EE HPC WG Liquid Cooling Controls Team Whitepaper. June 11, 2017

This paper defines data inputs for dynamic controls to manage high performance computing (HPC) facility and IT control systems. Each input includes parameters about measurement frequency and accuracy that are within a rough order of magnitude, but not an absolute limit. Each input also includes information about whether it would typically be provided by the facility or by the HPC system or whether its provision would have to be negotiated. This document is intended to be a guideline for data inputs to consider when designing a liquid cooling control system. It is not a design specification. Each site will develop their specific design based on their specific situation. We will be posting this whitepaper on the EE HPC WG website. We would really like to hear your feedback. We are especially interested in hearing from any sites that have deployed and/or are considering deploying liquid cooling control systems. Please send feedback, comments, questions to natalie.jean.bates@gmail.com.

TABLE: DATA INPUTS FOR DYNAMIC CONTROLS FOR LIQUID COOLED HPC:

Name	**Where**	**Typical Provider (facility, HPC system or negotiated)**	**Frequency of measurement**	**Accuracy/Units**
Water flow	System	Facility	Once every 30 sec	+/- 5% Liter/min (gal/min)
	CDU	Negotiated	Once every 30 sec	+/- 10%Liter/min (gal/min)
Thermal data	System or branch	Facility	Once every 60 sec	+/-1° C (1.8 °F)
	CDU	Negotiated	Once every 60 sec	+/-1° C (1.8°F)
	Rack	HPC system	Once every 60 sec	+/-1° C (1.8 °F)
	Node	HPC system	Once per sec	+/-2° C (3.6 °F)
	Component	HPC system	Once per sec	+/-2° C (3.6 °F)
Power	System	Facility	Once per sec	+/- 5% Watts
	Rack	HPC system	Once per sec	+/- 5% Watts
Dew Point Temperature	System	Facility	Once per 60 sec	+/-2° C (3.6 °F)
	Branch, rack or cabinet	Negotiated	Once per 60 sec	+/-2° C (3.6 °F)
Pump Speed	System	Facility	Once per 60 sec	+/-3 %(%of full speed)
	CDU	Negotiated	Once per sec	+/-3 %(%of full speed)
Pressure Differential	System	Facility	Once per 5 sec	+/-10 kpa (1.5 PSI)
	Branch, rack or cabinet	Negotiated	Once per 5 sec	+/-10 kpa (1.5 PSI)
Valve Position	System	Facility	Once per 60 sec	+/- 5% (%Open)
	Branch	Negotiated	Once per 60 sec	+/- 5% (%Open)
	Rack	HPC system	Once per 3 sec	+/- 2% (%Open)

The data inputs have the following characteristics:

- a name
- the unit of measurement
- where it is taken – hierarchically not physically- at what level in the system architecture
- whether the measurement capability is typically provided by the facility or the HPC system or whether its provision is negotiated
- the frequency with which the measurement
- the accuracy of the measurement capabilities

MOTIVATION AND OVERVIEW:

Liquid cooling is key to reducing energy consumption for this generation of supercomputers and remains on the roadmap for the foreseeable future. This is because the heat capacity of liquids is orders of magnitude larger than that of air. Once heat has been transferred to a liquid, it can be removed from the datacenter efficiently. The Energy Efficient HPC Working Group (EE HPC WG) has seen the transition from air to liquid cooling as an opportunity to work collectively to set guidelines for facilitating the energy efficiency of liquid-cooled High Performance Computing (HPC) facilities and systems.

The EE HPC WG has worked with the American Society of Heating, Refrigeration, and Air Conditioning Engineers (ASHRAE) to develop guidelines for warmer liquid-cooling temperatures in order to standardize facility and HPC equipment, and provide more opportunity for reuse of waste heat. The vision is to encourage non-compressor-based cooling, to facilitate heat re-use, and thereby build solutions that are more energy-efficient, less carbon intensive and more cost effective than their air-cooled predecessors. These guidelines were developed and adopted in 2011 and are summarized in the Table below.

2011 ASHRAE Liquid-Cooled Thermal Guidelines

Classes	Typical Infrastructure Design		Facility Supply Water Temp (C)	IT Equipment Availability
	Main Cooling Equipment	Supplemental Cooling Equipment		
W1	Chiller/Cooling Tower	Water-side Economizer Chiller	2 – 17	Now available
W2			2 – 27	
W3	Cooling Tower	Chiller	2 – 32	Becoming available, dependent on future demand
W4	Water-side Economizer (with drycooler or cooling tower)	Nothing	2 – 45	
W5	Building Heating System	Cooling Tower	> 45	Not for HPC

Since that time, there has been an increasing interest in exploring the opportunity for building solutions for all classes (W1-W5) that further increase energy-efficiency and cost effectiveness with increased control of the liquid cooling systems. This is important because the energy efficiency of the cooling system can be improved with dynamic controls and cooling system energy costs can be reduced with improved energy efficiency. Controls may also reduce capital costs, but they are primarily an operational cost improvement.

There are factors that can influence the return on investment of liquid cooling controls. These include:

- Initial site and design considerations (W1-W5)
- Variability in the environment
- HPC system load dynamics as well as the amount of power usage variation between the points of the systems that can vary flow (e.g. – are control valves at the system, branch, rack, node, or component level?)
- Additional capital investment required for the implementation of dynamic controls (e.g. – additional control valves, communications software/hardware, and sensors)
- Actual power draw verses worst case maximum (wall-plate)

Some of the opportunities for increased energy-efficiency and reduced costs arise from more control in the infrastructure side of the liquid cooling system. The following three examples describe the use of controls for improved energy efficiency. These case studies are meant to be illustrative, not exhaustive.

- Using variable frequency drives to minimize energy use in chilled water pumps.
- Interconnecting the chiller and cooling tower with controls to operate in distinct modes that leverage local environmental weather conditions and, thus, use the least power necessary to provide chilled water for the HPC systems.
- Controlling cooling based on measured electrical load. As an example, power meters in the electrical rooms measure server IT load. As load increases or decreases, they trim their cooling water set-points and modify their economization windows accordingly.

Other opportunities for increased energy-efficiency and reduced costs may arise from more control in the HPC system side of the liquid cooling system.

- Active connection feeding data from the rack valves to the BAS to help control some building valves.
- Controlling flow through a node based on liquid temperature. The way that is implemented depends somewhat on the server. One way is to take the temperature and then adjust the speed of the pumps.

Controls at the system level are typically provided by the facility, whereas controls at the node and component level are typically provided as part of the HPC system. It becomes less clear for levels in-between, such as the rack, CDU and branch. In these middle levels, provision for controls needs to be negotiated between the site and the supplier of the HPC system.

The EE HPC WG Liquid Cooling Controls Team has defined data inputs for dynamic controls to manage high performance computing (HPC) facility and IT control systems. This document is intended to be a guideline for data inputs to consider when designing a liquid cooling control

system. It is not a design specification. Each site will develop their specific design based on their specific situation. Costs may vary by site and should be carefully evaluated. This is a cross check. The document lists inputs that are generally considered important. Each input includes parameters about measurement frequency and accuracy that are within a rough order of magnitude, but not an absolute limit. Each input also includes information about whether it would typically be provided by the facility or by the HPC system or whether its provision would have to be negotiated. Again, this should not be considered a hard requirement, but rather a practical recommendation.

- These data inputs are focused on the compute system, not storage or network.
- These data elements and their characteristics are being described here for a particular use case; that of dynamic controls for liquid cooling. This data might be important for other use cases too. For example, power data can be used for a wide range of use cases; e.g., optimizing application energy to solution.
- The control system may use these data inputs with very different frequencies and accuracies from that provided in the table above.

The mission of the Energy Efficient HPC Working Group (EE HPC WG) - https://eehpcwg.llnl.gov/ - is to mobilize the HPC community to accelerate energy efficient HPC by peer to peer exchange, sharing best practices, exploring innovative approaches and taking collective action.

The Liquid Cooling Controls Team is comprised of active members from major supercomputing centers, system integrators and liquid cooling solution vendors - https://eehpcwg.llnl.gov/pages/infra_ctrls.htm .

The Team is expecting that a summary of the results of the whitepaper will be included in the EE HPC WG Energy Efficiency Considerations for HPC Procurement Document - https://eehpcwg.llnl.gov/pages/compsys_pro.htm .

FURTHER DISCUSSION:

An extensive review process was used to generate the list of data inputs and their attributes. While there was general consensus on the information included in the table above, the Team did struggle with whether or not to include thermal and power data for the node and component.

The Team did include thermal data at the node and component level because it could identify at least one example of a component level liquid cooling control system where pumps actually sit on the processors. This is internal to the HPC system. There weren't any examples identified that use node or component level thermal data for facility liquid cooling controls.

The Team did NOT include power data at the node and component level because it could NOT identify any examples where this data was used for liquid cooling control systems. Component and even node level power data is becoming a standard feature for HPC systems. It is possible that this level of power data could be used in future liquid cooling control systems as a controlling input or status or an alarm point.

Collecting power data at the system level for the facility may be difficult for legacy data centers. It requires an electrical distribution system with meters that are unique to HPC systems. Many HPC system solutions today are delivered with the ability to measure power at the system level, but not necessarily with the accuracy that can be attained with meters deployed in a facility.

Caveat- list is subject to change depending on technology changes.

Future predictions/Trends

- Moving from air to hybrid air/liquid cooling.
- Moving from liquid cooling with no controls to some degree of control
- Moving from stand-alone cooling control systems in the facility and HPC system to integrated facility/HPC cooling control systems
- Dedicated verses more distributed cooling systems
- Designing for future systems- flexibility

ADDITIONAL MATERIAL

Overview

Characterizing a HPC's system level load is very challenging given the number of variables involved. Often, system integrators/vendors can only provide the worst case power usage of a single cabinet. Understanding the total power usage of a large HPC system can improve an owner's total cost of ownership (http://www.missioncriticalmagazine.com/ext/resources/MC/Home/Files/PDFs/(TUI3011B)SimpleModelDetermingTrueTCO.pdf) if there is room for leverage in utility cost schedules (demand limiting strategies) and in the design of support infrastructure. Infrastructure design can be right sized and mechanical systems optimized for efficient operation. Design of reliable and efficient cooling systems require a number of design specifications that should be obtained before design work is began. Designing an efficient cooling system requires knowledge of how the load being cooled behaves and how the cooling system reacts to those behaviors as well as those changes that occur in the environment in which the heat is being rejected. This paper begins to discuss some considerations required to characterize an HPC system's power usage profile.

Goal

Determine the load characteristics of a HPC system's power usage to inform infrastructure design to ensure cooling systems meet the cooling load. The HPC system load includes the compute, switch, management, storage, and other cabinet loads. Characteristics of this load include but are not limited to:

1) HPC system design load
2) HPC system idle load
3) HPC system typical load
4) What are the maximum load oscillations at the typical and design load levels and what is the frequency of these changes?
5) What is the HPC system's tolerance to temperature and flow set-point excursions?

To begin to try to answer these questions, we must gather information from the HPC System Owner and the system integrator/vendor.

Questions for the Owner

1) Will the resource manager be biased in any way?
a. Will individual racks be loaded while others will not?
2) Will load or demand limiting or load shedding be utilized?
a. In what circumstances will this be done?
3) Will a power manager be used, and if so, to what will it control?
a. Will the power manager or other system be used to project the future power usage of the system?
4) Will the overall system be segmented into smaller independently acting systems?
5) What are the stranded and trapped capacity goals of the project?
6) What job types will be ran on the system?
a. Are the capability jobs more CPU or GPU intensive?
b. Will the job types be categorized prior to running them?
7) Will an interface be available to the facility's control system?
a. What information will be made available?
b. What communications protocols and connections will be used?

8) During high wet bulb periods and/or low available chiller capacity, is load shedding/processor throttling an option?

Questions for the System Integrator/Vendor

1) Does the system have internal thermal protections?
2) For each cabinet type, and for idle, typical, and design power usage, what do the 100% heat removal flow and inlet temperature curves look like?
3) What are the allowable flow and temperature excursion magnitudes and durations?
4) For idle, typical, and design power usages, how much heat load goes to the liquid cooling circuit verses the air cooling circuit?
5) Does the air cooled side exchange air with the data center?
6) What telemetry data is available for export from the system API?
7) If flow control is provided by your contract:

a. At what level (e.g. – CDU, rack, node, etc.) is flow controlled?
b. What is the fail position of your valves?
c. Are pressure independent maximum flow regulators provided?
d. What algorithms and sensor data control the valve position?

8) What is the maximum allowable inlet pressure?
9) What is the flow coefficient for each cabinet type between its system connection points?
10) What are the minimum and maximum flow rates for each cabinet type?
11) What are the connection types (e.g. – flange, quick connect, sanitary, etc.)?

==========

Stranded capacity is a physical constraint. What this essentially means is that the data center has to have sufficient power, space and cooling (PSC) resources to ensure that a system can operate at full its full utilization. As you alluded to, these resources result in tremendous installation costs to a data center.

Trapped capacity is a managerial constraint. Once the calculations have been completed to ensure the system can be supported by the PSC available, trapped capacity occurs when the workload does not use the intended PSC that has been allocated to it. As an example: if an HPC has a technical power requirement of 10MW and is only using 6MW the delta (4MW) is trapped capacity. This power could be used for another system. So as you say, sub optimization of the actual IT capability is the cause of trapped capacity and although the equipment is "on" it is not fully optimized and does produce higher operational and maintenance costs.

Shared capacity : Once a system has been designed for X amount of power /cooling capacity (based on design specification from said manufacture) carefully placed valves etc. breakers are included in the design (for flexibility and ease if installation) and installation , therefore if there is trapped capacity this design will allow the installation to become flexible enough to install other compute needs with additional loads and capacity to meet the total intent of the design without experiencing any outages or interruption to services. . This strategy will also allow to alleviate Stranded capacities.