
# PRACTICES AND INSIGHTS INTO LIQUID COOLING ON META'S AI TRAINING PLATFORMS

Authors:

Cheng Chen
Yin Hang
Noman Mithani
Chris Malone
Yueming Li
Wenying Zhang
John Fernandes
Kalpak Dhake
Jaret Wyatt
Jarrod Clow
Darron Young

## Executive Summary

This document introduces a series of liquid cooling case studies based on Meta's AI training platforms (Zion & Grand Teton). It shares our prior learnings on:

- Speculative cooling limits of various technologies, in certain assumed conditions
- Performance and learnings of cold plate solutions
- Contributing factors to the temperature gradients/cooling limits
- Extensive challenges for large scale implementation

By sharing those studies and context, we would like to raise the community's attention to collaborate on:

- Convergence/alignment of liquid cooling boundary conditions
- Advancement of package technologies and solutions
- Improvement of liquid cooling solutions' reliability and consistency
- Integration of coolant delivery loop into the IT monitoring & control system

Results shared in this document are intended to explain prior learnings and challenges observed, and may not represent the outcome of optimum solutions as of today. We believe some qualitative learnings in those studies would apply to multiple generations of products in the future. Please be open-minded while reading this document, and feel free to contact us directly for feedbacks and further discussion.

# Table of Contents

## Introduction

Liquid cooling has been a promising technology for high power data center grade chip cooling for almost a decade. The increasing demand of chip power, system density and operational efficiency, are driving the transition from traditional air cooling to liquid cooling of various forms. Such a decision has been made by a variety of users such as HPC customers, specialized users, and high-end gaming consumers.
For users with pre-existing natively air-cooled data centers, however, the move to liquid cooling is a very complicated decision to make, requiring coordinated trade-off analysis, judgment calls, and implementation plans. In the OCP Global Summit 2021 [1], OCP Regional Summit 2023 [2] and AI Infra at Scale talk [3], Meta has articulated our roadmap to enable liquid cooling facilities for AI hardware, for multiple generations of products in the foreseeable future.

In this article, we will walk through the liquid cooling practice based on the Zion platform, including design, performance and the opportunities/challenges observed from this study. Undoubtedly there are dependencies and uncertainties with our results/forecasts. The purpose of surfacing those studies is to share our vision and call out the common challenges which we hope the community can work together on.

All liquid cooling analysis presented in this article is based on single phase cold plate solutions, with 25% propylene-Glycol as the coolant unless called out differently.

# 1. AI Power Trend and Cooling Limits

The growing need of AI/ML applications have been driving the need of training modules (including GPUs, OAMs, and others alike) with higher compute capability, higher bandwidth, and therefore, higher power. Since the start of using GPU products for Machine Learning training in 2016 [4], training module power has been on an ever-increasing trend. Figure 1 shows the TDP's of various training modules that have been announced.

## 1.1. A glimpse of the Trend

Based on a simplified trendline forecast at the OCP Regional Summit panel discussion [2], the power offerings of the training module may reach 1kW in 2025, and 1.5kW before 2030. Multi-chip modules, or so-called super chips, could go further above the trendlines much sooner, as more functionalities and compute capabilities would be enabled on the module, together with larger package surface area.

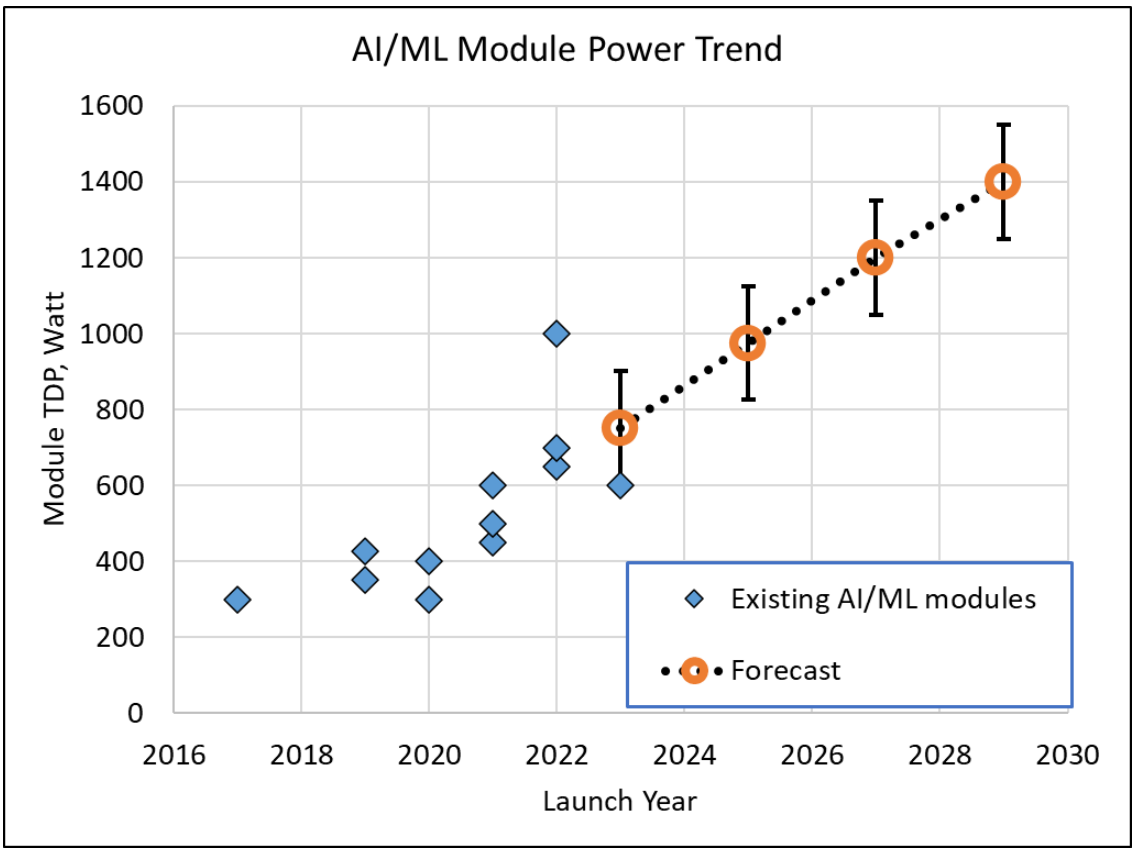


*Figure 1, AI/ML training module power trend, based on various products that were announced between 2016 ~ 2023, and the forecast for year 2023 ~ 2029 from the OCP Regional Summit panel discussion [2].*

## 1.2. Analysis for practical demands

Worried about the risk of not being able to cool higher powers, we derived cooling limit estimations under various conditions, and mapped that with the prediction of power utilization that is efficiency-oriented, instead of potential max TDP offerings. This leads to lower power forecast than Figure 1, however with stronger desire and higher probability to enable. The Meta AI/ML cooling roadmap 2021 [1] was derived based on a number of assumptions as follows, where some may still hold true but some may not match the status quo as of today:

- Single chip module, 2.5D packaging
- 1x ASIC die + 4x HBM die
- Large Training system of 8x modules per board
- Confined by a variety of efficiency, operational and deployment requirements

Figure 2 is a simplified version of the AI/ML cooling prediction that we published in OCP Global Summit 2021 [1], which illustrated how the growing module power would demand the evolution of facility offering generation over generation. The rooflines for AALC and Facility Water Cooling (FWC) in Figure 2 were derived with specific assumptions of:

- Lidless Package
- Optimized HBM stack resistance
- Performance-oriented Cold Plate Loop
- Thermal Interface at 0.2 °C·cm$^2$/W
- Temperature margins built in for uncertainties
- L10 boundary conditions at
    - Air Cooling - Dry Air, 30 °C supply, 0.145 CFM/W
    - AALC - PG25, 40 °C supply, 1.5 LPM/kW
    - FWC - PG25, 30 °C supply, 1.5 LPM/kW

Based on the analysis, the need to implement liquid cooling was apparent. Both module power and HBM power trends are breaching the rooflines of air cooling, and even liquid cooling with warmer coolant supply temperatures are not sufficient. Meta's strategy to move from Air Cooling to Liquid Cooling, with AALC as a transitional solution, has been introduced in the AI Infra at Scale talk in May [3].

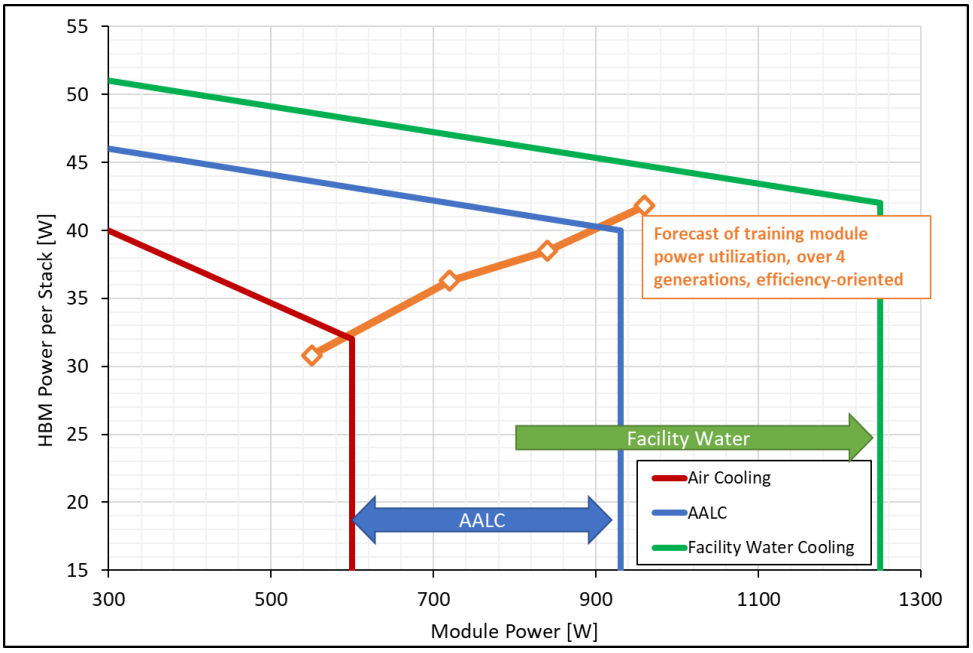


*Figure 2, Meta's AI Training Module Cooling vs. Power roadmap, 2021 version.*

In the past couple of years, we've observed changes in design, technology, and deployment considerations which favors module cooling capabilities. Looking into the future, from where we are today, users may be able to cool higher power products than demonstrated in Figure 2. On the other hand, however, we may also see products of higher power levels be offered earlier than this prediction, to maximize AI Training capabilities within each user's infrastructure limits.

## 1.3. Importance of Community Alignment

It is obvious that the coolant supply temperature could directly impact the acceptable TDP of short range products with a determined architecture, and determines how long range products achieve certain TDP targets with enough thermal robustness. It is beneficial to have consensus across users and module providers on the achievable coolant supply temperatures in both short range and long range, to avoid either apparent mismatch between cooling capability and deliverable power, or unnecessary over-preparation and risk taken in infrastructure design which may block the solution advancement from multiple perspectives.

Particularly, why start with 30 °C as the technical fluid supply setpoint? It's a result of following factors:

- Facility PUE and WUE Goals (sustainability metrics)
- Air supply loop constraints (condensation)
- Capability to support, or stretch to support, the foreseeable roadmap

Influenced by a variety of factors, 30 °C might not become the final setpoint as we learn more in the next few generations, but facilities with such capability would not lose the flexibility to accommodate slightly different scenarios either. We would like to have closer partnership across the community, to surface considerations and materials that help drive convergence towards a narrower temperature range.

# 2. Platform Practices

## 2.1. Tide 1.0 - Assembly and Reliability Oriented

Zion system [5], as Meta's flagship AI/ML platform announced in 2019, is also our main vehicle to explore liquid cooling solutions, setting future expectations and identifying risk areas. In a prior publication at ASME Interpack 2022 [6], we've introduced the learnings on Tide 1.0 solution which supports OAM-A in Zion platform.

Figure 3 shows the design concept of Tide 1.0 and performance curves across CFD modeling, TTV tests and real Zion system tests. Both CFD analysis and TTV preparation were completed with high quality, representing real performance with small discrepancies. There's a unique feature of this practice: an adapter plate covers the lidless package and all VR components of OAM-A; only one piece of TIM material exists between the top surface and cold plate. This design was oriented towards assembly and reliability, while performance was still strong enough to support future growth to some extent.

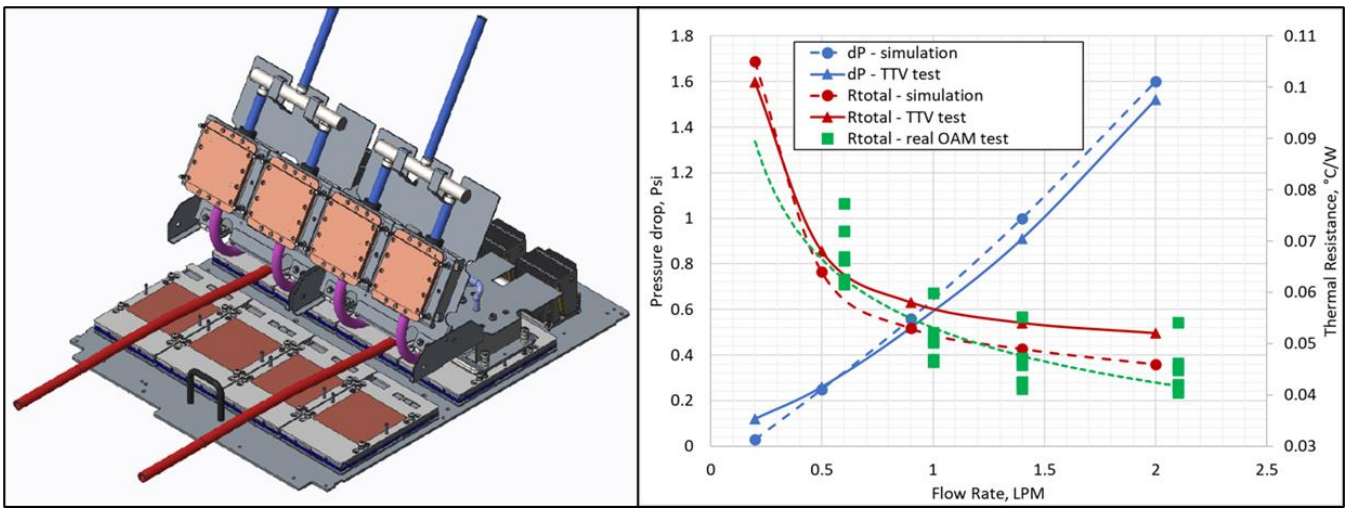


*Figure 3, Left - design concept of Tide 1.0, installed on Zion's accelerator module board, attached to 8x OAM-A modules. Right - junction_to_inlet thermal resistance and pressure drop of single coldplates, uniform flow distribution assumed.*

## 2.2. Tide 1.5 - Performance and Simplicity Oriented

Later, Tide 1.5 was developed for Zion, to support another module OAM-B of higher TDP. OAM-B was a lidded product, hence the solution design was adjusted towards simplicity and performance. Tide 1.5 cold plate assembly, as shown in Figure 4, was also contributed to the OAI group for coordinated studies of OAM liquid cooling guideline development [7].

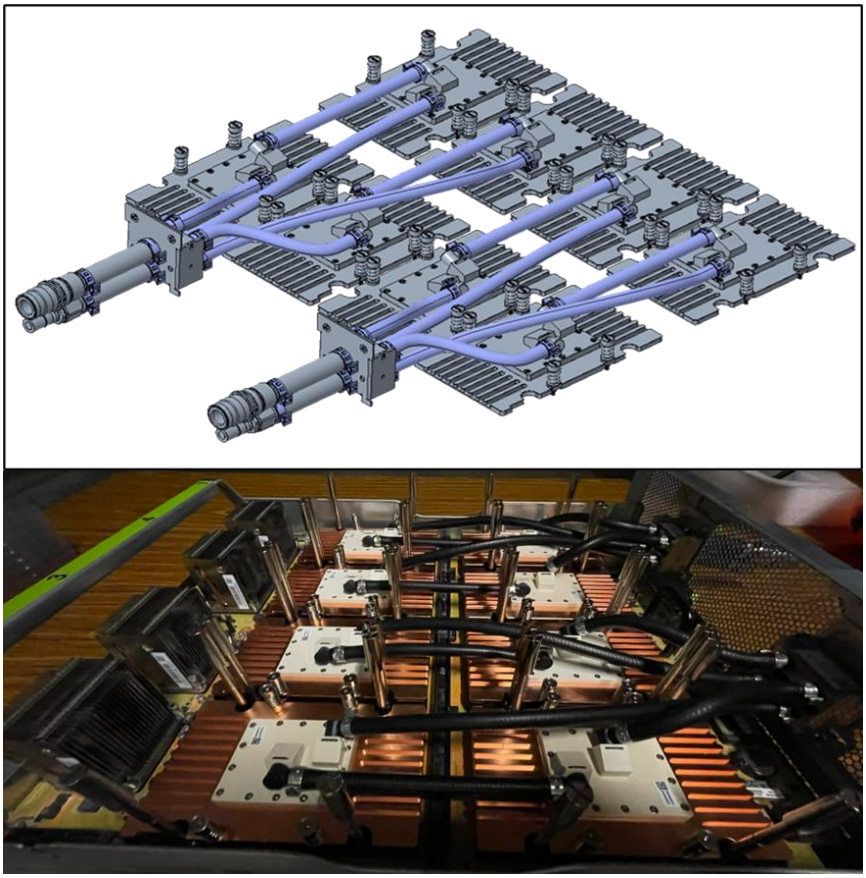

*Figure 4, Tide 1.5 cold plate loop assembly for Zion platform using OAM-B. The assembly was split into two identical portions, each with one pair of coolant supply/return using SCG06 QDs.*

Tide 1.5 cold plate solution is a good representation of single phase coldplates optimized for performance, at the given chip characteristics. Performance compared to air cooling, based on real system tests, is shown in Figure 5, demonstrating at least 60% case-to-inlet thermal resistance reduction. The lowest thermal resistance achieved was at 0.02 °C/W, where TIM contact resistance was estimated to be 0.1 °C·cm$^2$/W.

TTV continues to serve as an essential part of the solution validation. OAM1.0 TTV was adopted in this validation, whose design is shown in Figure 6 and performance shown in Figure 7. Though not a perfect product that can represent all thermal and power delivery characteristics of an OAM product, it captures the top side heat

dissipation under ASIC-centric load conditions with good accuracy. The comparison shown in Figure 7, shows a good match with the nominal thermal resistance of OAM-B. OAM1.0 TTV design and prototypes have been made available to various OCP partners for cooling solution benchmarking.

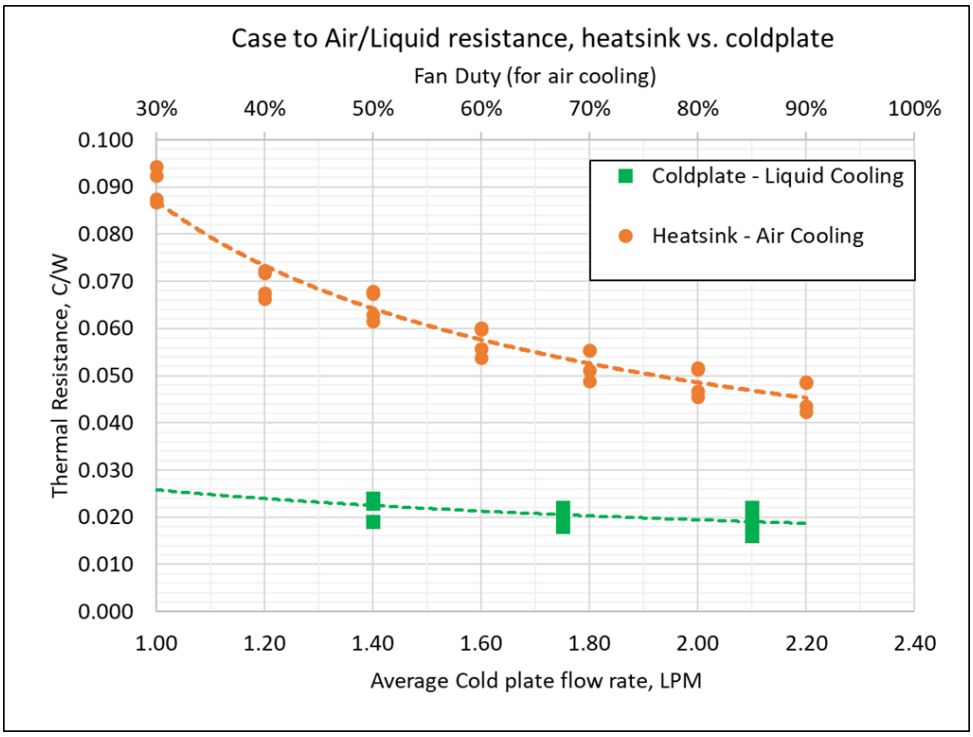


*Figure 5, case-to-air/liquid inlet thermal resistance, at single heatsink/cold plate level, across the fan duty / flow rate ranges. The impact of preheat was removed from the equations by assuming uniform flow distribution.*

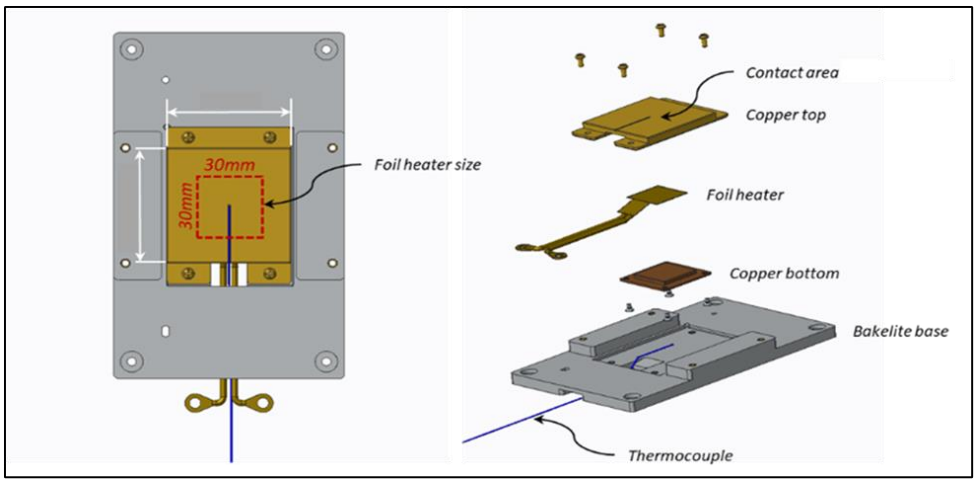


*Figure 6, OAM 1.0 Thermal Test Vehicle using thin foil heaters. Max achievable TDP is 1kW. Can be further modified to represent various power maps.*

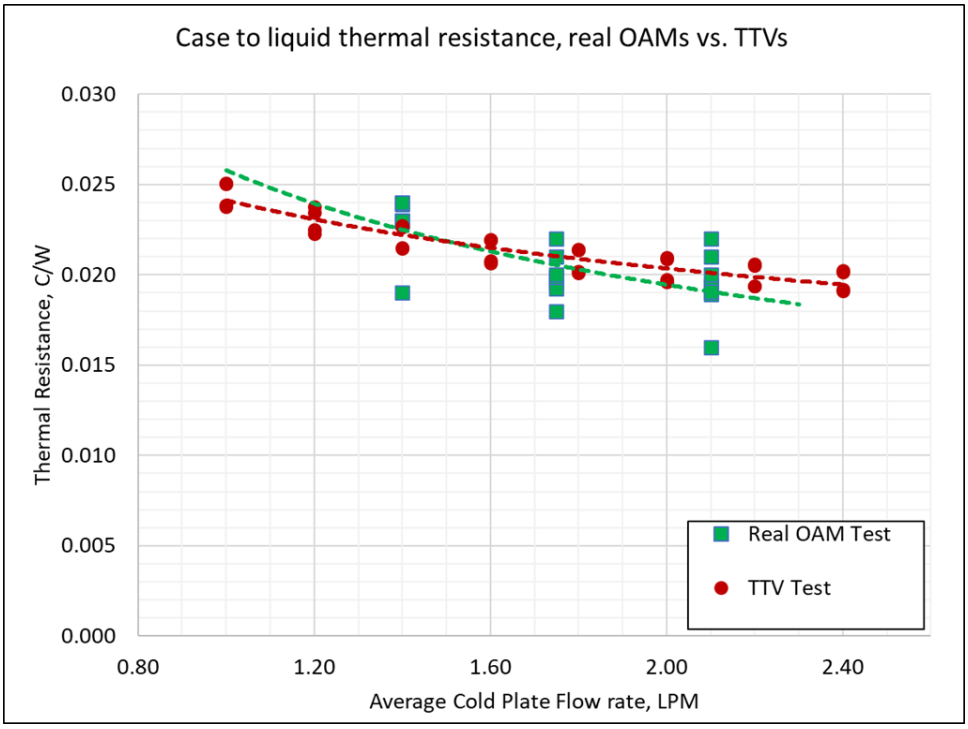


*Figure 7, thermal resistance comparison of Tide1.5 coldplate on OAM1.0 TTV vs. real OAM-B modules, in one Zion system. TTVs demonstrated even better part-2-part uniformity.*

## 2.3. Tide 2.0 - Forward Looking

The cold plate performance demonstrated by Tide 1.5 is considered as one of the necessities, however not enough, to sustain power growth that would soon reach north of 1kW. Tide 2.0 is the liquid cooling solution developed on Grand Teton system (Figure 8), with following features built in:

- Parallel GPU cold plate loop
- Leakage detection and communication mechanism
- Electrical Valve
- Assembly assistance kit

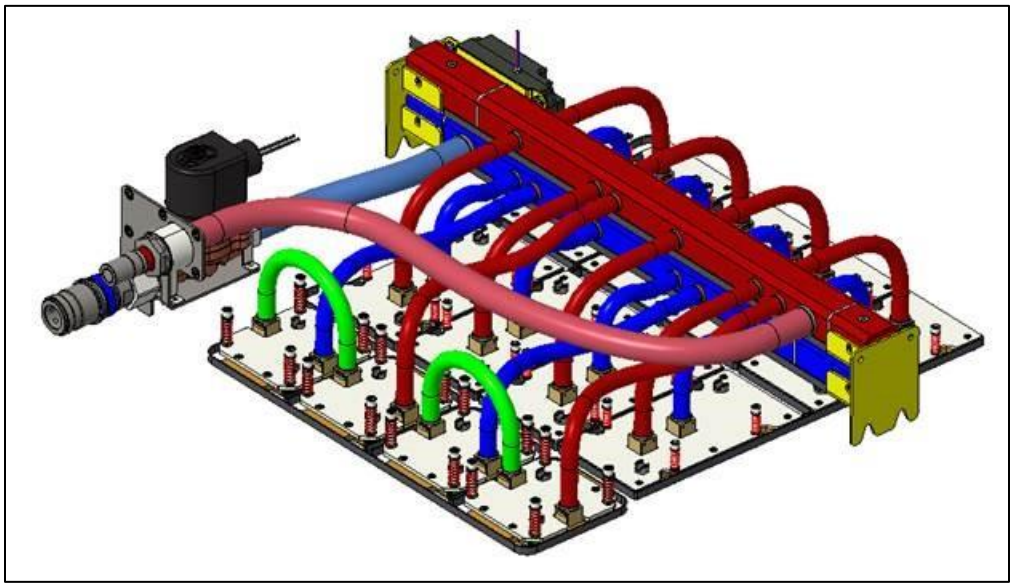

*Figure 8, Tide2.0 coldplate loop assembly design for Grand Teton Platform*

Tide 2.0 was developed with expectations across performance, assembly, reliability, and scalability axes. The proof-of-concept prototypes will be demonstrated at Meta's booth at OCP Global Summit 2023, with more details to be introduced in our further communications. The performance that Tide 2.0 demonstrated, along with the package characteristics assumptions, are closest to, and even slightly more capable than, the rooflines demonstrated in our roadmap forecast (Figure 2).

## 3. Impact of Packaging

What's preventing us from supporting higher power training modules? A simple way to answer this question is: all of us. Not surprisingly, all of us are also responsible for solving it, to some extent. Based on analysis of OAM-B (Tide 1.5) in the Zion system, we derived the thermal resistance stack chart as in Figure 9 [1], particularly for ASIC and HBM cooling. It shows that the Lid and TIM1 are estimated to contribute to half of the total resistance for ASIC, while heat crosstalk from ASIC and internal stack resistance contributed to the majority of HBM temperature gradients.

Many assumptions and dependencies were factored into the analysis behind results in Figure 8. The exact allocations could vary slightly depending on how we define each element, and much more significantly as the product changes (SCM vs. MCM., lidded vs. lidless, HBM generation and exact SKU, etc.). However it's still valid to consider the thermal and mechanical characteristics of the package being equally, if not more, as important as the solutions provided on top of it. Designers of chip, system, cooling solutions and facilities have shared

responsibility to support the AI roadmap. Such connection is much closer than any prior products, as margins are more constrained and all elements need to be pushed to extremes.

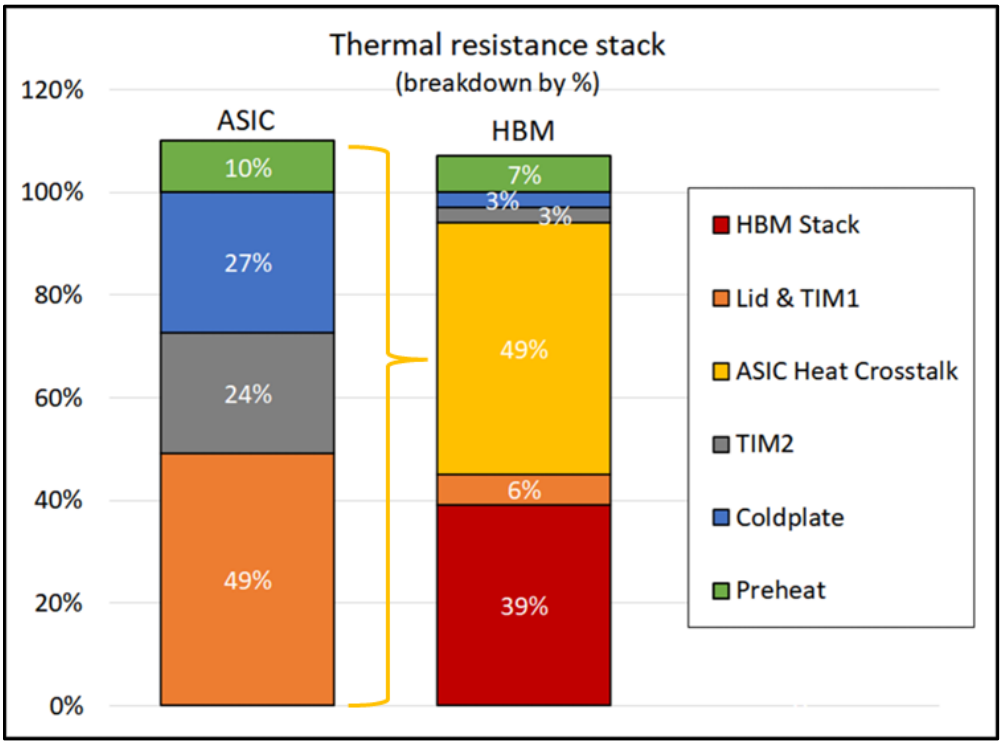


*Figure 9, thermal resistance stack up for ASIC and HBM cooling, based on OAM-B in Zion system. The contributors to ASIC thermal resistance are all contributing to the ASIC Heat Crosstalk portion in HBM thermal resistance.*

# 4. Other Challenges

Gaining the capability to cool an AI Training Product is just the starting point for a successful deployment. Large deployment scale, sensitivity to capacity loss, and small maintenance team sizes, are putting extra requirements on the robustness and risk containment of liquid cooled AI platforms for hyperscale users.

## 4.1. Performance Degradation

Thermal degradation is one of the major concerns for liquid cooling products at scale. Over the product's lifetime, the performance may degrade slowly due to factors originating from the cold plate design, TIM material, quality control, flow control, pumping unit, wetted materials, coolant chemistry, etc. Long term reliability validation is important to expose those potential risks and identify the right solution path, long before landing the 1st production rack.

Figure 9 demonstrates the cold plate performance degradation observed on an outlier TTV rack, through the accelerated reliability test. Over a duration of 90 days, the TTV case temperatures showed 1~3 °C temperature increase, with one extreme outlier reaching 5 °C increase. Coherent troubleshooting efforts across multiple partners were conducted to solve this issue, while a lot of time was spent to narrow down the factors, and decide the logistics flow.

Down the road, we foresee the following needed to account for the reliability axis:

- Thermal Margin reservation for performance degradation over lifetime
- Reliability validation as a standard process of product/component/material offering
- Coordinated diagnose and logistic mechanism in place

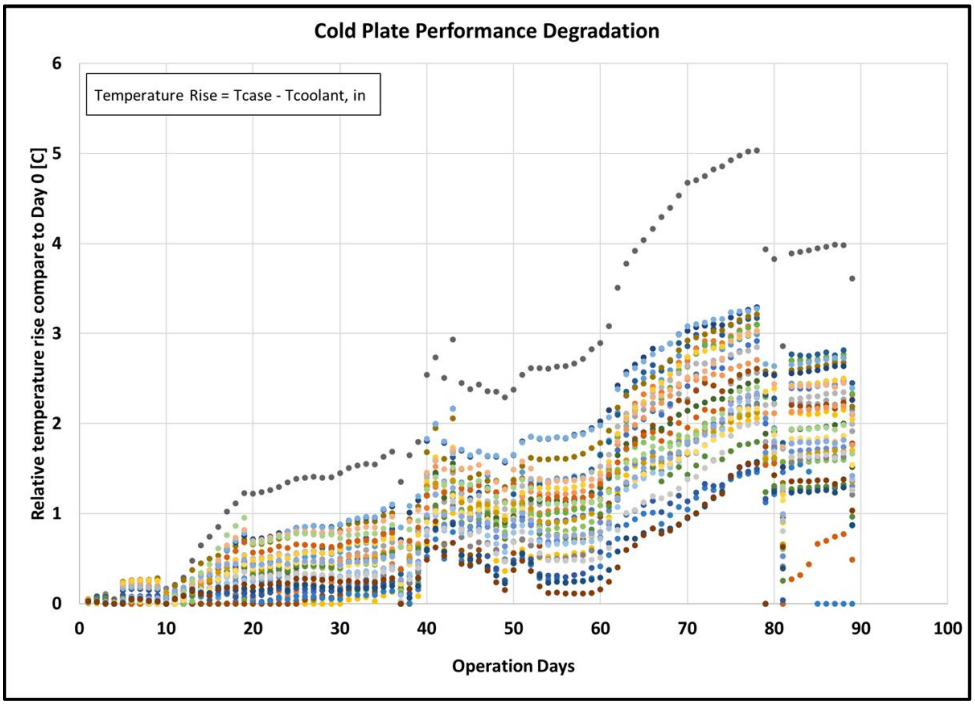


*Figure 10, cold plate performance degradation, in the form of TTV temperature increase, from an outlier rack, over the duration of accelerated reliability test. Tide 1.5 and OAM 1.0 TTV were adopted in this validation.*

## 4.2. Flow Variation

Compared to air flow loops, liquid flow loops usually have fewer active flow-driving units and control logics for a variety of considerations. This reduces the granularity of control structures and leads to coarser flow distribution management. Moreover, liquid flow channels have smaller hydraulic diameters in general, hence the relative part-to-part uncertainties of liquid cooling materials (cold plate, QD, hose, etc.) would be larger than air cooling materials (heatsinks, air ducts, etc.).

In a hypothetical flow network model, we studied the flow distribution across 4x Zion systems in one rack, with cold plate impedance variation, QD impedance variation, and manifold layout considered. The result is demonstrated in Figure 11, showing +/- 10% cold plate flow rate variation in the given scenario. Modeling at a larger scale would be necessary to characterize the statistical significance, yet the indication of noticeable flow rate variation in real clusters is valid. In addition, contribution from the QD variation is found to be most impactful, hence indicating the desire to have QD products of appropriate uncertainty range, particularly UQD products which could have more varying factors.

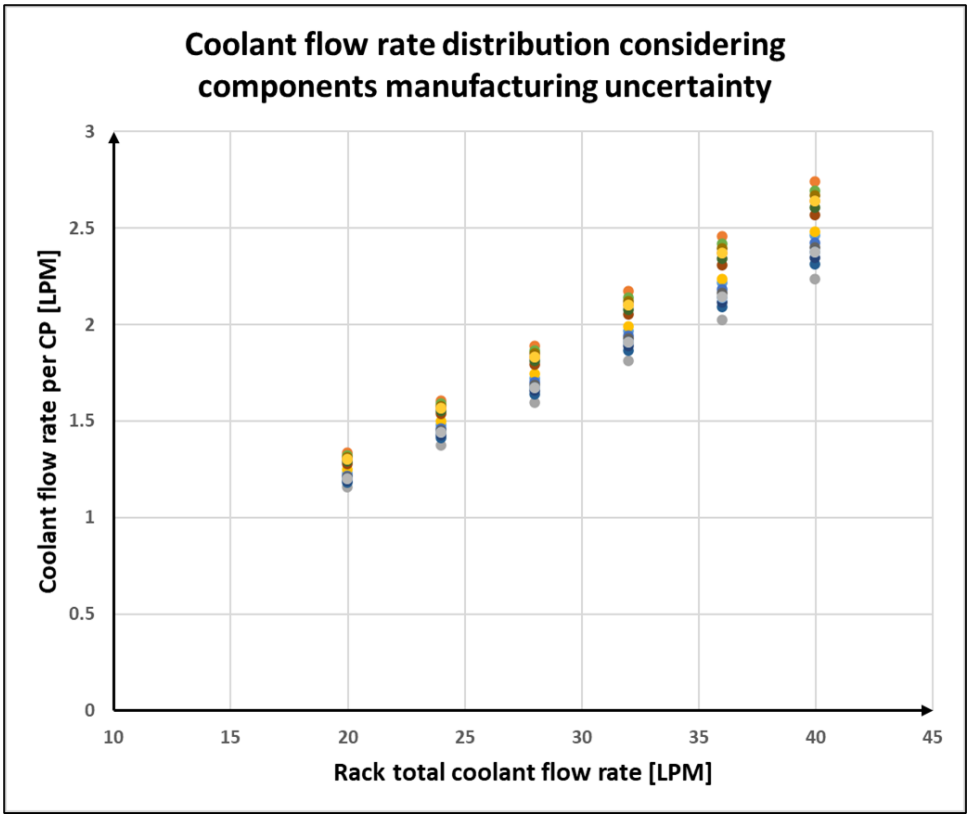


*Figure 11, cold plate flow rate variation, based on a hypothetical flow network analysis at single rack level. 4x Zion systems with Tide 1.5 solutions were assumed in this analysis. The variations originate from manifold layout, QD variation, and cold plate imped*

## 4.3. Part to Part Variation

Flow variation is not the only factor that creates discrepancies. As all elements are pushed towards optimum levels, the variation of chip/package/interface could become an even larger contributing factor. Figure 12 shows the junction-to-inlet thermal resistances of OAM-B modules using Tide 1.5 solutions, across 6x Zion systems.

Preheat was removed from the calculation by assuming uniform flow distribution. The resulting thermal resistances arrived between 0.05 ~ 0.06 °C/W, with an outlier at 0.047 °C/W.

Chip temperature variation is not a surprising phenomenon, but is more important for AI chip cooling where every degC would matter. It is hard to quantify exactly how much of the variation was contributed by each possible factor. To capture the variations in production, it seems more practical to get enough sample size for statistical modeling, if it is not possible or practical to dive deep into the physics of various underlying issues. A noticeably wide range of the distribution range would require a much larger design margin, that either puts extra burden on the infrastructure, or limits the product capability.

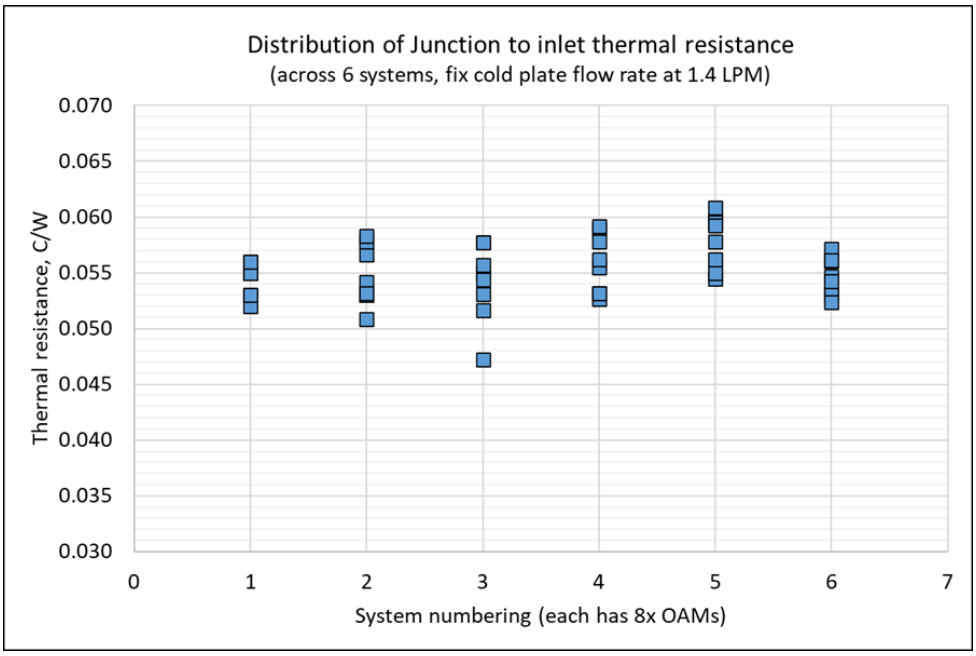


*Figure 12, comparing junction-to-inlet thermal resistances of 48x OAM modules in 6x Zion systems. Each system was tested independently with 5.6LPM, averaging 1.4LPM per cold plate (every 2 in serial).*

## 4.4. Communication and Control

For the air delivery loop from facility to chip, it consists of multiple tiers of flow delivery/management devices and control logics connected via pressure, temperature and power readings; the control algorithms are also able to function independently upon upstream/downstream failures, being either a device or a reading. When it comes to the liquid delivery loop, we are setting the expectations based on the same philosophy. The Reservoir and Pumping Unit specification [8] has articulated the control scheme and failure mode responses for AALC

solution, a lot of which will be leveraged/translated into the communication and control mechanisms for facility water cooling.

Translating those to the side of liquid cooled AI platform of robustness and scalability, we would be expecting:

- The L10 liquid cooling solutions shall have enough tolerance to operate with relatively coarse flow rate and supply temperature control.
- Leakage detection and flow halt mechanisms at some granularity (module/board/system/rack level) would be necessary to contain the blast radius immediately upon leakage.
- Minimize latency and SW dependency for leak detection. Enable HW protection for leakage events in the system with debuggability to identify the component causing the leak

Liquid cooling integration requires multi-level of communication, control and failure recovery plans, from cold plate, board, system design all the way to facility level integration and alerting. Options to deal with failures, exceptions and corner case conditions with enough maturity and compatibility is important for liquid cooling deployment at scale for hyperscale adopters

## 5. Call to Action

The future of liquid cooling AI Training platforms has indications on complexities beyond typical hardware level challenges. We want to bring the following understandings to the community's attention, and call to pursue alignment or take actions on those topics:

- **Alignment on the boundary condition** (i.e. coolant supply temperature) - a narrow coolant supply temperature agreement/expectation could help avoiding mismatch across chip design, hardware power & cooling solutions, infrastructure preparation and sustainability targets.
- **Package technology/solution advancement** - the interior resistance and thermal interface characteristics are becoming the dominant factor in temperature gradient. Proper package design (2.5D, 3D, SCM, MCM, etc.), warpage control and TIM implementation are speculated to play more important roles than the external solution and facility boundary conditions in the future. Lower interior and interface resistances could make cooling of 1kW+ modules much easier than past products.
- **Highly Reliable and Consistent Solutions** - The flow network of future AI training products would have larger scale and higher inter-dependencies compared to past air cooling solutions. The thermal elements (package, coldplates, TIMs, flow distribution structures, flow delivery units, etc.) need to meet higher requirements of reliability and consistency than past products, so smaller margins can be factored into design and enable higher design powers.

- **Make Liquid Cooling 'Smart' –** there's strong desire to integrate liquid cooling solutions into the data center monitoring, control, and failure management systems for large scale users. A scalable solution needs to meet many expectations related to leakage, communication, and control, which is a growing area at this moment. We would like to see more solutions with those features enabled, so that the liquid cooling solutions can be treated with fewer exceptionalities in the data center operation procedures.

## Acknowledgement

Work in this article was completed with help from Quanta Thermal Team, Wiwynn Thermal Team, CoolerMaster, and OAI partners.

## Terminology

AALC - Air Assisted Liquid Cooling

FWC - Facility Water Cooling

OAM - Open Accelerator Module

OAI - Open Accelerator Infrastructure

PUE - Power Usage Effectiveness

PG - Propylene Glycol

QD - Quick Disconnect

TDP - Thermal Design Power

TTV - Thermal Test Vehicle

WUE - Water Usage Effectiveness

## References

1. Liquid Cooling: Drivers, timelines and a case for industry convergence on coolant temperatures, https://146a55aca6f00848c565-a7635525d40ac1c70300198708936b4e.ssl.cf1.rackcdn.com/images/46c790c34e6f3c5444a57c82db2b2d1da422f29e.pdf
2. Panel: Coolant Temperatures for Durable Data Center Designs, https://drive.google.com/file/d/1CxcjdtdIznn_0AC2cJeVbhTLPar3E9P8/view?usp=share_link
3. Meta AI Infra at Scale, Next-Generation Data Center Design, https://atscaleconference.com/meta-ai-infra-scale/?tab=0&item=5#agenda-item-5
4. Accelerating AI with GPUs: A New Computing Model, https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/
5. Zion system specification 1.0, https://www.opencompute.org/documents/facebook-zion-system-spec-1-0-pdf
6. Liquid Cooling Practice on Meta's AI Training Platform, https://asmedigitalcollection.asme.org/InterPACK/proceedings-abstract/InterPACK2022/86557/V001T01A002/1153367
7. OAI Liquid Cooling Guidelines, https://www.opencompute.org/documents/oai-system-liquid-cooling-guidelines-in-ocp-template-mar-3-2023-update-pdf
8. Reservoir and Pumping Unit Specification, https://www.opencompute.org/documents/ocp-reservoir-and-pumping-unit-specification-v1-0-pdf

## License



## About Open Compute Foundation

The Open Compute Project Foundation is a 501(c)(6) organization which was founded in 2011 by Facebook, Intel, and Rackspace. Our mission is to apply the benefits of open source to hardware and rapidly increase the pace of innovation in, near and around the data center and beyond. The Open Compute Project (OCP) is a collaborative community focused on redesigning hardware technology to efficiently support the growing demands on compute infrastructure. For more information about OCP, please visit us at http://www.opencompute.org