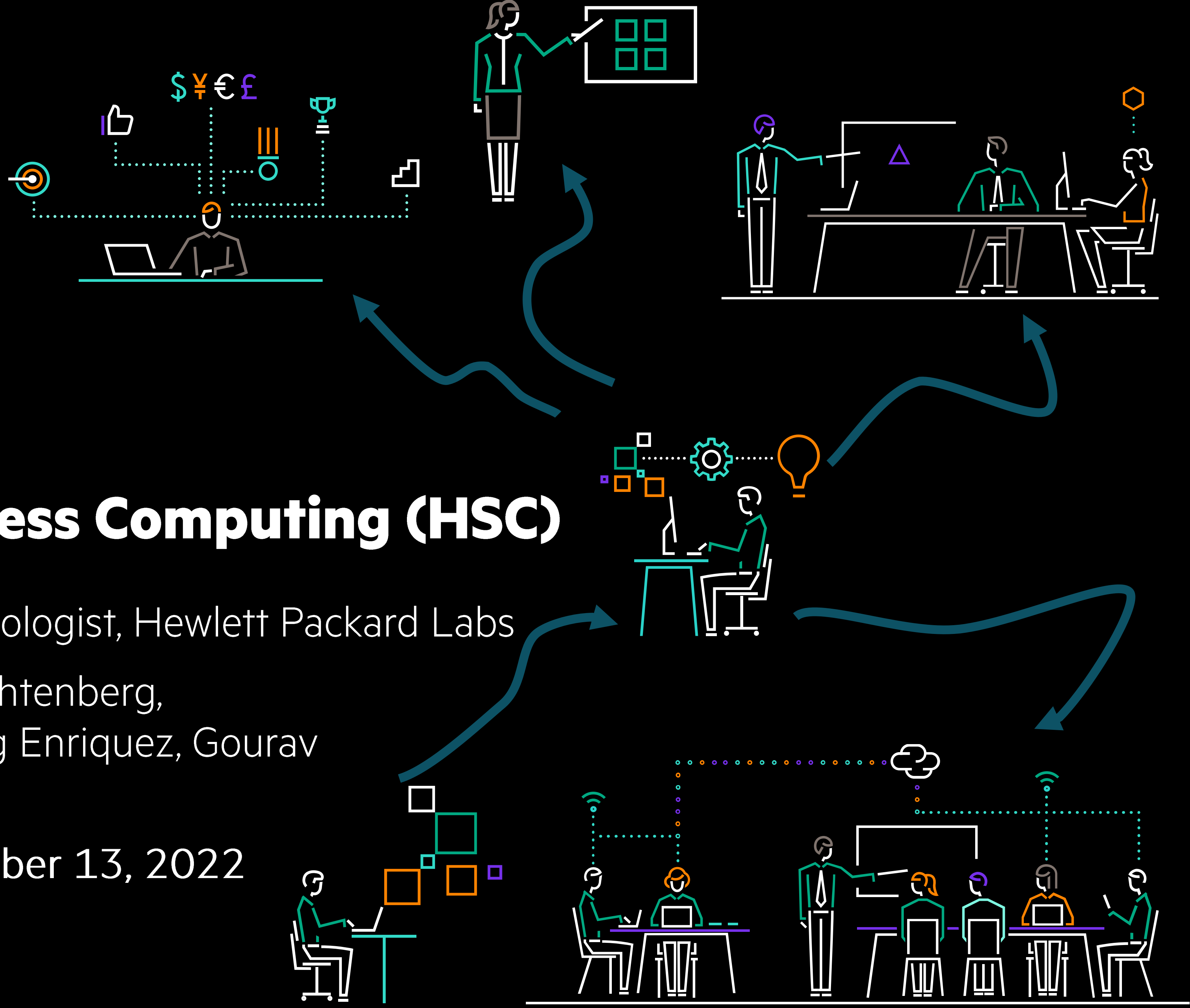
Hewlett Packard
Labs

Heterogeneous Serverless Computing (HSC)

Dejan Milojicic, Distinguished Technologist, Hewlett Packard Labs

Work with Aditya Dhakal, Eitan Frachtenberg, Ninad Hogade, Rolando Pablo Hong Enriquez, Gourav Rattihalli, Tobias Pfandzelter

ROSS Workshop at SC’22, November 13, 2022



OAK RIDGE NATIONAL LABORATORY'S FRONTIER SUPERCOMPUTER

- 74 HPE Cray EX cabinets
- 9,408 AMD EPYC CPUs, 37,632 AMD GPUs
- HPE Slingshot 11 interconnect
- 700 petabytes of storage capacity, peak write speeds of 5 terabytes per second using Cray ClusterStor Storage System

TOP500

ORNL's Frontier supercomputer is #1 on the TOP500.

1.1 exaflops of performance on the May 2022 Top500 list.

GREEN500 *

ORNL's Frontier's TDS supercomputer is #1 on the GREEN500.

62.68 gigaflops/watt power efficiency.

HPL-AI

ORNL's Frontier supercomputer is #1 on the HPL-AI list.

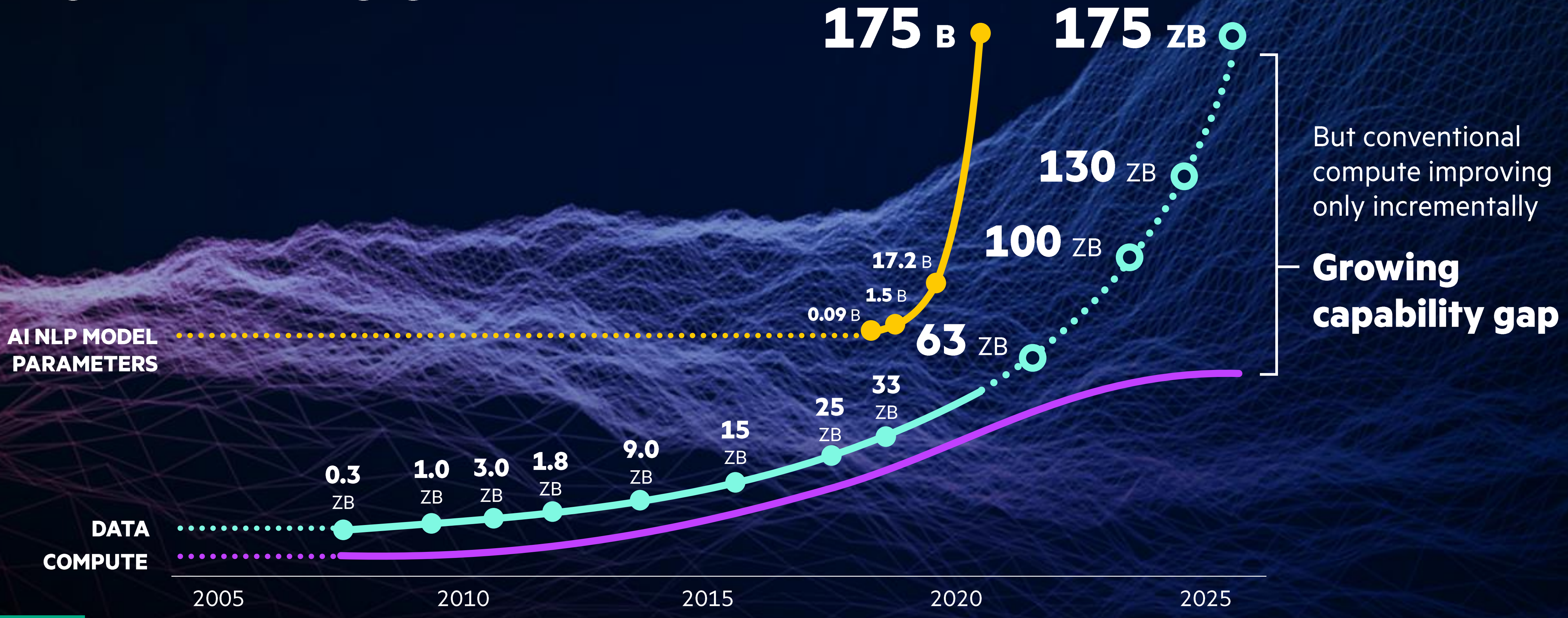
6.88 exaflops on the HPL-AI benchmark.

Sources: May 30, 2022 Top500 release

* Frontier is number 2 on the Top500 Green List

THE NEW NORMAL: COMPUTE IS NOT KEEPING UP

EXPONENTIALLY INCREASING DATA × EXPLODING DATA SOURCES × SHRINKING TIME TO ACTION = Massive advances in computing power **NEEDED EVERYWHERE**



But conventional compute improving only incrementally

Growing capability gap

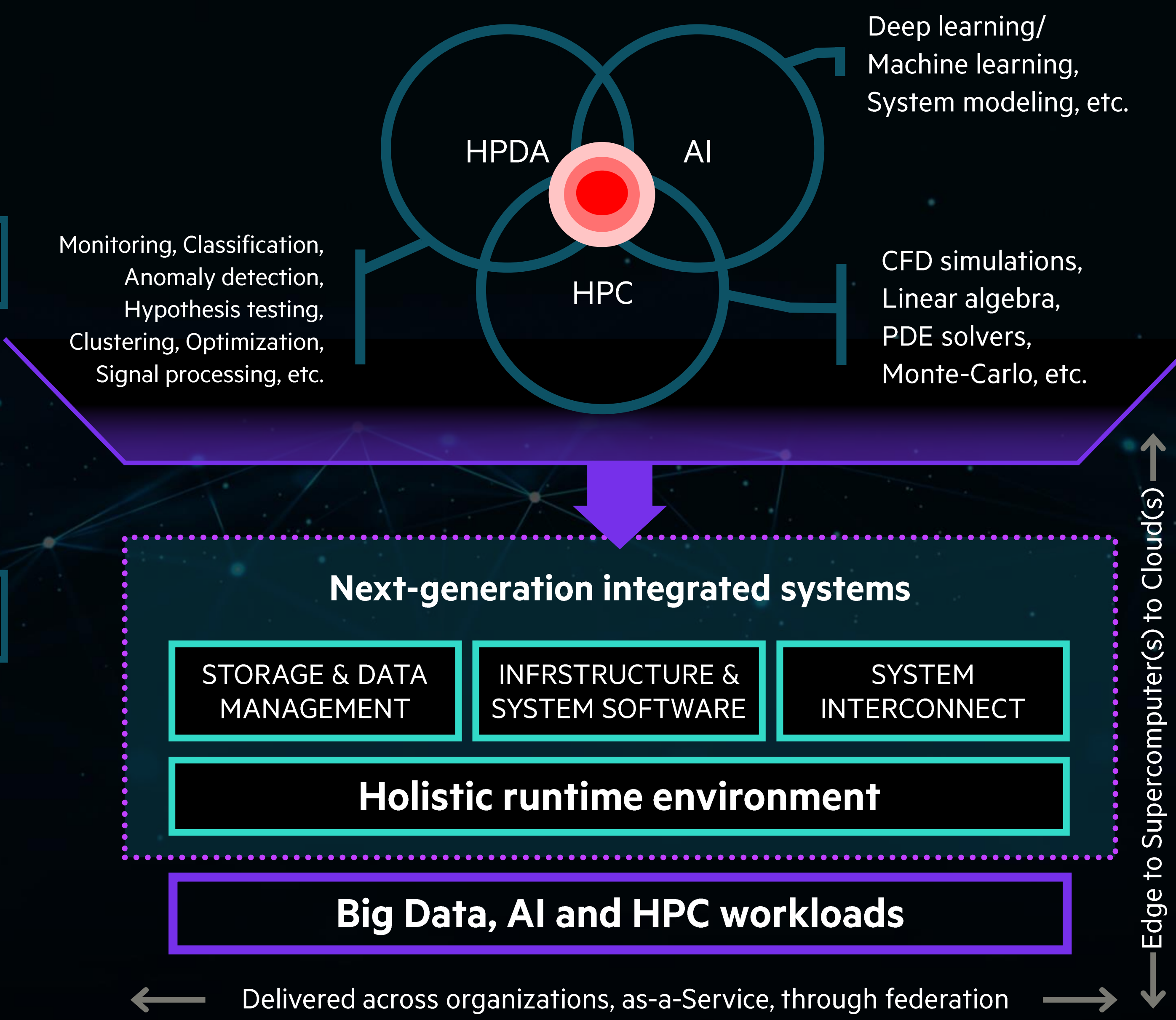
A HETEROGENEOUS FUTURE

Trends in Computing

- Rise of non-x86 processing…
- The “Cambrian explosion” – enhanced performance through specialization
- Efficient use of resources

Key Requirements to Enable Heterogeneity

- **FLEXIBLE -** System architecture that allows us to assemble and “compose for purpose”
- **STREAMLINED -** Easy to incorporate new silicon into our workflows
- **ABSTRACT –** Programming environment to abstract specialized ASICs to accelerate workload productivity



Deep learning/
Machine learning,
System modeling, etc.

HPDA

AI

HPC

Monitoring, Classification,
Anomaly detection,
Hypothesis testing,
Clustering, Optimization,
Signal processing, etc.

CFD simulations,
Linear algebra,
PDE solvers,
Monte-Carlo, etc.

Next-generation integrated systems

STORAGE & DATA MANAGEMENT

INFRSTRUCTURE & SYSTEM SOFTWARE

SYSTEM INTERCONNECT

Holistic runtime environment

Big Data, AI and HPC workloads

Delivered across organizations, as-a-Service, through federation

Edge to Supercomputer(s) to Cloud(s)

CONTINUOUSLY ADAPTING TO TOMORROW'S PACE OF CHANGE

RESOURCE COMPONENTS

MEMORY

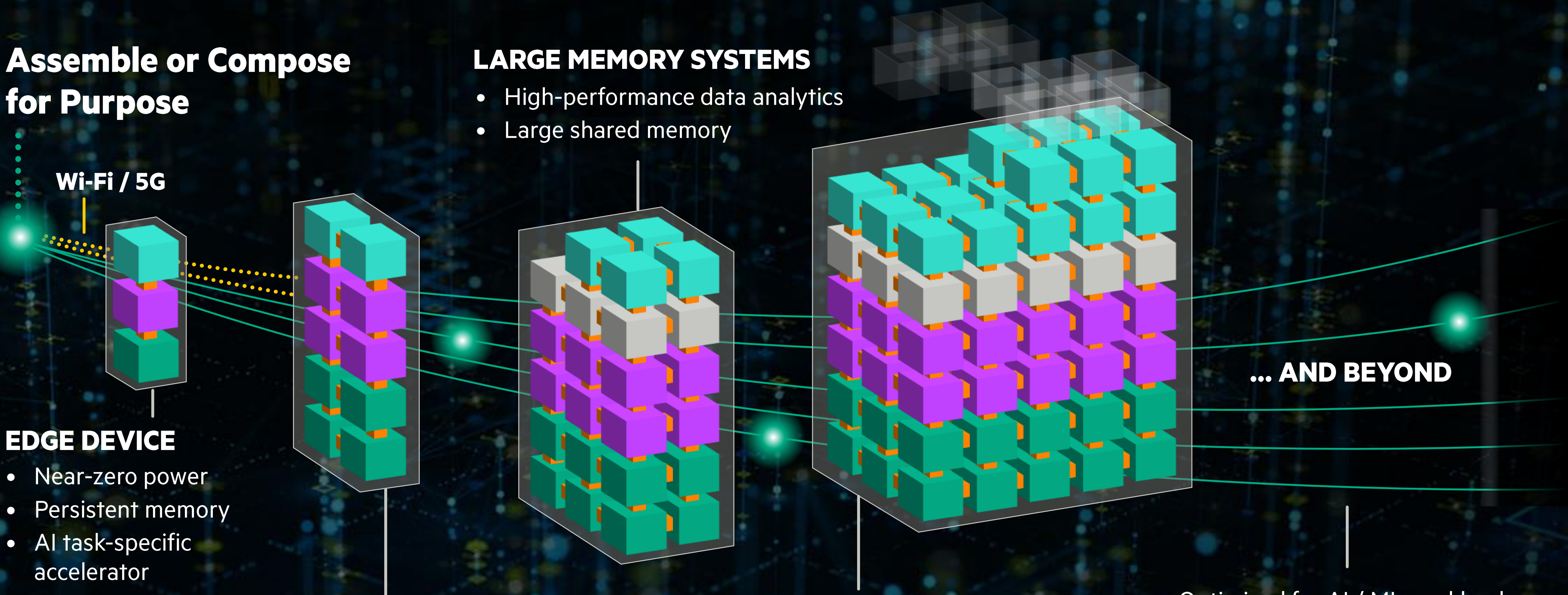
COMPUTE

DATA

ACCELERATOR

INTERCONNECT

Assemble or Compose for Purpose



LARGE MEMORY SYSTEMS

- High-performance data analytics
- Large shared memory

EDGE DEVICE

- Near-zero power
- Persistent memory
- AI task-specific accelerator

CLOUD INFRASTRUCTURE

- Composable infrastructure from every edge to any cloud
- Microservices in microseconds at massive scale

EXASCALE SYSTEMS

- 100,000+ components
- Ultra-fast message passing and checkpointing
- 20x more energy-efficient than state-of-the-art

... AND BEYOND

- Optimized for AI / ML workloads
- Quantum computing

PITTSBURGH
SUPERCOMPUTING
CENTER

DEMOCRATIZING AI TO SOLVE THE WORLD’S BIGGEST PROBLEMS

Neocortex high-performance AI system under development

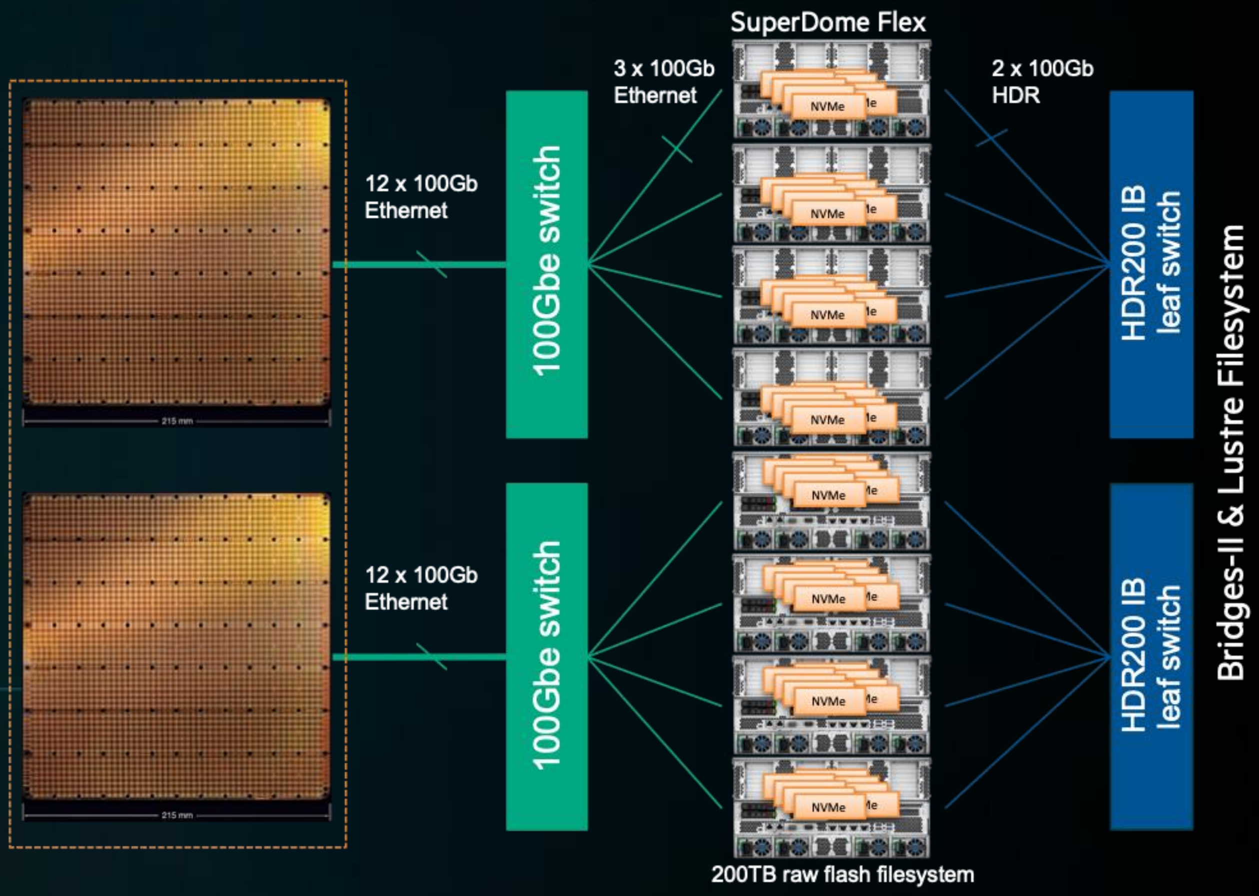
to democratize access for researchers to game-changing compute power for training

Pathfinding

a new approach to building an unconventional architecture consisting of a large system powered by extreme accelerators

cerebras

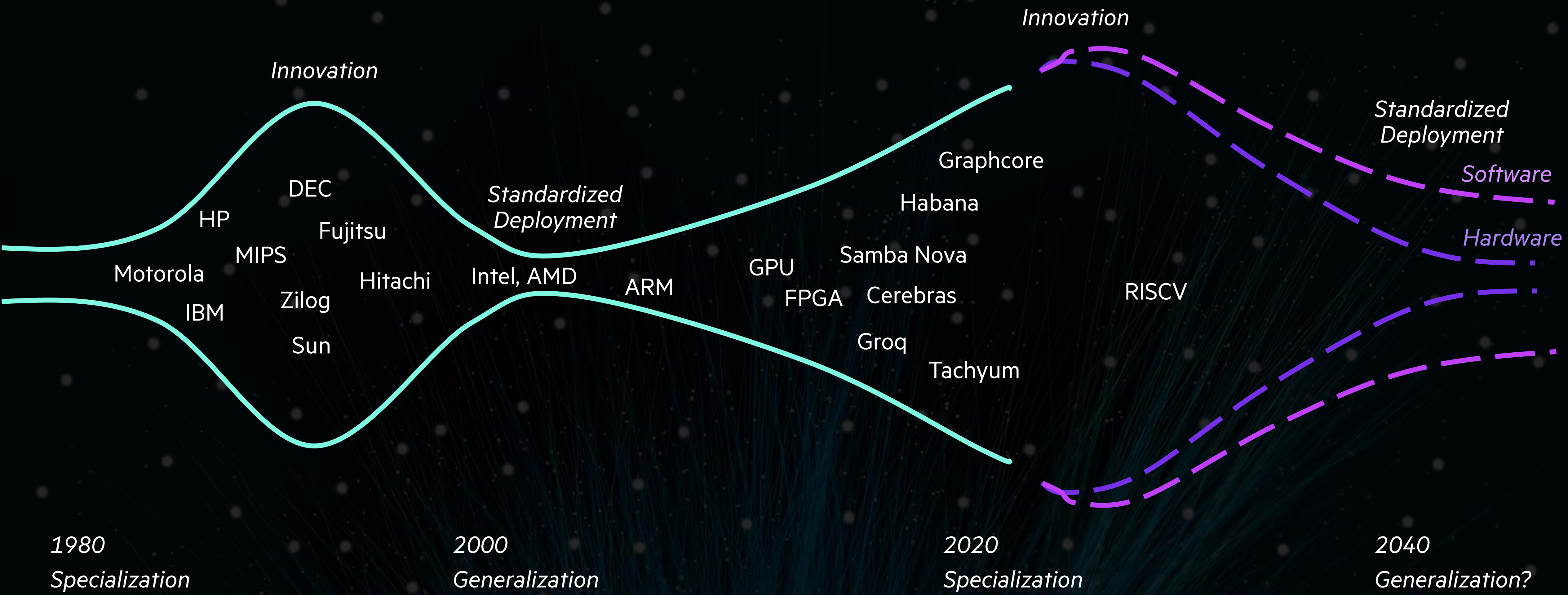
Hewlett Packard Enterprise



Aggregate throughput per stage	1200Gb each *(raw)*	1200Gb each *(raw)*	1600Gb *(raw)*

Courtesy of PSC
http://psc.edu

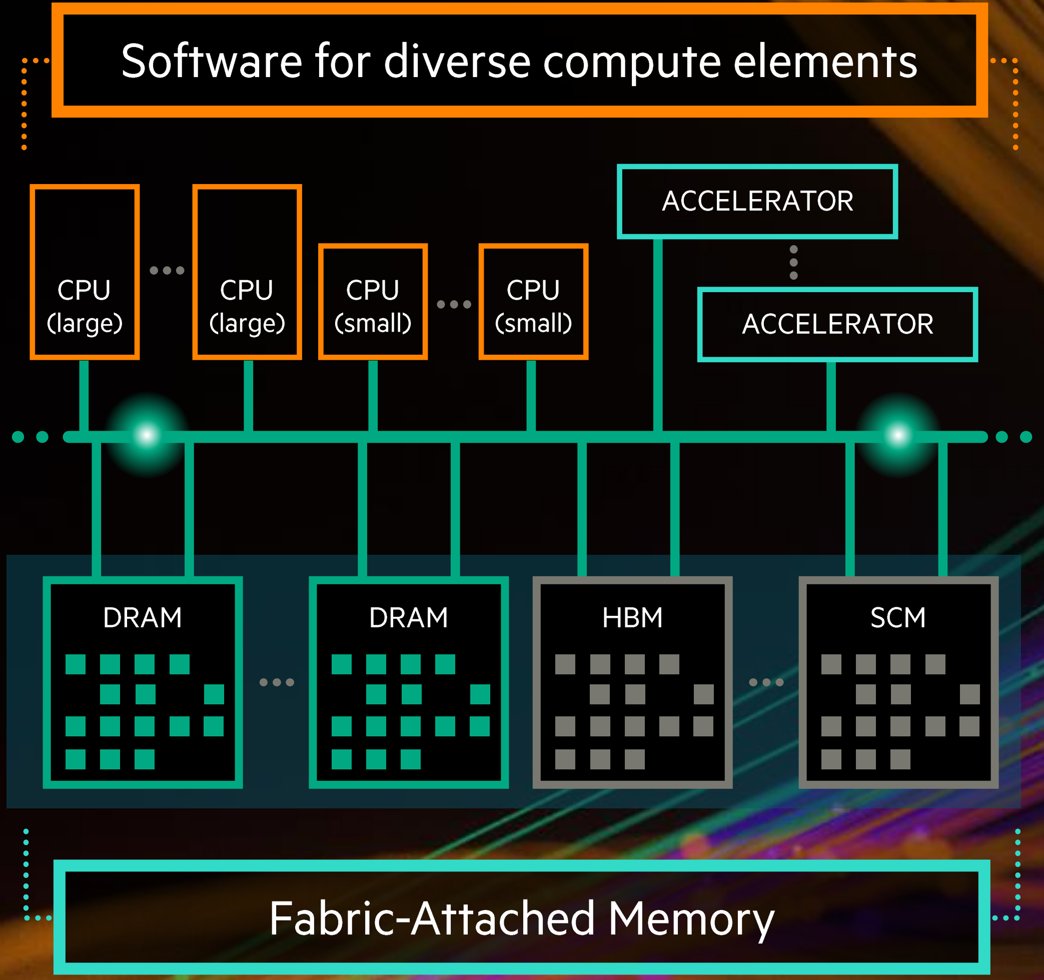
PENDULUM SWINGS BETWEEN HETEROGENEITY AND HOMOGENEITY DRIVEN FIRST BY INNOVATION THEN STANDARDIZATION



TRANSFORM PERFORMANCE WITH NOVEL PROGRAMMING

Today, applications are limited in reuse across different accelerators and systems vendors

The **software layer** can be **composed** and **customized** for right-sized solutions



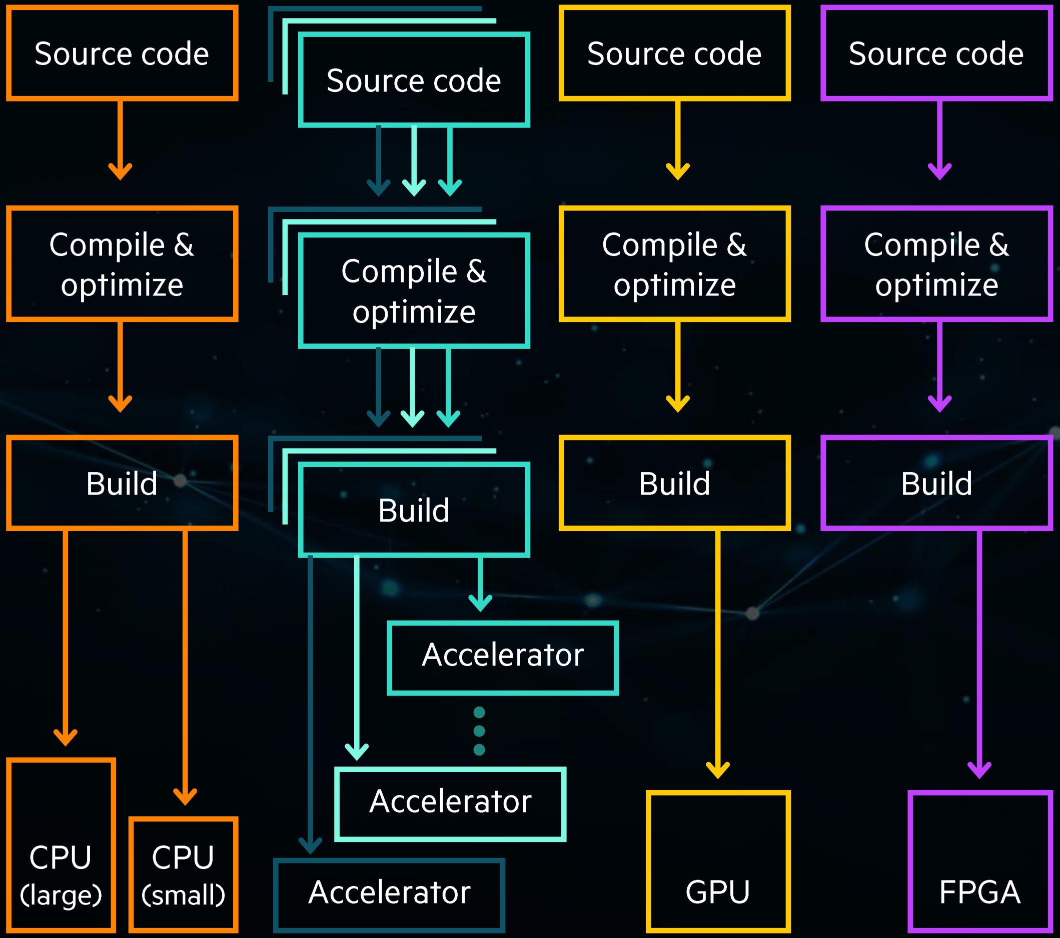
Creating standardized software that will work for diverse architectures so that everything can play together

PHOTONIC INTERCONNECT

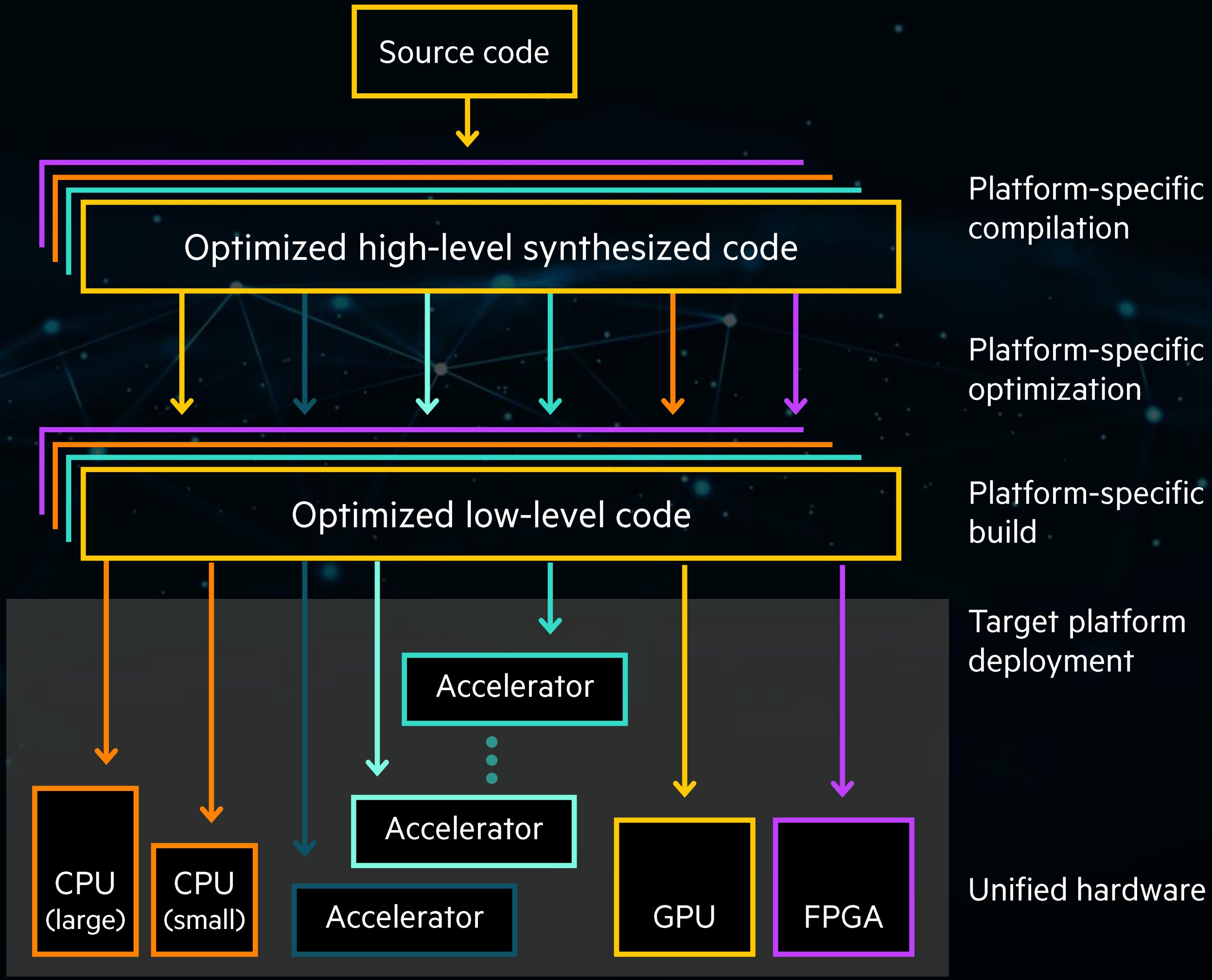
Allowing multiple compute nodes to simultaneously access a global memory pool

HETEROGENEOUS COMPUTING

TODAY



HEWLETT PACKARD LABS INNOVATION

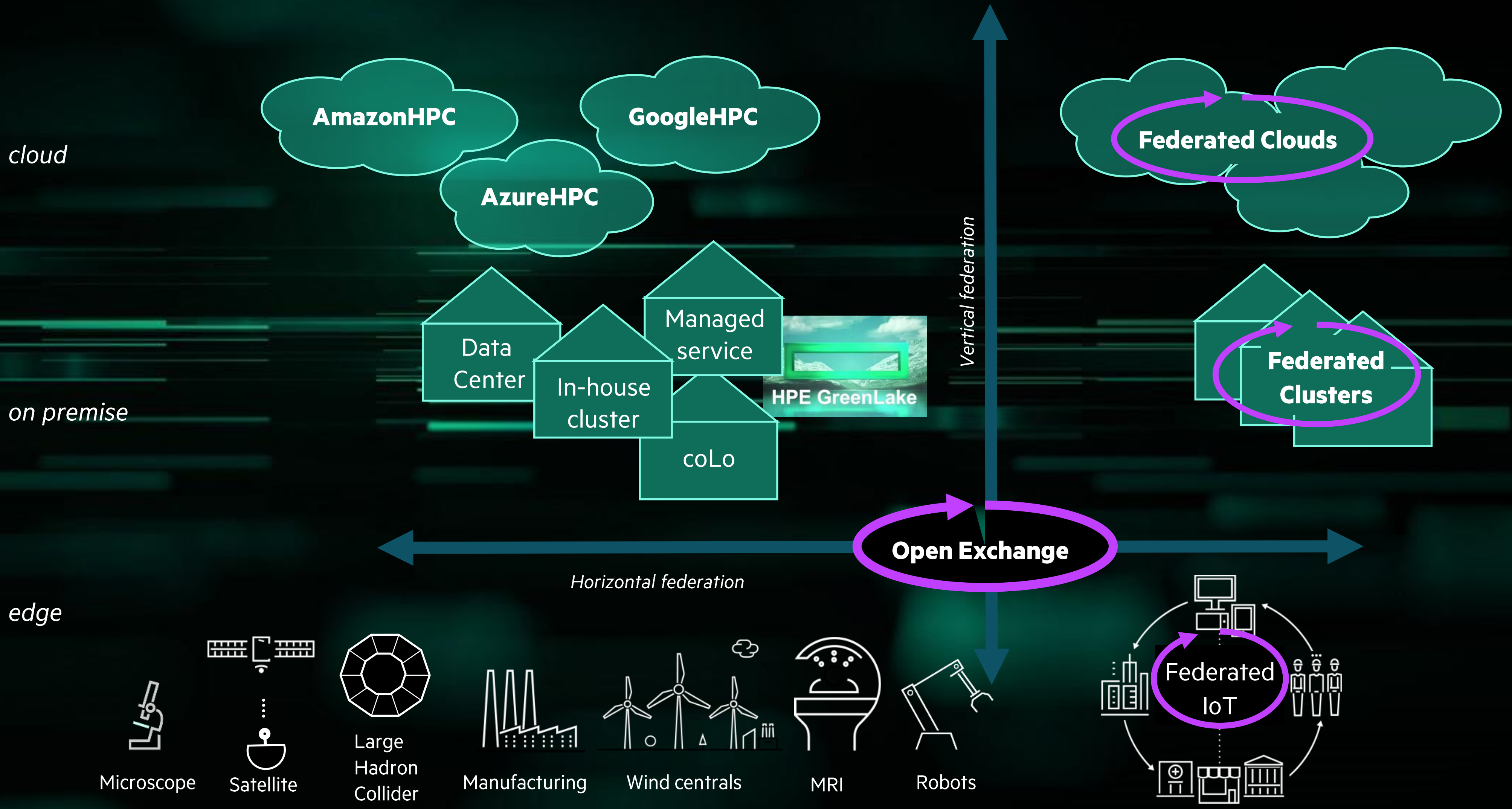


https://github.com/hst10/pylog

S. Huang et al, PyLog: An Algorithm-Centric Python-Based FPGA Programming and Synthesis Flow, IEEE Transactions on Computers, Dec. 2021

FEDERATION, VERTICAL AND HORIZONTAL

LEARN MORE:
Podcast: https://youtu.be/Aw5g6MBwGao
Report: Heterogeneous Computing: Point of View and Call for Action | HPE

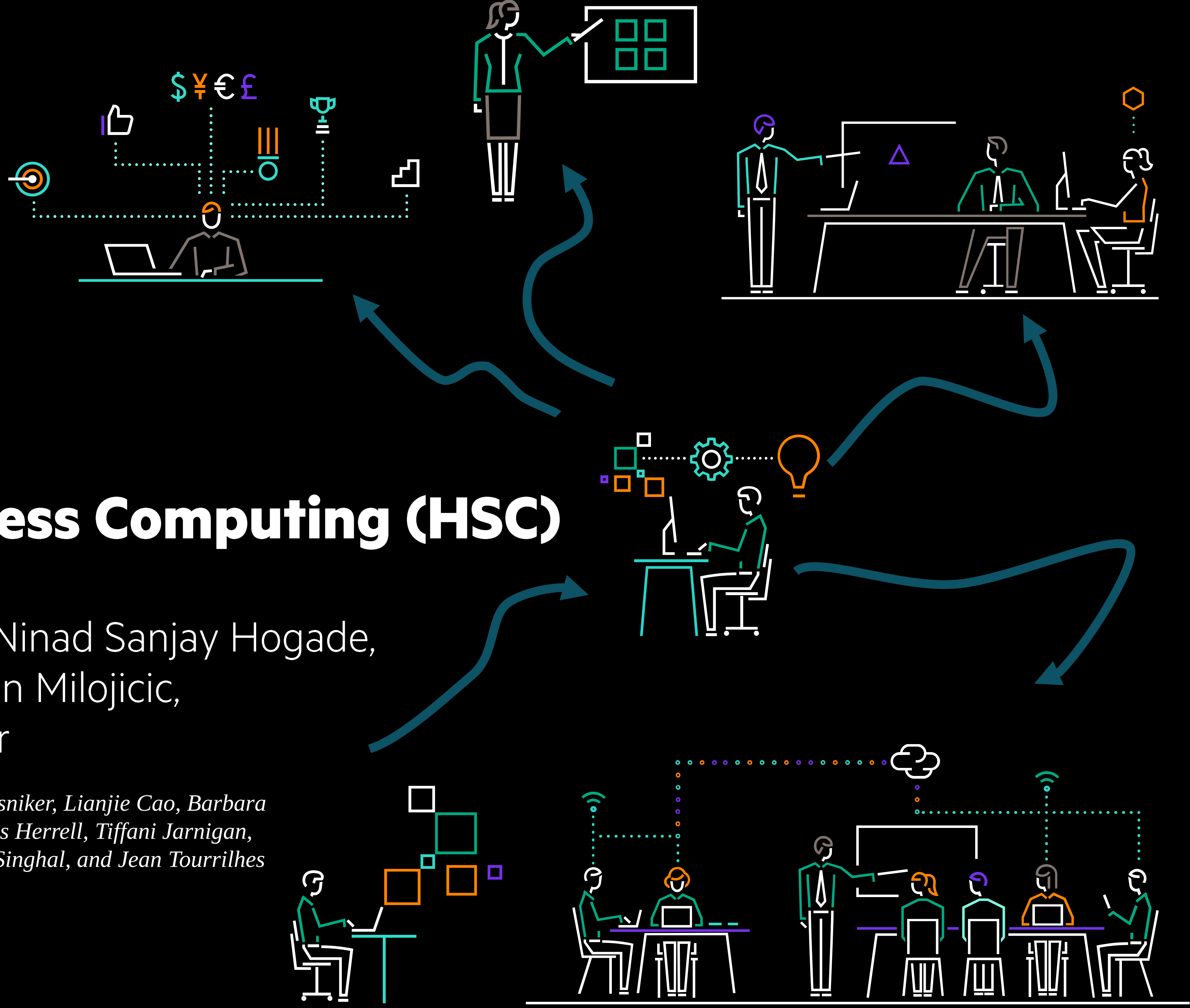


Hewlett Packard
Labs

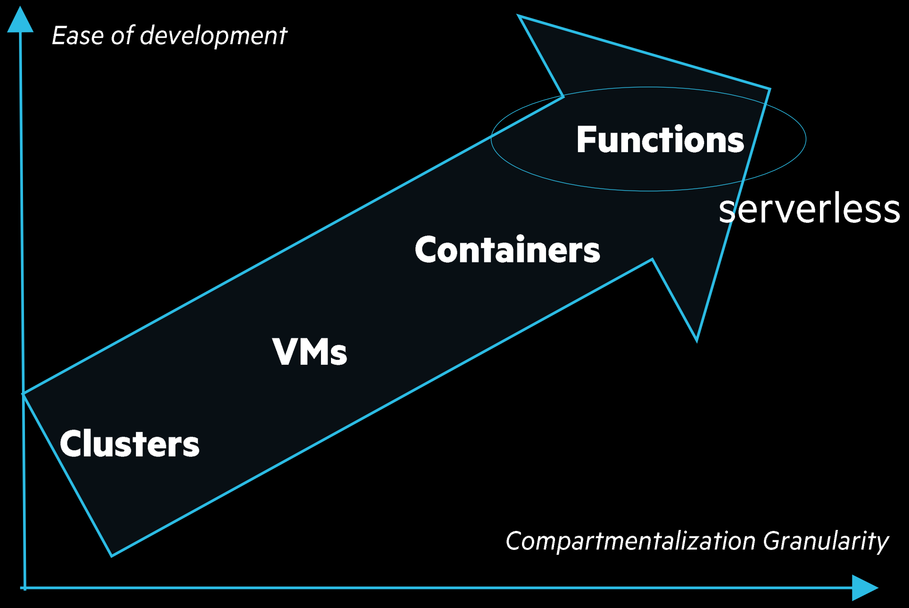
Heterogeneous Serverless Computing (HSC)

Aditya Dhakal, Eitan Frachtenberg, Ninad Sanjay Hogade, Rolando Pablo Hong Enriquez, Dejan Milojicic, Gourav Rattihalli, Tobias Pfandzelter

With support from: Max Alt, Cullen Bash, Kevin Boyum, Kirk Bresniker, Lianjie Cao, Barbara Chapman, Jeff Edlund, Paolo Faraboschi, Marco Fiorentino, Russ Herrell, Tiffani Jarnigan, Tom Phelan, Leo Popokh, Amit Sharma, Puneet Sharma, Sharad Singhal, and Jean Tourrilhes and Diman Zad Tootaghaj



Landscape Today

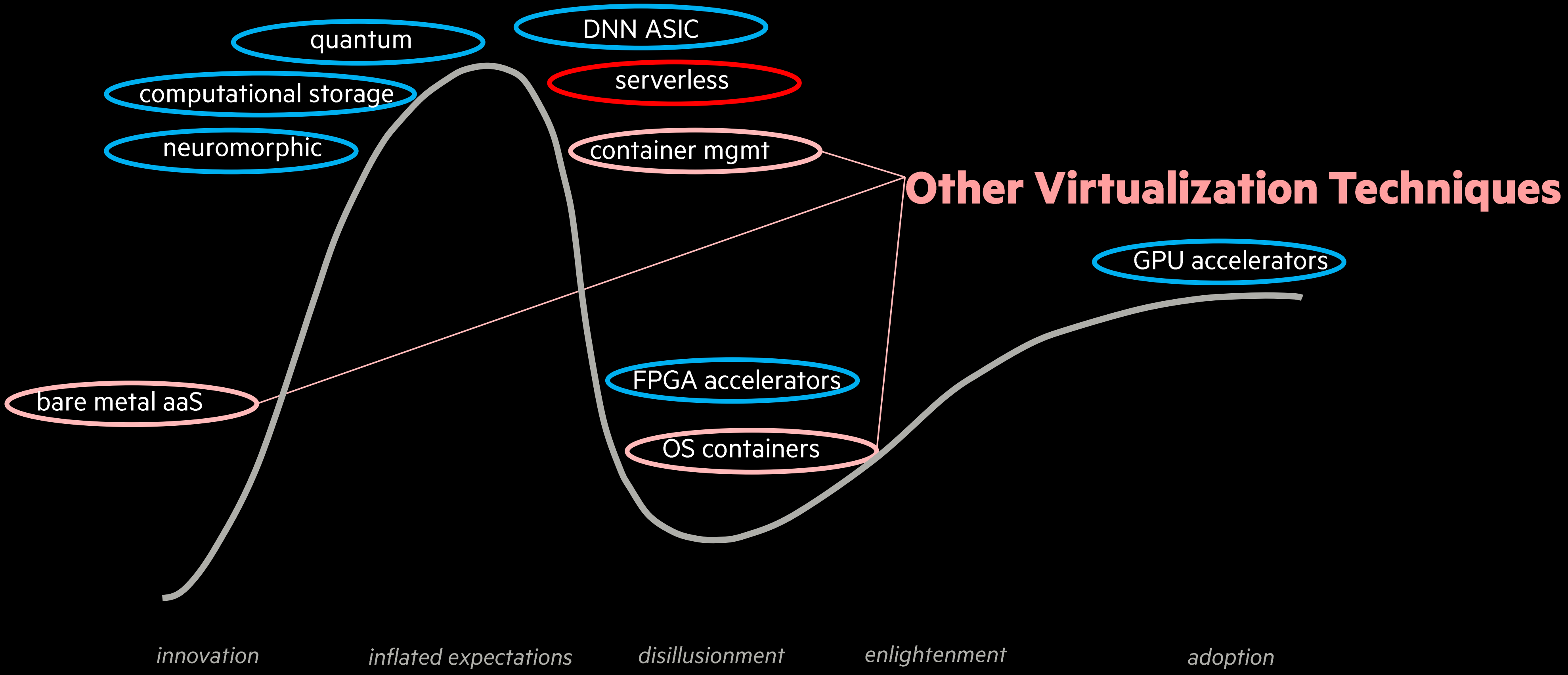


Moving Beyond Containers

Heterogeneity:
NVIDIA , AMD, Intel, and many startups introduce new heterogeneous accelerators

Serverless:
AWS Lambda, Azure Functions, Google Cloud Functions are growing in adoption

Heterogeneous & Serverless

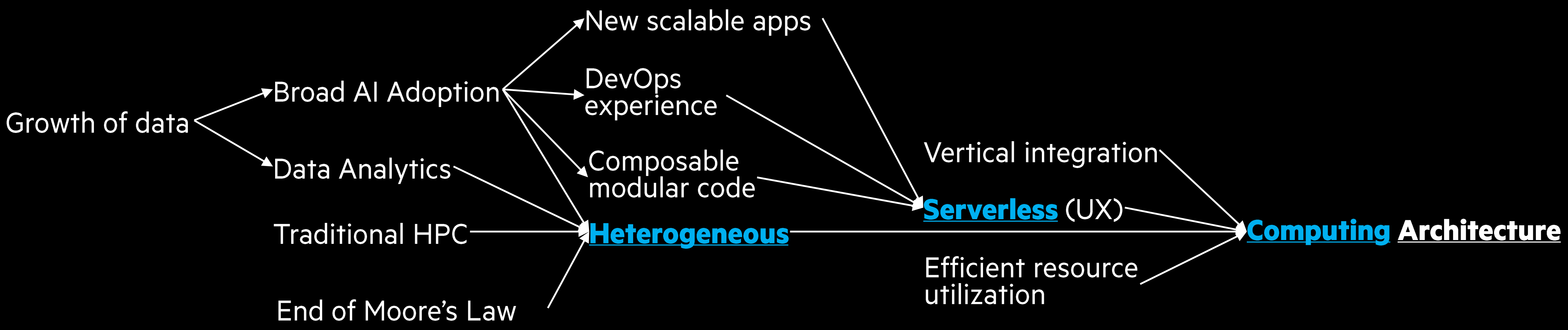


*By **2026**, more than 50% of global enterprises will have deployed serverless functions as a service (FaaS), up from less than 25% today (Gartner, June 2021)*

Heterogeneous Serverless Computing

What is it?

- HSC is workflow-optimized architecture inclusive of (compatible with) the public Cloud, for a set of workloads broader than what the public Cloud can support



Why Heterogeneous Serverless Computing

Why is it necessary?

- Traditional apps embracing AI, Data Analytics, and HPC. Delivery models expand from on-prem to Cloud
- Need for standardized solutions; Interoperability, aligns well with Internet of Workflows
- Reduction in operational costs via serverless. Higher level of abstractions: towards low-code/no-code
- Precision in time and resources: quantum of workload on quantum of resource

What will it enable for customers?

- End users: seamless scalability and fluidity of new applications
- Developers: increased programmer productivity
- Providers: performance efficiency to profitably run new applications

Business Landscape of the Future

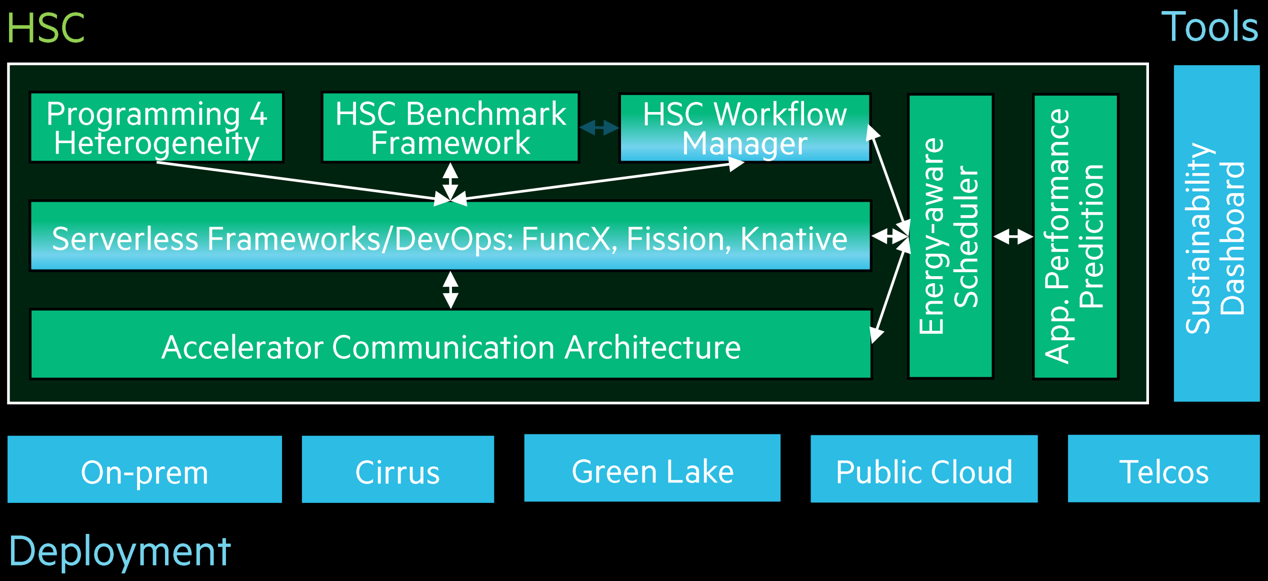
- Growing demand
  - Complex WW problems require timely solution: global warming, climate, pandemics, supply chain disruption, etc.
  - HPC/HPDA achieves broader adoption for engineering needs in HPC and enterprise markets
  - AI/ML transforms computing from Cloud to edge

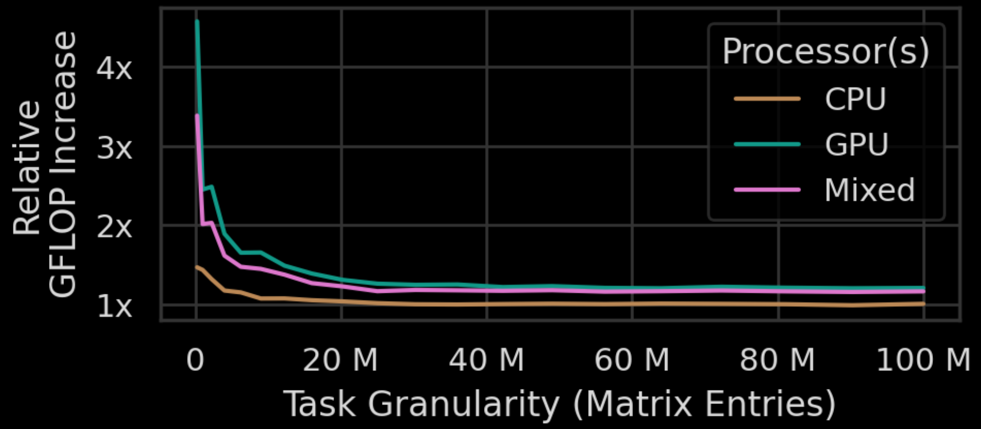
- Evolving Supply
  - Continued heterogeneity improves infrastructure utilization
  - Workflow-composition deploys modern workloads where resources are available
  - aaS delivery, secured through modernized end to end architectures, dominates over traditional apps

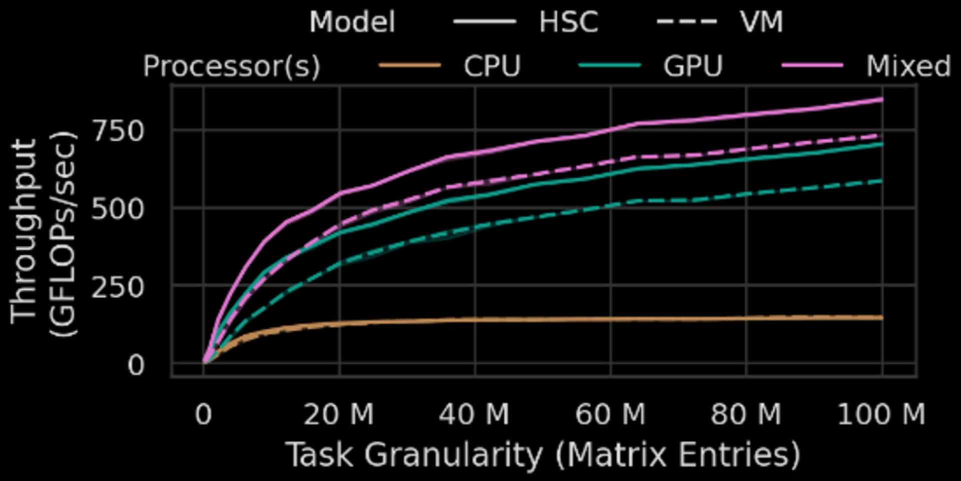
- Closed Innovation
  - Hyperscalers' vertical integration is a key threat to innovation
  - Major silicon vendors embrace hyperscaler lock-in model
  - Accelerator startups use non-standard I/F for platform enablement

Heterogeneous Serverless Computing

- Hypothesis: matching fine granularity of accelerators with that of serverless
  - Short time to execute, service lifetime
  - Size of deployed code



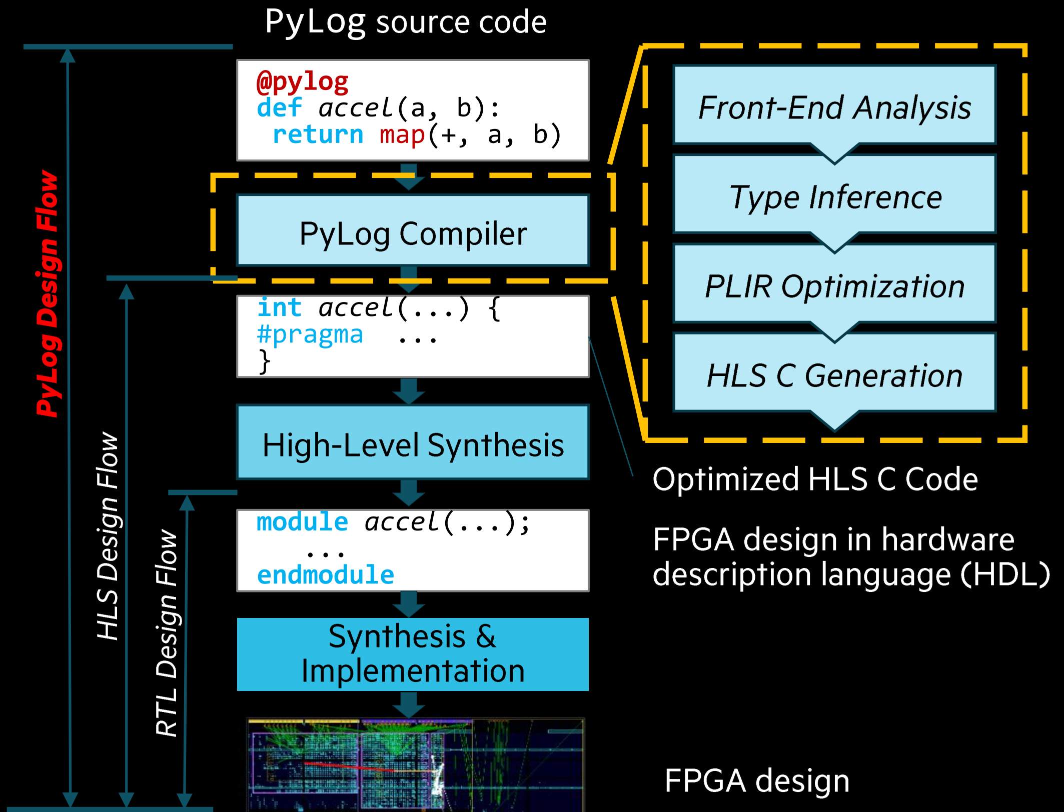




https://github.com/hst10/pylog

HSC Programming for Heterogeneity (with UC Irvine, Prof. Sitao Huang)

- **Why** is this a problem?
  - Huge gap between hardware and software abstraction levels
  - Applications/frameworks developed in high-level languages
  - Many applications need hardware acceleration
  - Challenging to accelerate high-level apps due to abstraction gap
- **How** are we approaching the problem?



```
@pylog
def accel(a, b):
  return map(+, a, b)
```

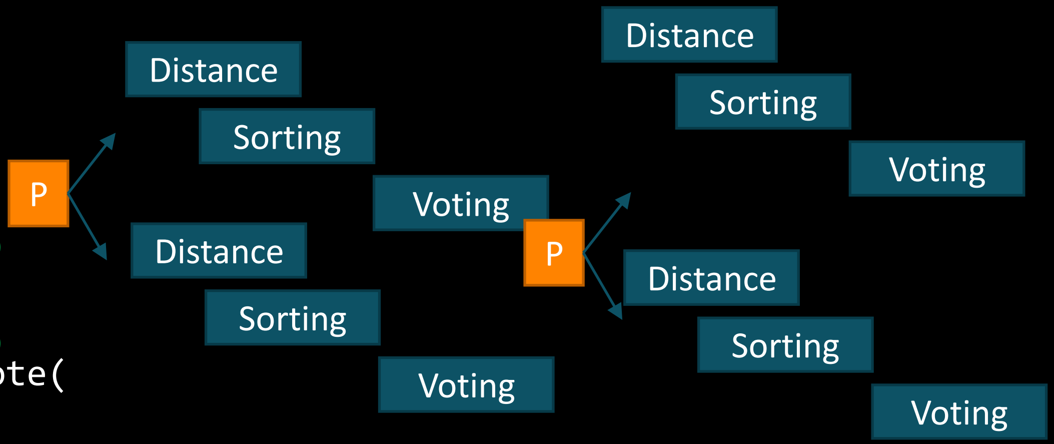
```
int accel(...) {
#pragma   ...
}
```

```
module accel(...);
  ...
endmodule
```

- **What** are we developing?
  - Start with Pylog compiler supporting collaborative computing
  - Extend for data placement and movement (PCIe, DDR, HBM,…)
  - Account for heterogeneity (across nodes, accelerators, hosts..)
  - PyLog + FaaS, enable accelerator support in serverless
- **So What**? What is the impact?
  - Improved developer productivity, fewer lines of code (LOC)
  - Easier system integration (e.g. cloud/edge FPGAs)
  - Easier design space search, easier design migration

```
@pylog
@config(dup = N)
def prep(a, b):
    ...
@config(dup = N)
def dist_sort_vote(
    @pylog
    def dist(c

    @pylog
    def sort(c, d):
```



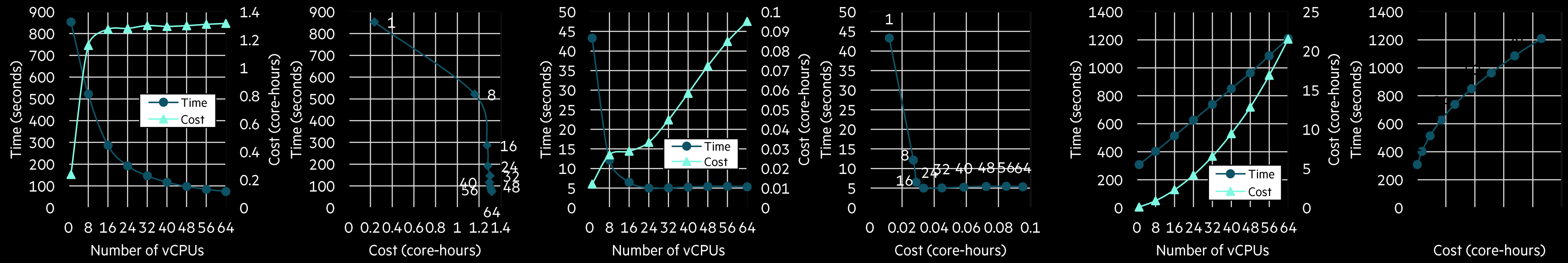
HSC Serverless Frameworks (with U. Chicago, Ian Foster)

- **Why** is this a problem?
  - Many serverless frameworks exist that could serve as a platform for HSC
  - State-of-the-art frameworks do not fully support HPC/HPDA/AI applications (e.g., heterogeneity, responsiveness)
- **How** are we approaching the problem?
  - Surveying open-source, academia, scientific computing, industry offerings
  - Testing functionality from an HSC perspective
- **What** are we developing?
  - Evaluation framework for serverless platforms
  - Qualitative evaluation of existing serverless software platforms (open-source, academia, industry)
- **So What**? What is the impact?
  - Identified **funcX**, **Fission**, and **Kubernetes** as forerunners
  - Candidates support many HSC requirements but not all
  - Suitable platforms can be used to implement our future solutions

🟢	Fully Supported
🟩	Supported w/ external components
🟡	Further investigation needed
🔴	Not supported

Framework	**Heterogeneity**				**Autoscaling**	**Multiple Container technologies**	**HPC Support**	**Open Source**	**Maturity**	**Responsiveness**
	CPU	GPU	FPGA	NIC						
funcX	🟢	🟩	🔴	🔴	🟡	🟢	🟢	🟡	🟩	🔴
Apache Airavata	🟢	🟩	🔴	🔴	🔴	🔴	🟢	🟢	🔴	🔴
Kubernetes	🟢	🟩	🟩	🟩	🟢	🟡	🔴	🟢	🟢	🟡
OSS Serverless Frameworks	🟢	🟩	🔴	🔴	🟢	🟡	🔴	🟢	🟩	🟡
Fission	🟢	🟩	🔴	🔴	🟢	🟡	🔴	🟢	🟡	🟩
rFaaS	🟢	🟡	🟡	🟡	🟡	🟡	🟩	🟢	🔴	🟢

HSC Application Performance Prediction (with AUB, Prof. Izzat El Hajj)

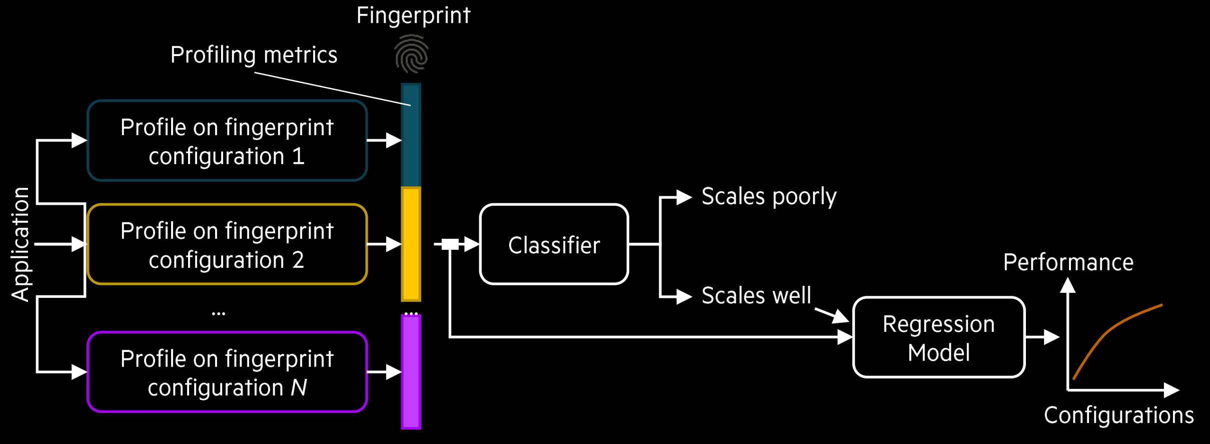


(a) 350.md execution time versus cost

(b) 376.kdtree execution time versus cost

(c) streamcluster execution time versus cost

- **Why** is this a problem?
- **How** are we approaching the problem?

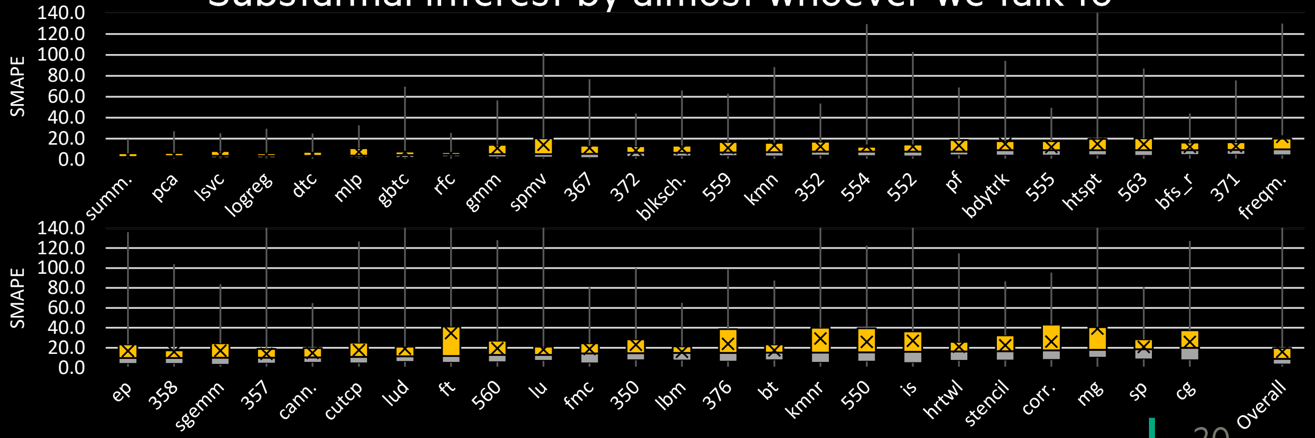


Step 1: Fingerprint Generation

Step 2: Classification by Scalability

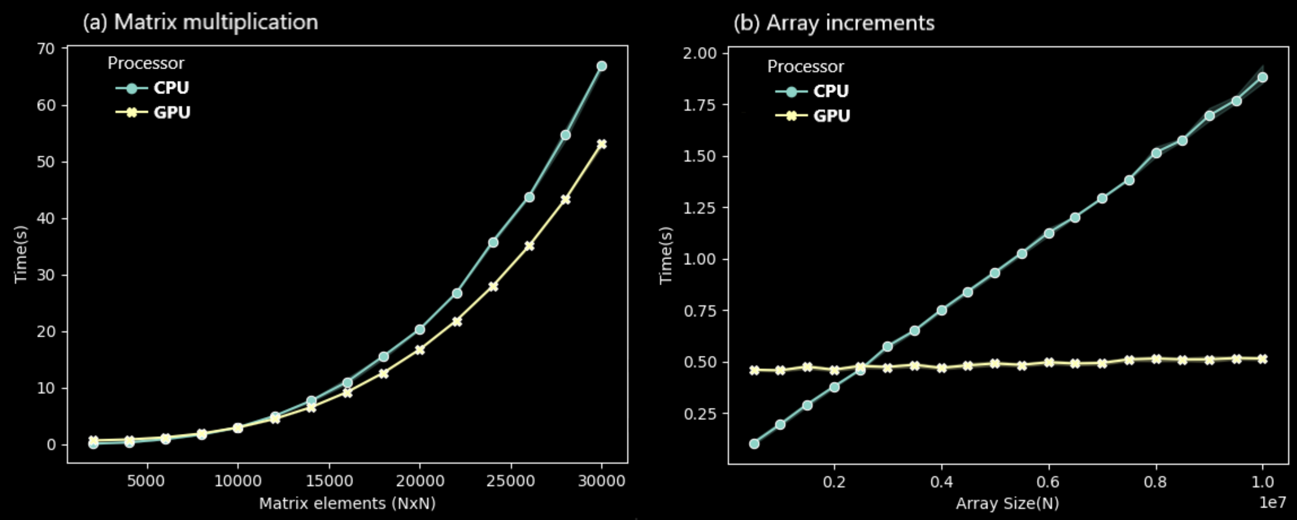
Step 3: Performance Prediction

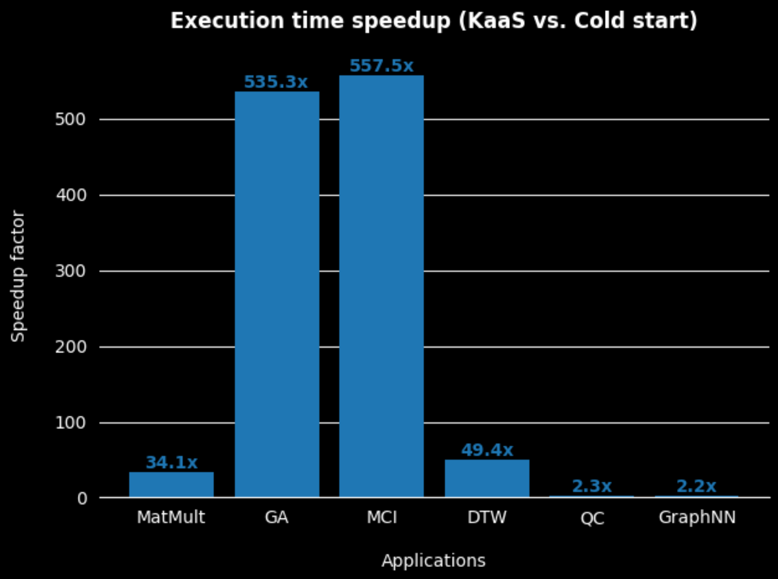
- **What** are we developing?
  - a) Service for schedulers, such as in Green Lake; b) Service for customers; c) Libraries for CPE
- **So What**? What is the impact?
  - Substantial interest by almost whoever we talk to



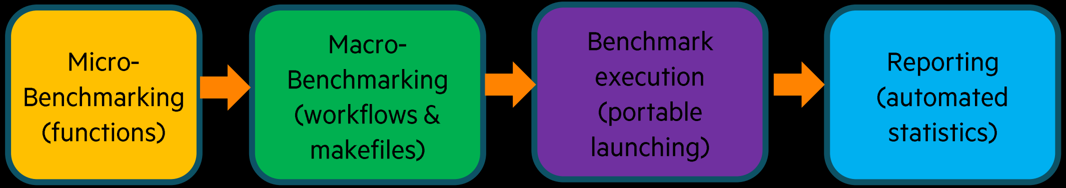
HSC Workflow Manager (HSC-WM)

- **Why** is this a problem?
  - The rise of computing heterogeneity & serverless will empower ever more complex workflows
  - A fundamental part of the HSC technology stack is to be able to manage these new workflows effectively
- **How** are we approaching the problem?
  - Review most likely use cases for HSC workflows
  - Explore capabilities of workflow managers built for other calculations (e.g., HPC)
  - Build HSC capabilities on existing workflow managers to create a fit-for-purpose HSC-WM
- **What** are we developing?
  - Integrate existing serverless frameworks and add granularity to the calculations.
  - Add performance prediction
  - Leverage energy aware capabilities.
- **So What**? What is the impact?
  - Simplify the way to create and run HSC workflows
  - Transfer HSC capabilities to other BUs (e.g., Green Lake, Cray)
  - Contribute new use cases and workflows to AI for science.

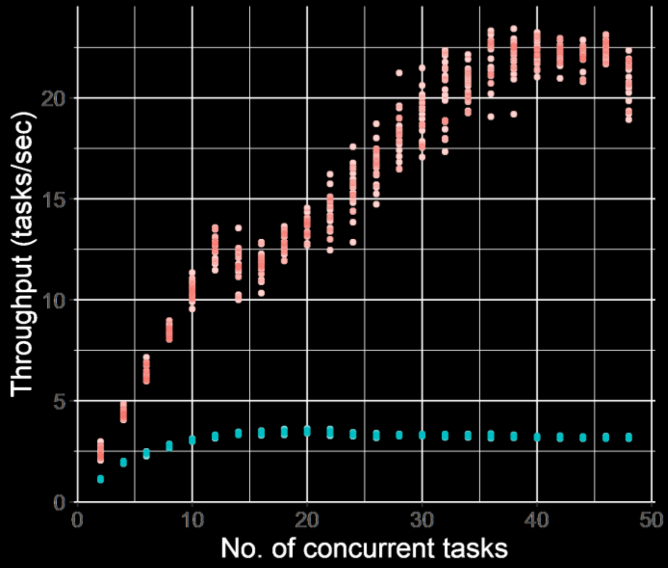


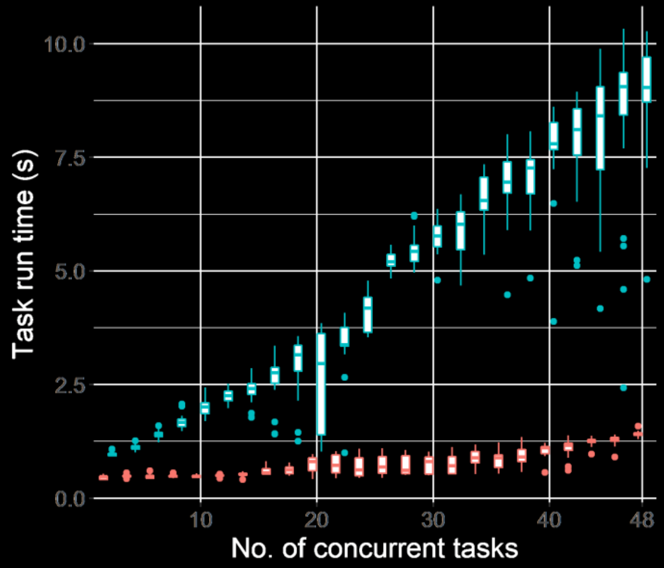


HSC Benchmark Framework


Micro-Benchmarking (functions)
Macro-Benchmarking (workflows & makefiles)
Benchmark execution (portable launching)
Reporting (automated statistics)

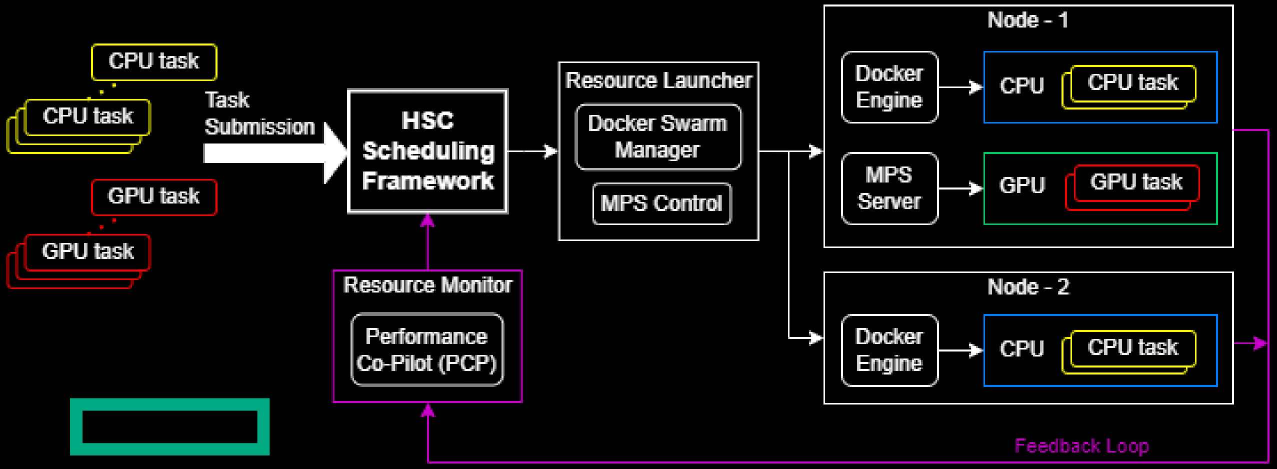
- **Why** is this a problem?
  - You can't improve what you can't measure: there exist no tools currently to measure performance of HSC platforms.
  - Reproducibility is vital: quantifying uncertainty and noise is not currently a priority in existing FaaS benchmarks.
- **How** are we approaching the problem?
  - Collection of microbenchmarks (functions) that can be composed into complex workflows to measure all performance aspects of an HSC (or HPC or cloud) platform.
  - Automated reproducibility controls.
  - Automated statistical analysis, graphing, and reporting of uncertainty and noise.
- **What** are we developing?
  - Collection of microbenchmark functions.
  - Portable launching framework for any FaaS/HPC platform.
  - Standard workflows converted to standard Makefiles
  - Statistical analysis and reporting
- **So What**? What is the impact?
  - Easier performance evaluation of platforms and applications.
  - Measure any HPC/FaaS/HSC platform.
  - Portable and reproducible performance tests.
  - Automated statistics and reporting.



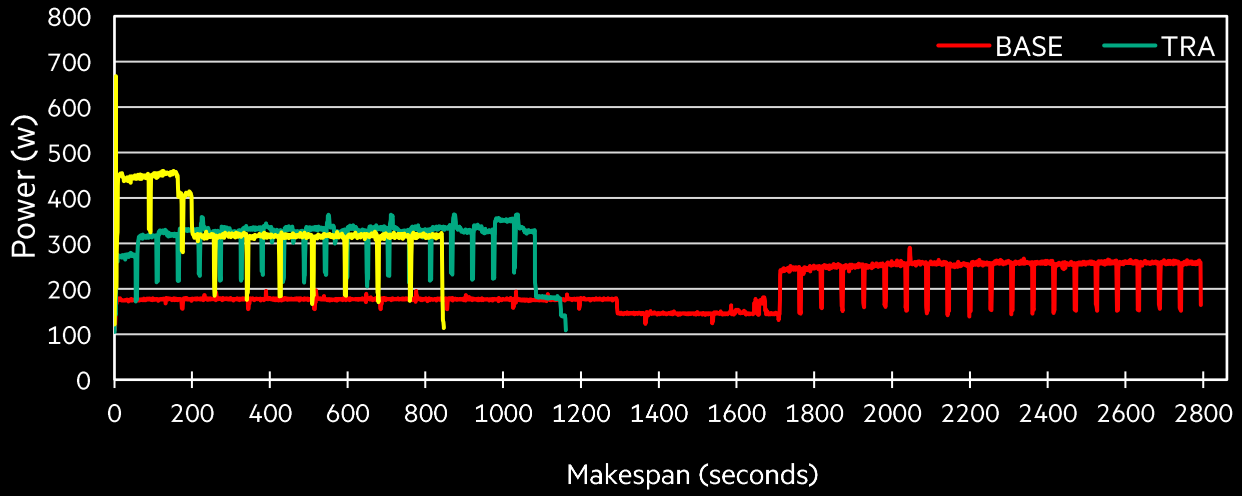


HSC Heterogeneity- and Energy-Aware Scheduler

- **Why** is this a problem?
  - HSC applications are diverse and have different hardware requirements
  - Inefficient scheduling of functions reduces throughput and increases energy consumption
  - Current solutions do not include hardware-specific features that support fine-grained scheduling, e.g., MPS with GPUs
- **How** are we approaching the problem?
  - Exploring the effects of fine-grained heterogeneous function scheduling on energy and throughput
  - Developing multiple test beds to study the problem



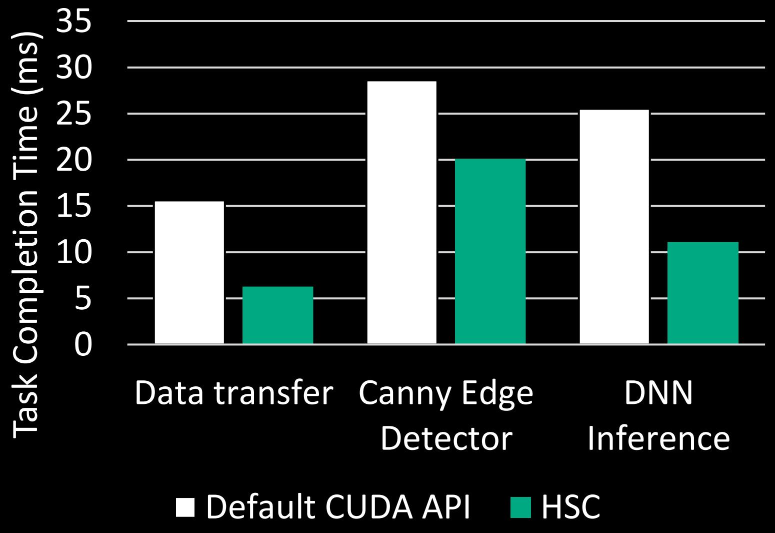
- **What** are we developing?
  - A heterogenous function and heterogeneity-aware scheduler
  - Enable special hardware/software features e.g., MPS for GPUs, InfiniBand, AI accelerators within the proposed scheduler
  - Customer or provider goal-specific algorithms for the scheduler
- **So What**? What is the impact?
  - The proposed scheduler will schedule heterogeneous functions on heterogenous hardware at run-time
  - Enable sharing hardware resources more efficiently



Cluster Power Measurement per Second

HSC Accelerator Communication Architecture (w/ UIUC, Prof Deming Chen)

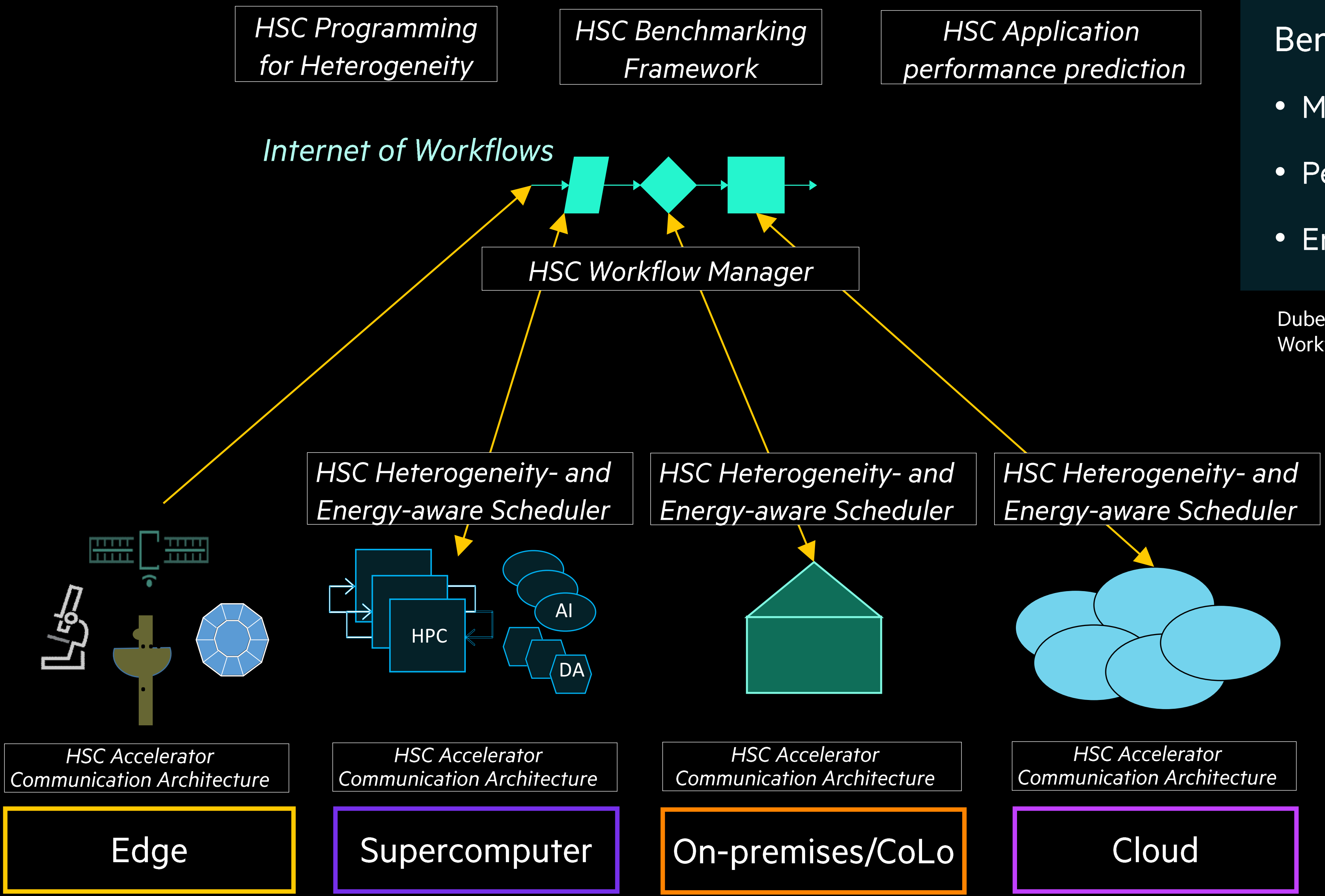
- **Why** is this a problem?
  - Accelerators heavily rely on CPU for communication (Overhead)
  - Different types of communication (Broadcast/All-gather/reduce) not well defined for heterogeneity
  - Universal memory addressing not available across accelerators
- **How** are we approaching the problem?
  - Focus on understanding workflow communication pattern
  - Create lightweight communication library to interact with runtimes of heterogenous accelerators



- **What** are we developing?
  - Low overhead communication library that works with devices from different vendors
  - Communication scheduler that masks communication with compute
  - Universal virtual memory addressing across het. devices
- **So What**? What is the impact?
  - Less communication overhead increases system throughput
  - Communication architecture will facilitate adoption of heterogenous accelerator framework

Lower CPU overhead (HSC) showing lower task completion time for different application compared to default (relying on CPU side API)

Internet of Workflows



HSC Programming for Heterogeneity

HSC Benchmarking Framework

HSC Application performance prediction

Internet of Workflows

HSC Workflow Manager

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

Edge

Supercomputer

On-premises/CoLo

Cloud

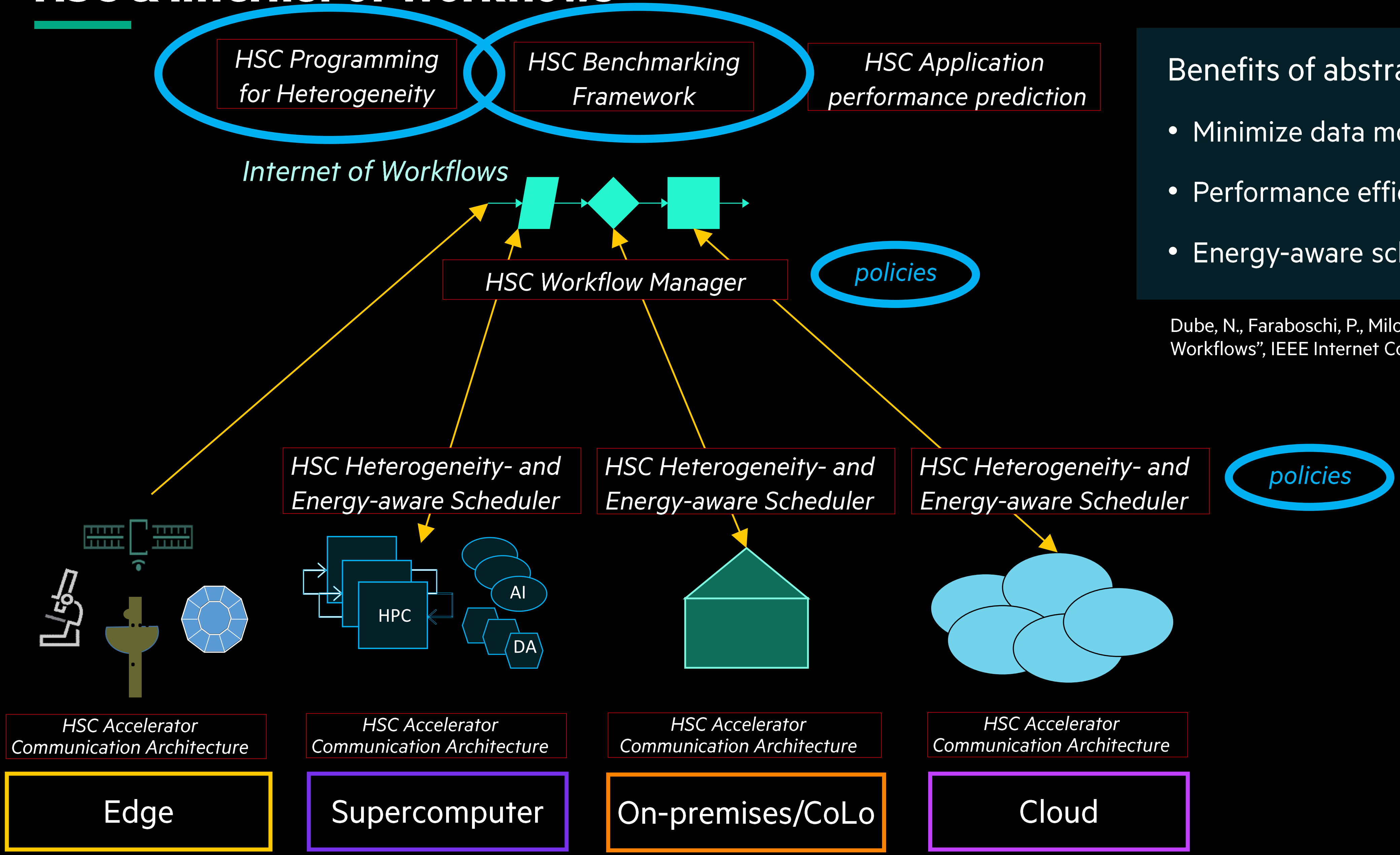
Benefits of abstracting workflows

- Minimize data movement
- Performance efficiency
- Energy-aware scheduling

Dube, N., Faraboschi, P., Milojicic, D., Roweth, D., "Internet of Workflows", IEEE Internet Computing, Sept/Oct 2021

HSC & Internet of Workflows

DEVOPS



HSC Programming for Heterogeneity

HSC Benchmarking Framework

HSC Application performance prediction

Internet of Workflows

HSC Workflow Manager

policies

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

policies

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

Edge

Supercomputer

On-premises/CoLo

Cloud

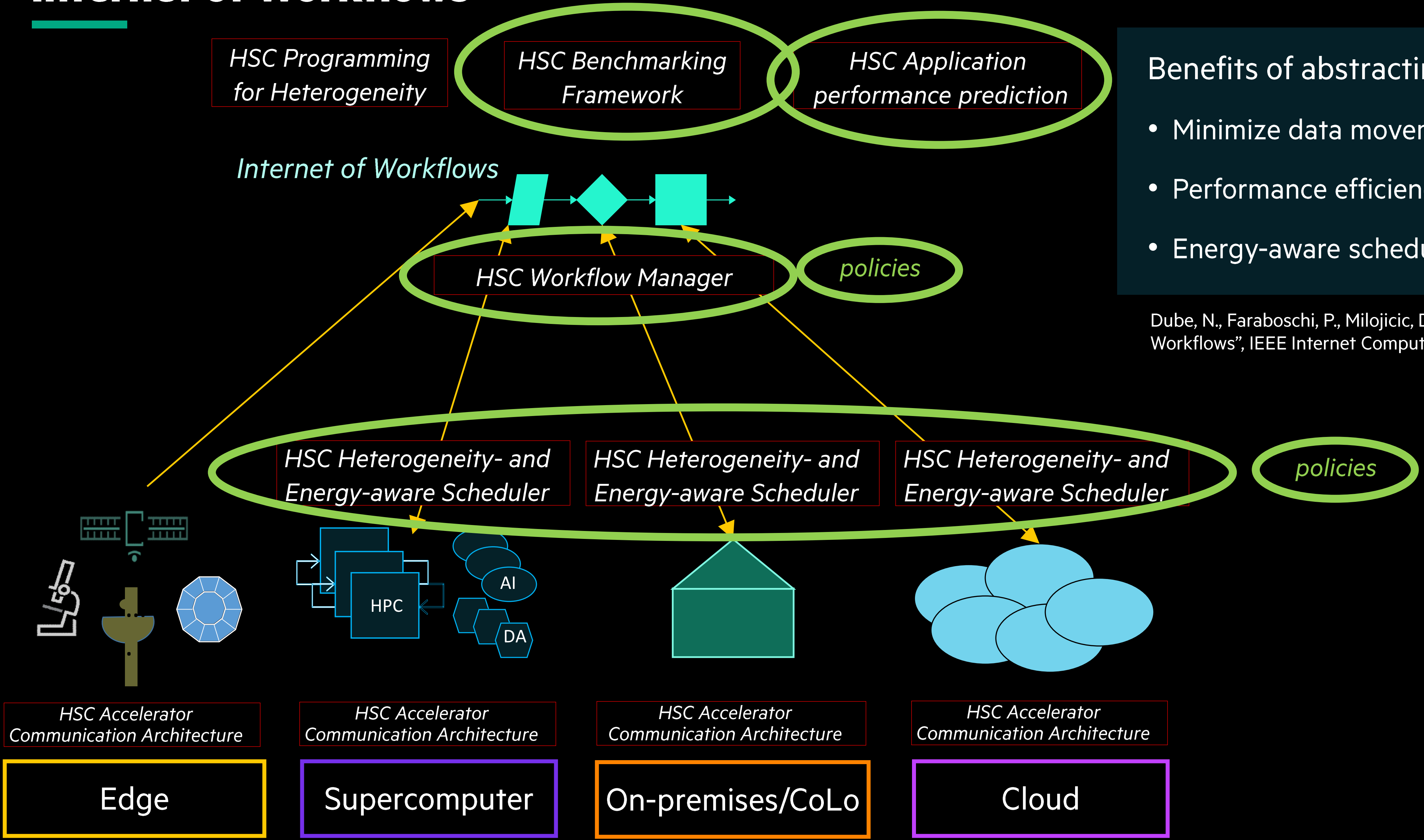
Benefits of abstracting workflows

- Minimize data movement
- Performance efficiency
- Energy-aware scheduling

Dube, N., Faraboschi, P., Milojicic, D., Roweth, D., “Internet of Workflows”, IEEE Internet Computing, Sept/Oct 2021

Internet of Workflows

SUSTAINABILITY



HSC Programming for Heterogeneity

HSC Benchmarking Framework

HSC Application performance prediction

Internet of Workflows

HSC Workflow Manager

policies

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

HSC Heterogeneity- and Energy-aware Scheduler

policies

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

HSC Accelerator Communication Architecture

Edge

Supercomputer

On-premises/CoLo

Cloud

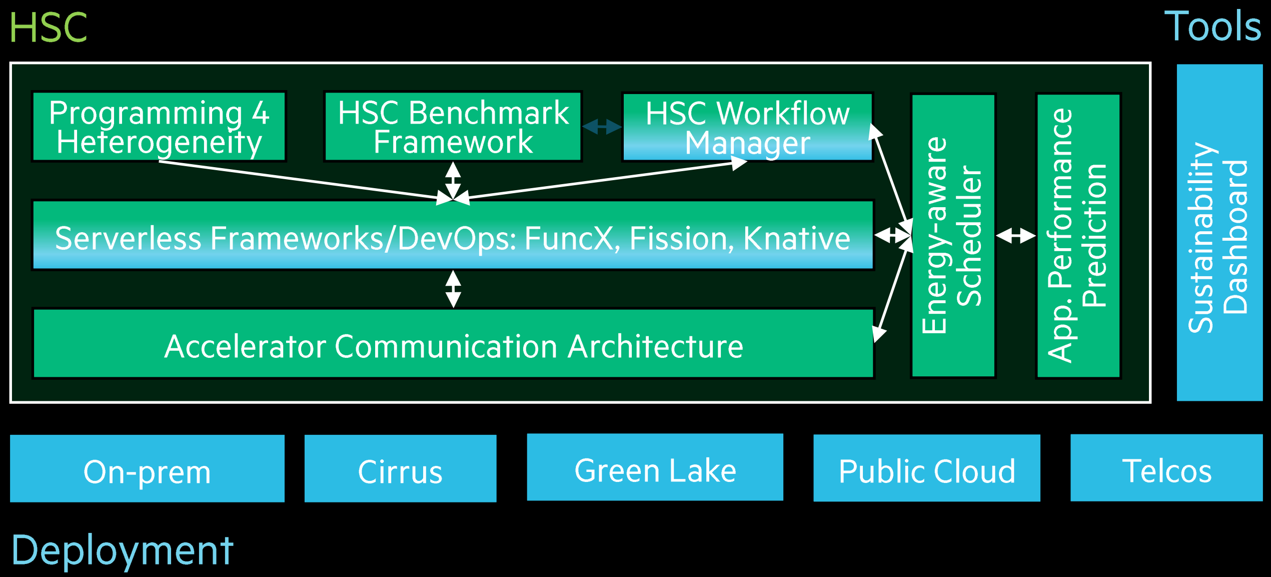
Benefits of abstracting workflows

- Minimize data movement
- Performance efficiency
- Energy-aware scheduling

Dube, N., Faraboschi, P., Milojicic, D., Roweth, D., “Internet of Workflows”, IEEE Internet Computing, Sept/Oct 2021

Summary

- HSC is workflow-optimized architecture inclusive of (compatible with) the public Cloud, for a set of workloads broader than what the public Cloud can support
- We have proved HSC hypothesis of matching fine granularity of accelerators with that of serverless
- We are working towards productization of individual HSC components



Thank you

Questions?

✉ dejan.milojicic@hpe.com
twitter.com/dejanm
www.linkedin.com/in/dejanm
https://dejan.milojicic.com
www.facebook.com/dejan.milojicic
www.facebook.com/DejanHPE
www.instagram.com/dejanmilojicic