Hewlett Packard Enterprise

CUG2023
CRAY USER GROUP | SUSTAINABLE EXASCALE
Helsinki, Finland

Power Capping of Heterogeneous Systems

Andrew Nieuwsma, Principal Cloud Developer
Dr. Torsten Wilde, Master System Architect

Cray User Group 2023

Agenda

- Background
  - Changing Landscape
  - Customer Concerns
- Algorithm & Solution
- Methodology & Results
- Conclusion & Future Work

Changing Landscape

- Raising energy prices
- Expected increase in system power consumption
  - Frontier at ORNL consumes 21MW running Linpack
- Regulatory concerns around data center sustainability
  - Reduction of carbon footprint
  - Heat re-use

Why is heterogeneous power capping complex to solve?

- The equation looks simple.

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

- The implementation is complex because of heterogeneous systems with heterogeneous node architecture.

	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

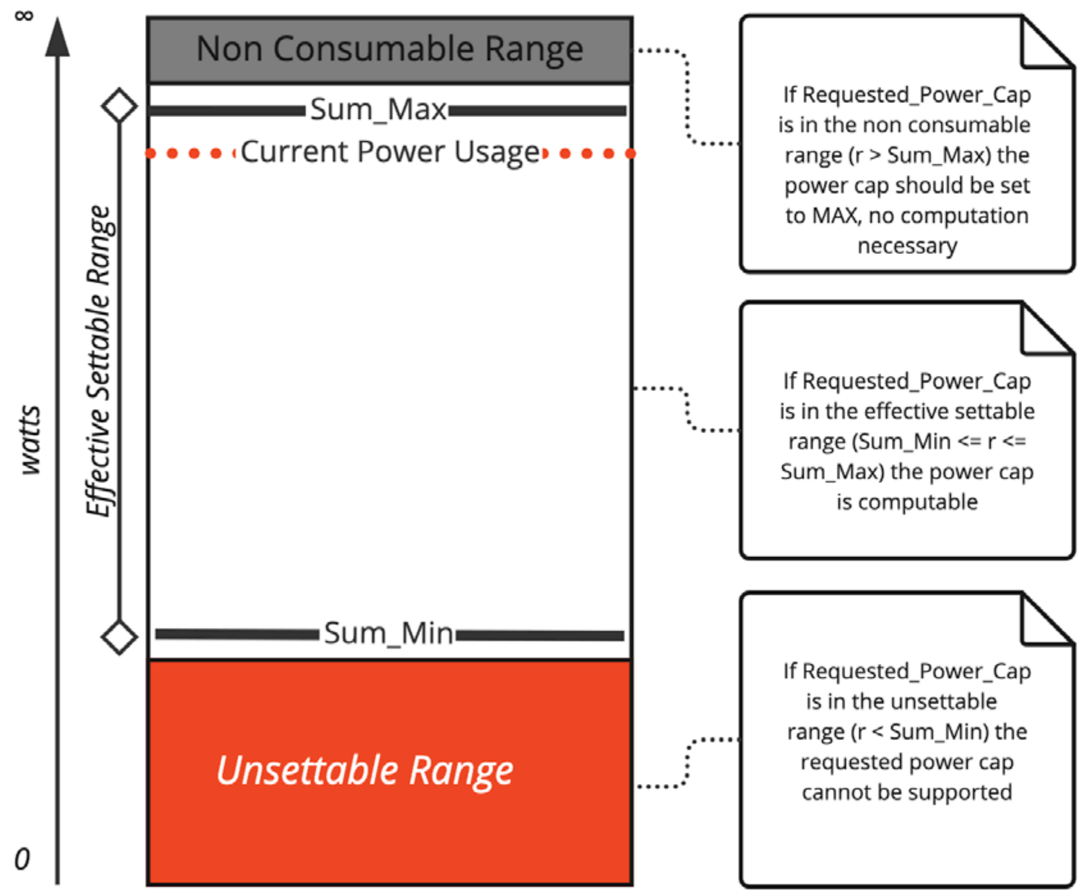
TABLE I
EXAMPLE SYSTEM: HETEROGENEOUS HARDWARE POWER CAPPING RANGES

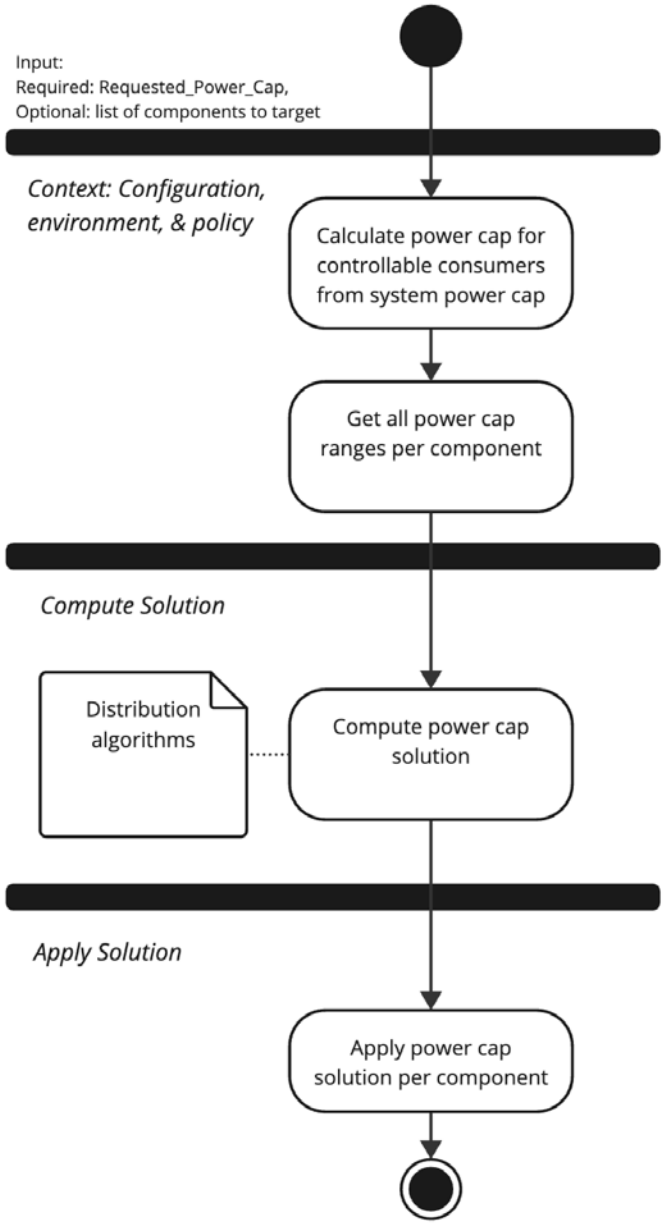
Total Node Count	4,096
Sum_Max	8,471,040 watts
Sum_Min	2,493,440 watts

TABLE II
EXAMPLE SYSTEM: SUMMARY

- Different node types have very little overlap, which means that a uniform power distribution is not appropriate.

Solution Space and Algorithm



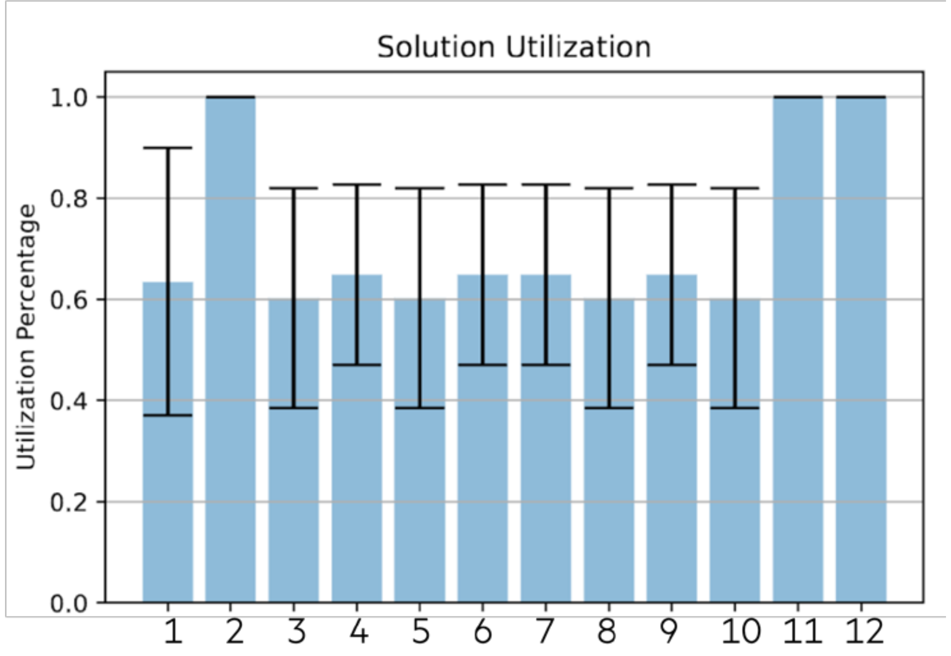


Methodology & Results

- Implementation of 8 distinct example distribution algorithms to set a system power cap.

- Using the example system simulated setting power caps across 70 different input values (from below SUM_MIN to above SUM_MAX)

- **Conclusion**:
  - <u>If Solution Utilization is the deciding criteria for a power distribution algorithm to apply, three algorithms provide 100% of the available power.</u>



Comparison of Solution Utilizations		Node Type 1 (W)	Node Type 2 (W)
Index	Algorithm		
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

Algorithms

Name	Description
base_solution	This algorithm mirrors the compute solution decision graph. It determines if any solution is possible (is in range between Sum Min and Sum Max).
even_split	This algorithm take the difference between Requested Power Cap and Sum Min and divides it evenly among all nodes.
equal_percentage	For each node type calculate the range (max – min) and split it up into 10,000 discrete steps. Then starting from Max for each node, decrease all nodes values by 1/10,000th until the sum of the power caps is less than or equal to Requested Power Cap. It is likely the value will be a decimal, which is then truncated to an integer, which is required for the hardware setting.
count_down	For each node, decrease power cap value by 1W from Max until the sum of the power caps is less than or equal to Requested Power Cap.
delete_by_*	A collection of algorithms that group the nodes by power capping characteristics and then systematically set each group to minimum until an overall solution is found.

Status

- Prototype for managing heterogeneous & homogeneous systems power cap
- Static solution
- Various compute power cap algorithms to choose from
  - Additional algorithms can be added

High Level Power and Energy Management Concerns

- Making stranded power available – actual power usage is less than ‘name plate’ power
- Demand / response – involves shifting or shedding electricity demand
- Time of use costs – energy costs may vary based on the time of day
- Energy efficient system operation – reduces OPEX and carbon impact
- Regulatory efforts – various regulatory efforts to improve IT system sustainability

Multi-datacenter support

(N-Level Power Capping Hierarchy)

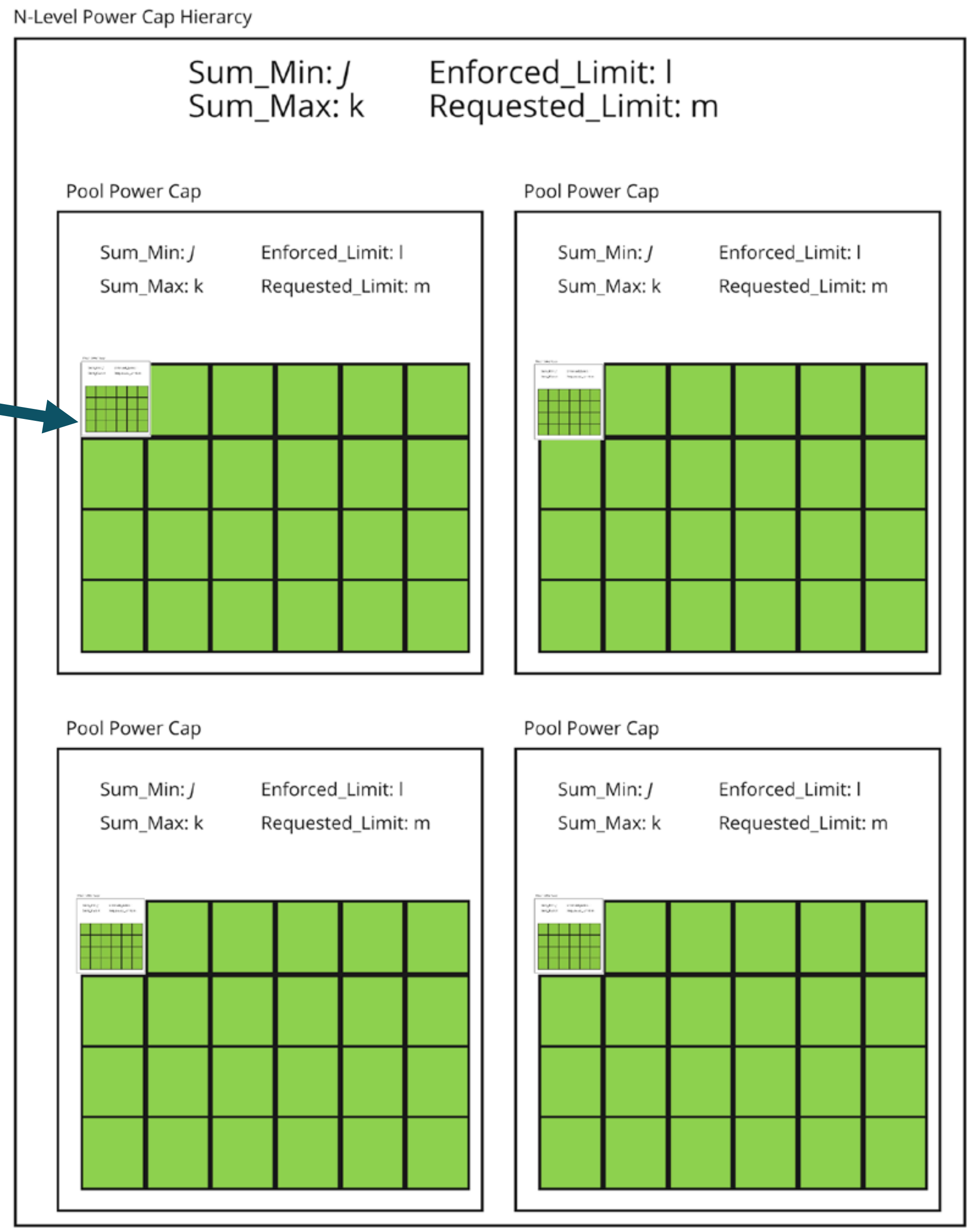
- Concept can apply across n-depth data center hierarchy
  - System of systems

- Concept of pools to manage system power semi-static and dynamically

$$FacilityPower = \sum_{i=1}^{C} notControllableConsumers_i + \sum_{j=1}^{S} SystemPower_j)$$

$$SystemPower = \sum_{i=1}^{C} notControllableConsumers_i + \sum_{j=1}^{N} ComputeNodePower_j)$$

$$ComputeNodePower = \sum_{i=1}^{C} notControllableConsumers_i + \sum_{j=1}^{U} ComputeUnit_j)$$



Future Work

- Reducing stranded power/cooling capacity
  - Reliable dynamic system power capping (e.g supporting hardware over-provisioning)
- Supporting dynamic system power management based on jobs
  - Per pool (job) power cap
- Power management algorithms per pool
  - Power can be distributed based on hardware and job
- Pool priorities
  - Power Rationing and Power Starvation of pools

Thank you

Andrew.Nieuwsma@hpe.com
Wilde@hpe.com