Hewlett Packard Enterprise

2023

SUSTAINABLE EXASCALE

HPE'S HOLISTIC SYSTEM POWER AND ENERGY MANAGEMENT (HPM) VISION

Dr. Torsten Wilde
Andy Warner
Larry Kaplan

With support from: Dr. Christian Simmendinger, Andrew Nieuwsma, Jan Maeder, Marcel Marquardt

Mai, 2023

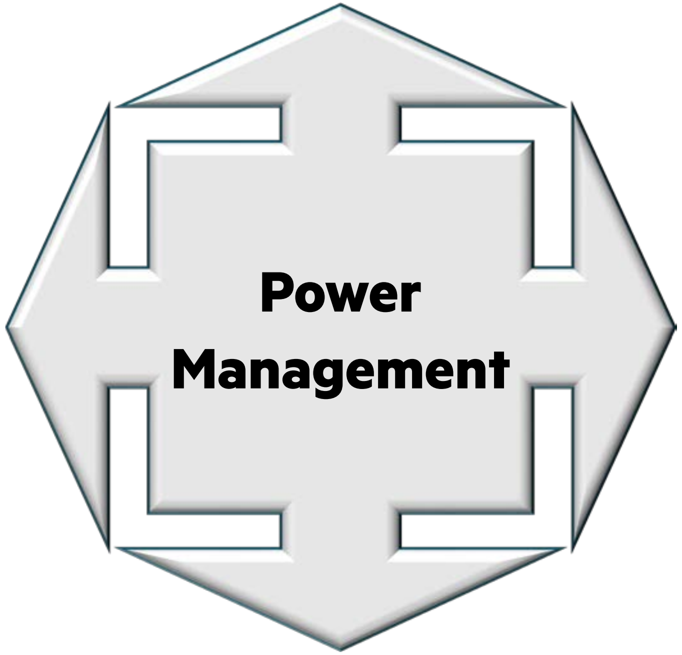
CUSTOMER CONCERNS

Optimized Job Performance under resource constraints

- Customers want to run hardware over-provisioned systems with for better overall system performance
- Need: Balance between available power/cooling and workload performance

Minimize Energy Consumption

- Customers want to reduce OPEX due to increased power needs of new technologies and increased energy prices
- Need: reduce energy consumption of workloads according to a TtS / EtS tradeoff metric



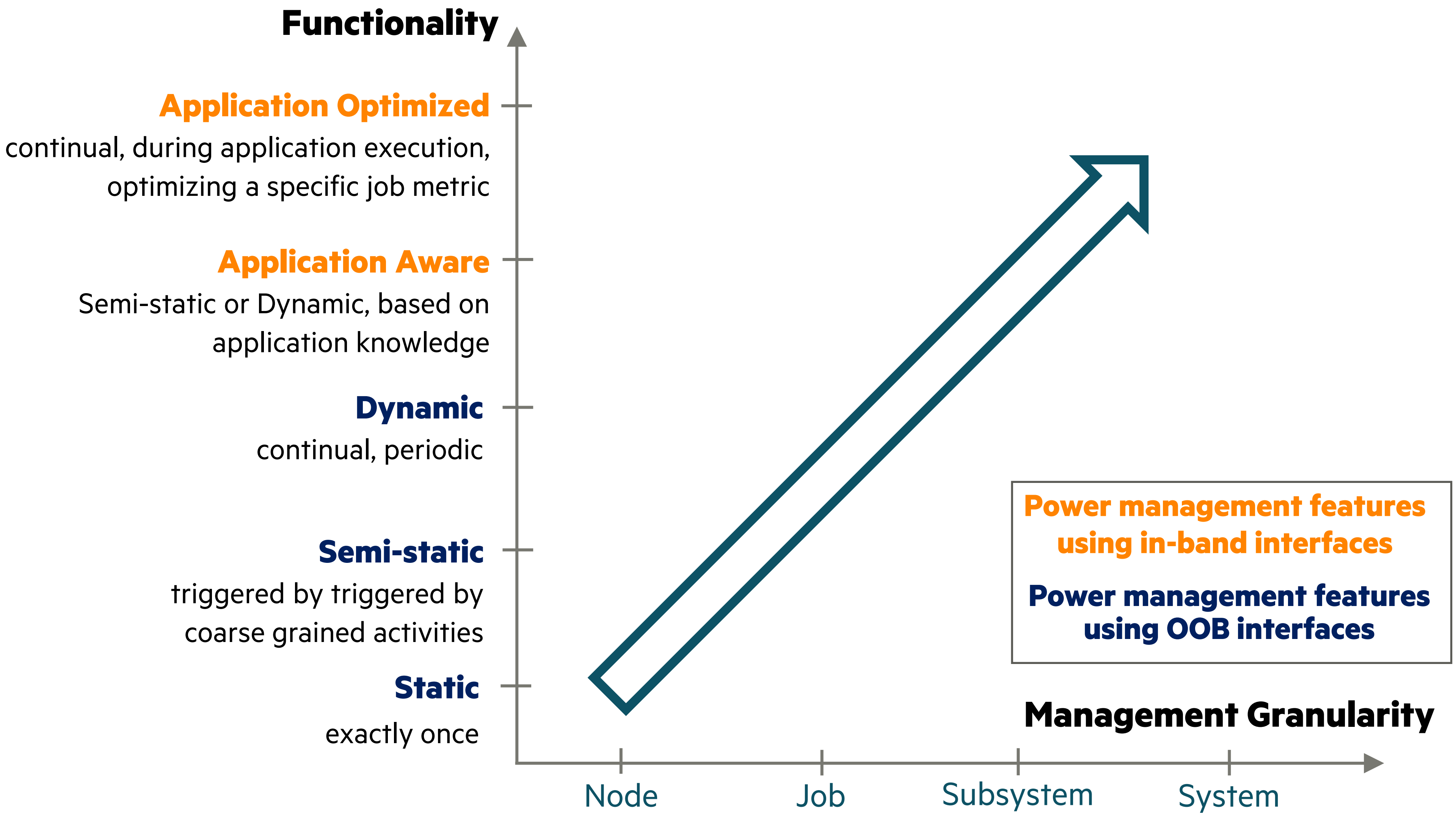
Data Center Sustainability

- Governments (US, EU) are developing mandates for our customers to address sustainability aspects
- Need: bring down current energy usage and carbon footprint, optimize system operation according to data center TCO (balance facility efficiency with system operation)

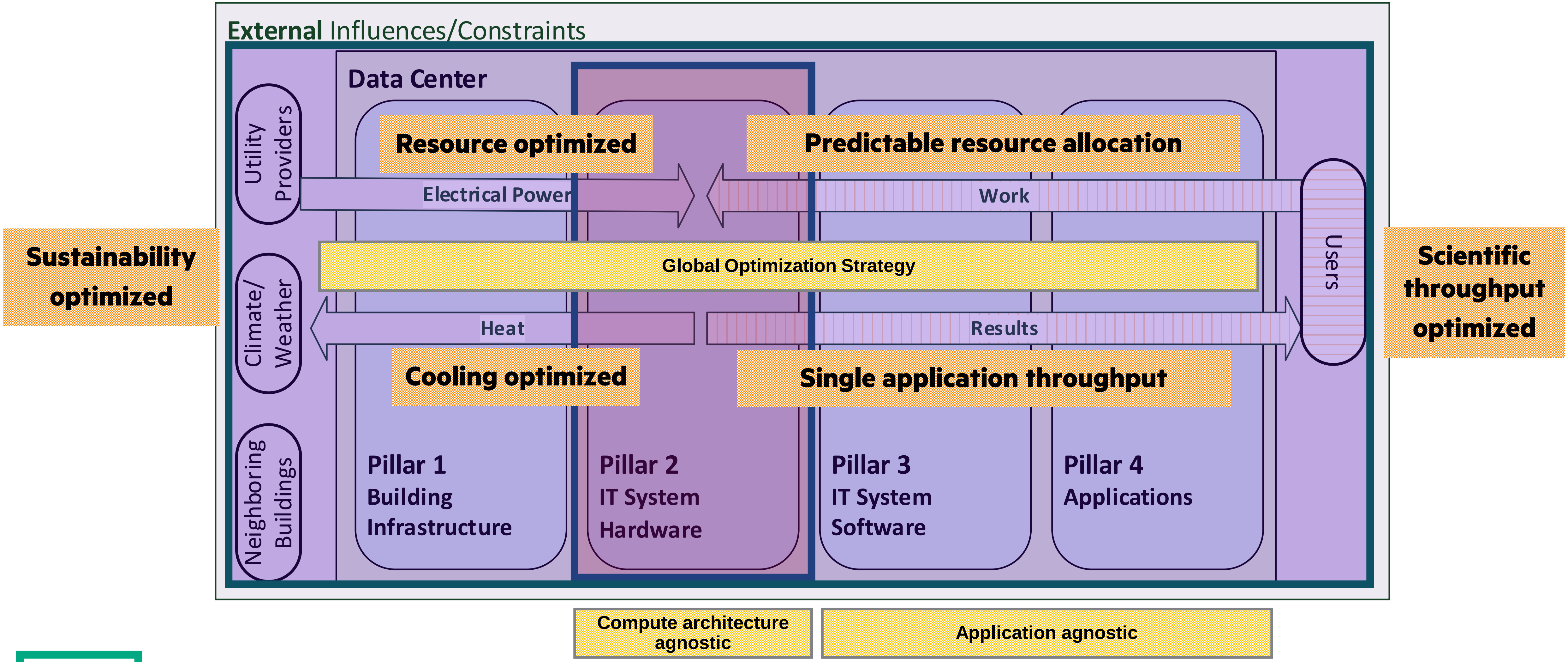
Maximize Resource Utilization

- Customers need to optimize power and cooling needs to support sustainable HPC efforts
- Need: optimized use of available resources, minimize stranded capabilities (power, cooling) in datacenter and HPC system

POWER/ENERGY MANAGEMENT CAPABILITIES



HOLISTIC POWER MANAGEMENT APPROACH – BREAKING SILOS



AUTOMATED POWER AND ENERGY MANAGEMENT

Customer Defined Preference:

Least Resource Usage ⟷ **Max Performance**

Customer

Automation

Constraint:

- **Power**
- **Cooling**
- **OPEX**
- **Sustainability**

HPE automatic workload control delivers desired outcome

- **Monitoring**
- **Data**
- **Decisions**
- **Reports**

CHALLENGES GOING FORWARD

- **Substantial increase in device power consumption (>1kW)**
  - Optimal power cap for best efficiency (like flop/s/W) depends on specific compute device and application

- **Substantial increase in core counts**
  - Serverless computing
  - Asymmetric and asynchronous applications and programming models
  - Addition of none-bulk synchronous workloads
  - More granular controls needed

- **Creation of heterogeneous compute devices**

- **Move to chiller-less cooling (also referred to as ‘free cooling’) environments**
  - Inlet cooling temperature changes with weather (potentially 40C or above)

- **Multitenancy and per-tenant power management**

RESEARCH AND DEVELOPMENT ACTIVITIES

PowerSched

Dr. Christian Simmendinger, Marcel Marquardt, Jan Maeder, Dr. Torsten Wilde

System Power Capping

Andrew Nieuwsma, Dr. Torsten Wilde

EE-HPC Project

Dr. Torsten Wilde, David Brayford, Dr. Christian Simmendinger, Marcel Marquardt

POWERSCHED

ONE SYSTEM TO PROTOTYPE IDEAS

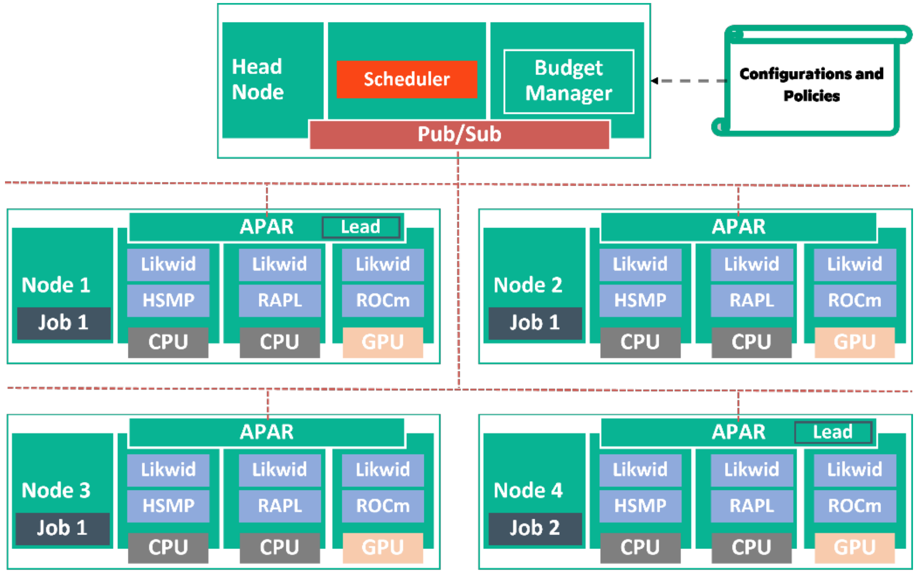
Overprovisioned Systems
- Holistic view on cluster
- Shift budget between nodes
- Scheduler Integration
- Distribution policies

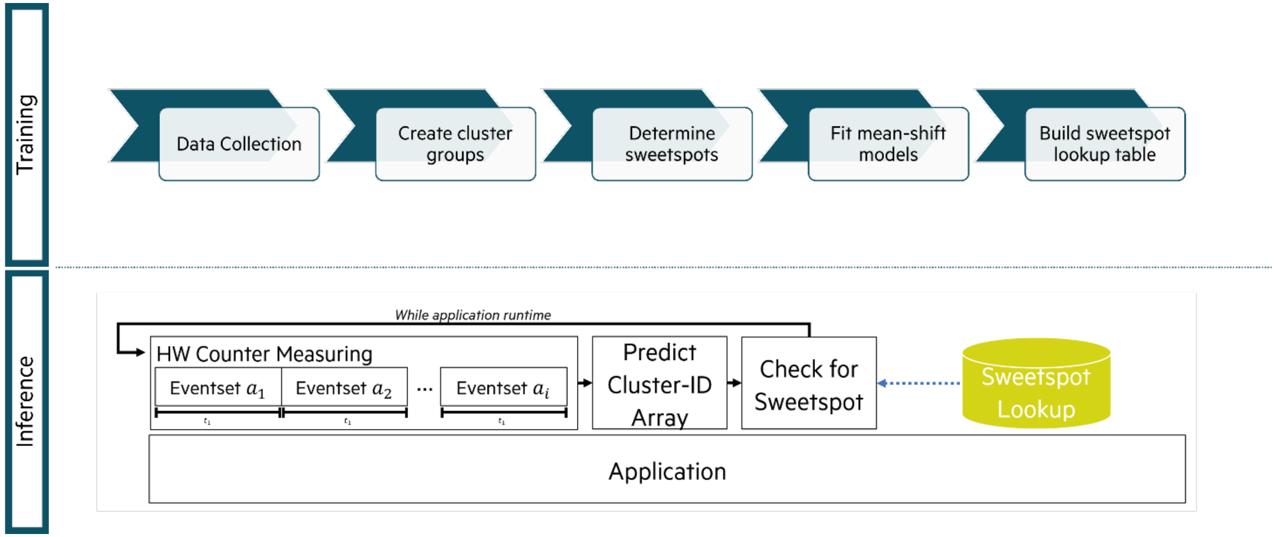
Heterogeneous Systems
- Multiple components in modern systems
- Support diverse landscape

Energy Optimization
- Use system metrics only
- Assign power level with best energy efficiency
- React to different app phases

Extensible framework that supports multiple user-defined components

Clustering approach that can **save up to 14% energy** with only 2% runtime increase



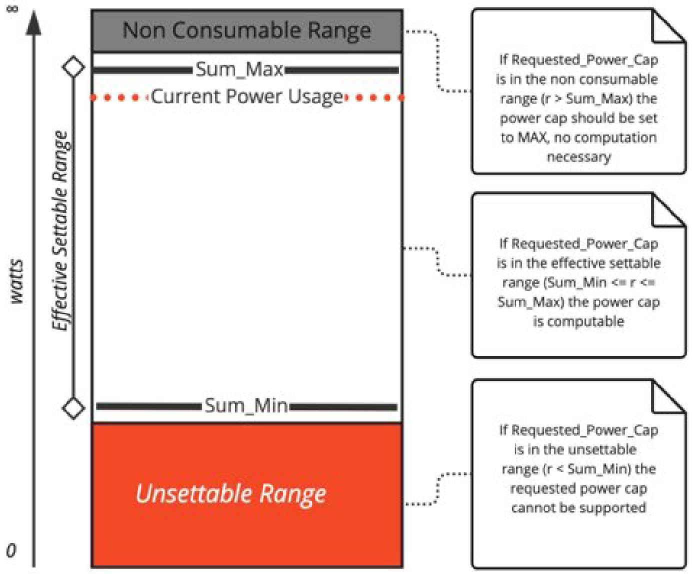


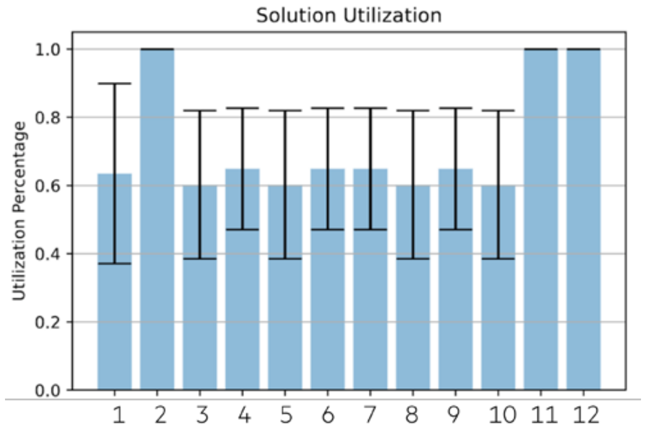
SYSTEM POWER CAPPING TOOL

ONE SYSTEM FOR OOB POWER CAPPING

	Node Type 1	Node Type 2
Node Architecture	Homogeneous	Heterogeneous
Node Composition	2 CPU, 0 GPU	1 CPU, 4 GPU
Min Power Cap in Watts	350	764
Max Power Cap in Watts	925	2754
Max - Min Power Cap (Delta) in Watts	575	1990
# nodes in system	1536	2560

- Set a system power cap
- Allocate node power based on different distribution algorithms and customer allocation policy
- Simplifying the process of setting power caps on a heterogeneous system

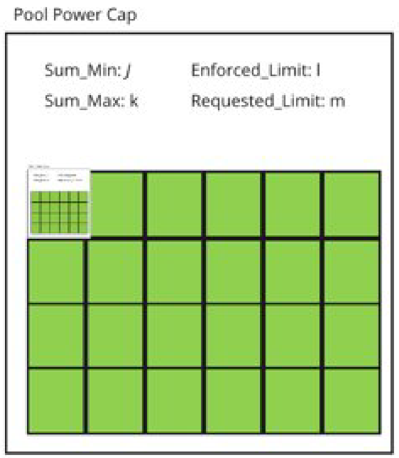




Comparison of Solution Utilizations Index	Algorithm	Node Type 1 (W)	Node Type 2 (W)
1	base_solution	350	764
2	count_down	350	1444
3	delete_by_component_count_least-to-most	925	764
4	delete_by_component_count_most-to-least	350	764
5	delete_by_delta_largest-to-smallest	350	764
6	delete_by_delta_smallest-to-largest	925	764
7	delete_by_max_power_cap_largest-to-smallest	350	764
8	delete_by_max_power_cap_smallest-to-largest	925	764
9	delete_by_min_power_cap_largest-to-smallest	350	764
10	delete_by_min_power_cap_smallest-to-largest	925	764
11	equal_percentage	517	1343
12	even_split	775	1189

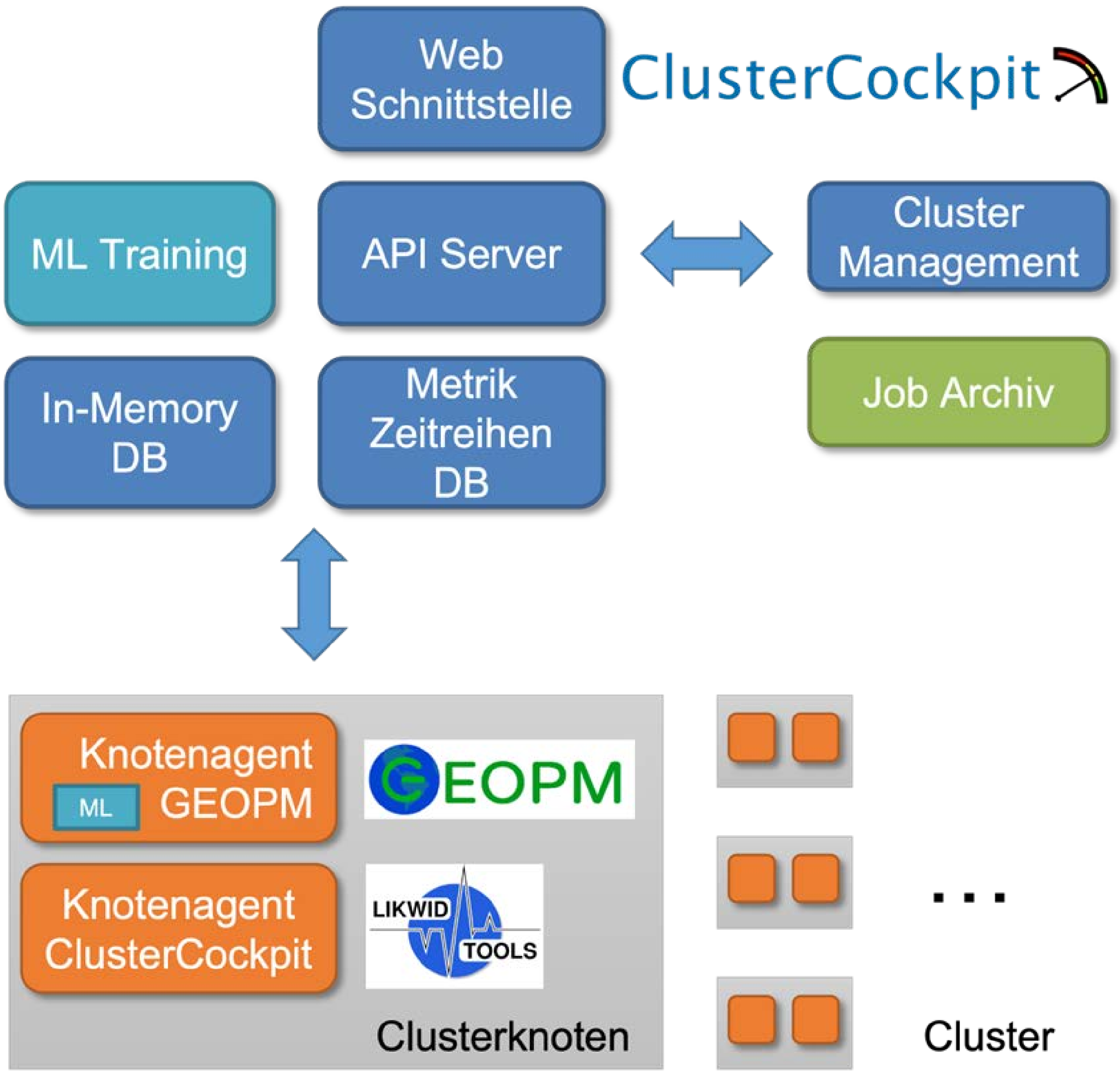
$$System\ Power = \sum_{i=1}^{S} no\ control\ consumers_{Nameplate\ Power} + \sum_{j=1}^{N} \left(NodePower_{Base\ Power} + \sum_{k=1}^{C} CPU_k + \sum_{l=1}^{A} Accel._l \right)$$

- Investigate how System Power Capping can be recursively applied (from data center to an individual node)
- Rationing – consider managing system power via pools (transient grouping of components that should share a powercap) and shifting power between those pools.



EE-HPC

ONE SYSTEM TO PROVIDE AN OPEN-SOURCE FOUNDATION



German funded project (start Q4/2022)

Friedrich-Alexander-Universität Erlangen-Nürnberg

DKRZ Deutsches Klimarechenzentrum

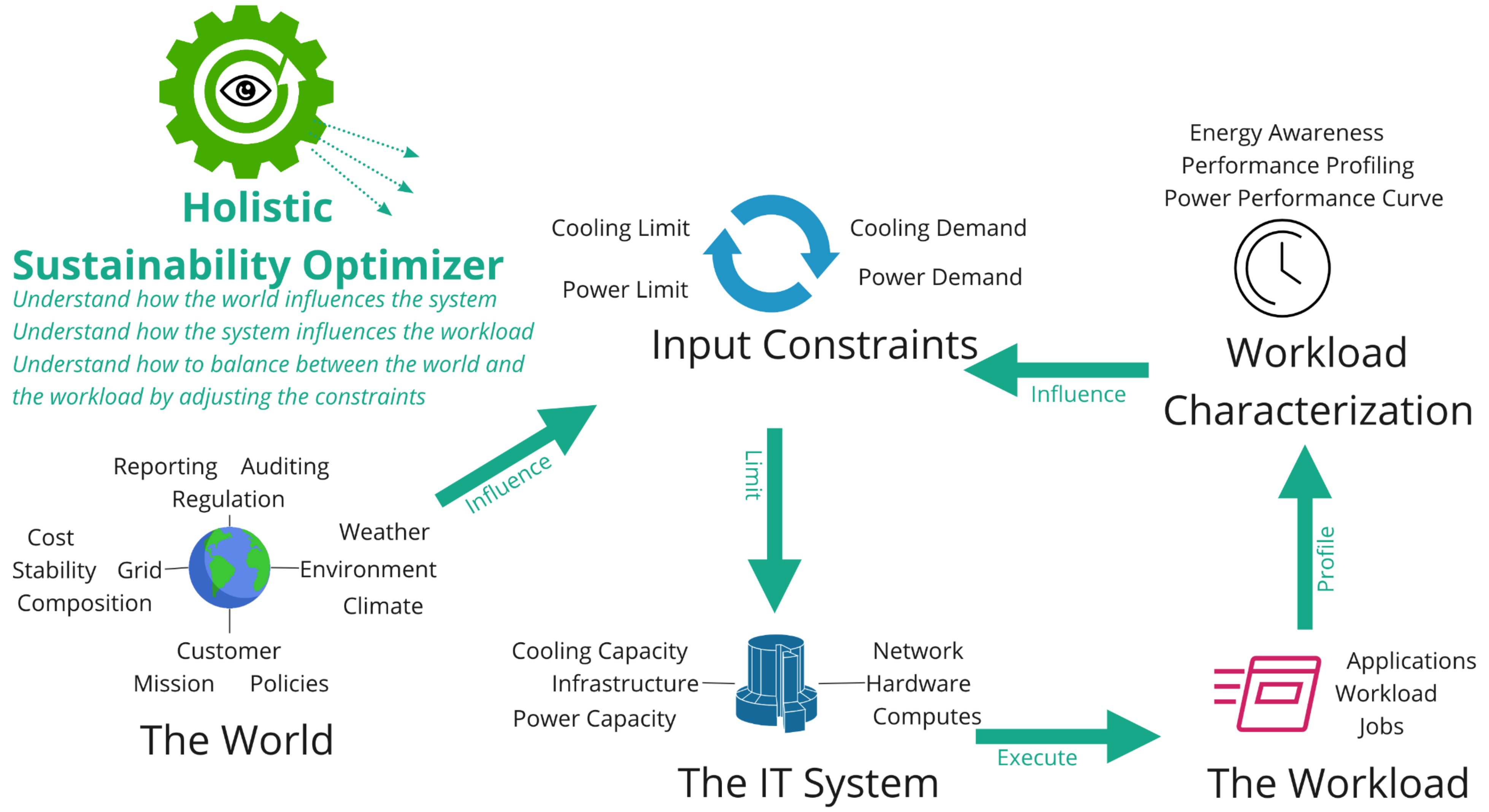
IT Center | RWTH Aachen University

HLRS High-Performance Computing Center Stuttgart

Hewlett Packard Enterprise

HOLISTIC SUSTAINABILITY OPTIMIZER

ONE SYSTEM TO RULE THEM ALL AND IN DARKNESS BIND THEM



THANKS

WILDE@HPE.COM