OPEN POSSIBILITIES.

Liquid Cooling: Drivers, timelines and a case for industry convergence on coolant temperatures

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

Server

Liquid Cooling: Drivers, timelines and a case for industry convergence on coolant temperatures

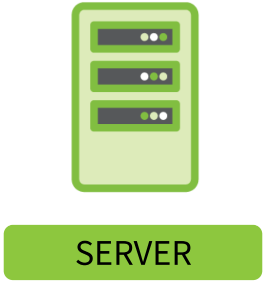
Chris Malone
Datacenter Systems Engineer, Meta

OPEN POSSIBILITIES.

OPEN PLATINUM™

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

TL;DR



- Our industry is on a steady course to break the limits of heavily air-economized cooling, starting with ML Training chips and high-speed switch fabric ASICs. The solution is liquid cooling.
- Our industry (chipmakers, facility builders, operators) will benefit from converged liquid cooling practices, as we've done previously with economized datacenters.
- 30C coolant temperature is compatible with efficient data center design principles.
- We can probably stretch 30C through 2035, with industry investments around packaging, layout, and TIM improvement.
- 30C likely won't last forever. Whatever forum & consensus we establish now can be used to better govern future changes, and rates of change.

Sequence

Meta DC Operating Conditions

ML/Training Module Cooling Trends

Fabric Switch Cooling Trends

Challenge & Proposal

Supporting Material (if time available, or in Q&A)

Meta DC Operating Conditions

Our Datacenter Operating Conditions

Economization has worked very well for Meta for a decade

- Reference boundary conditions for hardware thermal design
  - Supply air temperature: 65°F to 85°F (18.3°C to 29.4°C)
  - Air-side deltaT: 22°F (12.2°C)
  - Altitude: Sea-level to 6000 ft

- Validation and deployment of IT gear
  - Thermal validation of hardware conducted up to 35°C (corner cases), with specific focus on operation at 30°C
  - Power and airflow requirements at 30°C used as reference for deployment (cluster planning)

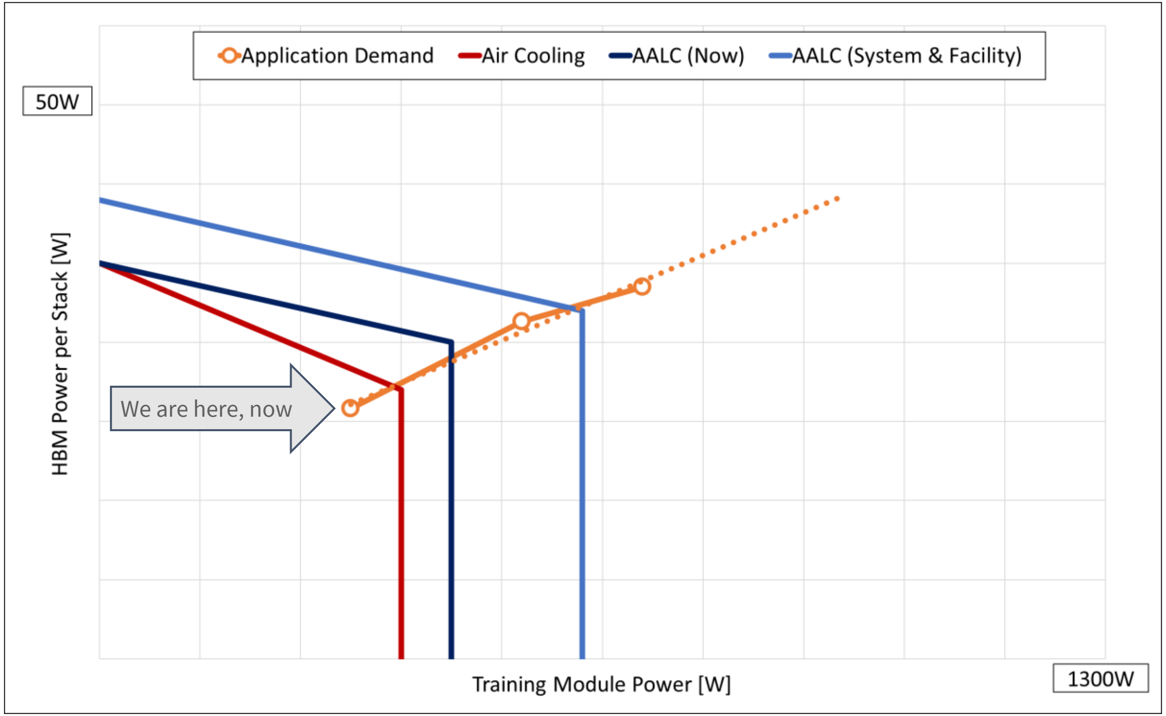
~30°C is used as the nominal operating / ambient temperature for air-side economization.

(Charts in this deck with projections of air-cooling limits are based upon these conditions).

Chip, System, Rack, and Facility Designers:

If you design for air-side or water-side economized environments, *our projections likely resemble yours*.

The Emergence of Need for Liquid Cooling in ML



- Application Demand trendline is a *composite* of trends in chips / modules.
  - This includes estimated effective utilization of chips.*
- These are *per training module* estimates, not *per system*.
- TDP-only plotlines of specific chips / modules (not shown here) fall both above and below the Application Demand line.

- AALC - Air Assisted Liquid Cooling

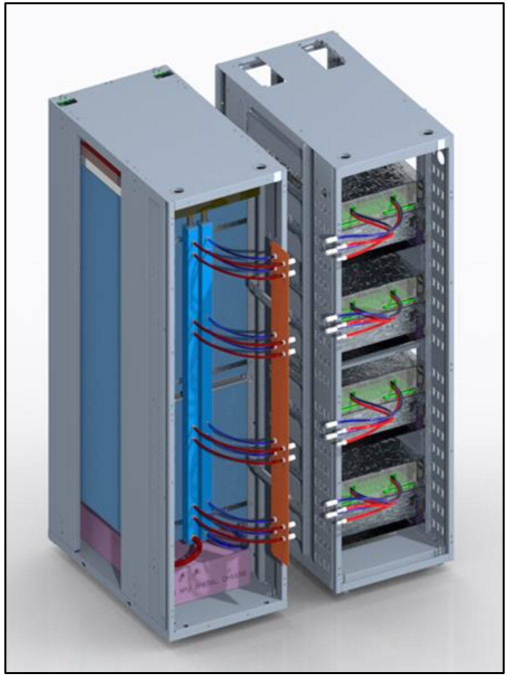
OPEN POSSIBILITIES.

* This varies some by architecture, data model, and software application.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

The Bridge: Air Assisted Liquid Cooling (AALC)

Transitional solution to enable liquid cooling



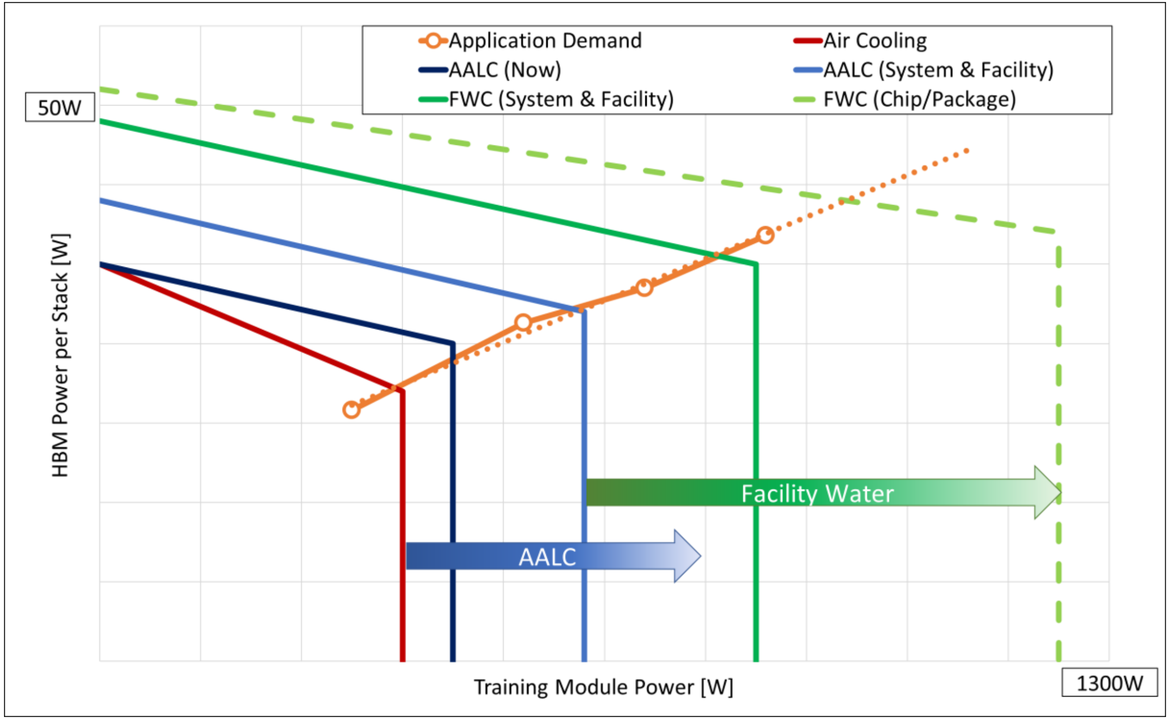
AALC: Open Rack v3 IT rack with adjacent ORv3 HX Rack

- IT rack + Cooling Rack Bundle
- Enables liquid cooling in free air-cooled data centers

- More info can be found at:
  https://www.youtube.com/watch?v=S3hNlZGj4UM

The Limits of AALC as a Bridge

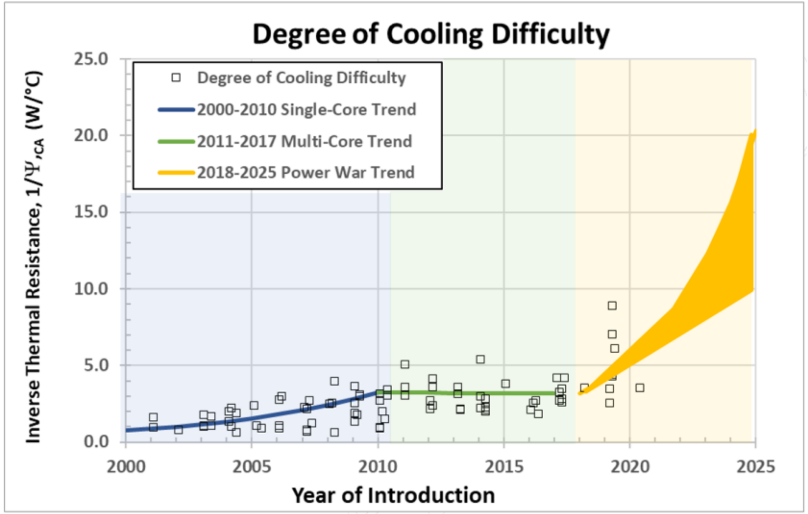
Likely good for 1 - 2 generations depending upon other optimizations (more on this, later)



- AALC - Air Assisted Liquid Cooling
- FWC - Facility Water Cooling

ASHRAE Describes Similar Challenges

A broad range of coolant temperatures, for a broad range of chip power levels and facility designs



Class	Typical Infrastructure Design		Facility Water Supply Temperature °C
	Primary Facilities	Secondary Facilities	
W17	Chiller/Cooling Tower	Water-side economizer	17
W27			27
W32	Cooling Tower	Chiller / District heating system	32
W40			40
W45	Cooling Tower	District heating system	45
W+			> 45

This range of products, conditions, and coolant temperatures make it very difficult to optimize.

This affects performance, scale, and predictability.

A Durable Thermal Interface Approach is Needed

What strikes the best balance of durability and efficiency, for new facilities and systems?

Specialized

Practical

Not Durable

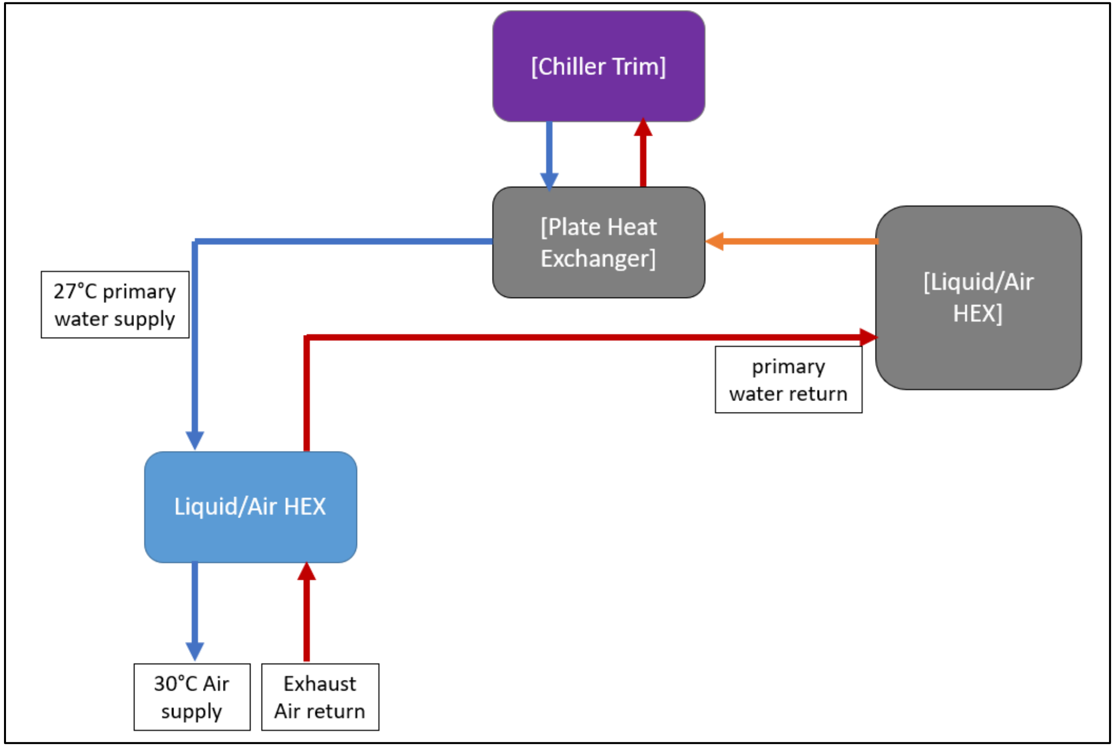
Class	Typical Infrastructure Design		Facility Water Supply Temperature °C
	Primary Facilities	Secondary Facilities	
W17	Chiller/Cooling Tower	Water-side economizer	17
W27			27
W32	Cooling Tower	Chiller / District heating system	32
W40			40
W45	Cooling Tower	District heating system	45
W+			> 45

Building a new facility with a goal to support emerging IT equipment for a decade, efficiently?

Which class should you choose?

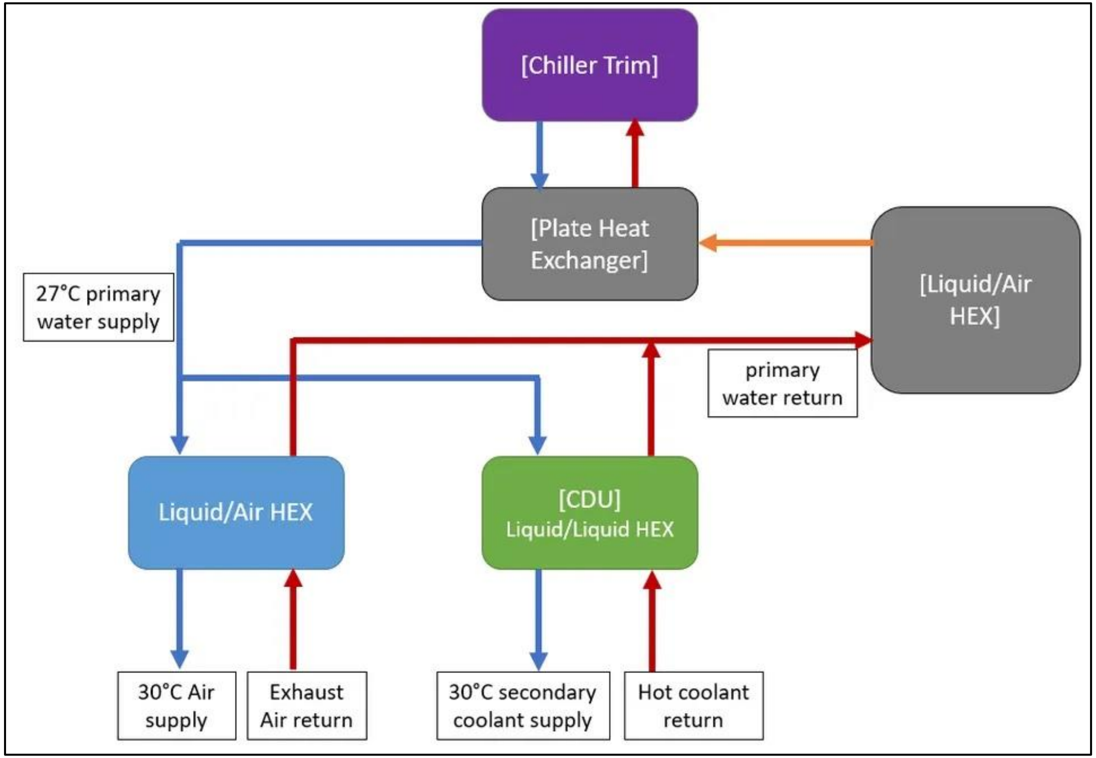
An Air & Liquid System Designed for 30C

Air cooling only: 30°C is optimal



An Air & Liquid System Designed for 30C

Air + Liquid Cooling: optimal in multiple aspects



- **At 30℃:**
- Efficiency - PUE & WUE are manageable
- Simplicity - Same primary loop supply for air & liquid
- Durability - 30°C will be sufficient for many generations

ML/Training Module Cooling Trends

Assumptions We Make About Training Systems

Analysis based on UBB style 8x modules training system layout

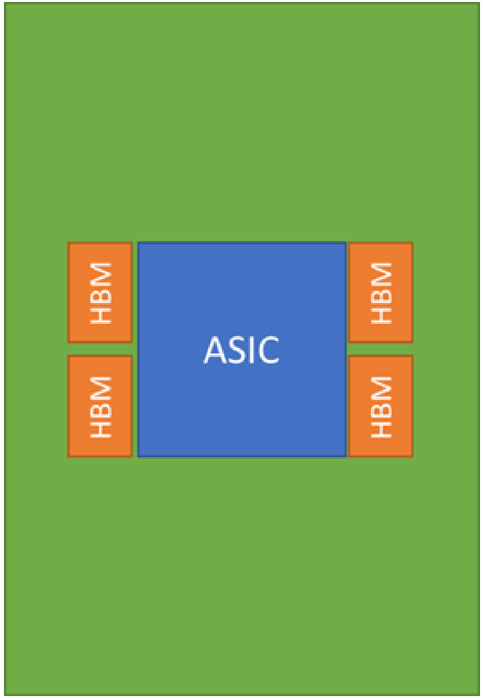
Boundary:

- 8x OAMs per system
- PG25% based coolant
- AALC ⟶ coolant supply above 30°C (more on this, later)
- FWC ⟶ coolant supply at 30°C (more on this, later)

Package assumptions:

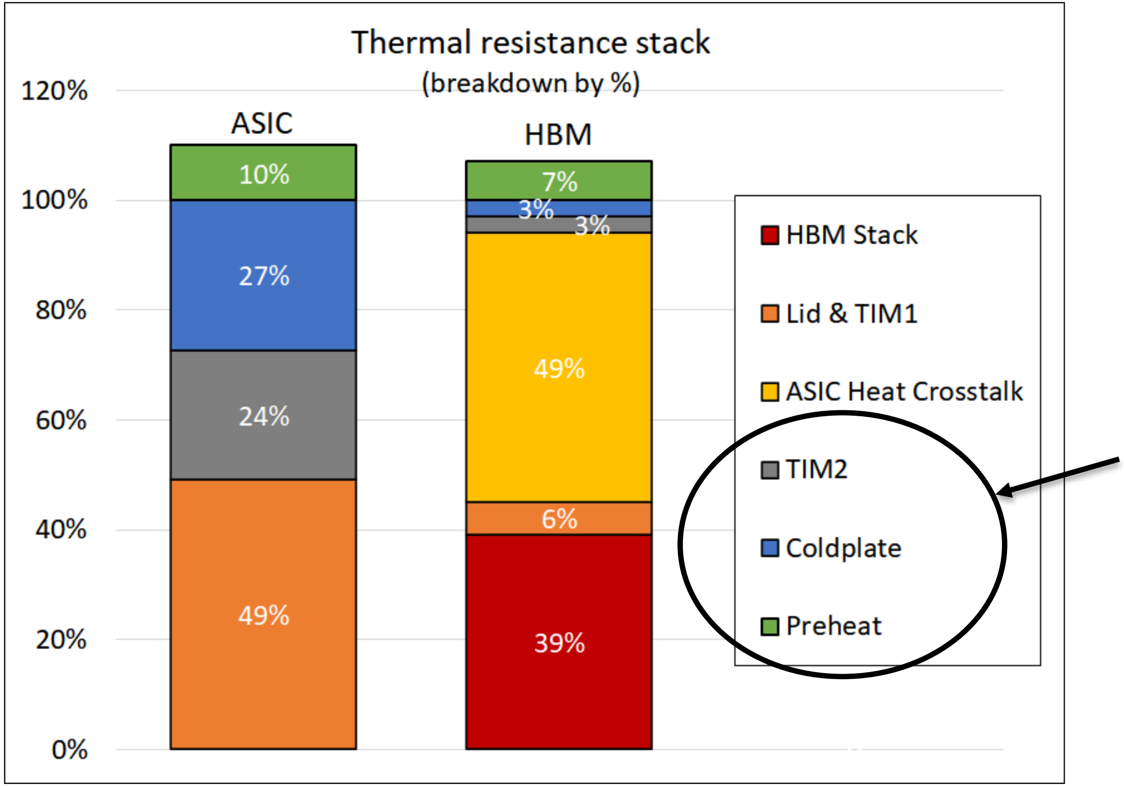
- Average ASIC die temp limit at 80 °C
- Improved HBM stack resistance

Hypothetical heat source map



Many Points to Optimize in a Thermal Stack

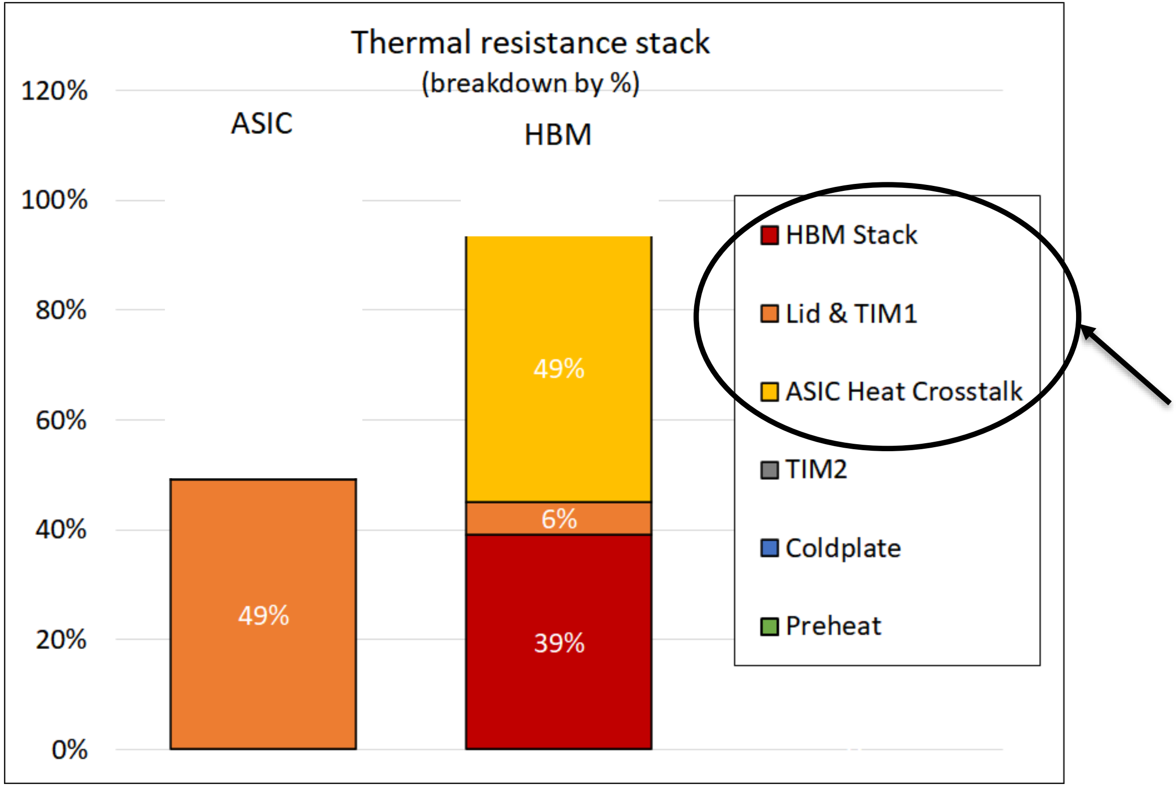
Some optimizations addressable by HW system designers.



Hardware design optimization opportunity

Chipmakers and Packagers Can Make a Big Difference

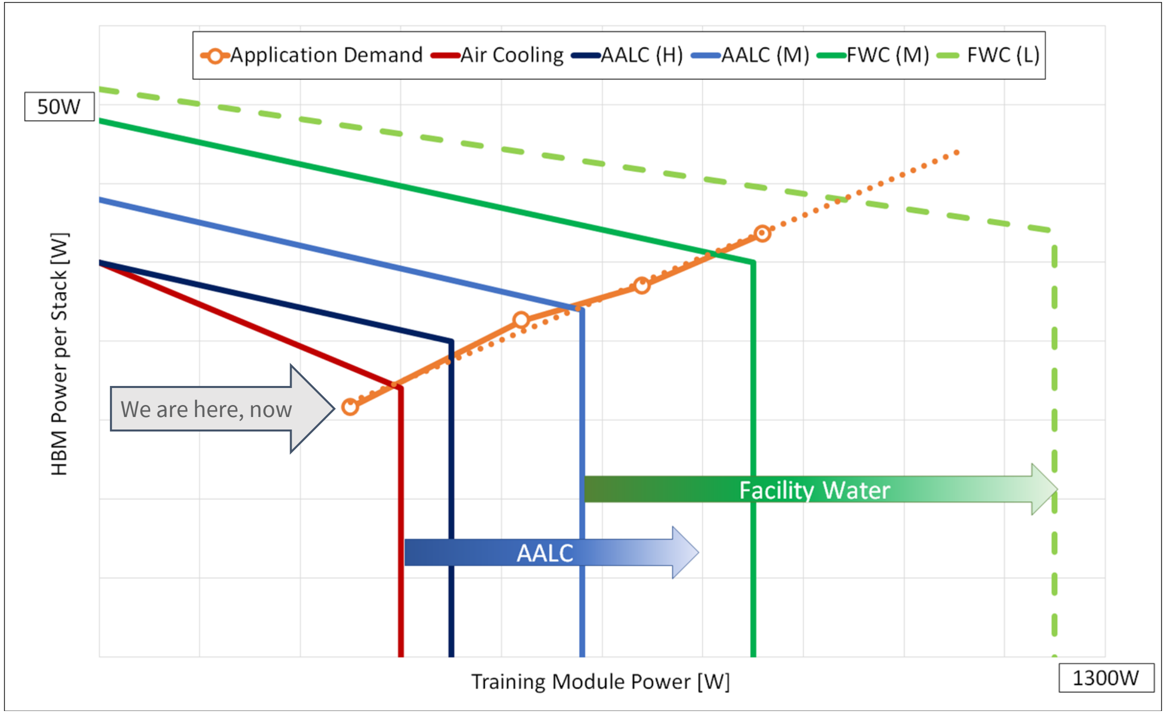
Improvements here benefit whole industry, enhance potential of all coolant set points & thermal solutions



Chip & package optimization opportunity

ML Power Trend vs. Cooling Limits

Both AALC and FWC can be stretched with optimizations in the thermal stack



- AALC serves as transitional solution in air-cooled facilities
- 30 °C facility water can support future ML demand over 5~10 years, and sustainable if package/cooling technology further advance.

- AALC - Air Assisted Liquid Cooling
- FWC - Facility Water Cooling

OPEN POSSIBILITIES.

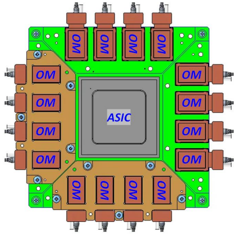
OCP GLOBAL SUMMIT NOVEMBER 9-10, 2021

* This varies some by architecture, data model, and software application.

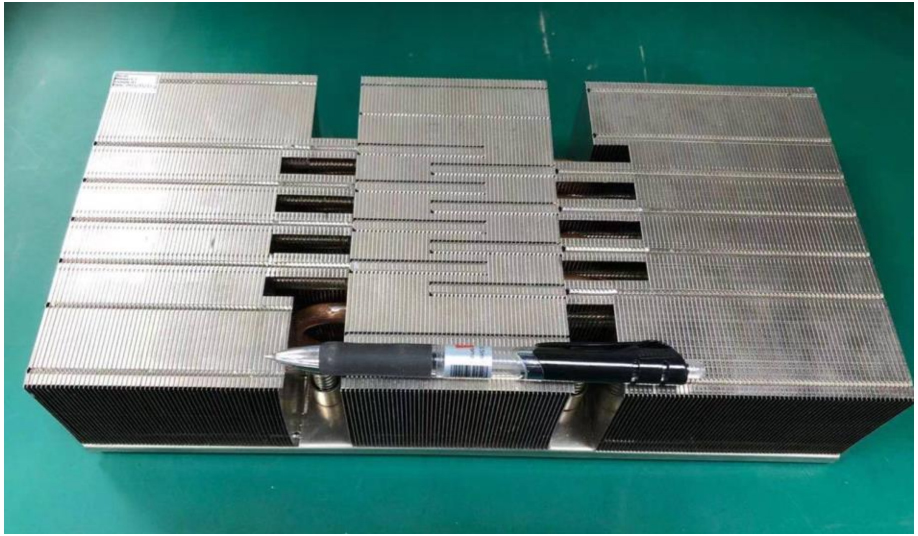
Fabric Switch Cooling Trends

This is not just an ML Problem

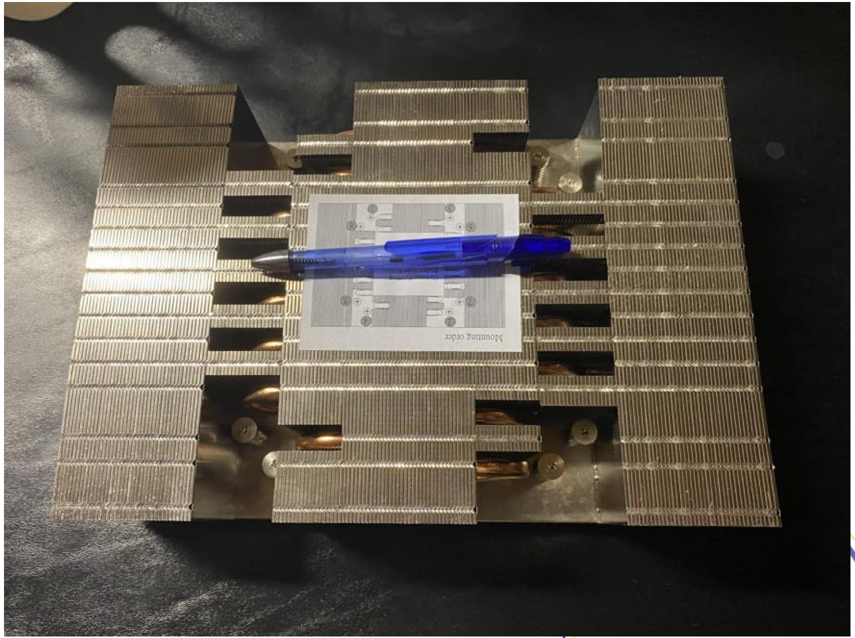
Impact of stretching air cooling for NW chips



2021 - 2.8 kg



2025- 3.7 kg



Challenge & Proposal

Challenge: Lack of community alignment

Current situation with facility water supply temperature

- Range is too wide. Difficult to optimize.
- No concrete boundary conditions to plan / constrain range of optimization. Applies to facility, system, chip thermal sol'n.
- Mature, efficient (PUE and WUE) cooling paradigms for ~30C air.
- Same paradigms durable with facility water at ~30C ranges.

Narrow the range. Benefit from common investment in…

- Infrastructure Design
- Chip design, packaging, thermal solution
- Component/Platform Solution
- Material/Coolant Standards and Supply
- Streamlined design/validation/manufacturing & quality standards

Specialized → W17–W27

Practical → W32–W40

Not Durable → W45–W+

ASHRAE Environmental Specification for Liquid Cooling:

Class	Typical Infrastructure Design		Facility Water Supply Temperature °C
	Primary Facilities	Secondary Facilities	
W17	Chiller/Cooling Tower	Water-side economizer	17
W27			27
W32	Cooling Tower	Chiller / District heating system	32
W40			40
W45	Cooling Tower	District heating system	45
W+			> 45

https://www.ashrae.org/file%20library/technical%20resources/bookstore/supplemental%20files/referencecard_2021thermalguidelines.pdf

Proposal

Better harmonize efforts across OCP to narrow down the boundary

- What supply temperature?
  - 30C liquid & air systems maintains PUE & WUE efficiencies + a durable entry point for liquid cooling
- What HW cooling technology to expect, for chipmakers and IT equipment designers?
  - 30C coolant temps can buy time for further optimization of TIMs, packaging, layout
  - Optimizations at the chip level can, in turn, extend the life of 30C-based facilities
    - Likely stretches air-cooled system viability for more mainstream applications

Lay a foundation to address key areas and gaps.

- How should the chip/platform evolve to maintain continuous perf growth along with efficiency and sustainability improvements?
- How can we reduce the ambiguities/uncertainties for every community member?

Call to Action

- Let’s converge as an industry on a facility water temperature. How about 30C?
  - If not 30C, bring your data and your rationale. You might have a better approach.

- If we can agree on 30C, let’s invest in an ecosystem (from chip to data center) that leverages our infrastructure investment for as long as possible without sacrificing efficiency or performance.

- Let’s form an industry community to amplify this approach
  - Large scale operators, chip suppliers, and facility builders should converge and partner with other industry standards groups (e.g., ASHRAE) for broader influence.

Call to Action

Want to get involved?

- How to get involved - Meta POCs
  - OAI Group Lead [Meta] - Whitney Zhao, whitneyzhao@fb.com,
  - OAI Cooling Lead [Meta] - Cheng Chen, chengchen@fb.com
  - Meta Thermal Lead - John Fernandes, jfern@fb.com

- OAI Group:
  - Where to find additional information: https://www.opencompute.org/wiki/Server/OAI
  - Mailing list: https://ocp-all.groups.io/g/OCP-OAI

- OCP Rack & Power - https://www.opencompute.org/projects/rack-and-power
- OCP Advanced Cooling Solution - https://www.opencompute.org/wiki/Rack_%26_Power/Advanced_Cooling_Solutions

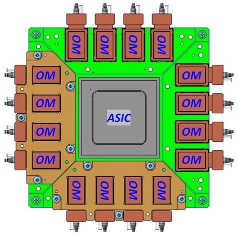
Supporting Material

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

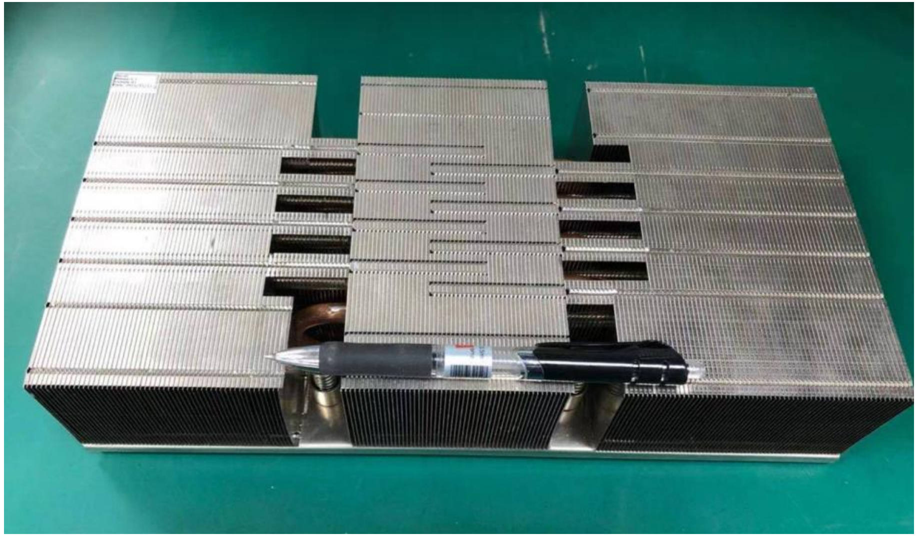
Fabric Switch Cooling Trends

This is not just an ML Problem

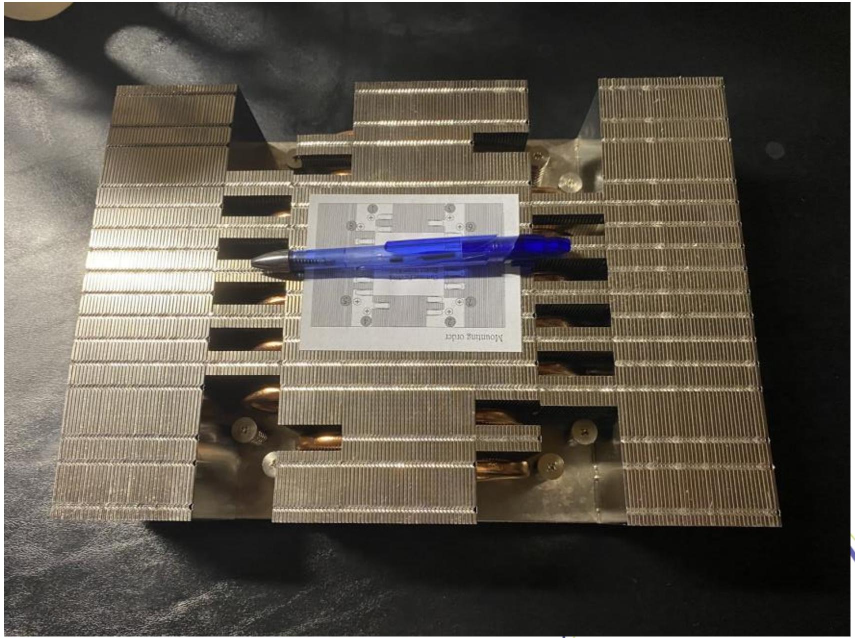
Can Stretch Air Further with Single Chips, Big Heatsinks



2021 - 2.8 kg



2025- 3.7 kg

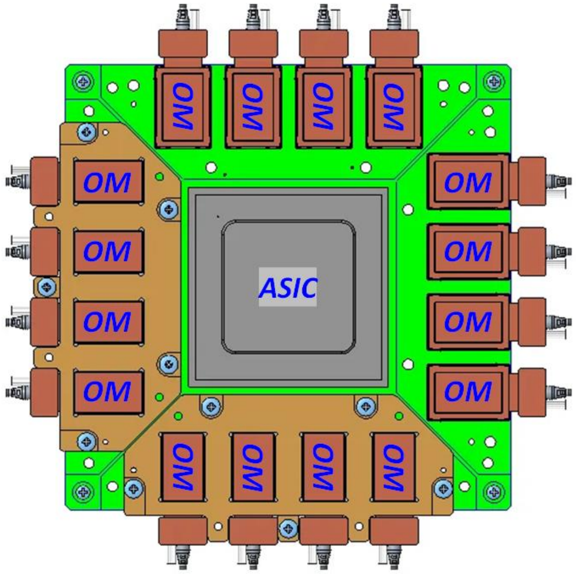


Fabric Switch ASICs. Power Efficient vs. Modular.

The most *power efficient* switch ASICs break free-air-economized limits faster.

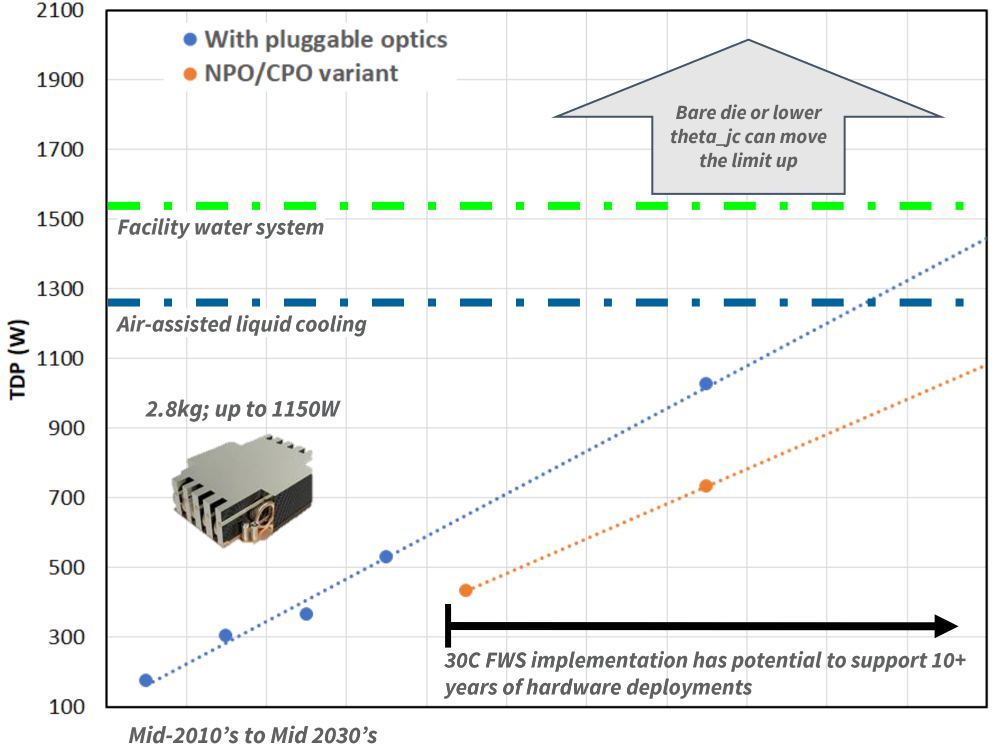
- **Pluggable optics** (less power efficient)
  - Separate cooling solutions for optics and switch ASIC
  - Air-cooling of pluggable optics still the preferred approach
- **Near/co-packaged optics** (more power efficient)
  - Combined cooling solution for ASIC and optics assembly

Hypothetical top side of near-packaged optics assembly



Switch ASIC only

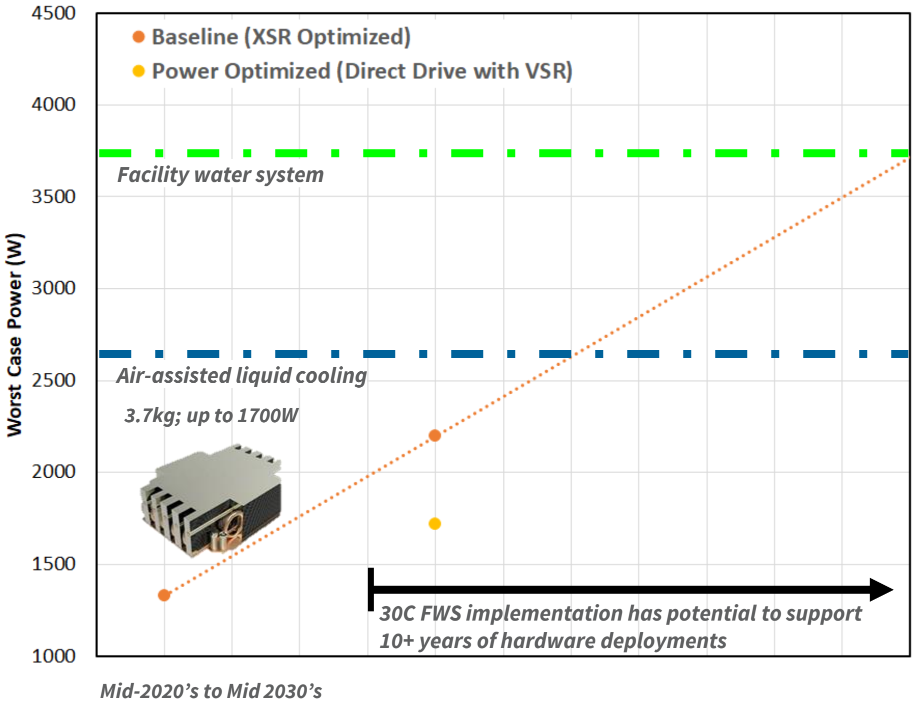
Power trends and cooling limits



- Heat sink solution (in a 4RU chassis) influenced by weight, airflow, fan power and noise limitations
- Cold plate solution should enable long-term support for future switch ASIC packages, but does not alleviate air-cooling concerns for pluggable optics
- System-level cooling (25% PGW mixture)
  - For Air-assisted liquid cooling, coolant supplied at 40°C
  - For Facility Water cooling, coolant supplied at 30°C
  - Coolant side ΔT = 10°C

Co-packaged optics assembly (Switch ASIC + OMs)

Power trends and cooling limits



- Heat sink solution limited by cooling capacity (in addition to factors outlined in the last slide)
- Cold plate solutions may cover us to 3900W. Power-efficiency of direct drive may enable support further out than projected by XSR optimized.
- System-level cooling (25% PGW mixture)
  - For Air-assisted liquid cooling, coolant supplied at 40°C
  - For Facility Water cooling, coolant supplied at 30°C
  - Coolant side ΔT = 10°C

Thank you!