

# **EXASCALE COMPUTING FOR NWP AND CLIMATE SCIENCE**

**Ilene Carpenter**

September 21, 2021





Simulations produce  
data

Models inferred from  
data

Insights from data

**Modeling and  
Simulation**

+

**Artificial  
Intelligence**

+

**Big Data  
Analytics**

=

**Exascale  
Era**

Running mission-critical workflows  
Exploiting diverse, high-volume, real-time data

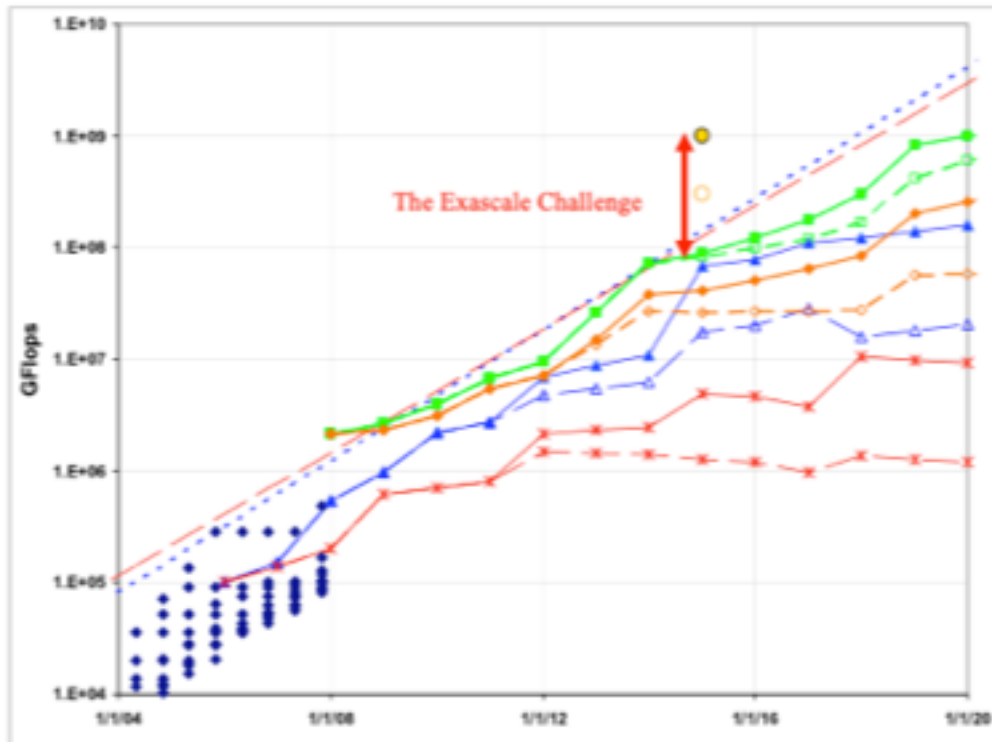
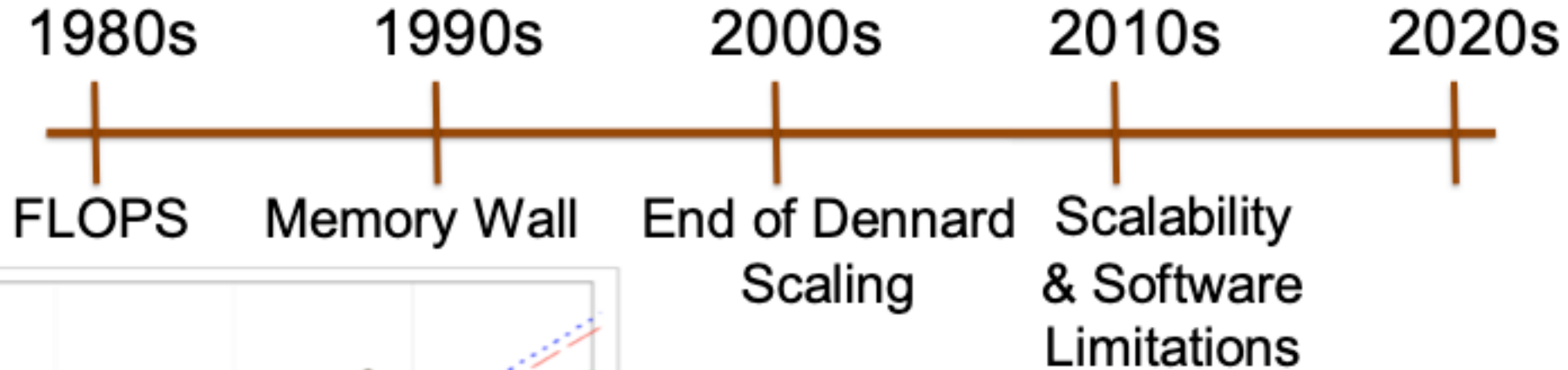




# Recent Eras of Computing

Slide from Mary Hall, "How Programming Systems Meet the Needs of New Architecture Classes", 2021 Salishan Conference

## Technology Trends Driving Each Decade of Innovation



Sarkar et al., Exascale Software Study, DARPA, 2009.

## Architecture Solutions

- Specialization to achieve efficiency
- Throughput-oriented systems
- Accelerators
- Memory and storage technologies

## Programming System Solutions

- Domain-specific languages/libraries/tools
- GPGPU
- Autotuning
- Compiler integration with libraries

# AGILITY

The key to success in the Exascale era

- Diversity, specialization, heterogeneity in processors and accelerators
- Hybrid (physics + AI) models and workflows
- Increased power consumption of systems that deliver more sustained performance
  - Exceeding many on-prem data center capabilities
- Significant changes in technology are rapid compared to traditional procurement cycles



**Need agile applications**



**Need agile procurement, aaS  
consumption**



# EARTH SCIENCES: WHY HPE?

## Experience

- Over 40 years of unrivalled experience in the industry
- Systems at a majority of the world's weather centers.
- Industry-unique deep bench of environmental science experts.

## Vision

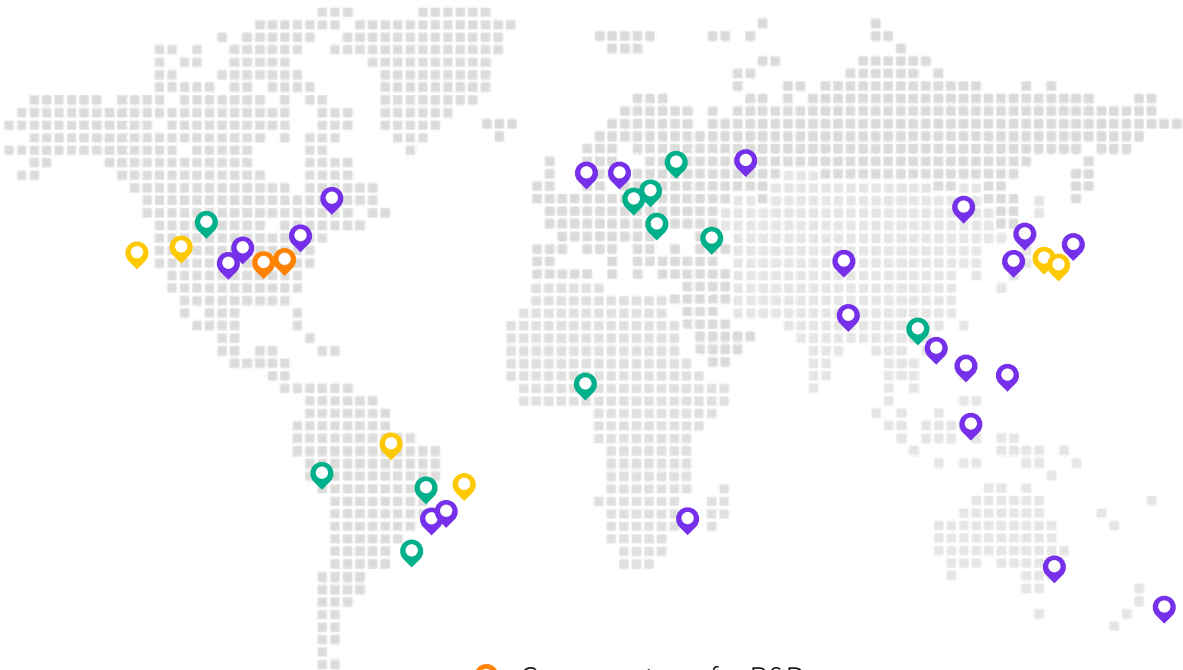
- The most advanced supercomputing technology & leading solution portfolio for the Exascale era
- Long-term customer partnerships





## Focus on TCO

- End-to-end solutions from HPE: hardware, software, interconnect, storage, cooling & services
- Range of consumption models

## Reliability & Performance

- World-leading software, accelerating time to results, reproducibility and resilience
- Advanced networks for **predictable runtimes, high throughput and excellent scalability**

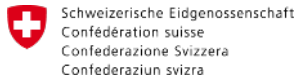
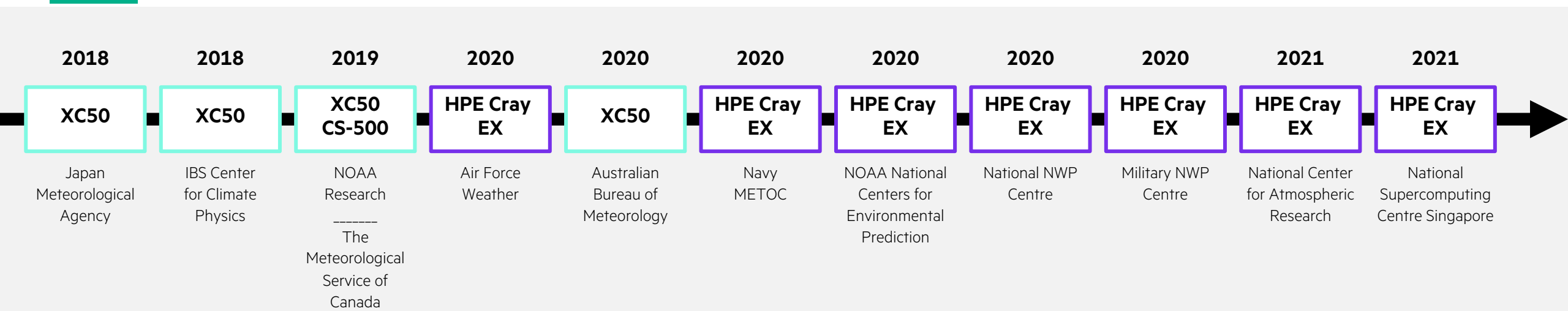


-  Cray - systems for R&D
-  HPE/SGI – systems for R&D
-  HPE/SGI – weather forecasting
-  Cray – weather for forecasting



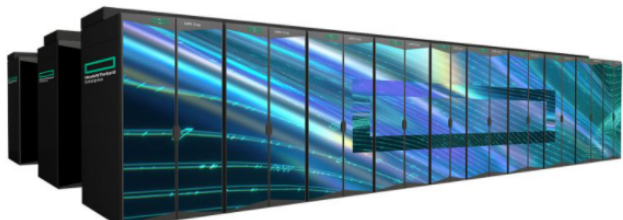


# NEW HPE SYSTEMS IN WEATHER, CLIMATE & OCEANOGRAPHY





# HPE CRAY EX SYSTEMS FOR WEATHER AND CLIMATE RESEARCH IN EUROPE



LUMI

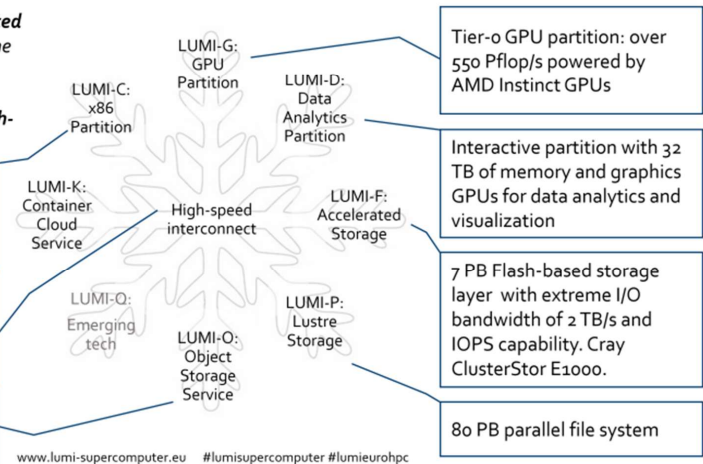
## LUMI, the Queen of the North

LUMI is a Tier-0 GPU-accelerated supercomputer that enables the convergence of high-performance computing, artificial intelligence, and high-performance data analytics.

- Supplementary CPU partition
- ~200,000 AMD EPYC CPU cores

Possibility for combining different resources within a single run. HPE Slingshot technology.

30 PB encrypted object storage (Ceph) for storing, sharing and staging data



AMD EPYC CPUs, AMD Instinct GPUs



## CSCS, HEWLETT PACKARD ENTERPRISE AND NVIDIA ANNOUNCE WORLD'S MOST POWERFUL AI-CAPABLE SUPERCOMPUTER

"Alps" system to advance research across climate, physics, life sciences with 7x more powerful AI capabilities than current world-leading system for AI on MLPerf.

April 12, 2021 - by CSCS

The Swiss National Supercomputing Centre (CSCS), **Hewlett Packard Enterprise** (HPE) and **NVIDIA** today announced that they are creating what is expected to be the world's most powerful AI-capable supercomputer.

Planned to come online in 2023, the 'Alps' system infrastructure will replace CSCS's existing Piz Daint supercomputer and serve as a general-purpose system open to the broad community of researchers in Switzerland and the rest of the world.

It will enable breakthrough research on a wide range of fields, including climate and weather, materials sciences, astrophysics, computational fluid dynamics, life sciences, molecular dynamics, quantum chemistry and particle physics, as well as domains like economics and social sciences.

Alps will be built by HPE based on the new HPE Cray EX supercomputer product line, which is a next-generation high performance computing (HPC) architecture designed from the ground up to efficiently harness insights from vast, ever-increasing amounts of complex data. It features the HPE Cray software stack for a software-defined supercomputing experience, as well as the NVIDIA HGX™ supercomputing platform, including NVIDIA GPUs, the NVIDIA HPC SDK and the new Arm-based NVIDIA Grace™ CPU, also announced today.

NVIDIA Grace CPUs and NVIDIA GPUs



# THE BEST OF BOTH WORLDS LEADS TO BEST-IN-CLASS ETHERNET

## Traditional Ethernet networks

- Ubiquitous and interoperable
- Broad connectivity ecosystem
- Broadly converged network
- Native IP protocol
- Efficient for large payloads only
- High latency
- Limited scalability for HPC
- Limited HPC features

## HPE Slingshot

- Standards-based/interoperable
- Broad connectivity
- Converged network
- Native IP Support
- Low latency
- Efficient for small to large payloads
- Full set of HPC features
- Very scalable for HPC and big data

## Traditional HPC Interconnects

- Proprietary (single vendor)
- Limited connectivity
- HPC interconnect only
- Expensive/slow gateways
- Low latency
- Efficient for small to large payloads
- Full set of HPC features
- Very scalable for HPC and big data

- Industry-leading performance and scalability
- Open Ethernet standards and protocols, plus optimized HPC functionality

- Innovative hardware-based congestion management, adaptive routing, and quality of service
- Application and fabric performance that competes with traditional HPC networks

# HPE CRAY EX SLINGSHOT: FEATURES FOR NWP WORKFLOWS

## NWP Characteristics

- Operational weather forecasting relies on **repeatable run times**
- **Complex workflows**, with many interrelated steps, are required to produce final forecast products
- Large HPC systems run multiple jobs at the same time, which may result in unpredictable run times due to use of **shared resources** like high-speed interconnect and storage systems
- Many NWP codes can be sensitive to **network congestion** and experience variable run times when network traffic exceeds certain levels

**HPE Cray EX  
Slingshot  
was  
designed to  
eliminate  
this  
hardware  
problem.**



## HPE Slingshot Congestion Management

- Hardware automatically tracks *all* outstanding packets
  - Knows what is flowing between **every** pair of endpoints
- Quickly identifies and controls causes of congestion
  - Pushes back on sources... *just enough*
  - Other traffic not affected and can pass stalled traffic
- Fast and stable across wide variety of traffic patterns
  - Suitable for dynamic HPC traffic*
- Performance isolation between apps on same QoS class
  - Applications much less vulnerable to other traffic on the network
  - Predictable runtimes
  - Lower mean *and tail* latency – a big benefit in apps with global synchronization

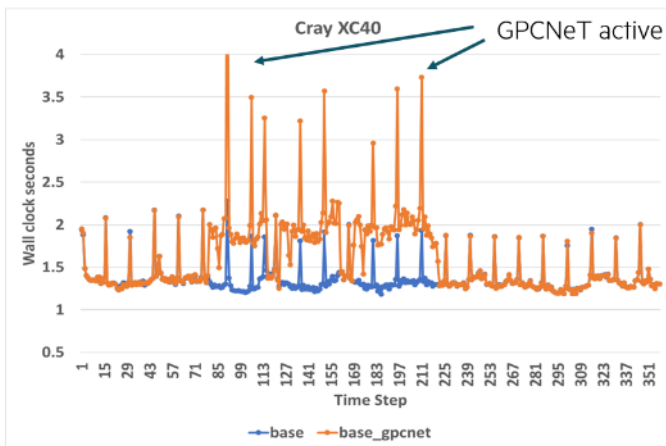


# HPE SLINGSHOT ADVANTAGE

The UM model and GPCNeT were run on two systems to compare network congestion effects on Unified Model (UM, UK Met Office code)

## Cray XC40 system

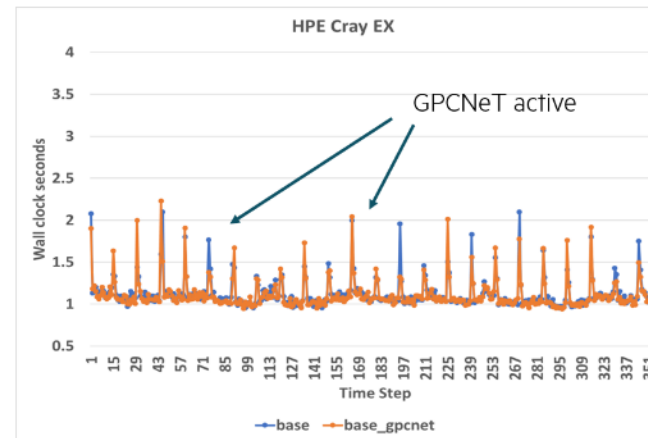
- 484 nodes, Intel Broadwell
- Cray Aries interconnect with Dragonfly topology
- 128 Gbytes memory per node



Network congestion is quite apparent on the XC40 system during the period when GPCNeT was running

## HPE Cray EX

- 800 nodes, AMD EPYC Rome
- HPE Slingshot interconnect with Dragonfly topology
- 256 Gbytes memory per node



**High rates of small messages does not cause network congestion and does not interfere with “normal” application communications on HPE Cray EX systems.**

# US AIR FORCE 557<sup>TH</sup> WEATHER WING

## First Shasta Supercomputer to Power Weather Forecasting

Hosted by Oak Ridge National Laboratory (ORNL):

- ORNL provides supercomputing-as-a-service on the HPC11 Shasta system to the Air Force 557th Weather Wing.
- The system enhances the Air Force's capabilities to create higher fidelity weather forecasts and weather threat assessments so that U.S. Air Force and Army missions can be carried out more effectively worldwide.

“The system's performance will be a significant increase over the existing HPC capability and will provide Air Force Weather operators with the ability to run the next generation of high-resolution, global and regional models, and satisfy existing and emerging warfighter needs for environmental impacts to operations planning.”

*Steven Wert, Program Executive Officer Digital, Air Force Life Cycle Management Center at Hanscom Air Force Base in Massachusetts*

### About the new system

#### 2 HPE Cray EX Supercomputers

8 cabinets in a Hall A/ Hall B configuration



- Each with 800 AMD Rome (64c) nodes and 256 GB/node

#### with HPE Slingshot interconnect

- 200 Gb/s bandwidth per port per direction, for congestion control



**HPE modular software stack:** HPE Cray OS, HPE Performance Cluster Manager + HPE Cray Programming Environment

Additional resources:



[Read the press release](#)



# NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR)

## 3.5 Faster and 6 x More Energy Efficient Supercomputer for Extreme Weather Research

NCAR needs more supercomputers powerful enough to support both complex physics-based modeling and machine learning algorithms to:

- Improve predictions of seasonal water supply, drought risk and flooding through detailed modeling and forecasting tools to inform water management experts, public utilities and farmers.
- Manage wildfire risk by simulating representations of physical processes in a region.
- Foresee hazards and impacts of climate change from extreme weather conditions, such as thunderstorms, tornadoes and hurricanes.
- Understand the dangers of solar storms using three-dimensional simulations to enable predictions of potential disruptions to the earth's atmosphere triggering space weather events that threaten communications systems and power grids.

“The resulting research will lead to new insights into potential threats ranging from severe weather and solar storms to climate change, helping to advance the knowledge needed for improved predictions that will strengthen society’s resilience to potential disasters.”

*Anke Kamrath, director, NCAR Computational and Information Systems Laboratory*

### About the new system

#### HPE Cray EX Supercomputer 19.87 Petaflops



- 2488 nodes with AMD EPYC CPUs
- 82 nodes with NVIDIA A100 GPUs

#### with HPE Slingshot v11 interconnect

- 200 Gb/s bandwidth per port per direction, for congestion control



- **HPE modular software stack:** HPE Cray OS, HPE Performance Cluster Manager + HPE Cray Programming Environment
- **Open Container initiative standard** – secure and flexible allocation of compute resources

#### Cray ClusterStor E1000



- 60 PB of storage

Additional  
resources:



[Read the press  
release](#)



[Watch  
video](#)

# UK MET OFFICE

## The World's Most Powerful Supercomputer is in the UK ... and ... in the Cloud

HPE, starting with Cray, has been servicing the Met Office, one of the world's largest weather forecasting agencies, for generations. Now Microsoft has partnered with HPE to take the Met Office's new **HPE Cray EX supercomputer to the cloud**.

HPE and Microsoft have had a supercomputing alliance through Cray in place for years, working towards this **supercomputing-as-a-service model**.

It's now coming to fruition – and at scale - in the Azure cloud, enabling the Met Office to work and act on the insights of their data using **AI, modeling, and simulation**.

"This investment by the UK government is a great vote of confidence in the Met Office's world-leading status as a provider of weather and climate science and services as well as in our national commitment to build a more resilient world in a changing climate, helping build back greener across the UK and beyond."

*Penny Endersby, Chief Executive, Met Office*



**HPE Cray EX Supercomputer**



**Cray ClusterStor E1000**

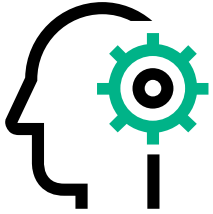
Additional resources:



[Read the press release](#)

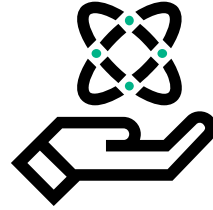


# AI FOR WEATHER FORECASTING



Rapid increase in the use of Machine Learning in the weather enterprise all through the forecast chain – where should I run it?

AI has different needs – Storage, software stack, sometimes processors



SmartSim

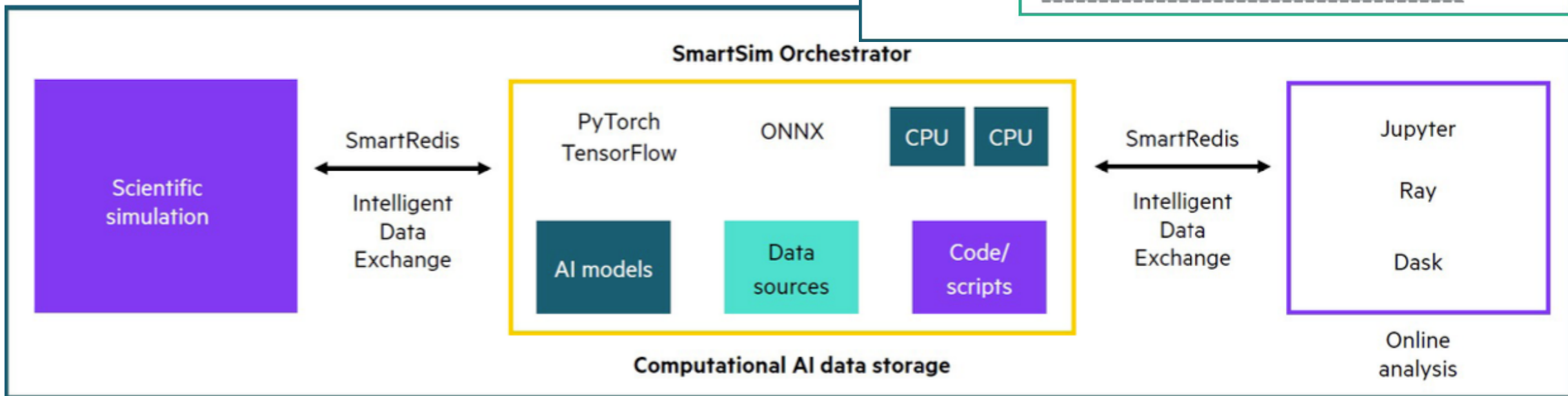
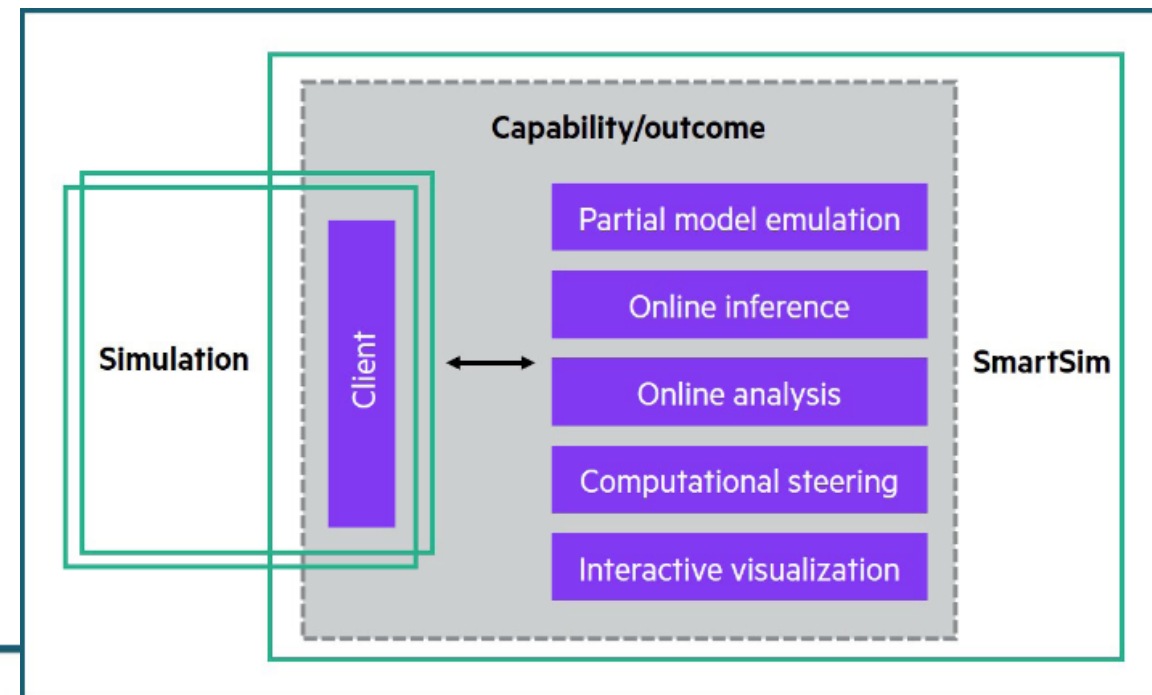
An open-source library to accelerate the convergence of AI, analytics and data science with HPC simulations. It enables the use of artificial intelligence (AI) within existing traditional high-performance computing (HPC) simulations by providing a new communication paradigm between simulations and AI methodologies.

- connect models written in Fortran, C, and C++ to the modern data science stack online
- write AI models with PyTorch, TensorFlow, scikit-learn, and such, and use them immediately inside scientific simulations at runtime
- scale to thousands of processors, all without writing to the file system.

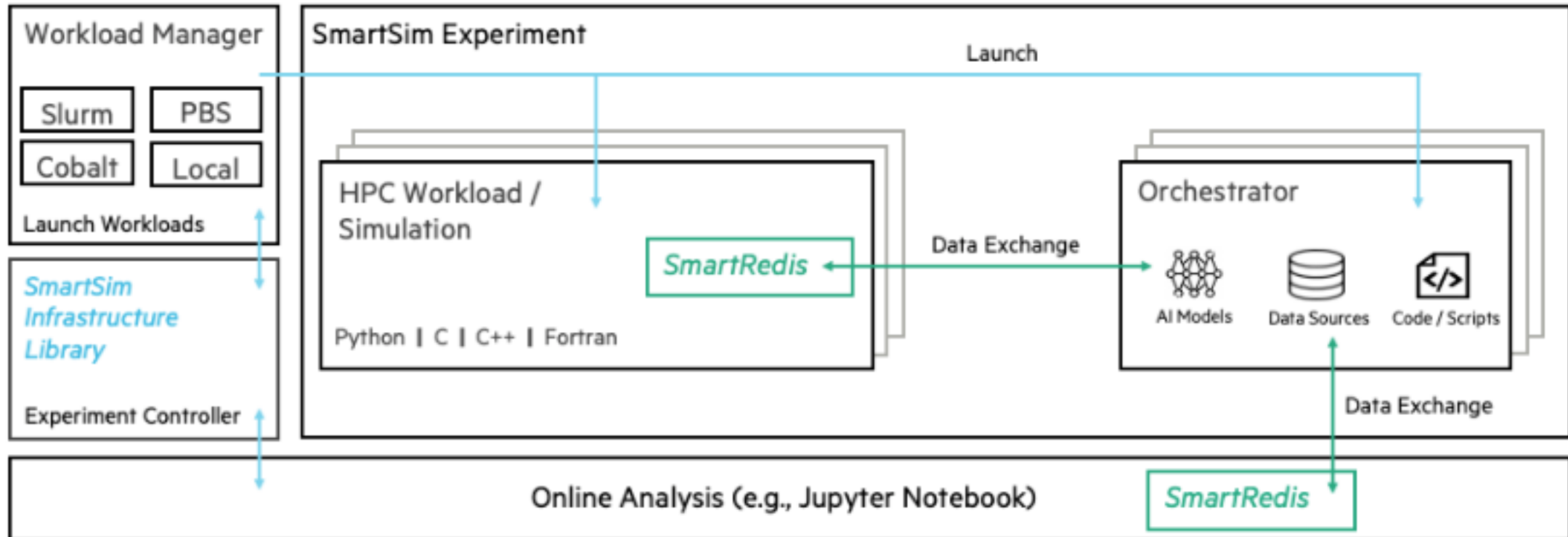
# DATA EXCHANGE BETWEEN SIMULATION AND IN-MEMORY DATABASE

Uses an in-memory database to transfer the data out of your simulation, it enables capabilities that can be used on the fly, at HPC scale, such as:

- AI model execution
- Online analysis
- Computational steering (start, stop, and reset of a simulation)
- Near continuous training of AI models (training and AI model as new data from the simulation comes in)
- Interactive visualization



# SmartSim

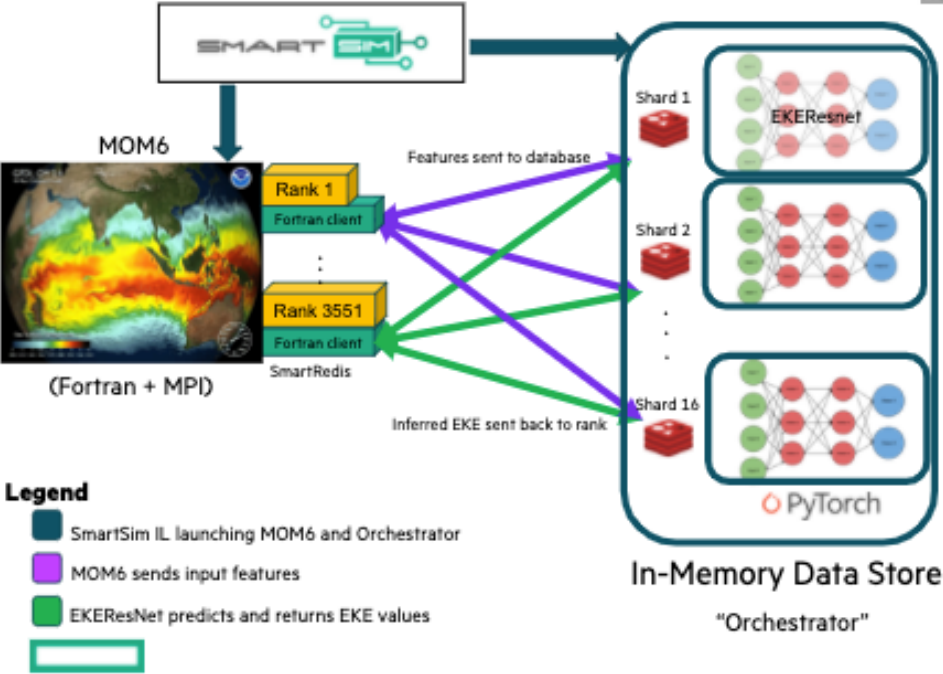


- execute, monitor and analyze simulations in a Jupyter notebook
- execute ML models hosted within in-memory storage (DRAM) for online inference, on CPU- or GPU-enabled nodes adjacent to or co-located within the simulation
- no communication via a filesystem
- ML-library agnostic - enables simulations written in Fortran to use tools in the rapidly evolving data science ecosystem



# SMARTSIM WITH MOM6

## COLLABORATION WITH NCAR



### Replacing the MEKE Parameterization

- **Goal:** Augment MOM6 MEKE parameterization with Machine Learning (ML) surrogate model
- Augmented Simulation:
  - **Modular Ocean Model 6**
  - GFDL's OM4 1/4 configuration (Adcroft et al., 2019)
- Training Data simulation
  - 1/10° nominal resolution (tx0 1, same grid used in POP)
  - Ocean (MOM6) and sea ice (CICE5) components
- AI Model
  - Modified ResNet – **EKEResnet**
- Input Features
  - *MKE\_sfc, Slope\_z, Rd\_dx\_z, Rel\_vort\_sfc*
- Inference
  - 16 GPUs – 16 copies of EKEResnet
  - MOM6 – 3551 ranks per model

18

### Using Machine Learning at Scale in HPC Simulations with SmartSim: An Application to Ocean Climate Modeling

Sam Partee  
Hewlett Packard Enterprise  
Seattle, WA  
spartee@hpe.com  
<https://orcid.org/0000-0001-6005-5116>

Matthew Ellis  
Hewlett Packard Enterprise  
Seattle, WA  
matthew.ellis@hpe.com  
<https://orcid.org/0000-0002-5782-5447>

Alessandro Rigazzi  
Hewlett Packard Enterprise  
Switzerland  
alessandro.rigazzi@hpe.com  
<https://orcid.org/0000-0003-2132-7726>

Scott Bachman  
National Center for Atmospheric Research  
Boulder, CO  
bachman@ucar.edu  
<https://orcid.org/0000-0002-6479-4300>

Gustavo Marques  
National Center for Atmospheric Research  
Boulder, CO  
gmarques@ucar.edu  
<https://orcid.org/0000-0001-7238-0290>

Andrew Shao  
University of Victoria  
Victoria, CA  
ashao@uvic.ca  
<https://orcid.org/0000-0003-3658-512X>

Benjamin Robbins  
Hewlett Packard Enterprise  
Seattle, WA  
benjamin.robbins@hpe.com

**Abstract**—We demonstrate the first climate-scale, numerical ocean simulations improved through distributed, online inference of Deep Neural Networks (DNN) using SmartSim. SmartSim is a library dedicated to enabling online analysis and Machine Learning (ML) for traditional HPC simulations. In this paper, we detail the SmartSim architecture and provide benchmarks including online inference with a shared ML model on heterogeneous HPC systems. We demonstrate the capability of SmartSim by using it to run a 12-member ensemble of global-scale, high-resolution ocean simulations, each spanning 10 complete nodes, all communicating with the same ML architecture at each simulation timestep. In total, 970 billion inferences are collectively served by running the ensemble for a total of 120 simulated years. Finally, we show our solution is stable over the full duration of the model integration, and that the inclusion of machine learning has minimal impact on the simulation runtimes.

#### 1. INTRODUCTION

Advances in machine-learning (ML) algorithms have spurred research and development for combining data-driven approaches and traditional numerical simulations to improve both efficiency and accuracy. The codebases of these numerical models are typically written in Fortran/C++ and run on high-performance computing platforms (HPC) via OpenMP and/or MPI parallelization. New software solutions are thus needed to connect these compiled language codebases to rapidly evolving ML and data analytics libraries, typically written in Python. Currently, the diversity of programming languages,

<sup>1</sup>All of the source code, datasets, and models for experiments described in this work are open source and publicly available for download at <https://github.com/CSG-CityLab/ML-ML-ML>

dependence on file input/output (IO), and large variance in compute resource requirements for scientific applications makes it difficult to perform analysis, training, and inference with most ML and data analytics packages at the scale needed for HPC numerical simulations.

On its surface, the problem of being able to interface HPC applications with ML libraries is one of language interoperability and software interface design. However, for the full convergence of these two disparate paradigms, the true difficulty (and opportunity) in bridging these workloads needs to be reformulated in terms of data exchange. That is, how is data passed between a simulation and ML model at scale while making efficient use of heterogeneous computational resources? Current approaches to addressing this problem can be roughly broken down into two categories: offline (the ML and numerical components of a simulation do not exchange data directly) and online (the ML component is called while the simulation is running). Note that in this work, this definition of "online" pertains to the process of inferring from a trained machine learning model, not continuously updating ML model parameters which is sometimes referred to as "online learning".

To illustrate the differences between online and offline approaches, we review recent studies that couple ML and numerical models with a focus on computational fluid dynamics (CFD) and climate modeling domains due to the application presented in this work.

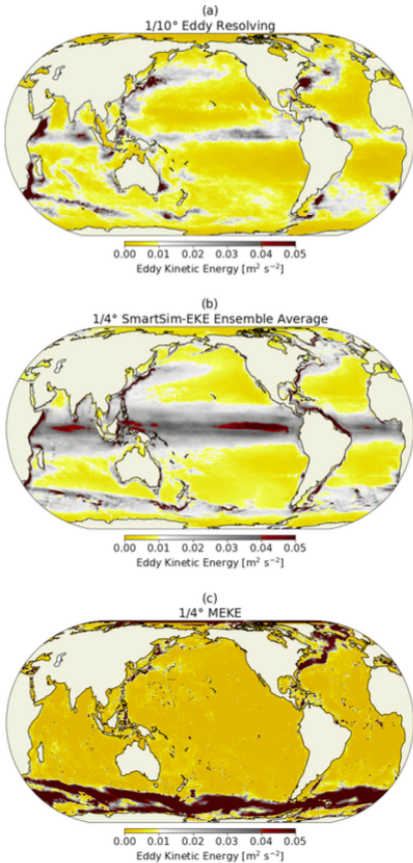


Fig. 7. Eddy kinetic energy (EKE), averaged over the last year of each simulation, calculated from the eddy-resolving (ER) 1/10° simulation (a), inferred online using EKEResNet (referred to as SmartSim-EKE), averaged over all 12 ensemble members (b), and the current state-of-the-art MEKE parameterization (c). Both SmartSim-EKE and MEKE use a 1/4° grid which is slightly coarser than the factor of 2 coarsening shown in (a); the ER EKE thus represents a lower-bound on what the 'true' EKE should be at that resolution.

arXiv:2104.09355v1 [cs.CE] 13 Apr 2021

# HPE CRAY PROGRAMMING ENVIRONMENT ADVANTAGES

Gold standard in HPC—Technology for real-life applications, not just benchmarks

| Performance and programmability  | Fully integrated heterogeneous optimization capability   | Integration with debuggers for performance optimization   | Focus on application portability and investment protection  |
|--|--|---|---|
| <ul style="list-style-type: none"><li>• Automatic optimizations deliver performance for a new target through a simple recompile.</li><li>• Automatically exploits the scalar, vector, and multithreading hardware capabilities of the systems.</li><li>• Compiler optimization feedback for application tuning</li></ul> | <ul style="list-style-type: none"><li>• Providing consistency across all HPE Cray systems</li><li>• Supporting x86-64 (both Intel and AMD) processors, Arm-based processors, and NVIDIA accelerators</li></ul> | <ul style="list-style-type: none"><li>• Parallel debuggers: Rogue Wave TotalView and Arm DDT</li><li>• For advanced debugging, performance analysis and optimization tools additional insights into compiler needs.</li></ul> | <p>Focus on compliance and language support:</p> <ul style="list-style-type: none"><li>• Languages: Fortran, C/C++, UPC and PGAS</li><li>• Programming models: OpenMP and OpenACC</li></ul> |

HPE-supported compiler for AMD-based systems.

# Where are we heading?

## Exascale Capability

- + **Compute Heterogeneity**
- + **Workload Diversification**
- + **Data-centric design**
- + **As-a-service**