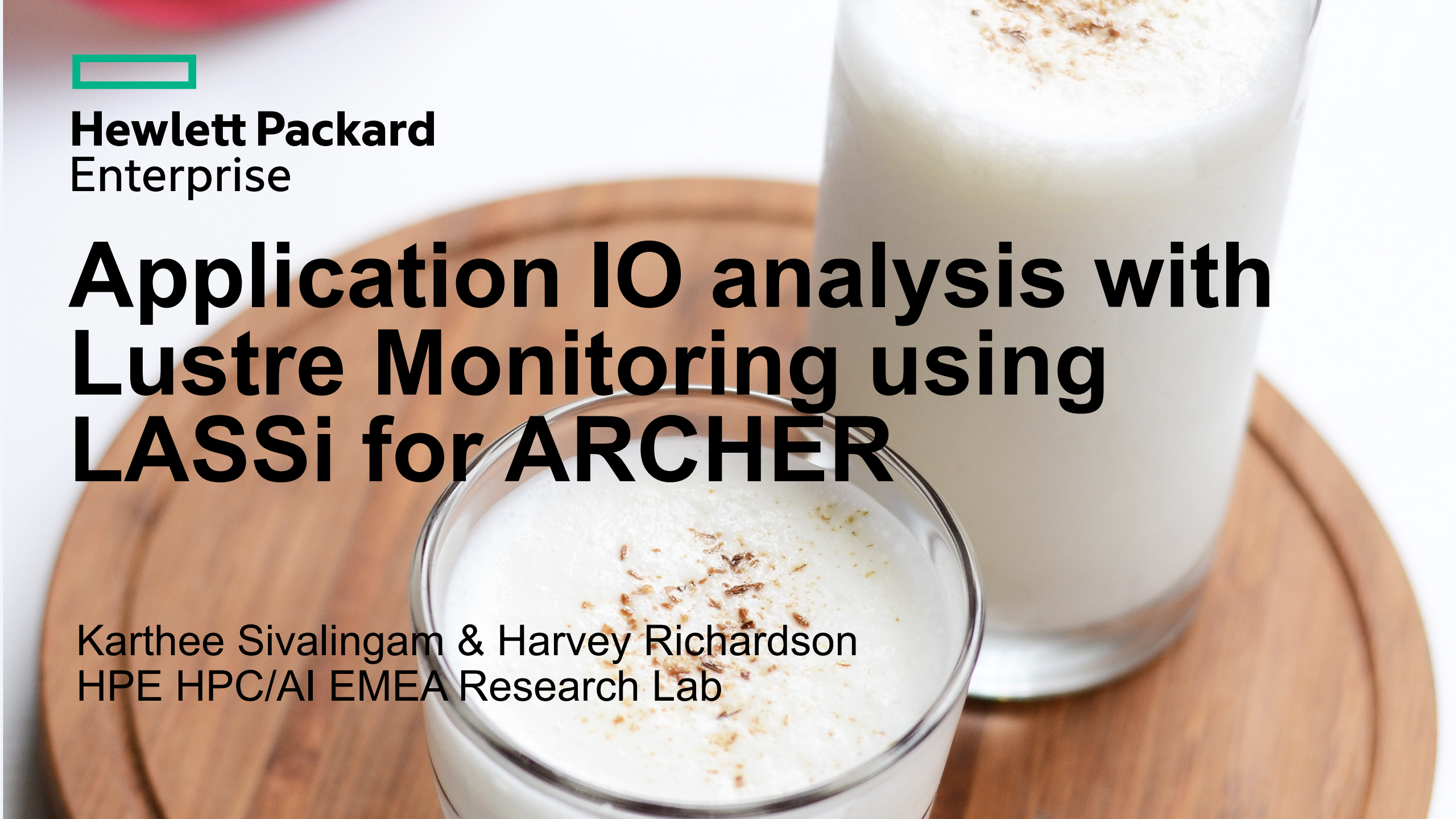


Hewlett Packard Enterprise

Application IO analysis with Lustre Monitoring using LASSi for ARCHER

Karthee Sivalingam & Harvey Richardson
HPE HPC/AI EMEA Research Lab

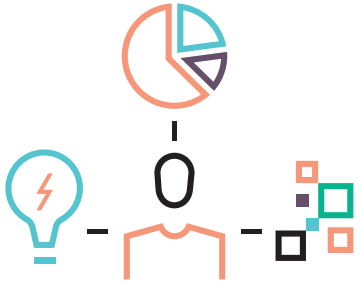
HPE HPC/AI EMEA Research Lab

Deep Technical Collaboration

- HPE & Customers work together
- Focus on new technologies
- Drive future HPE products
- Long term technical relationship

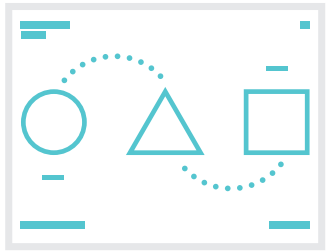
Research Interests

- Memory hierarchy
- Data Movement and Workflow
- Compiler and mathematical optimisation
- HPC in Cloud, AI and Big Data



Models

- Center of Excellence (CoE)
- Value Add projects
- EU H2020 research projects



Hewlett Packard Enterprise

Centers of Excellence in EMEA

ARCHER, UK

- LASSi - IO Monitoring and Analytics
- Application tuning (XC30)
- IO performance Optimisation

KAUST, KSA

- Deep Research investigation
- Asynchronous tasking
- Deep Learning for Bio-Science

GW4

- ARM system tuning
- ARM ecosystem development
- Joint ARM, Cavium partnership

Hewlett Packard Enterprise

Current H2020 Projects

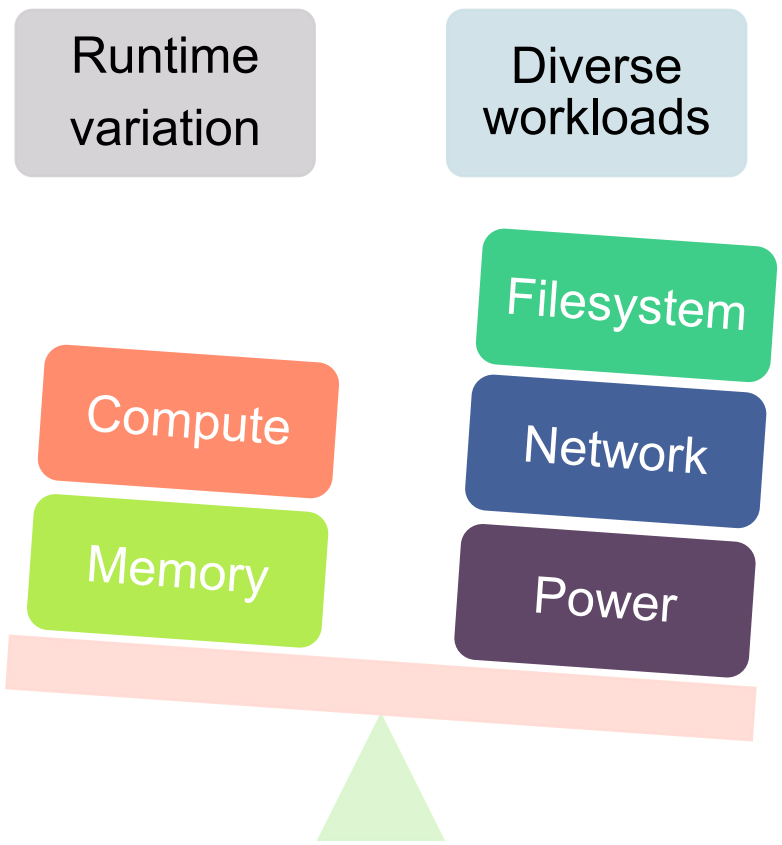
Imperial College London
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
HLRS High-Performance Computing Center | Stuttgart
MAVEL
POLITECNICO DI TORINO
ECMWF
SEAGATE
CEA
appentra
CSCS
JÜLICH FORSCHUNGSZENTRUM
epcc
ETH Zürich
Fraunhofer ITWM
KTH VETENSKAP OCH KONST
HLRS High-Performance Computing Center | Stuttgart
ADAPTANT
XLAB
IBM
POLITECNICO MILANO 1863
iti
CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS
Atos
JADS
Imperial College London
RWTH AACHEN UNIVERSITY
ZIB
ICOOR
EDF
SIEMENS Ingenuity for life
EuroPLEx
H2020
MINOA
MIXED-INTEGER NON-LINEAR OPTIMISATION: ALGORITHMS AND APPLICATIONS

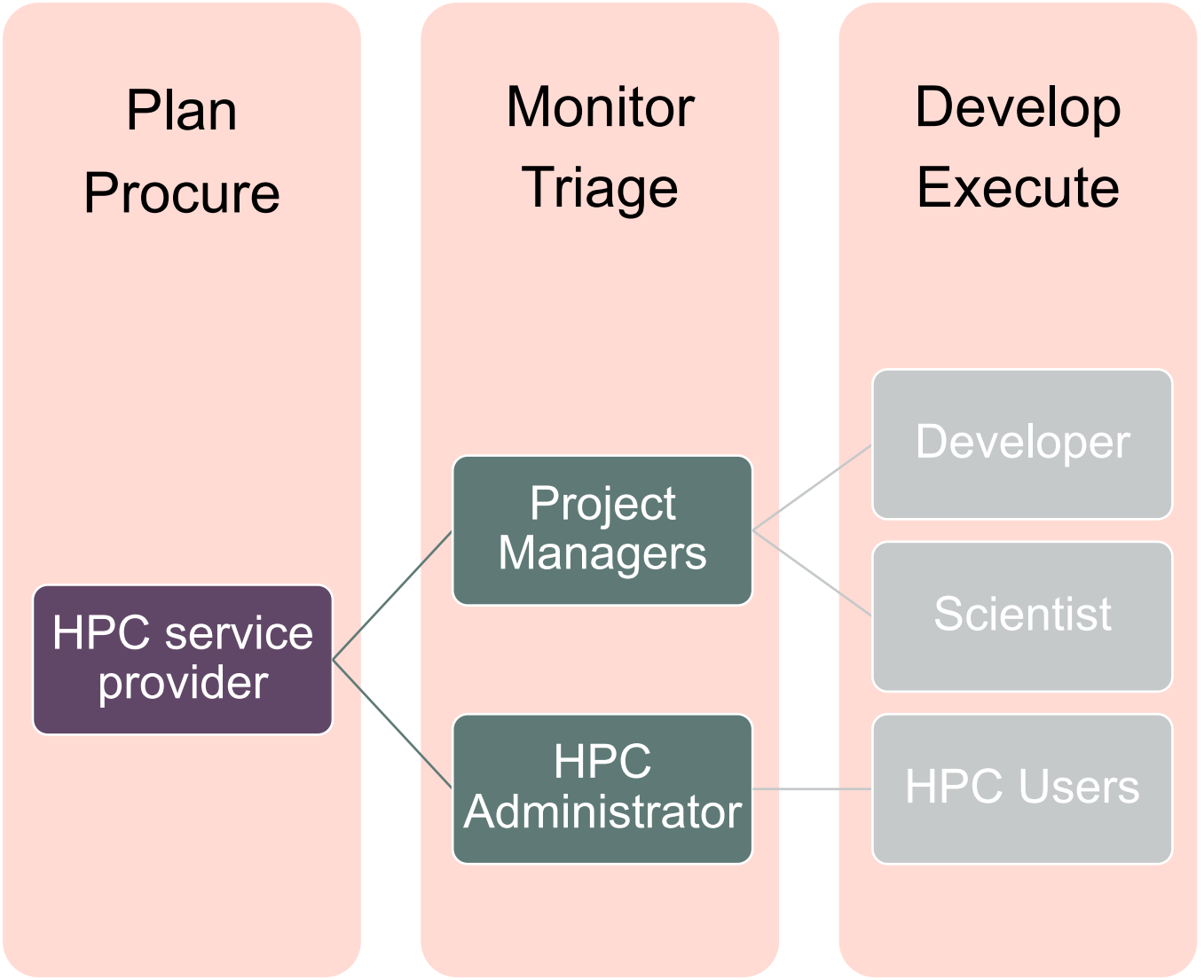
Hewlett Packard Enterprise

LASSi: an introduction

Gain better understanding of performance issues in a complex workload for a shared HPC system

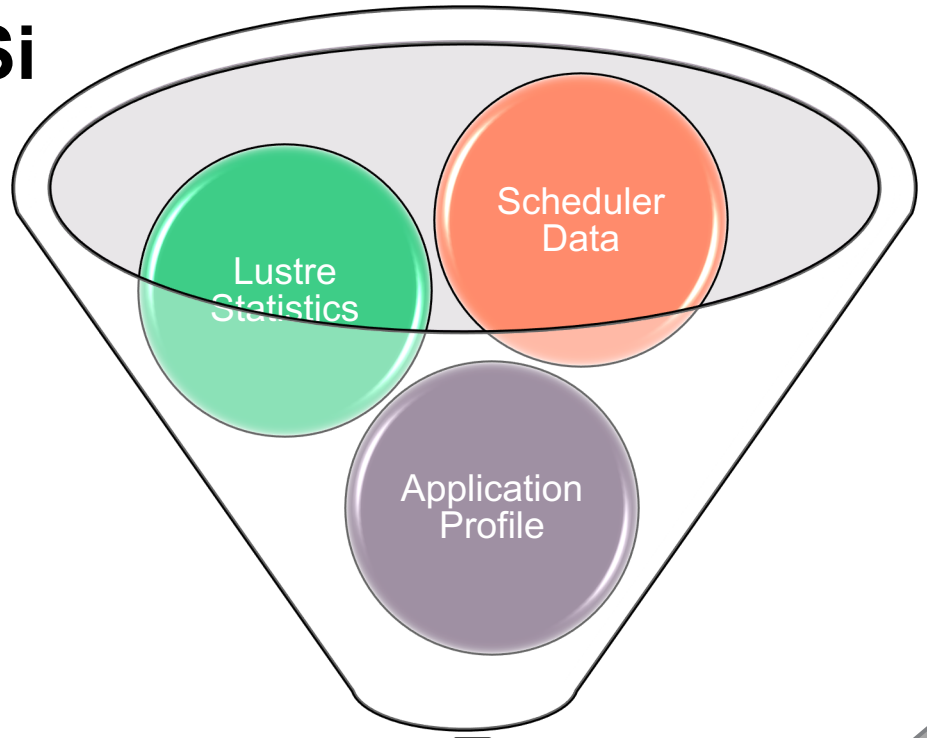
Why LASSi?

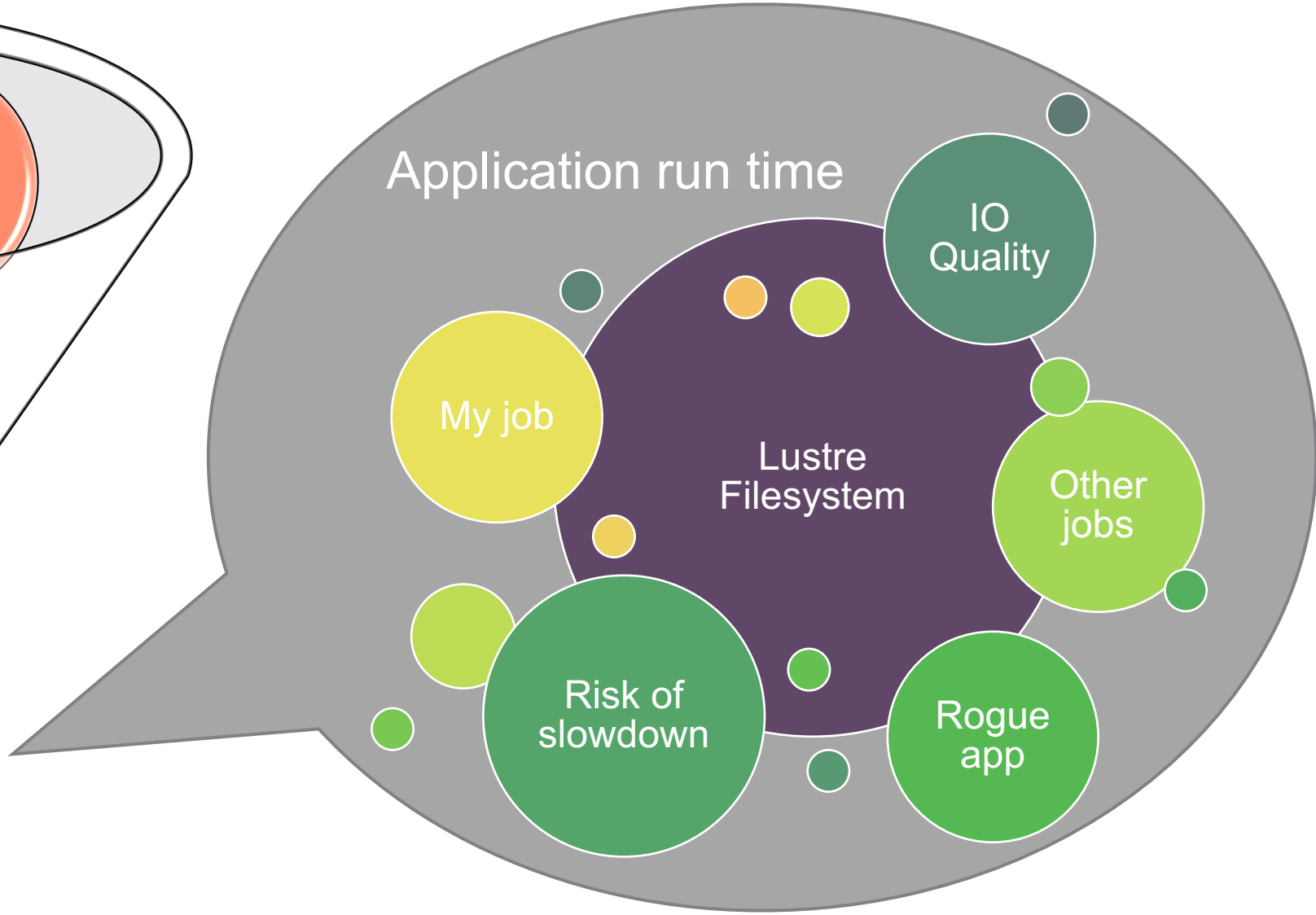




Hewlett Packard Enterprise

LASSi





Hewlett Packard Enterprise

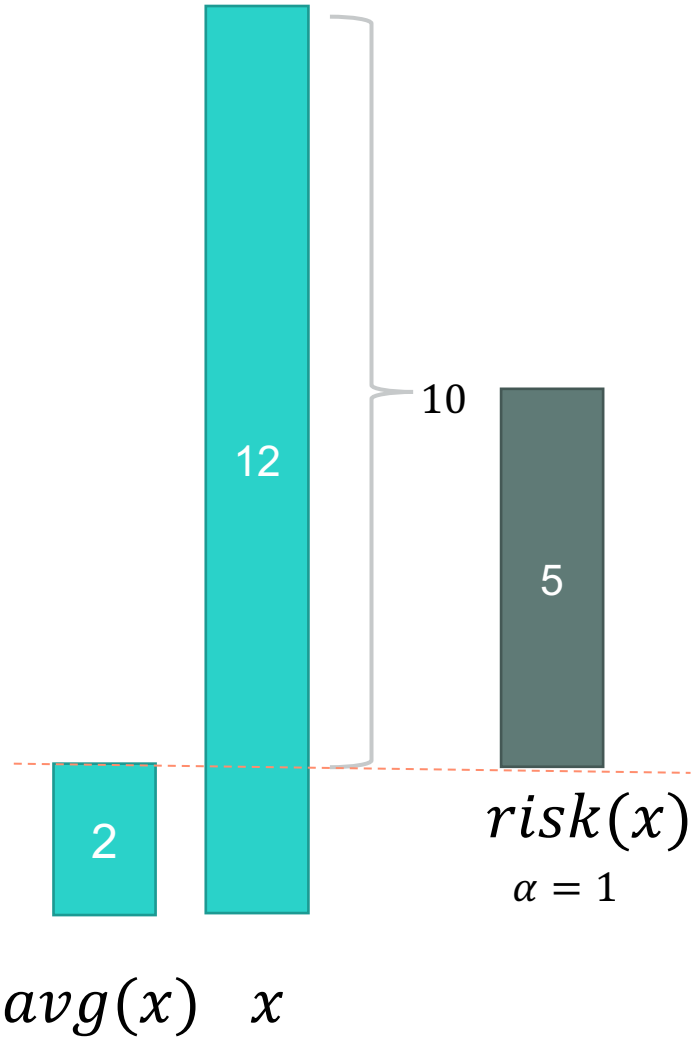
A different approach based on risks

- The simplest way to look at risks is perhaps:
- In isolation, slowdown will happen only when an application does more IO than expected (for example due to a configuration or code change)
- Also users will report slowdown only when they encounter more IO in a filesystem than expected
- We will use this idea as a metric for risks

Risk metrics

$$risk_{fs}(x) = \frac{x - \alpha * avg_{fs}(x)}{\alpha * avg_{fs}(x)}$$

- x is any IO operation OSS or MDS
- Risk is calculated for each application run
- We use averages for IO operation for each filesystem
- We calculate risk as
  scale of deviation from α times the avg on a filesystem
- Higher value of risk denotes a higher risk of slowdown



Metrics for IO

Risk

$$risk_{oss} = risk_{read\_kb} + risk_{read\_ops} + risk_{write\_kb} + risk_{write\_ops} + risk_{other}$$

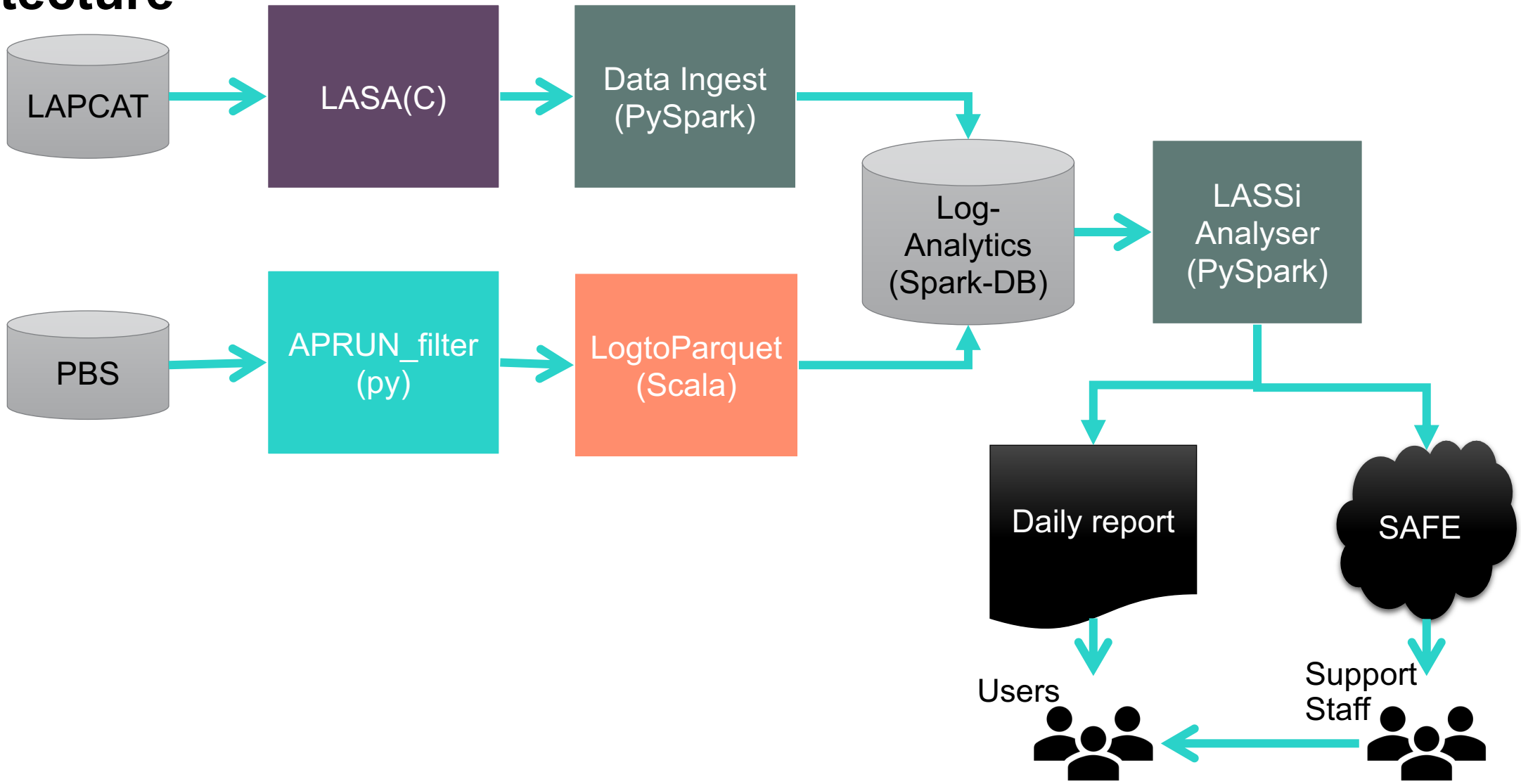
$$\begin{aligned} risk_{mds} = {} & risk_{open} + risk_{close} + risk_{getattr} + risk_{setattr} + risk_{mkdir} \\ & + risk_{rmdir} + risk_{mknod} + risk_{link} + risk_{unlink} + risk_{ren} \\ & + risk_{getxattr} + risk_{setxattr} + risk_{statfs} + risk_{sync} + risk_{cdr} + risk_{sdr} \end{aligned}$$

Quality

$$read\_kb\_ops = \frac{read\_ops * 1024}{read\_kb}$$

$$write\_kb\_ops = \frac{write\_ops * 1024}{write\_kb}$$

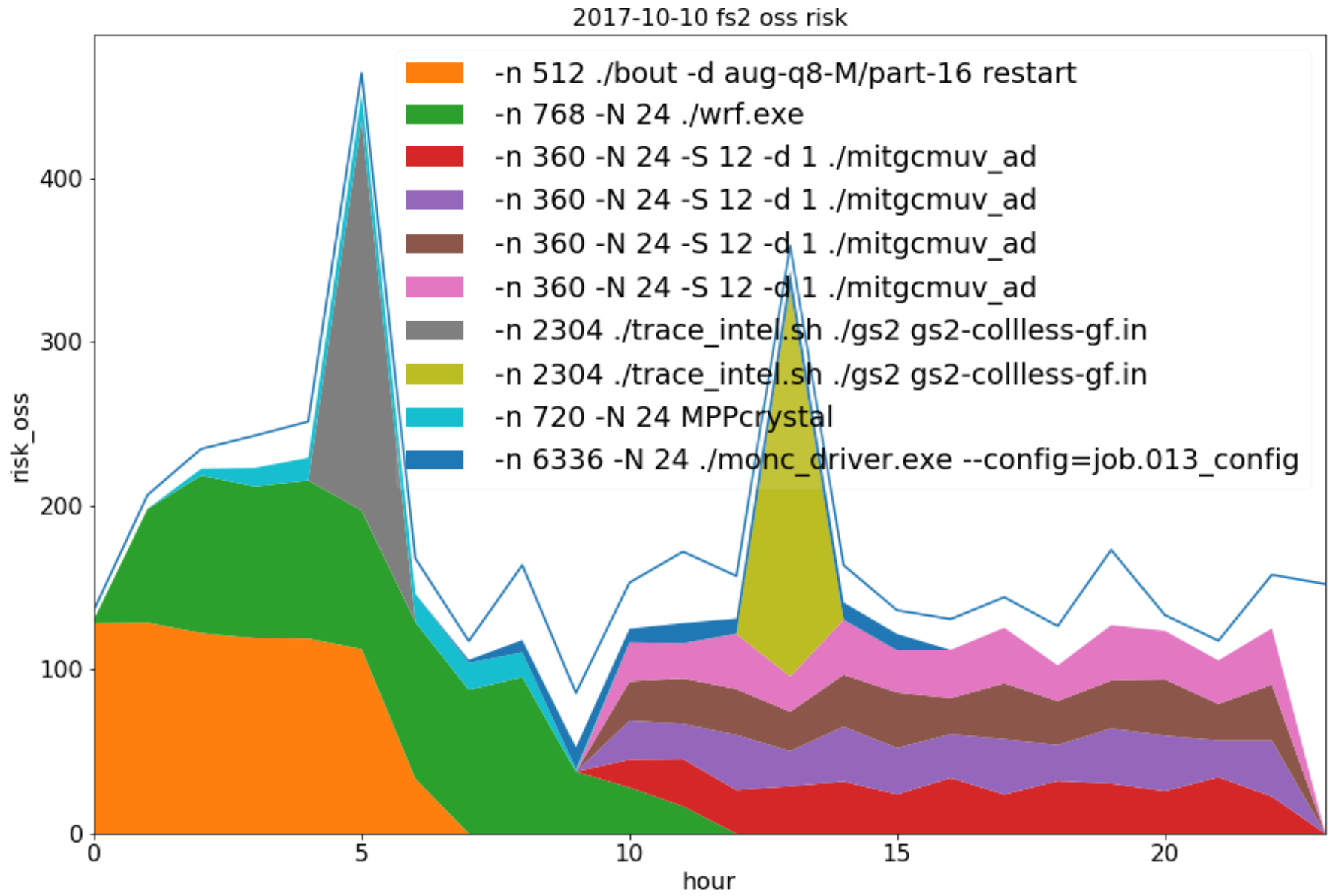
Architecture



LASSi – a tool for real-time analysis.

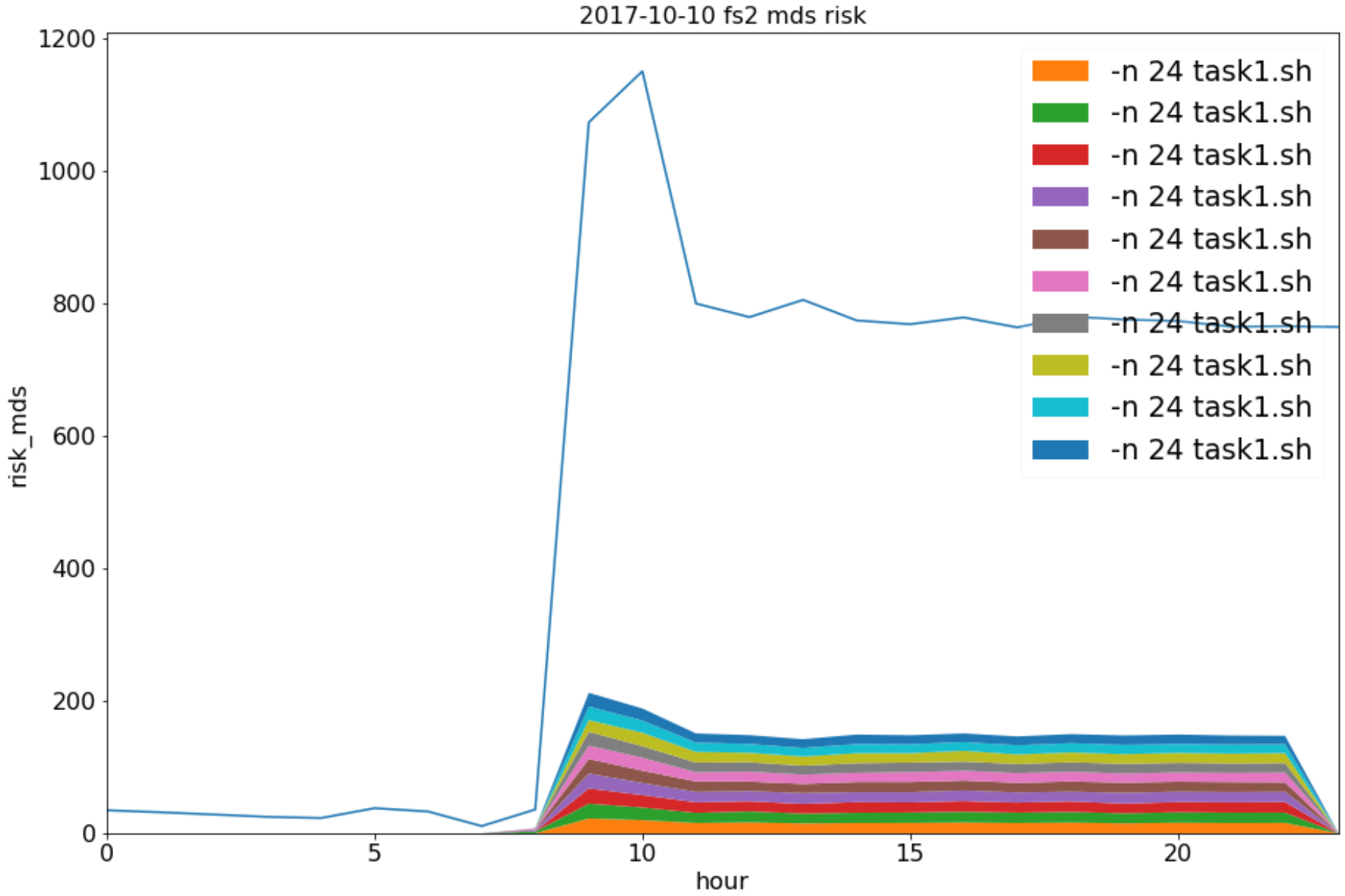
- Provides an automated metric based analysis of IO
- Risk model has been validated by comparison with actual reported slowdown incidents
- Quality model for studying efficiency of application IO
- LASSi offers
  - A coarse IO profile of each application running
  - Identification of abnormal filesystem and application IO usage
  - Identification of exact times when the filesystem is at risk of slowdown
  - Identification of exact applications causing the risk of slowdown

Example of displays for helpdesk - daily risk to oss



Hewlett
Enterpris

Example of displays for helpdesk - daily risk to mds



ARCHER Analysis

Based on data from April 2017 to November 2019

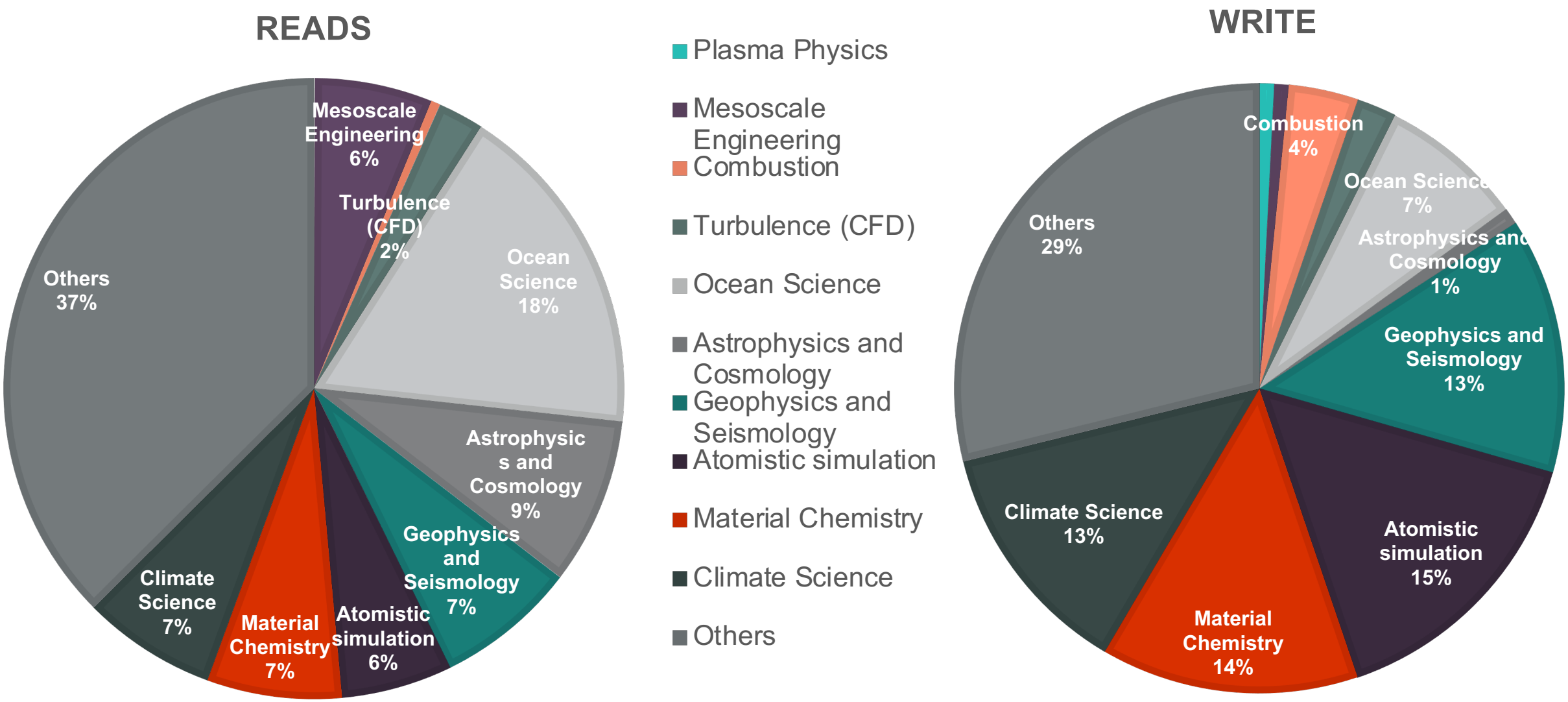
So what about Exascale?

- Technology
  - Initial Exascale systems will still use Lustre (gperformant with NVRAM)
  - We are likely to move to object store (key-value store as backend)
  - On top of this will use standard APIs like MPI-IO, NetCDF, HDF5
  - Will still want a POSIX layer (with its scaling limitations)
  - Projects like DAOS are interesting with increasing AI, Big Data applications
- Instrumentation and analysis still important
- Can we spot trends in applications/science as we move forward?
- Are we seeing changes today on ARCHER?

ARCHER Projects

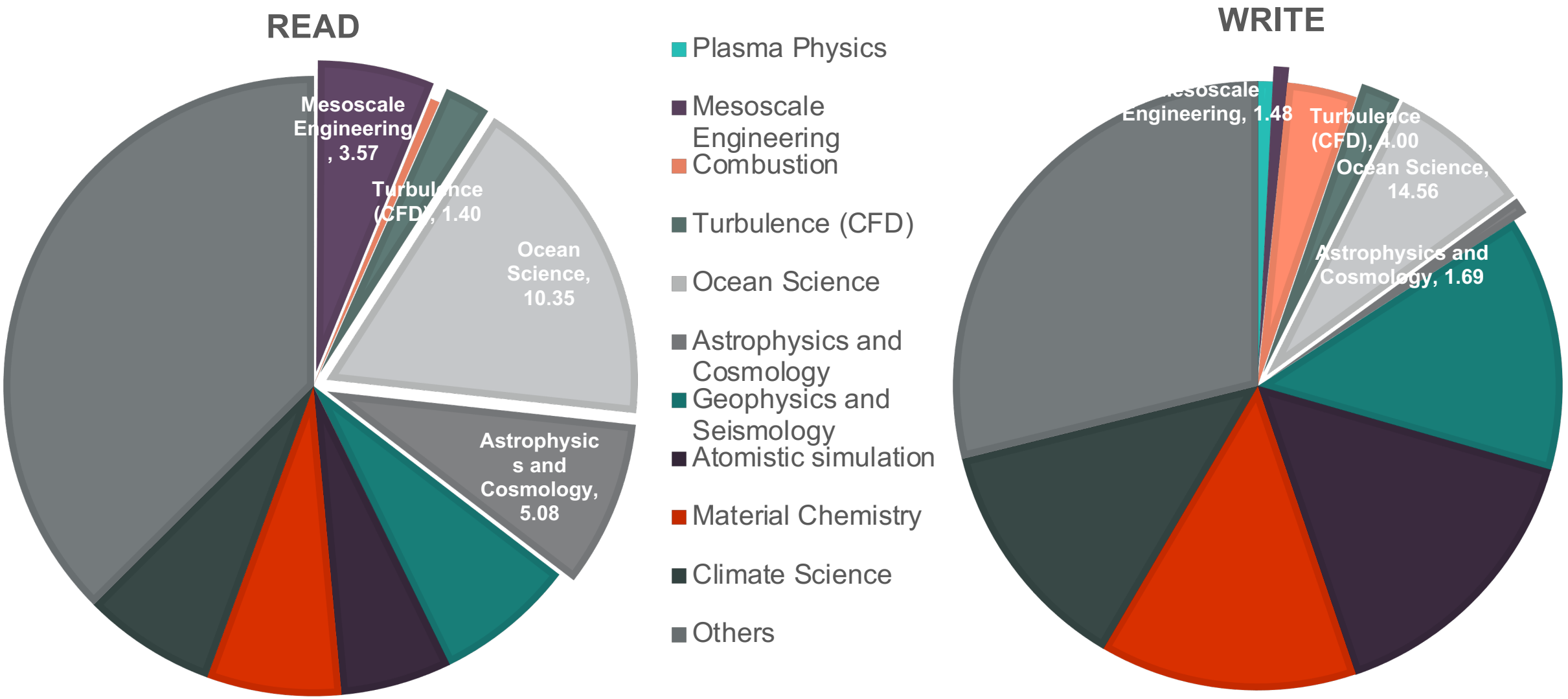
Mesoscale Engineering	: lammps, Foam
Turbulence	: HYDRA, incompact3D, solver
Combustion	: boffin, senga, Foam
Ocean Science	: OPA, Nemo, mitgcmuv
AstroPhysics and Cosmology	: UKRmol
GeoPhysics and Seismology	: vasp, buildcell, axisem3d, wein2k
Atomistic Simulation	: castep, vasp, elk
Material Chemistry	: aims, vasp, nwchem
Climate Science	: UM_atmos, mitgcmuv, nemo, wrf

ARCHER: Read ~ 59 PB, Write ~ 192 PB



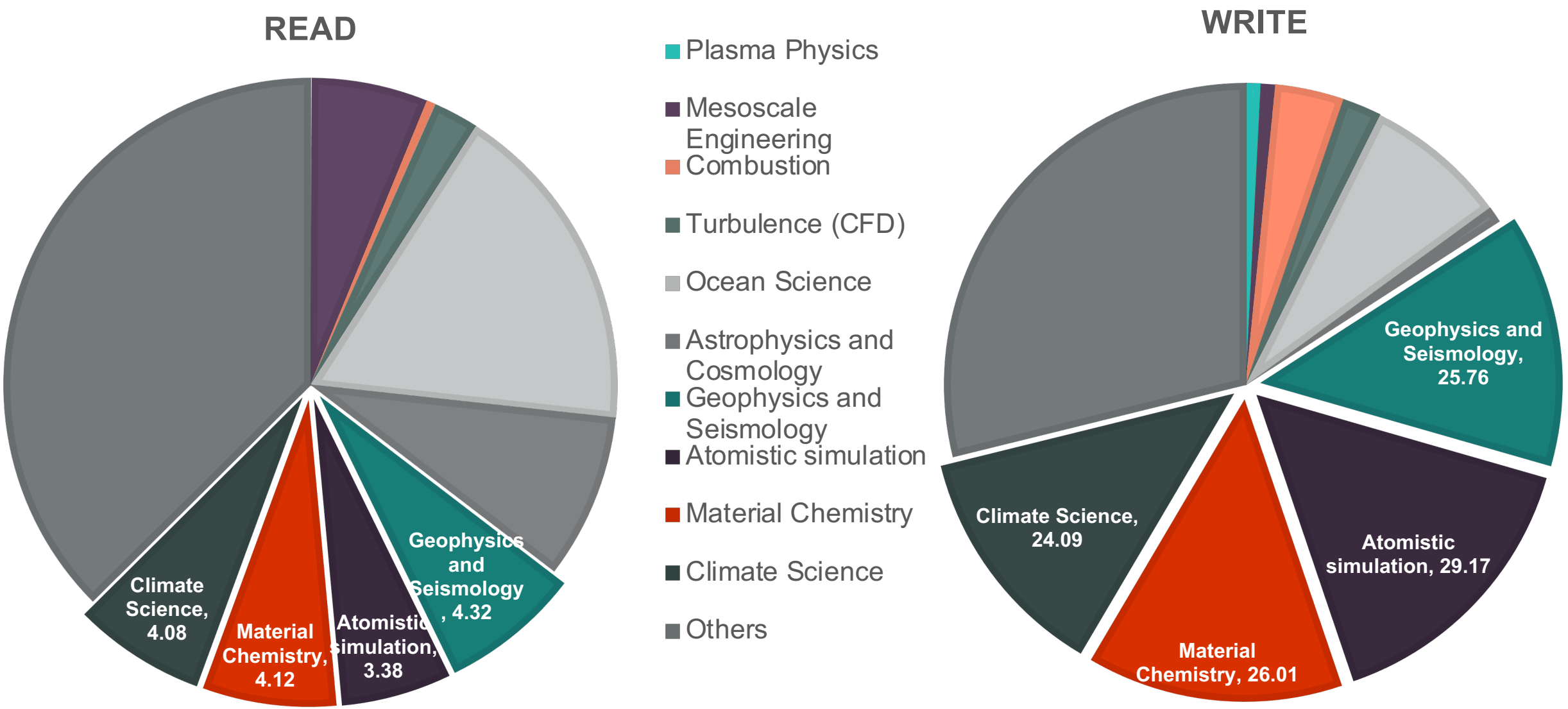
Hewlett Packard Enterprise

ARCHER: Read ~ 59 PB, Write ~ 192 PB



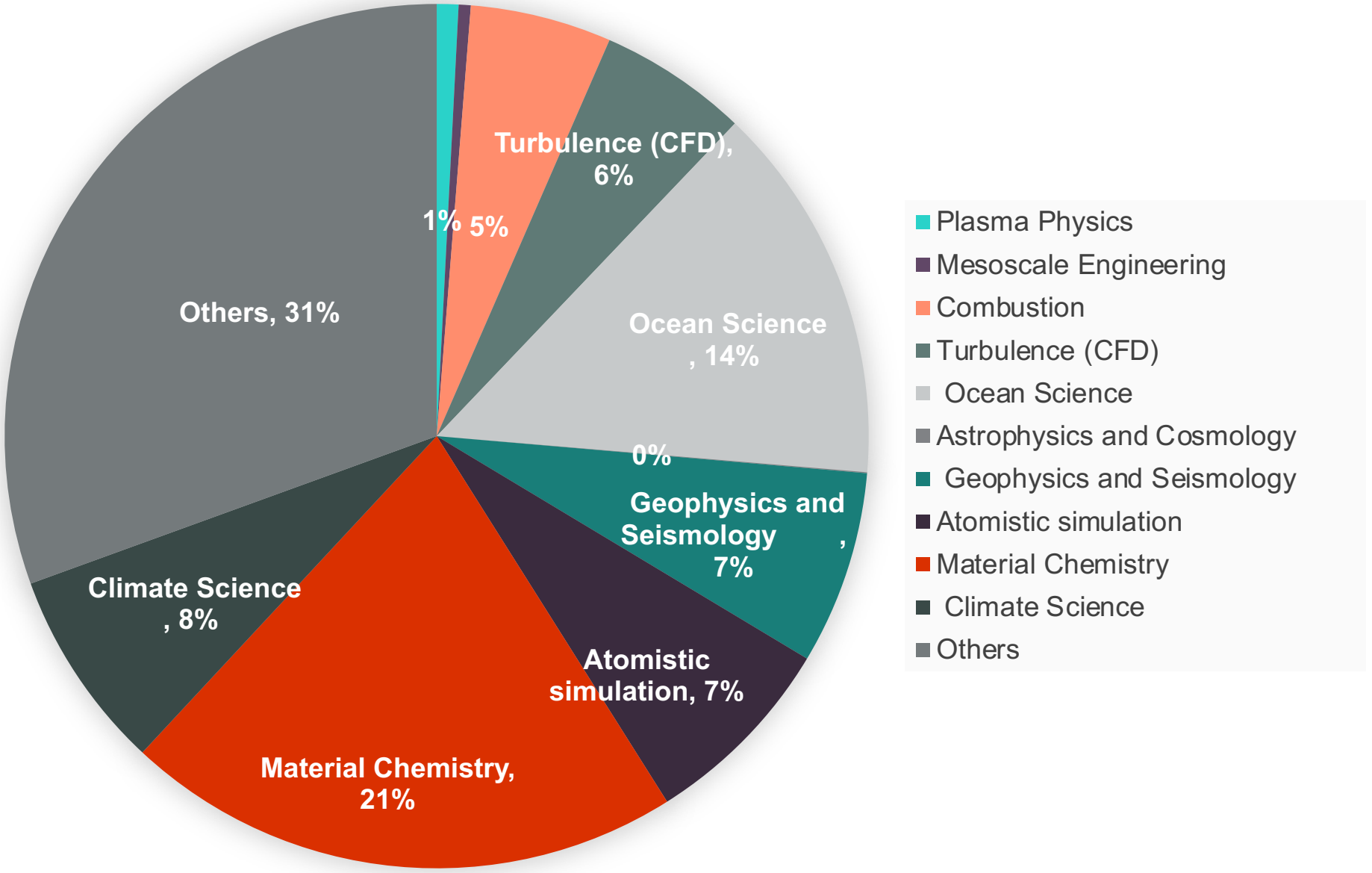
Hewlett Packard Enterprise

ARCHER: Read ~ 59 PB, Write ~ 192 PB

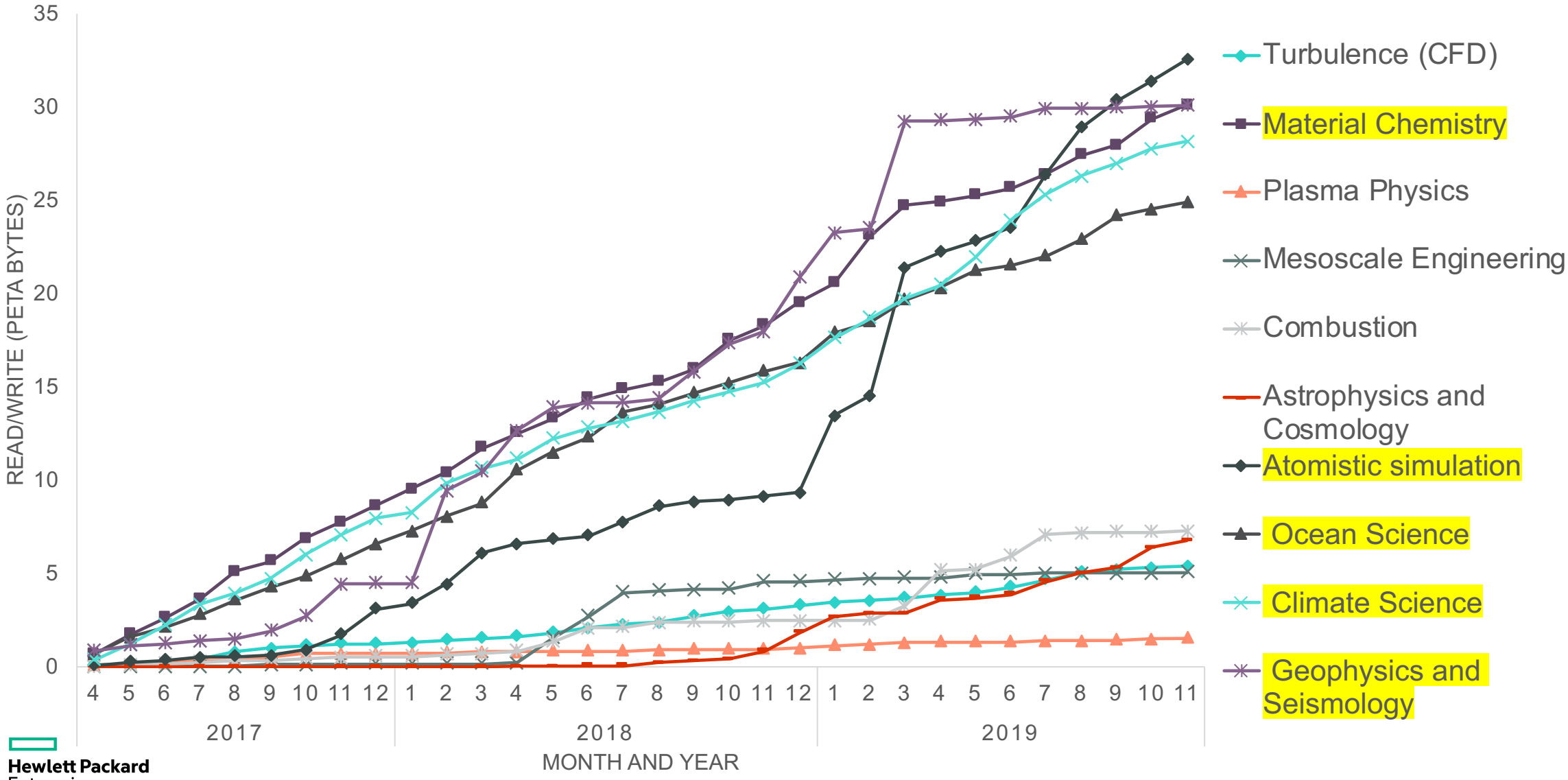


Hewlett Packard Enterprise

ARCHER MDS

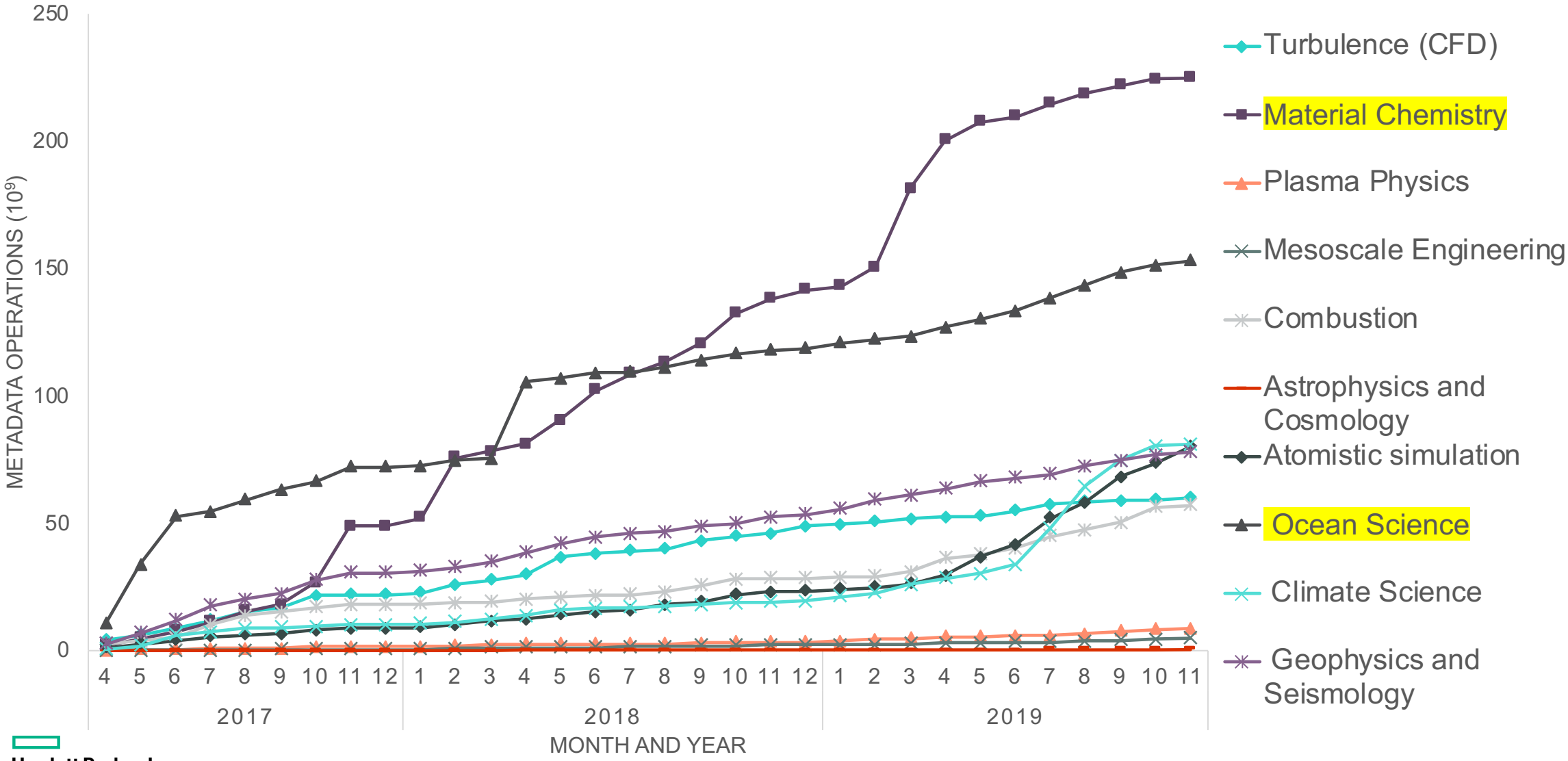

Turbulence (CFD), 6%
1%
5%
Others, 31%
Ocean Science , 14%
0%
Geophysics and Seismology , 7%
Climate Science , 8%
Atomistic simulation, 7%
Material Chemistry, 21%
Plasma Physics
Mesoscale Engineering
Combustion
Turbulence (CFD)
Ocean Science
Astrophysics and Cosmology
Geophysics and Seismology
Atomistic simulation
Material Chemistry
Climate Science
Others

Read/Write in ARCHER



Hewlett Packard Enterprise

Metadata operations in ARCHER

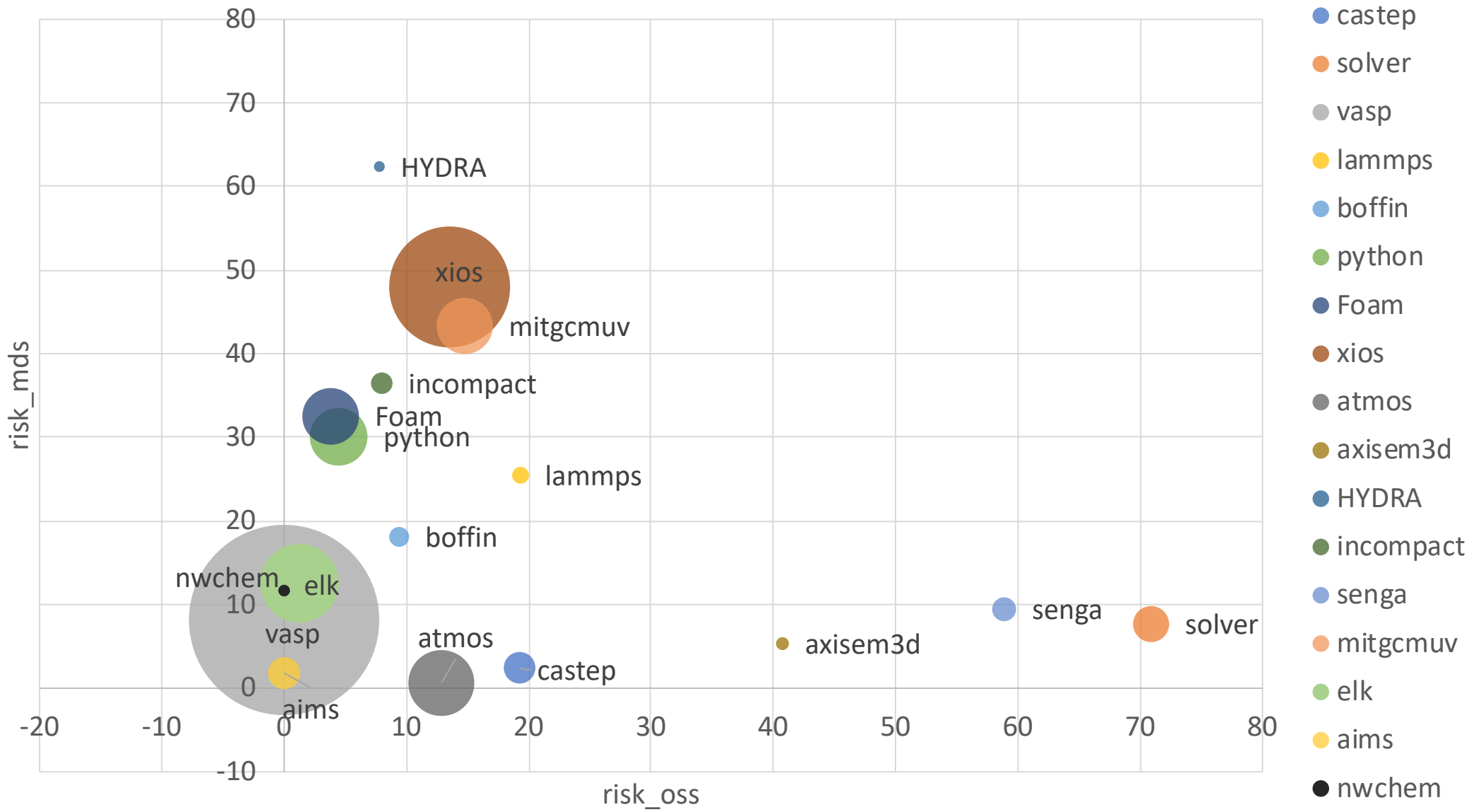


Hewlett Packard Enterprise

Applications

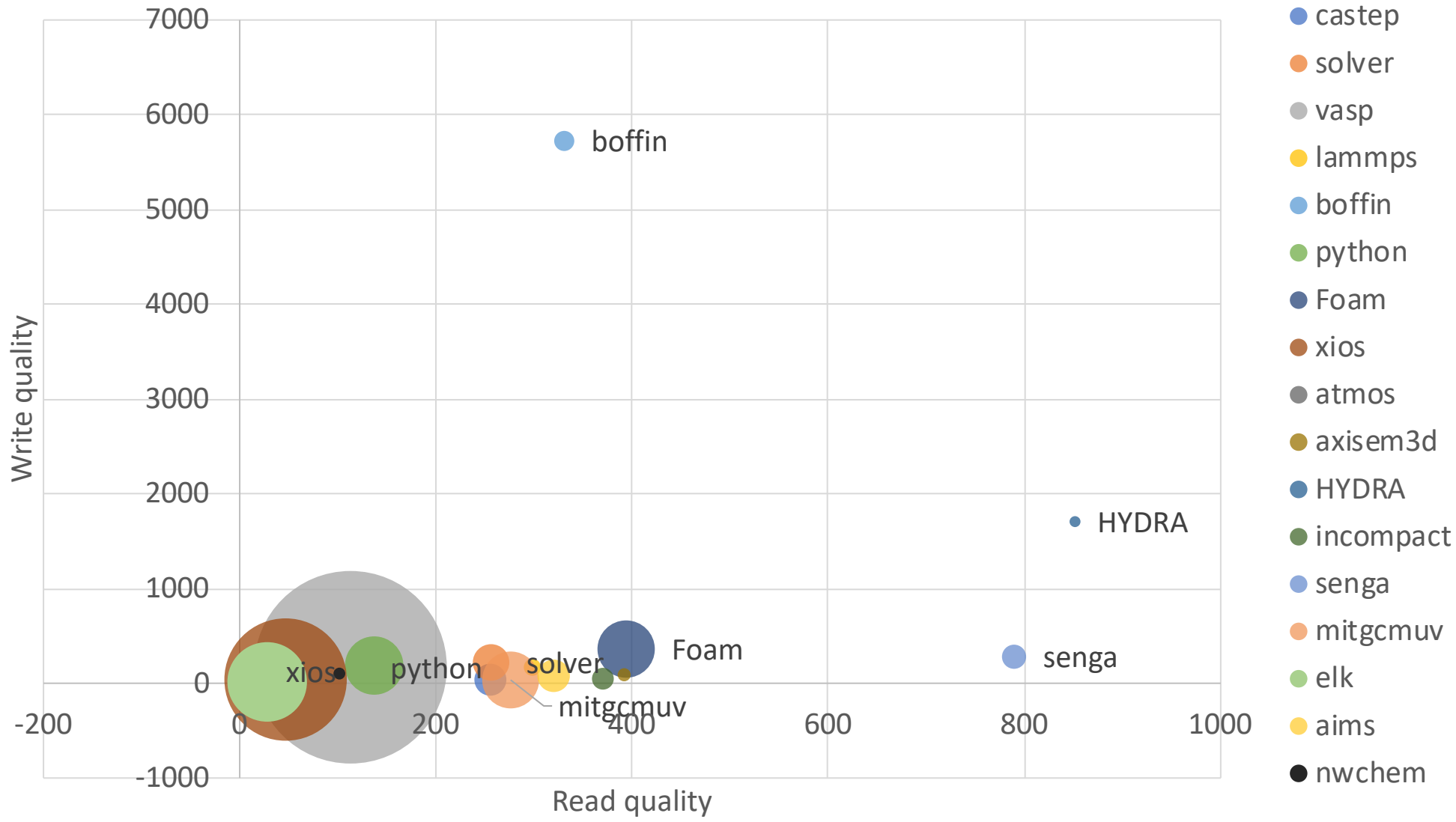
Code name	Application description
castep	Calculating properties of materials from first principles
solver	Flow Solver, https://www.ukturbulence.co.uk/flow-solvers.html .
vasp	Vienna Ab initio Simulation Package, https://www.vasp
boffin	Large Eddy Simulation(LES) code
lammps	Large-scale Atomic/Molecular Massively Parallel Simulator
python	Python based codes
Foam	Open Source Computational Fluid Dynamics (CFD) Toolbox
xios	XML-IO-Server - I/O management in climate codes
atmos	Numerical model of the atmosphere
axisem3d	Simulation of Seismic wave propagation
HYDRA	A Multi-physics Simulation Code
incompact	A high-order finite-difference flow solvers
senga	Direct Numerical Simulation (DNS) of turbulent combustion
mitgcmuv	MIT general circulation model
elk	all-electron full-potential linearised augmented-plane wave (LAPW) code
aims	ab initio molecular simulations
nwchem	Open Source High-Performance Computational Chemistry

Risk to OSS vs MDS



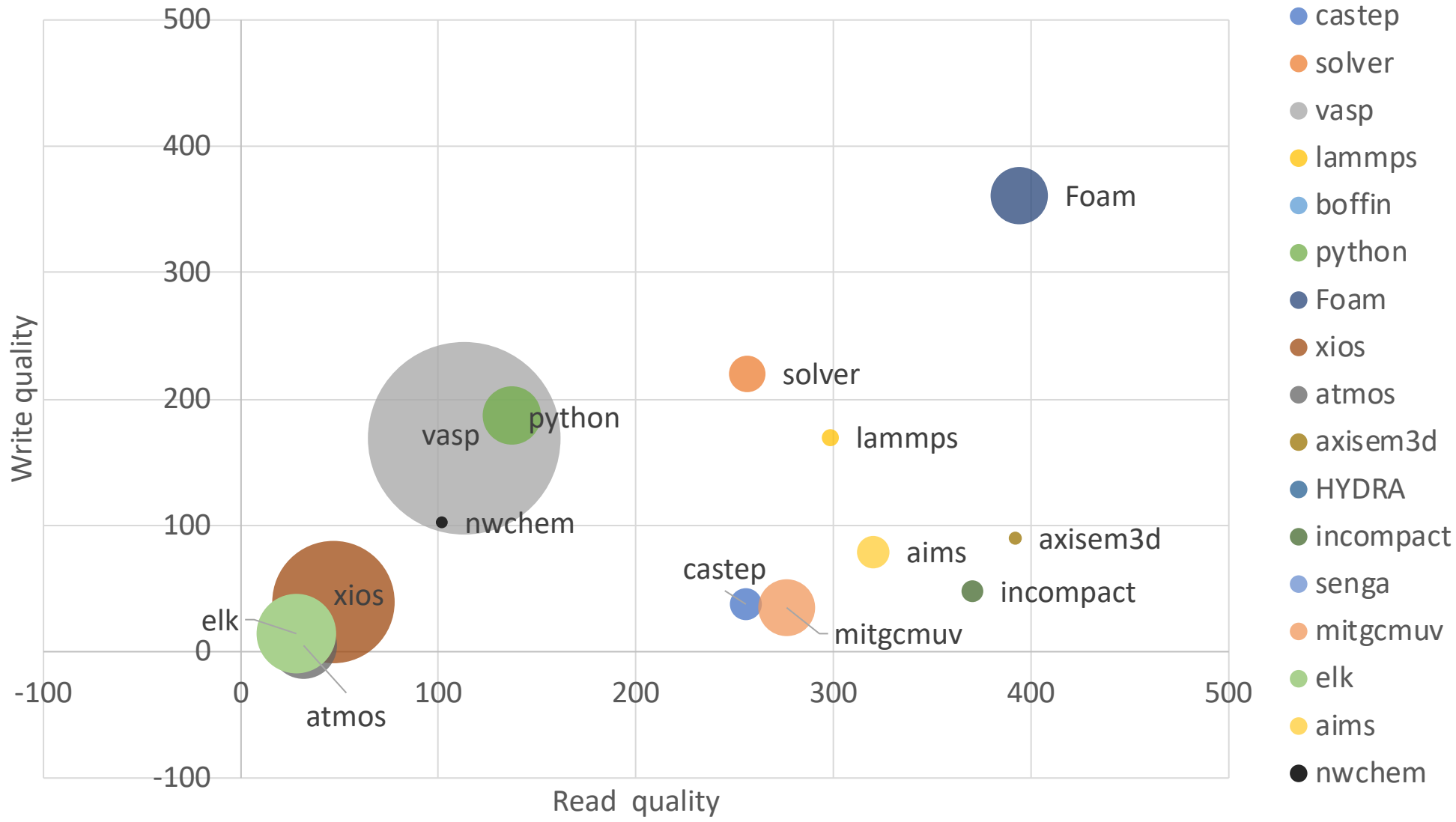
Hewlett Packard Enterprise

Read vs Write quality



Hewlett Packard Enterprise

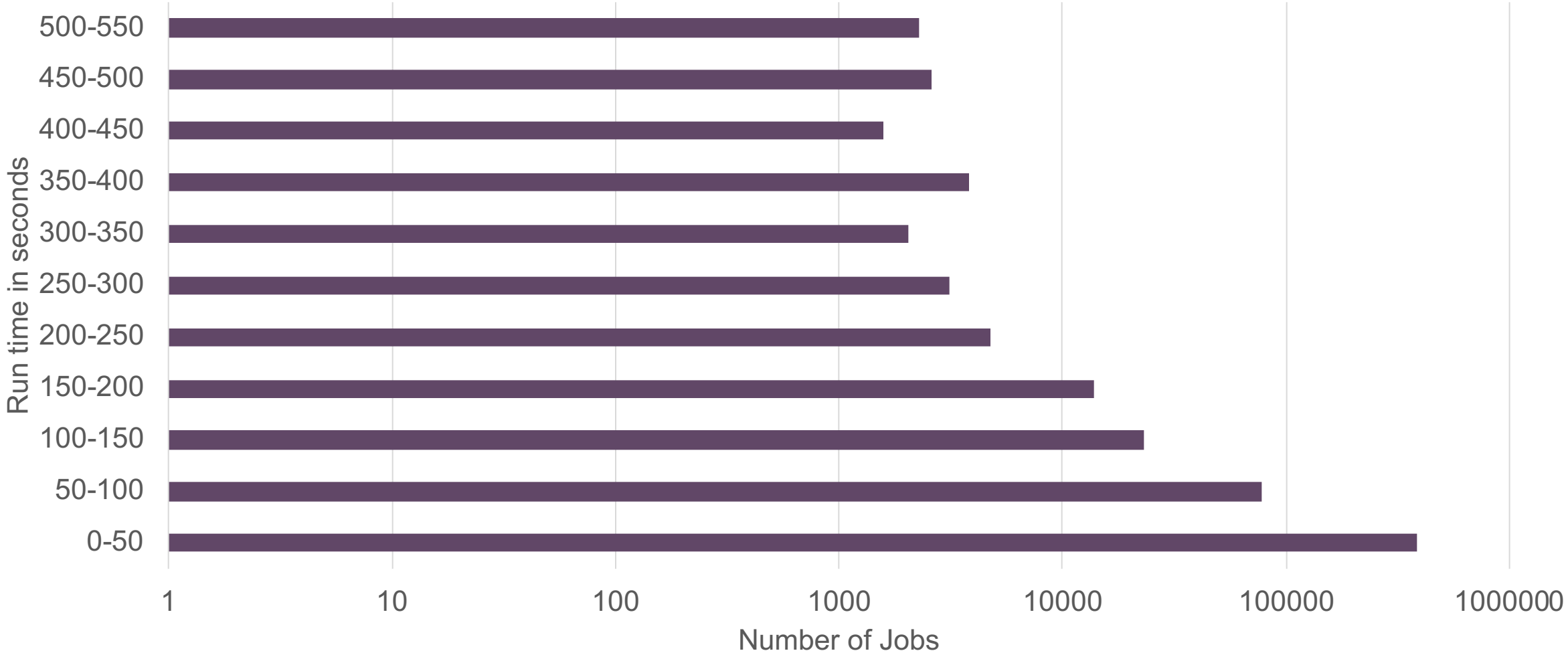
Read vs Write quality



LASSi in MetOffice

Analyse IO workloads of MetOffice collaboration machine

MetOffice job profile



Hewlett Packard Enterprise

MetOffice jobs profile

– Histogram of runtime shows 91% of the jobs run less than 6 minutes and 68% of the jobs run less than 50 seconds.

– Major proportion of jobs use shared nodes

– LAPCAT with 3 mins of data aggregation and limited/no shared node usage information is not suitable for MetOffice integration

– Cray View for ClusterStor is better suited for this workload.

Cray View for ClusterStor

– **HPC Storage System Monitoring (to be extended to full system monitoring in Shasta)**

– a monitoring and metrics software package

– collects and persists at a fine resolution of 30 seconds

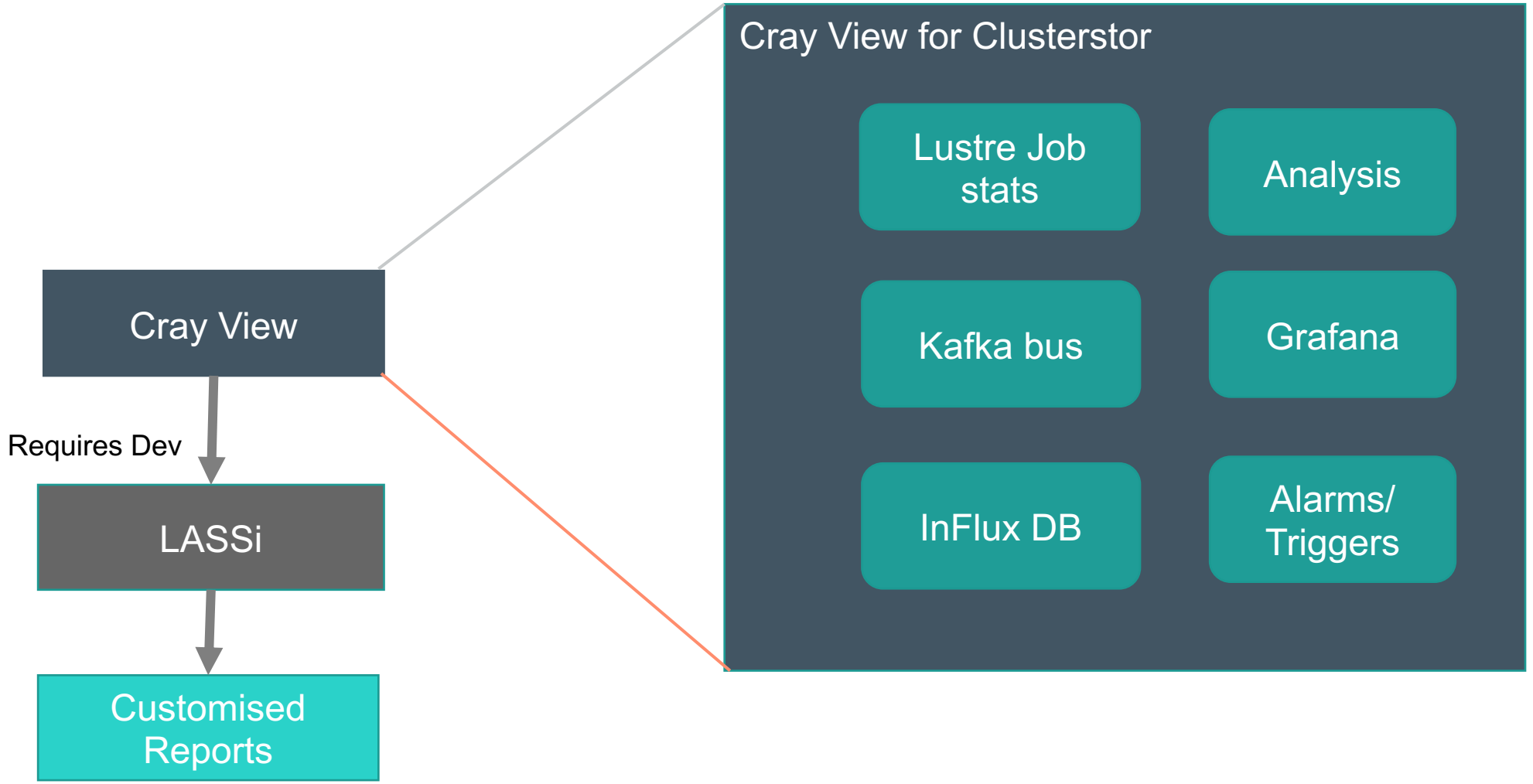
  – Lustre performance metrics

  – job metrics

– Monitors system events specific to storage

– Grafana based UI for continuous monitoring

MetOffice LASSi framework



Summary

- LASSi provides an application-centric, non-invasive approach based on metrics to analyse slowdown due to IO.
- Valuable in understanding application I/O behaviour on ARCHER. (ARCHER2 will have a much newer filesystem so the challenges and possibilities for optimisation will change)
- Different communities/applications stress the filesystem in different ways. For some communities these requirements are changing rapidly as the scale up
- Need to work with Project managers, Scientists and application developers to manage IO requirements and demands
- Continuous monitoring and analysis important in Exascale resource management.

Acknowledgements

We are done!
Any Questions

Karthee.Sivalingam@hpe.com

linkedin.com/company/cray-inc-/