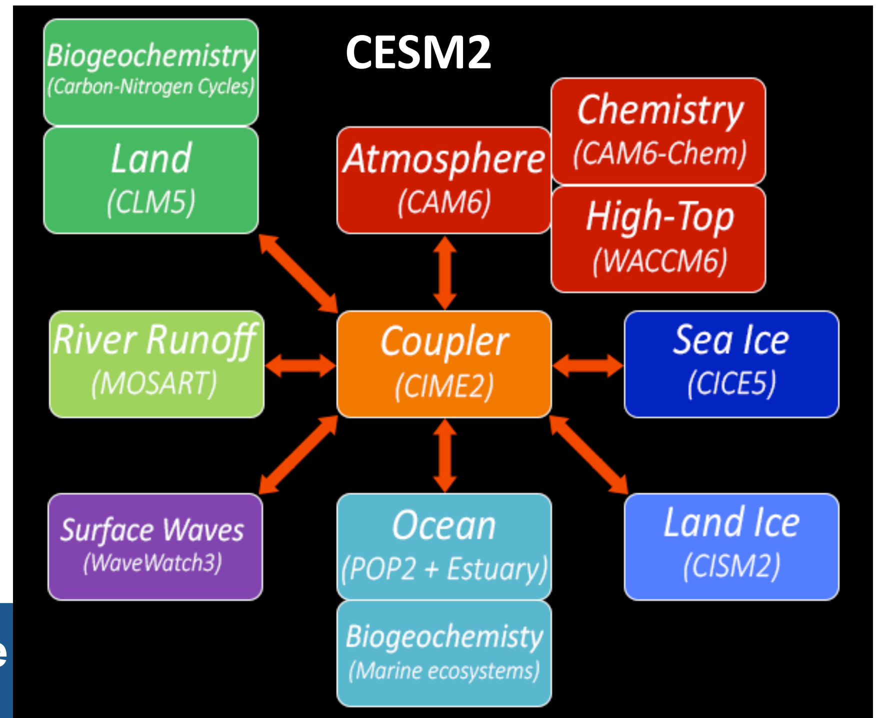
Application Strategy

- **Performance**
  - Must be able to track rapidly-changing future architectures
- **Portability**
  - Need to maintain flexibility in the choice of hardware (validate codes with several compilers)
- **Power savings**
  - TCO matters. Power efficiency of HPC-scale systems is a significant operational cost (consider GPUs where possible)

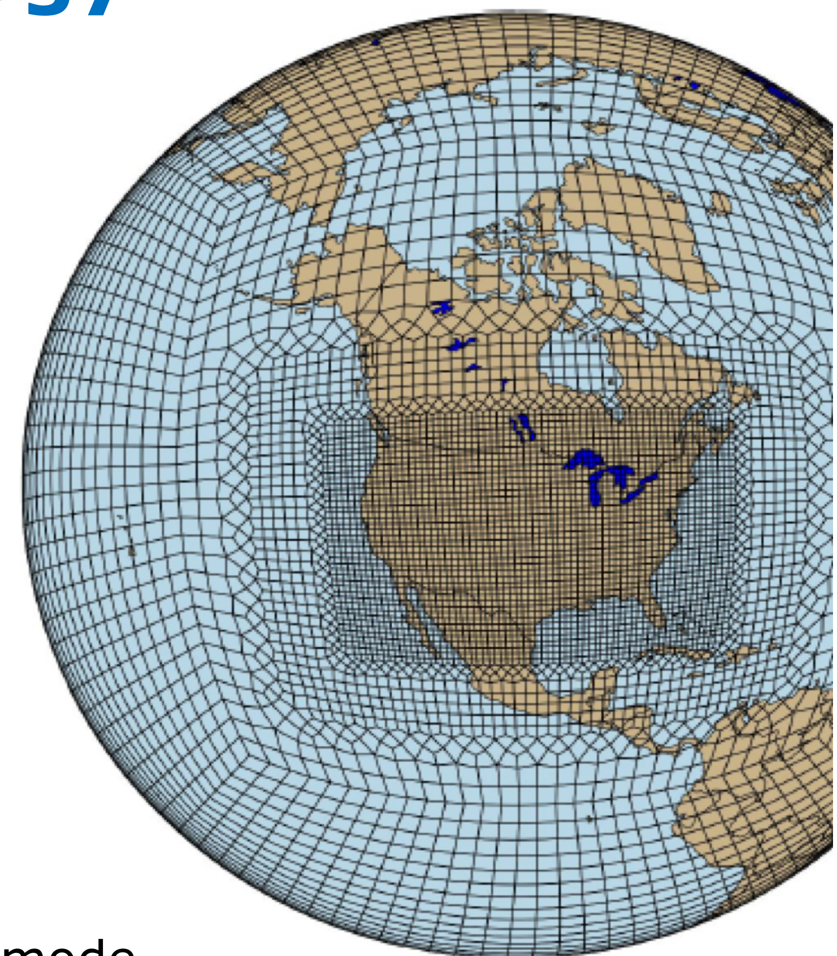
MPAS
Model for Prediction Across Scales



WRF
Weather Research & Forecasting



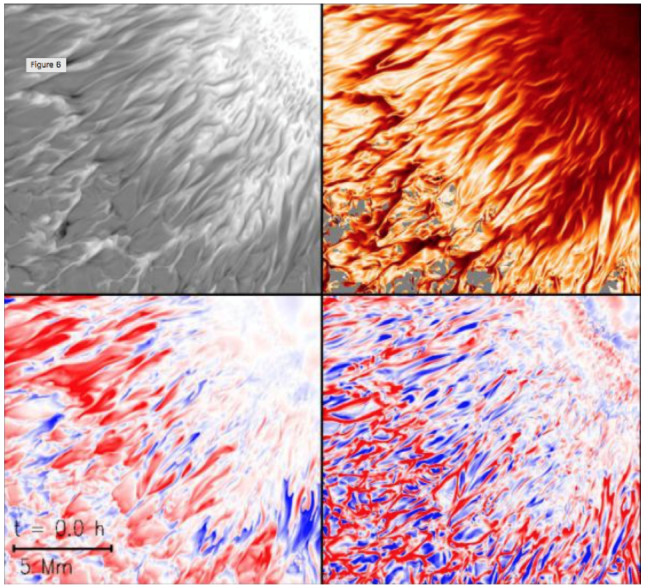
CESM Strategy



- **CESM2 code frozen during CMIP6**
  - CESM2 –1.6M LOC, 10,570 Subroutines
- **Compilers – ARM HPCC, ARM AOCC, Gnu, Cray, PGI, XLF and Intel**
  - Work with vendors to resolve port, validate, performance issues
  - **Compiler flexibility = procurement flexibility**
- **GPUs**
  - Strategy - Single Source Programming Model
    - CPU: OpenMP directives
    - GPU: OpenACC directives
  - Integrating MPAS DyCore into CESM
  - Considering port of MOM6 to GPUs
  - Ultimately expect to run CESM in “hybrid” mode (part of code on traditional nodes and part on accelerators).

Growing GPU-enabled workload

- **MPAS (MMM)**
  - Meteorological GCM
  - Fortran + OpenACC multi-GPU code
  - Showing 2:1 Xeon node to GPU device ratio
- **MURaM (HAO)**
  - A MHD radiative solar physics model.
  - C + OpenACC multi-GPU code.
- **Fast Eddy (RAL)**
  - LES model for microscale meteorology.
  - CUDA-based multi-GPU code + visualization



MURaM is used to study magneto-convection in sunspots. Courtesy Matthias Rempel, HAO

Influence on procurement: A portion of NWSC-3's HPC compute will likely have GPU accelerators.

Growing Machine Learning Portfolio
(ML that impacts HPC… "Big ML")

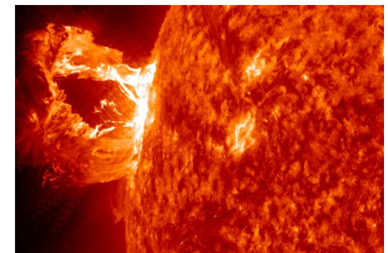
Why machine-learned emulation? The *per-core performance* of conventional computer architectures has stagnated, and models are getting *increasingly complex*. Replacing human-crafted modeling workflow components with machine learning algorithms may simplify, accelerate and improve them.

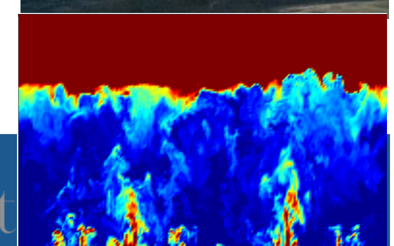
- **Interplanetary coronal mass ejection (CME) - Drs. Gibson & Flyer**
  - space weather prediction

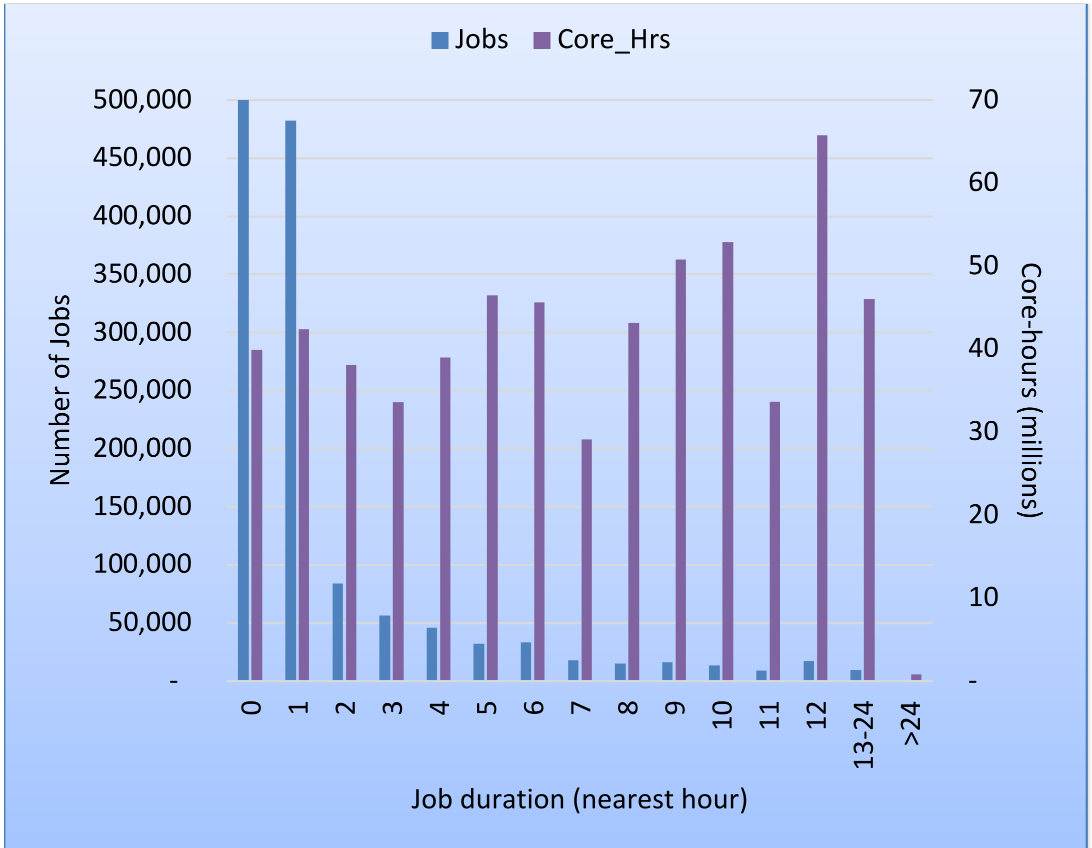
- **Seasonal weather patterns - Drs. Sobhani & DelVento**
  - Seasonal prediction of dangerous hot weather in the Eastern U.S.

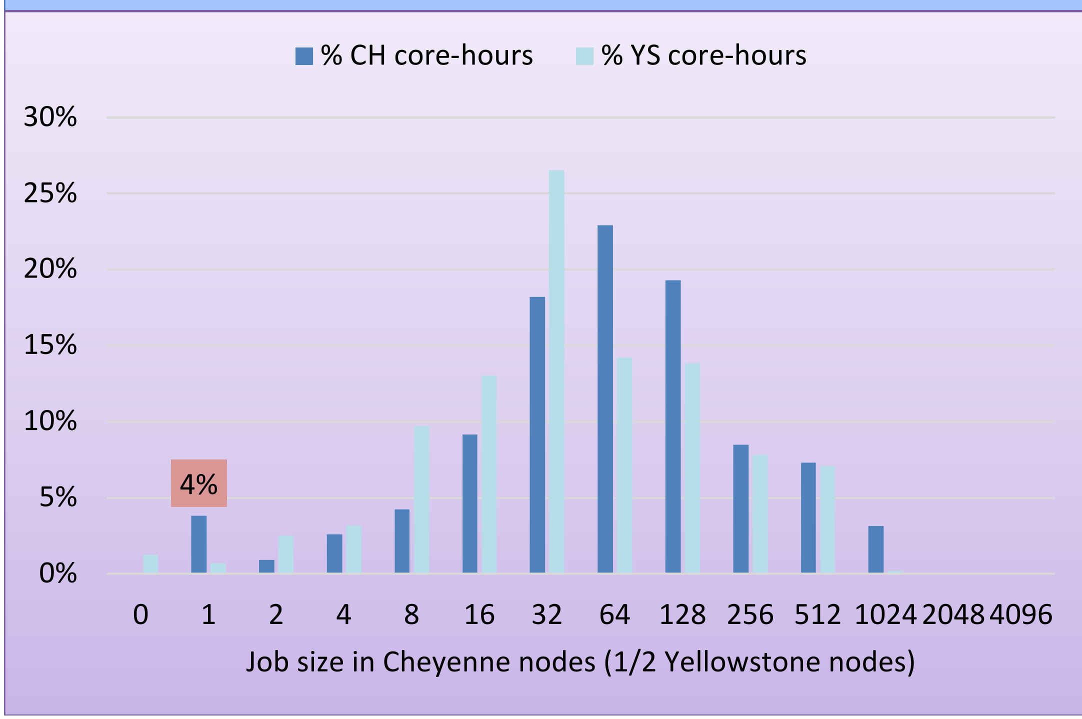
- **Cloud microphysics - Drs. Gettelman, Gagne & Sobhani**
  - improved weather and climate modeling

- **Sub-grid-scale turbulence -Drs. Kosovic, Haupt & Gagne**
  - improved representation of the surface layer in meteorological models







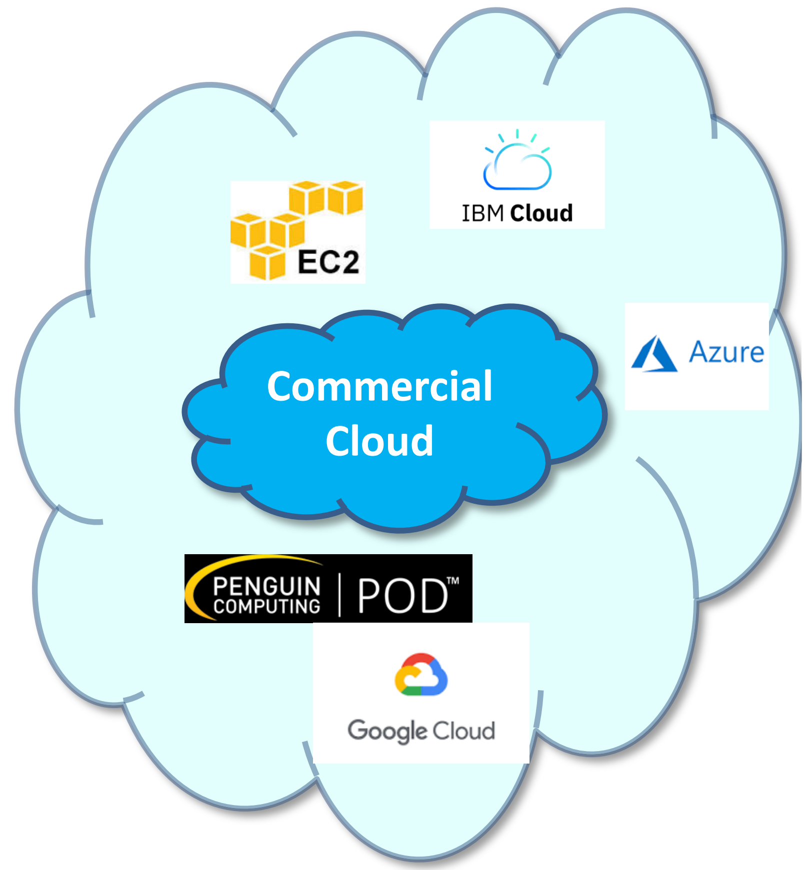


Growing Portfolio of Small Jobs

- **Pure “Capacity” Workload**
- **10.8M jobs (93% of all jobs) run <30mins**
  - but 7% of the core-hours
  - 3 projects responsible for 85% of these core-hours
- **4% of jobs single node**
- **Small jobs are creating jitter and noise (system instability) for larger runs**

Growing and Emerging Cloud Workload

- **Use Cases**
  – High Availability (Antarctic Forecast on Cloud when HPC down)
  – Bursting to cloud for urgent runs (e.g., Hurricane forecasting) via scheduler
  – General small job overflow for users willing to use more expensive allocation (e.g., WRF job swarms)
  – Larger role in future? (CESM2 1° running on AWS EC2 at 10 SYPD with I/O)
- **Developing Sample Containers to go with our applications**
  – On- or Off-Prem
  – Capturing complex environment needs
- **Data Discovery, Curation, Analytics and Storage – more to come**



NWSC-3 Benchmarking Strategy

- **Reduce Barrier of Entry for Vendors**
  - Lighter weight and easier-to-use process
  - Reduce number of full applications
  - Move to I/O benchmarking and important models (CESM) to Acceptance phase
- **Increase Use of 'mini-apps'**
  - Simplified benchmarks derived from important NCAR models
  - Reduced time to build and run
  - Representative of application performance characteristics
- **Benchmark Components**
  - Kernels and Synthetic (hardware characteristics)
  - Capability (limited number of "scalable science" apps)
  - Capacity (throughput – workload test)
  - Multi-GPU benchmark (new!)
  - Data Science and Deep Learning (*new!*)
- **Collaborate on mini-apps and synthetics - expect overlap with benchmarks at other centers (e.g., ECMWF)**

Technology Risk Mitigation for NWSC-3

- **Technology Partnerships:**
  - ARM Compiler, Chip and System (ARM HPC, Cavium, U. Bristol)
  - Application optimization (e.g. with Intel and nVIDIA)
  - Cloud-based access to benchmark systems (e.g. ReScale, AMD)
- **Recent Production Deployments**
  - SSD on Super
  - Campaign Store
  - CASPER (heterogeneous hardware, self integrated)
- **Systems Software**
  - SLURM deployment and testing new features
  - On-prem Cloud (Cumulus) coming soon
  - Containerization

Summary
NWSC-3 Procurement Strategy

- **Increase Vendor Pool**
  - Keeping pricing honest in procurements through competition: lightweight benchmarks, portable codes
  - Lower barrier of entry with easier-to-use benchmark suite.
  - Best value procurement
- **Create Environment to consider many solutions**
  - Retire as much technical risk as possible (hardware and system software)
  - Prepare Applications to handle wide-range of systems
  - Single or multiple-vendor options in RFP
- **Keep an eye on TCO (Total Cost of Ownership)**
  - Energy costs could be a deciding factor (consider GPUs where possible)

Thank you

Questions?

Procurement and Technology POC – Irfan Elahi (Irfan@ucar.edu)