



350W AI 晶片U型散熱模組技術

Auras Technology Inc.

R&D III director

陳恆隆 Ph.D

6 Nov., 2019

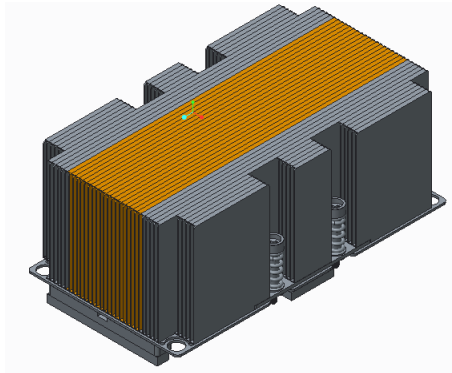




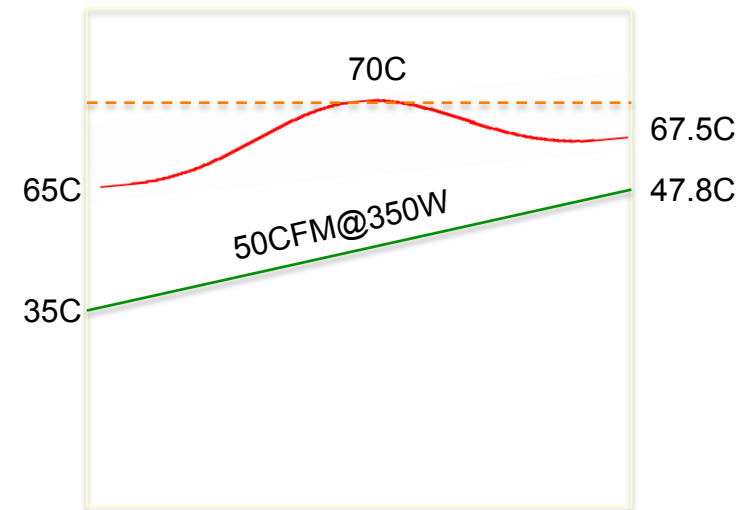
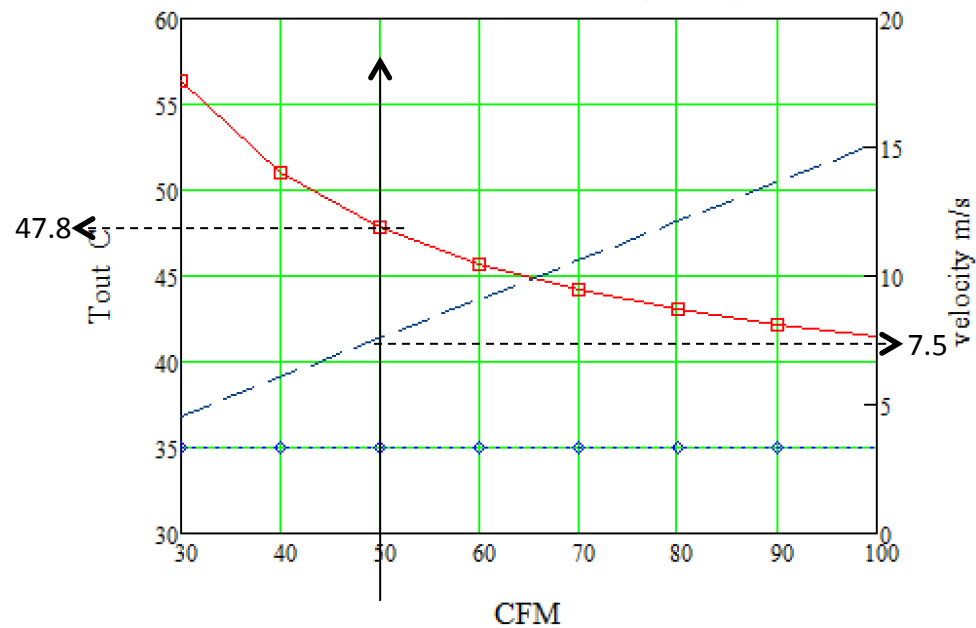
CONTENTS

- I. Original Nvidia 350W AI standard solution
- II. Enhanced Nvidia standard standard solution
- III. Auras U shape 350 W AI chip thermal solution
- IV. Auras U shape 350 W x 2 shadowed AI chip thermal solution

Outlet temperature for 350W GPGPU



CFM v.s. Tout v.s. Vf (350W)



Nvidia 350W VC version

Size: 140Lx78x60H

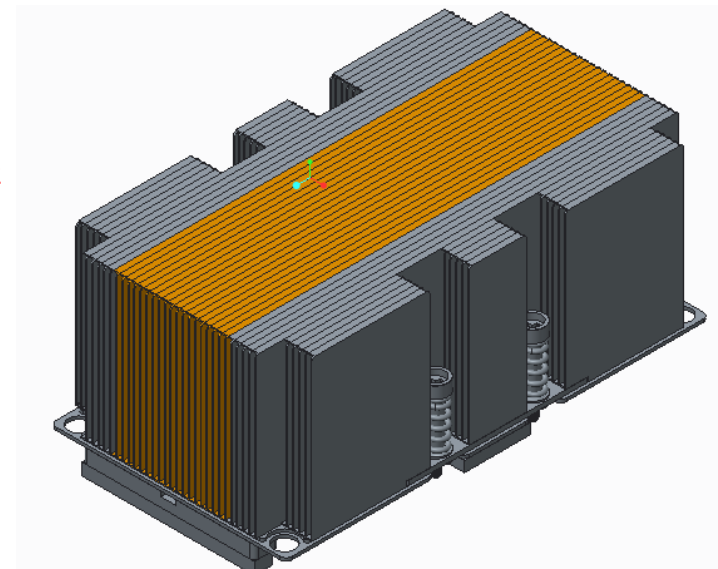
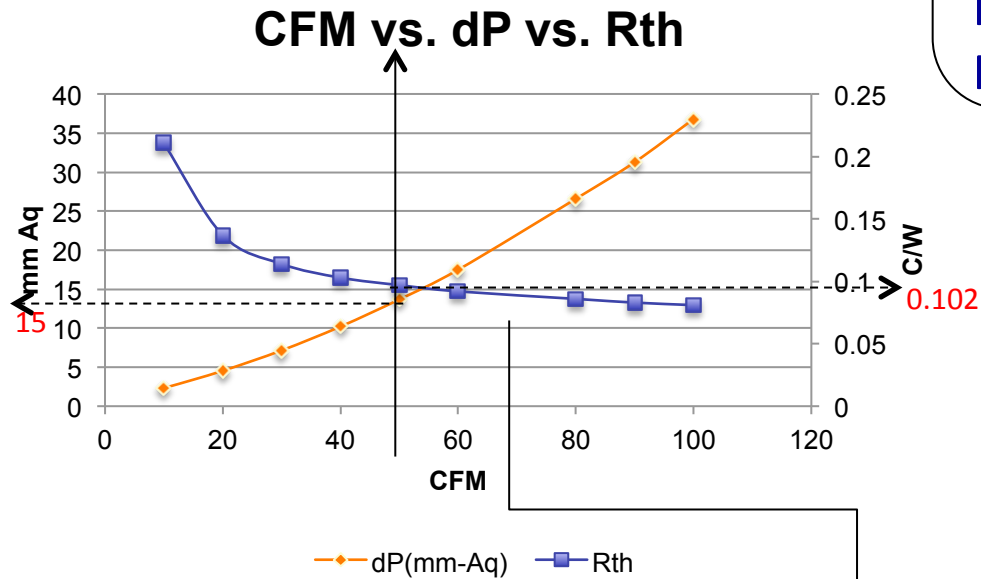
Base: VC

Fin: 40f x 0.4t (19 Cu + 21 Al)

Pitch: 2.0

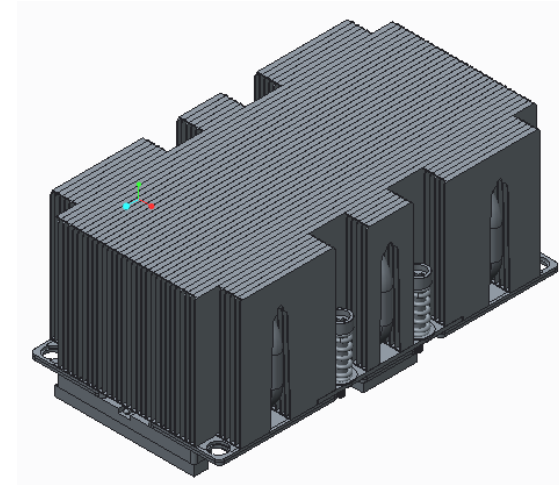
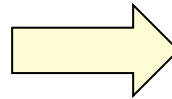
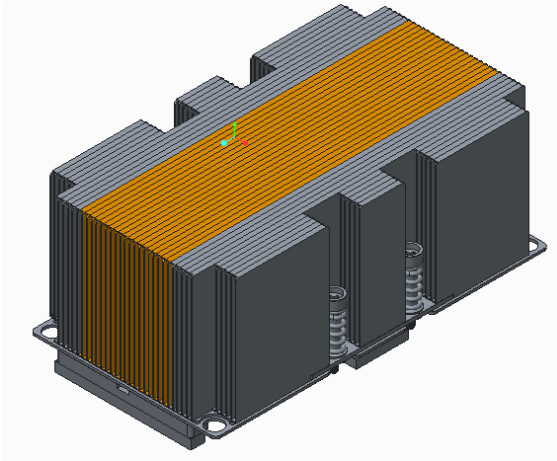
Power: 350W

Flow rate: 50 CFM

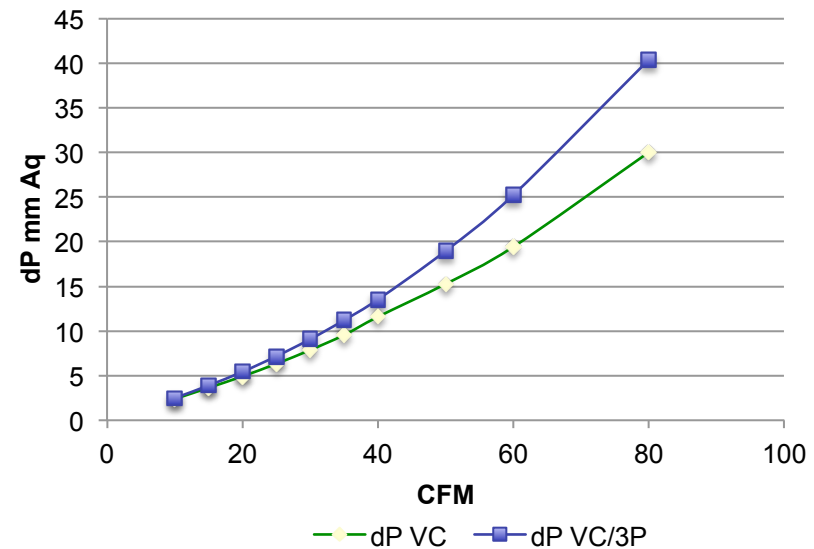
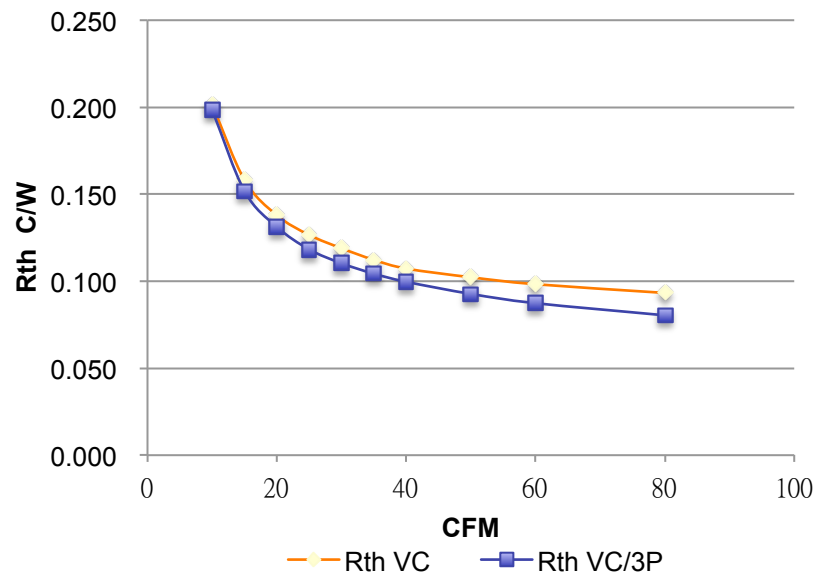


Performance saturated after 60 CFM!

Enhancement of GPGPU > 350W by heat pipes



VC vs. VC/3HP



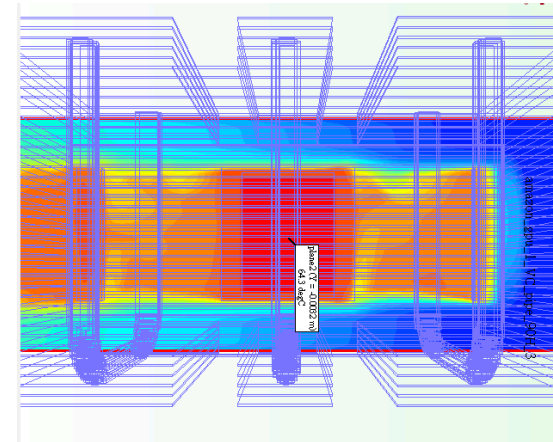
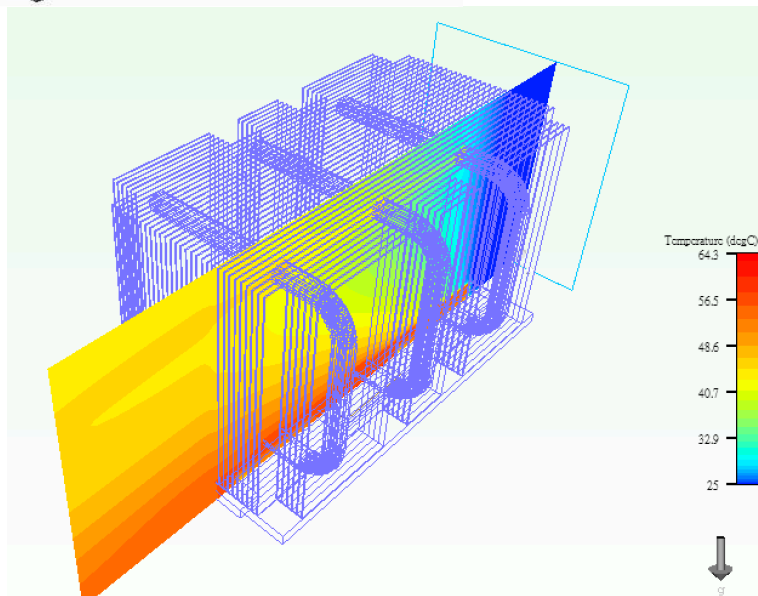
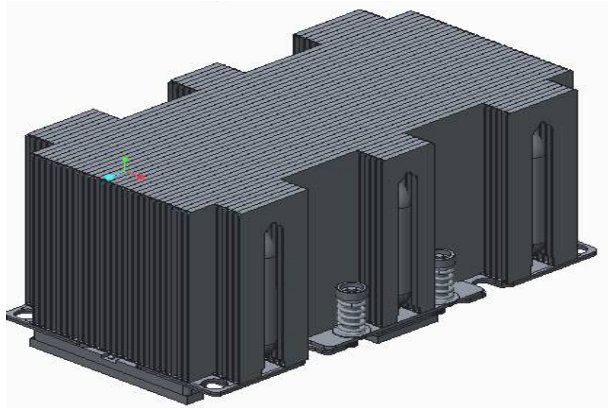
Thermal architecture review of 350W GPGPU

Performance bottle neck:

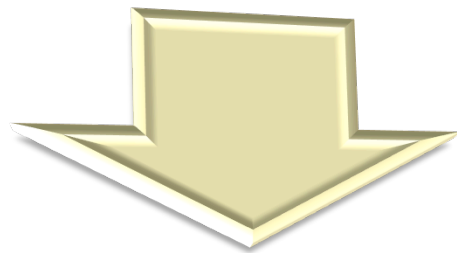
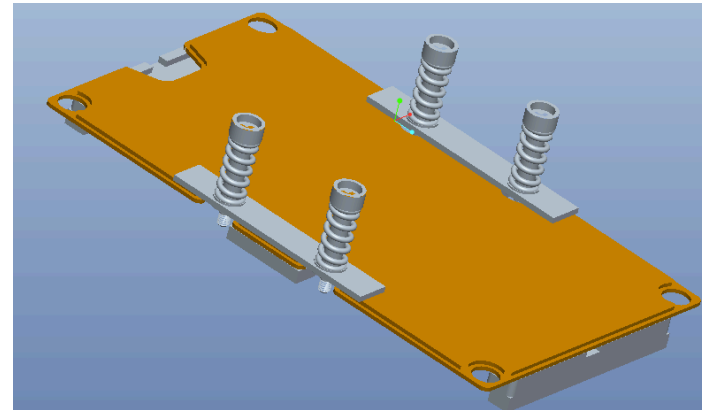
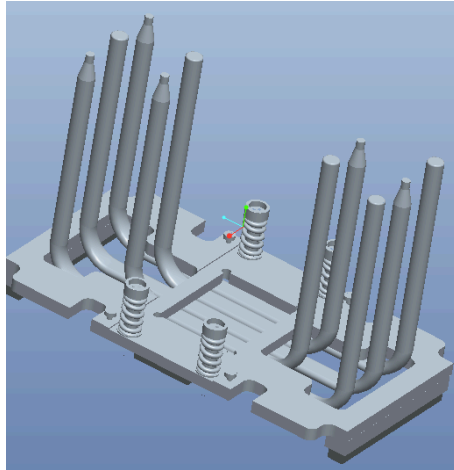
140mm length -> enormous friction

-> saturation of h value

-> air stream heated up in front



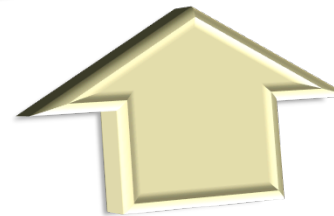
Redesign 350W GPGPU for performance $\uparrow$ & cost $\downarrow$



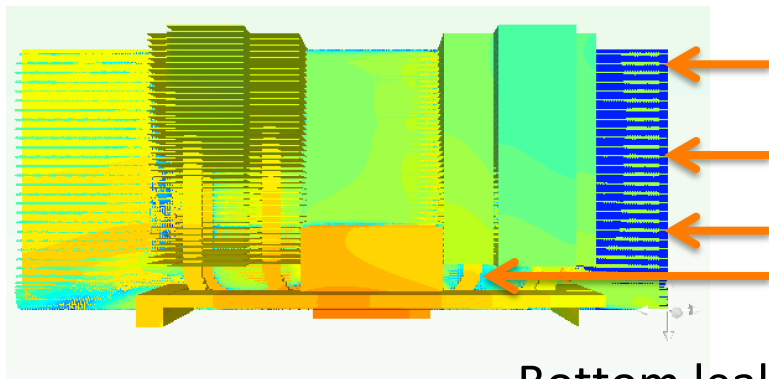
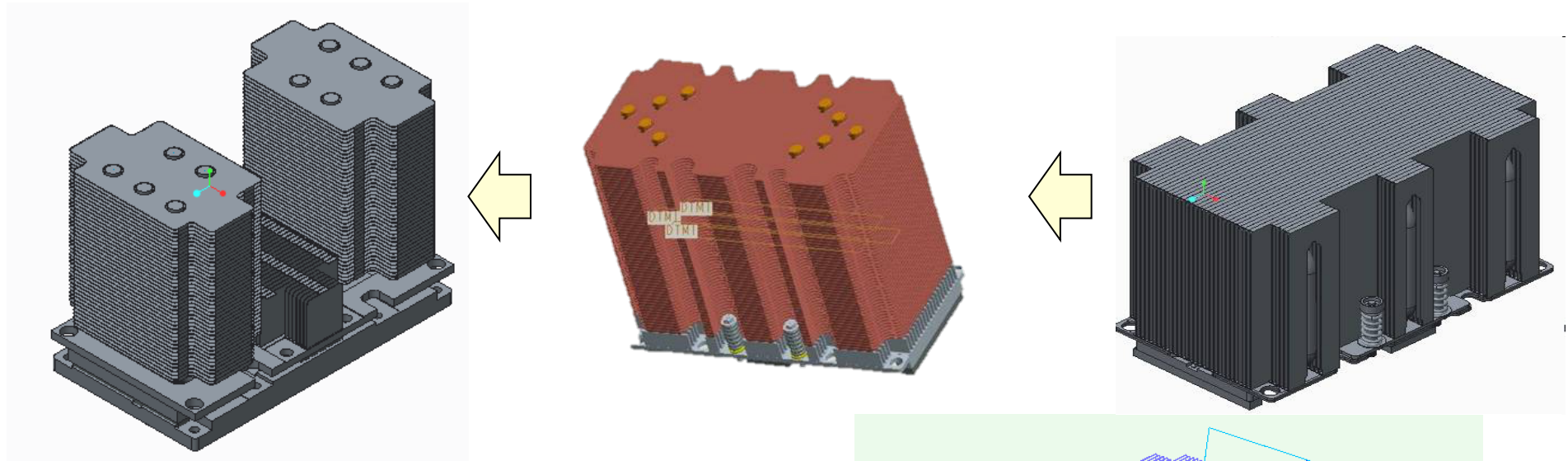
VC cost



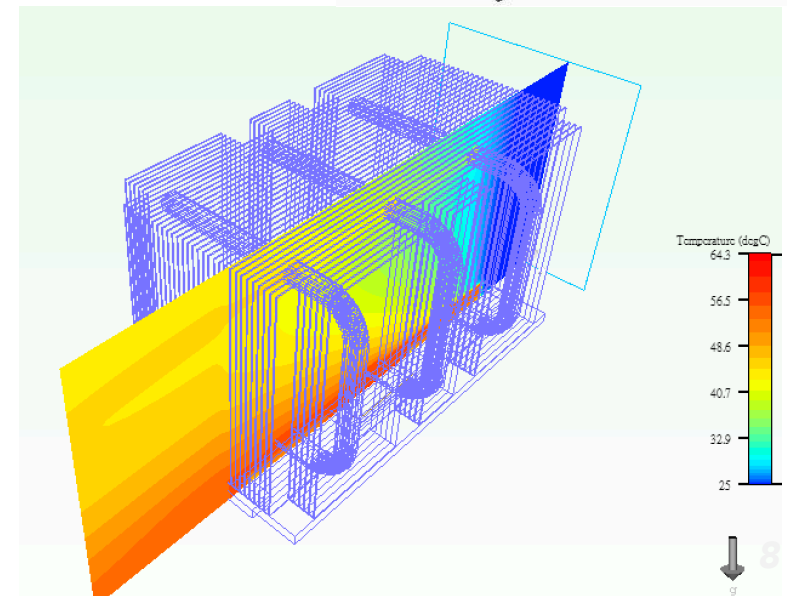
Heat pipe
cost



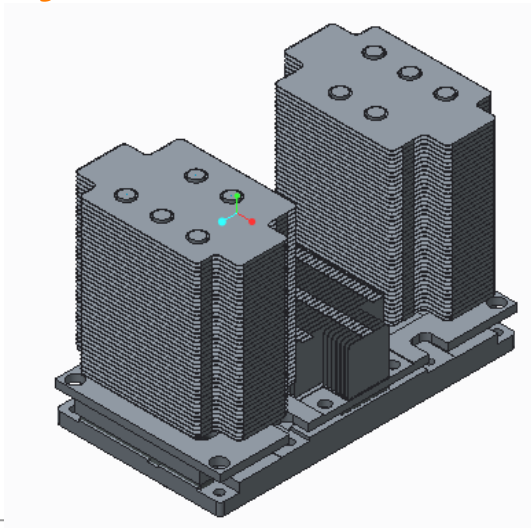
Revolutionary U shape heat sink for *lower impedance and better performance!*



Bottom leakage design

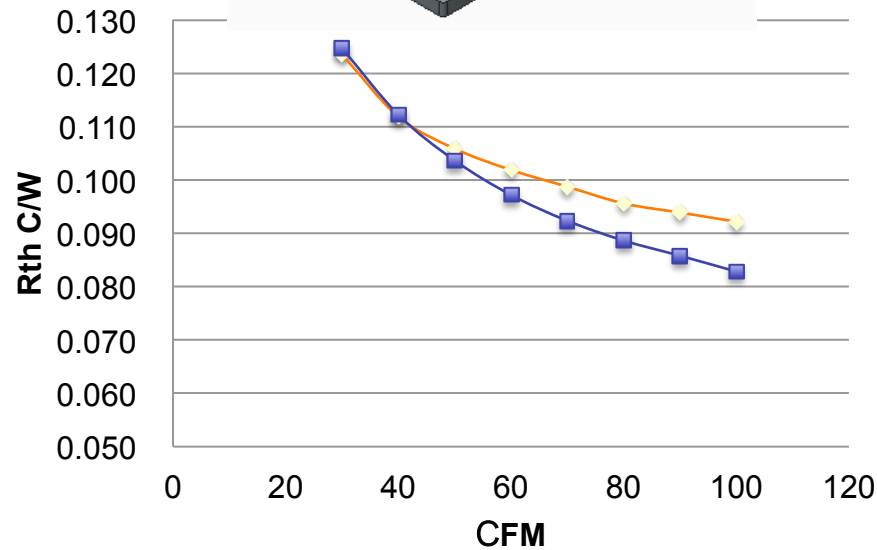


Revolutionary U shape heat sink for *lower impedance and better performance!*

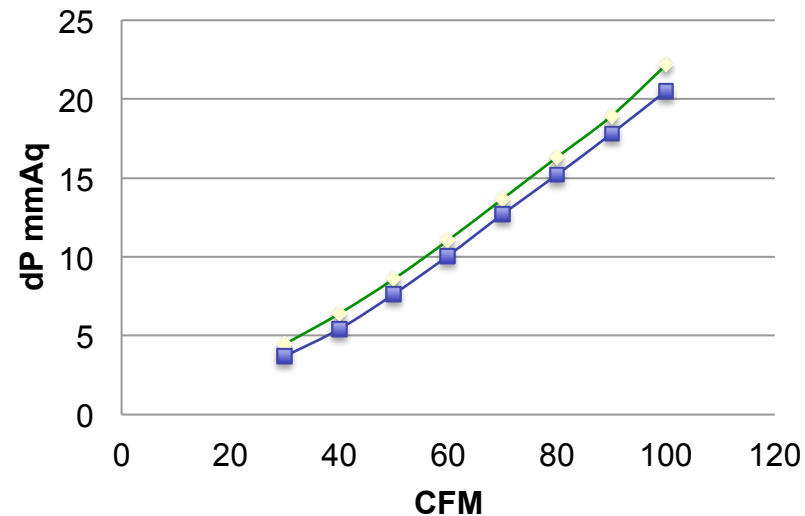


Product description:

1. Dimension: 140*78*68/105H
2. AL FIN (F/R) 49f/0.3t/1.8p
3. CU FIN (Mid) 40f/0.3t/2.0p
4. heat pipes: $\phi 6 \times 5$
5. AL-base
6. 350W



—●— Rth (VC+3P) —■— Rth (5P)



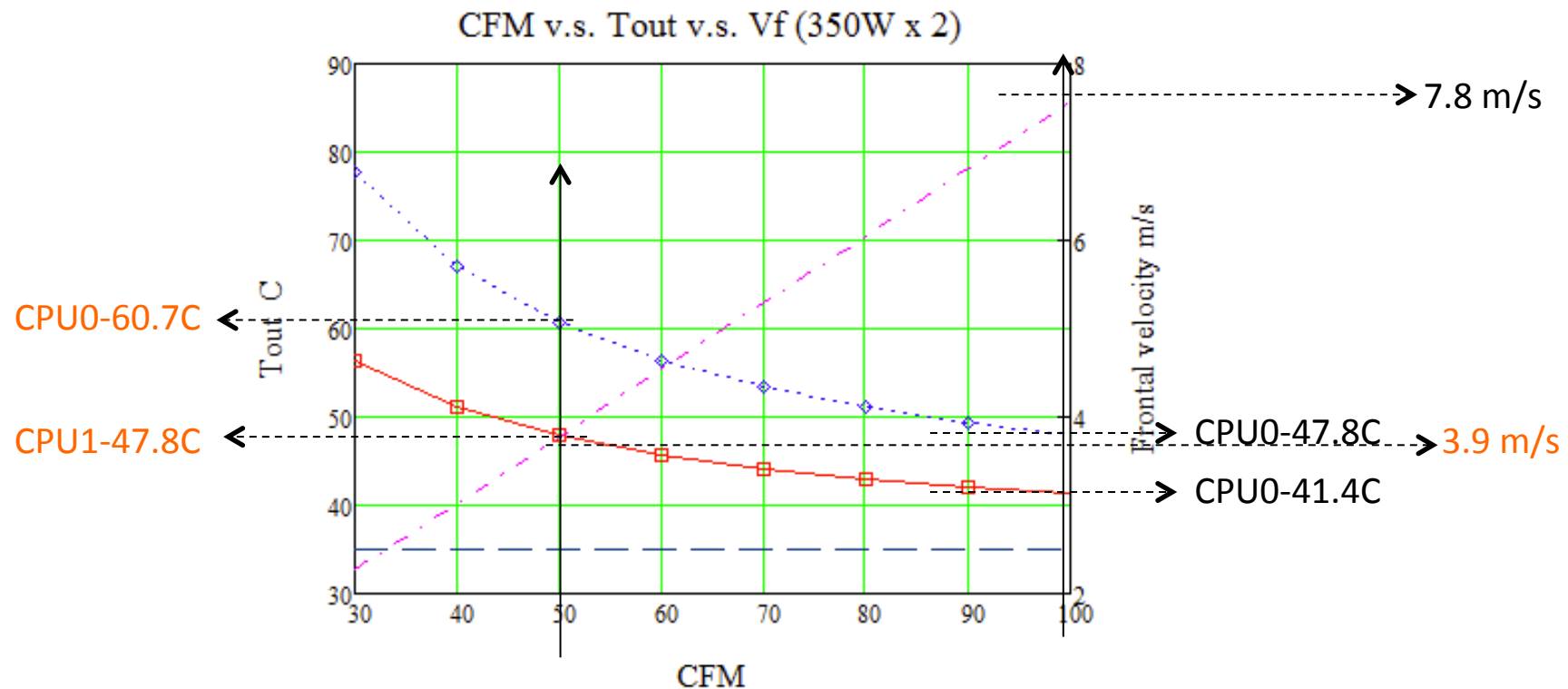
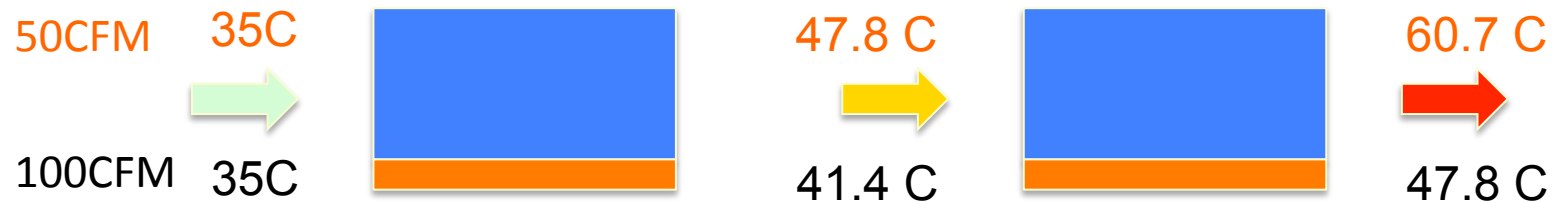
—●— ΔP VC(mm-aq) —■— ΔP 5P(mm-aq)



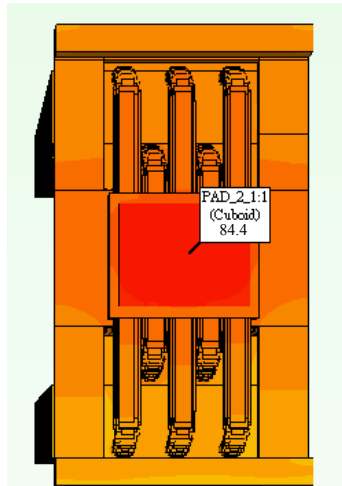
350Wx2 shadowed placement



Discussion of outlet temperature for 350W GPGPU

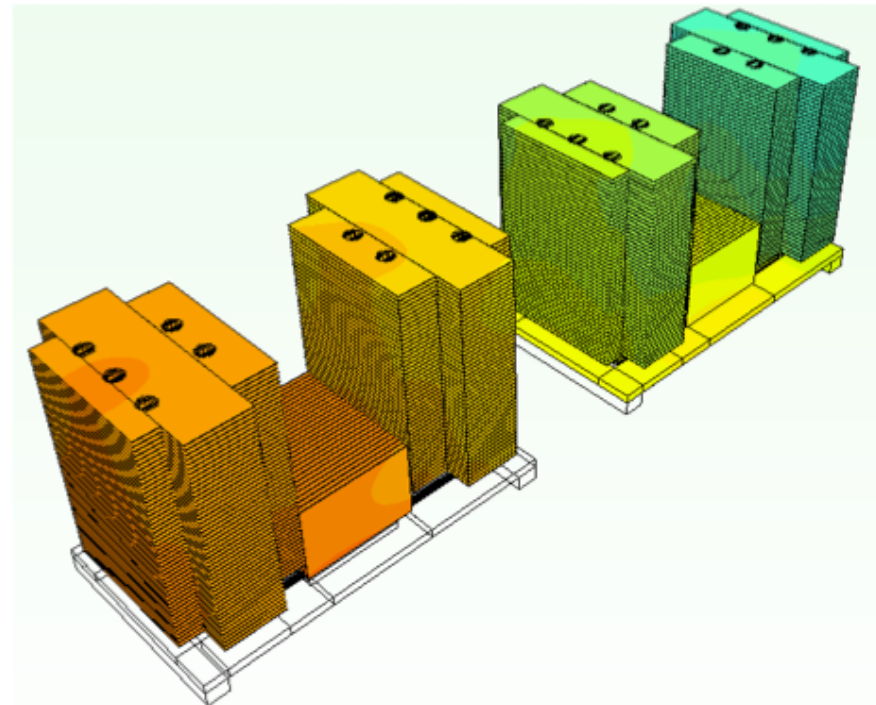
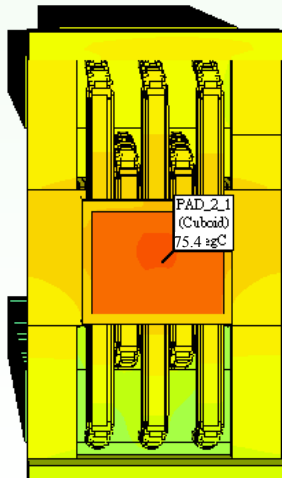


100mm(H)/front 49f/rear 49f

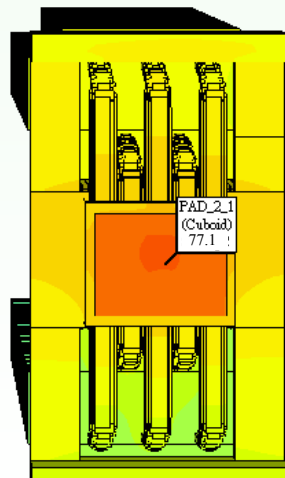
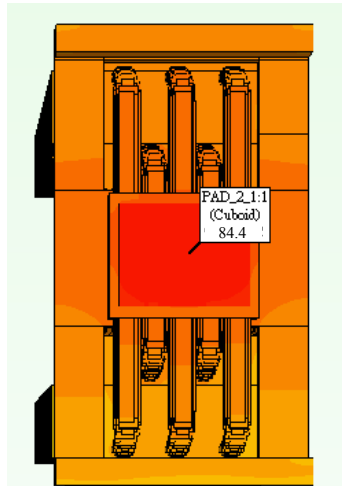


Front module 49f – 75.4°C

Rear module 49f – 84.4°C

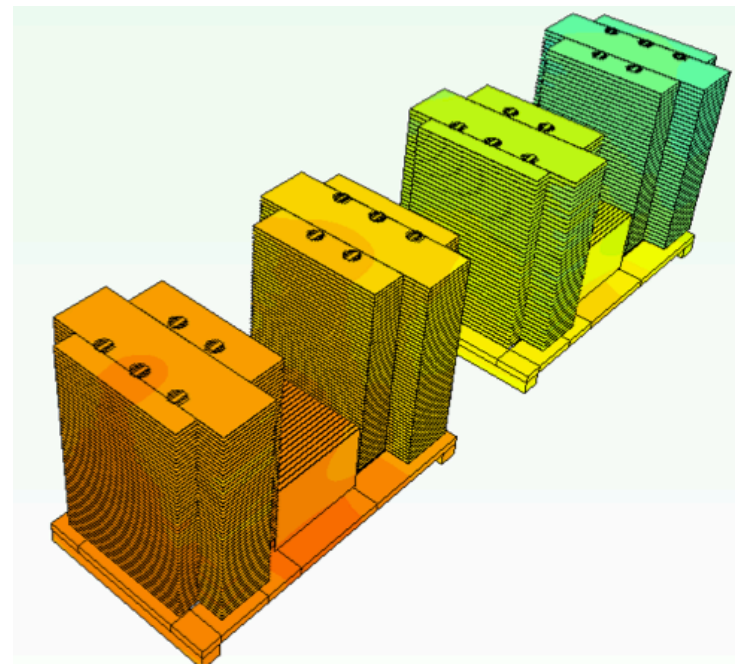


100mm(H) /front 49f/rear 40f



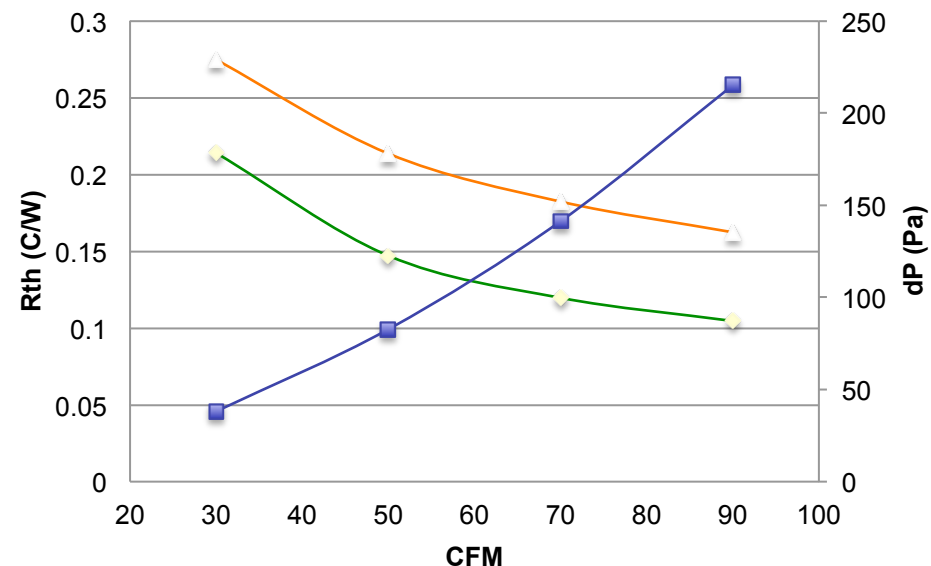
Front module 40f – 77.1°C

Rear module 49f – 84.4°C



- Front module becomes worse due to less fins
- Rear module keeps the same

Keep **the same height**, rear module is always 8~10C higher than front module in all CFM!

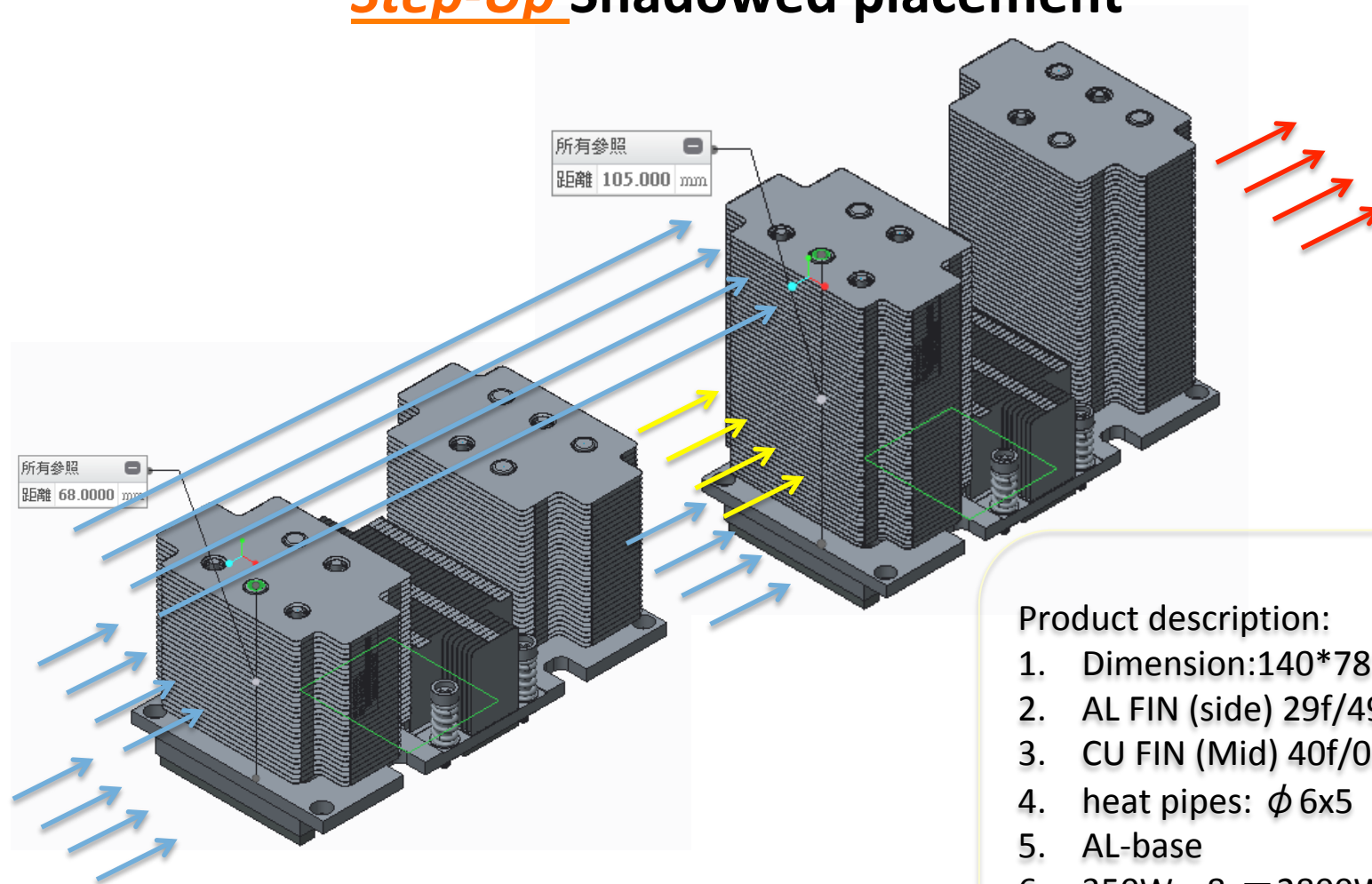


— F die : Rth (W/C) — R die : Rth (W/C) — Front : Pressure (Pa)



Change to **different height**, rear module 1/3 higher Than front module – **Step-Up** design

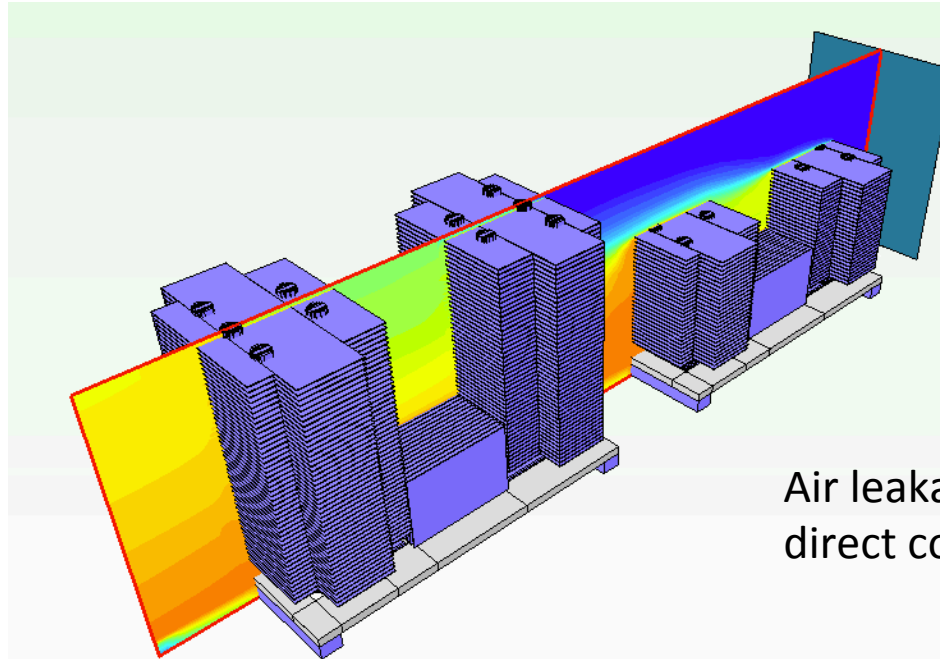
Step-Up Shadowed placement



Product description:

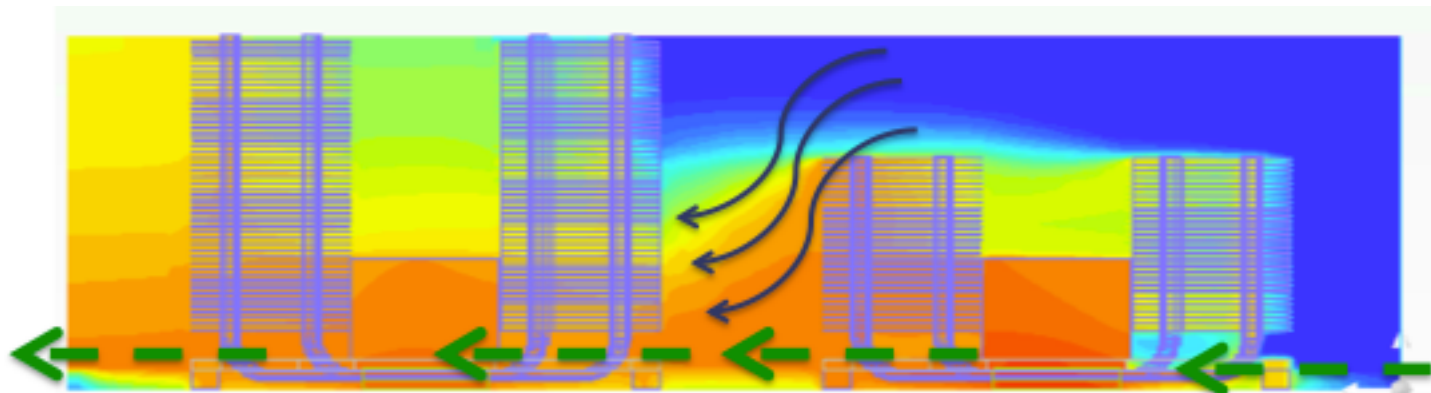
1. Dimension: 140*78*68/105H
2. AL FIN (side) 29f/49f/0.3t/1.8p
3. CU FIN (Mid) 40f/0.3t/2.0p
4. heat pipes: $\phi 6 \times 5$
5. AL-base
6. $350\text{W} \times 8 = 2800\text{W} / \text{rack}$

Flow field of Step-Up design

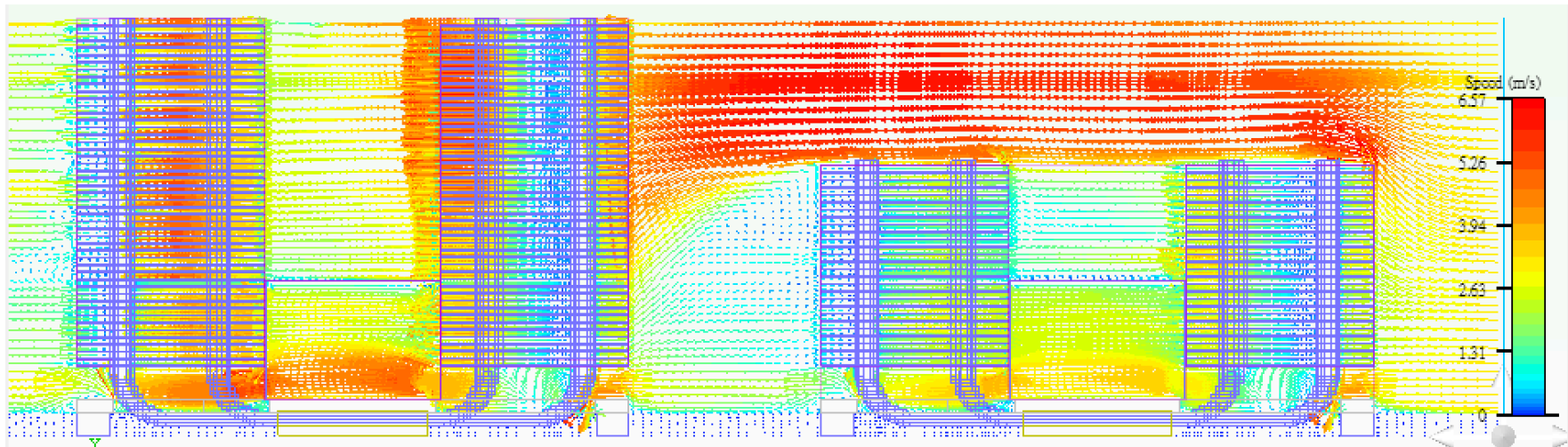
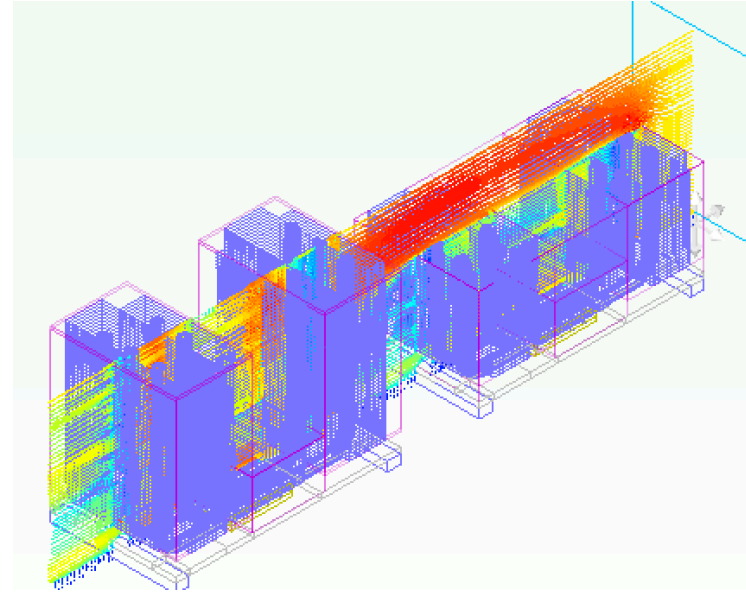
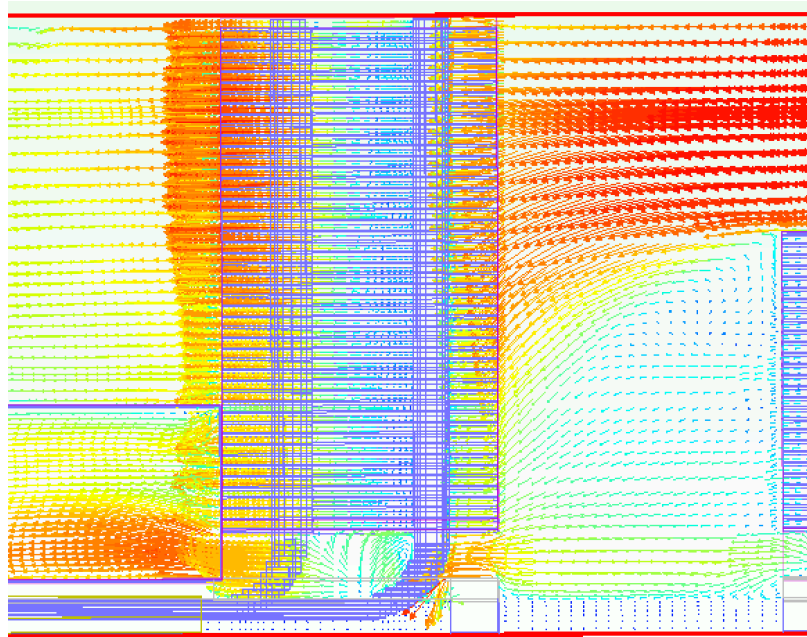


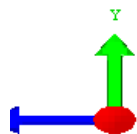
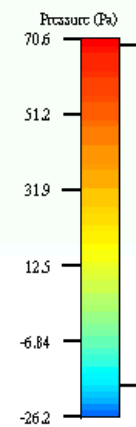
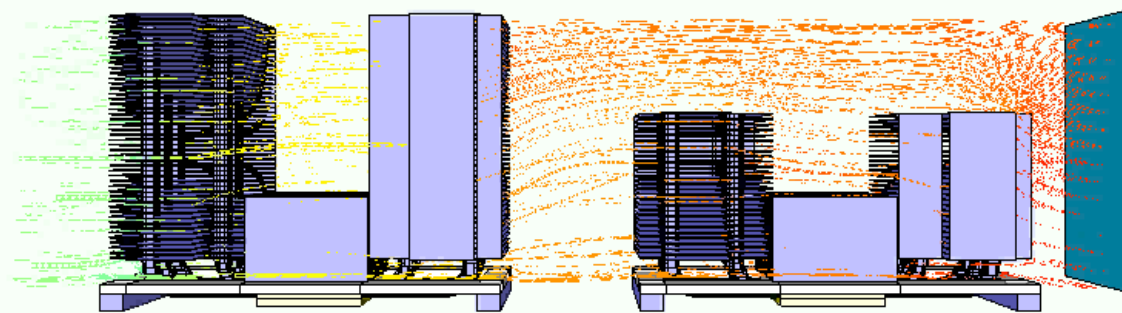
Air leakage on top to
Induce cool air flow downwards!

Air leakage on bottom to
direct cool air flow through hottest chip area!

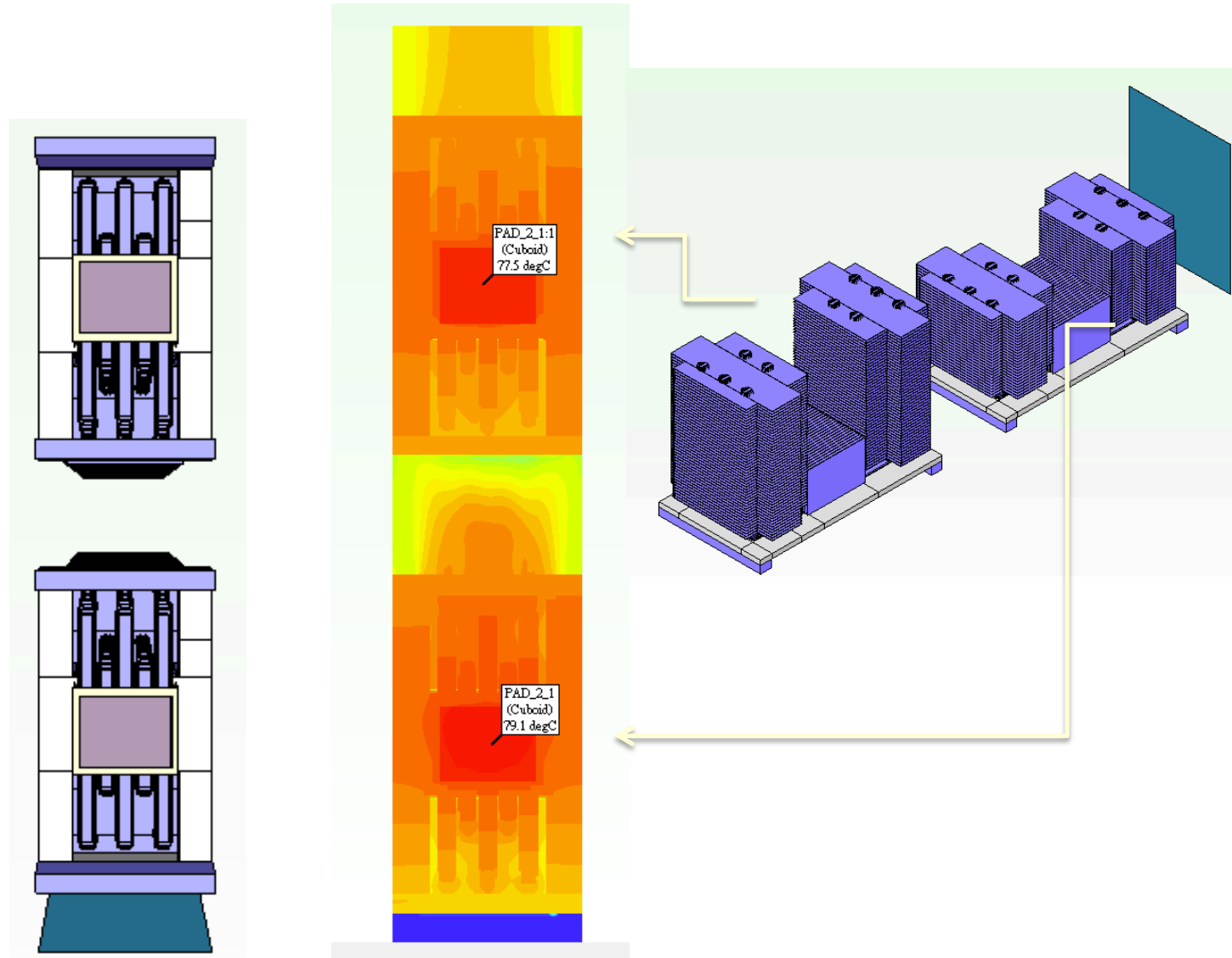


Flow field of StepUp dedign



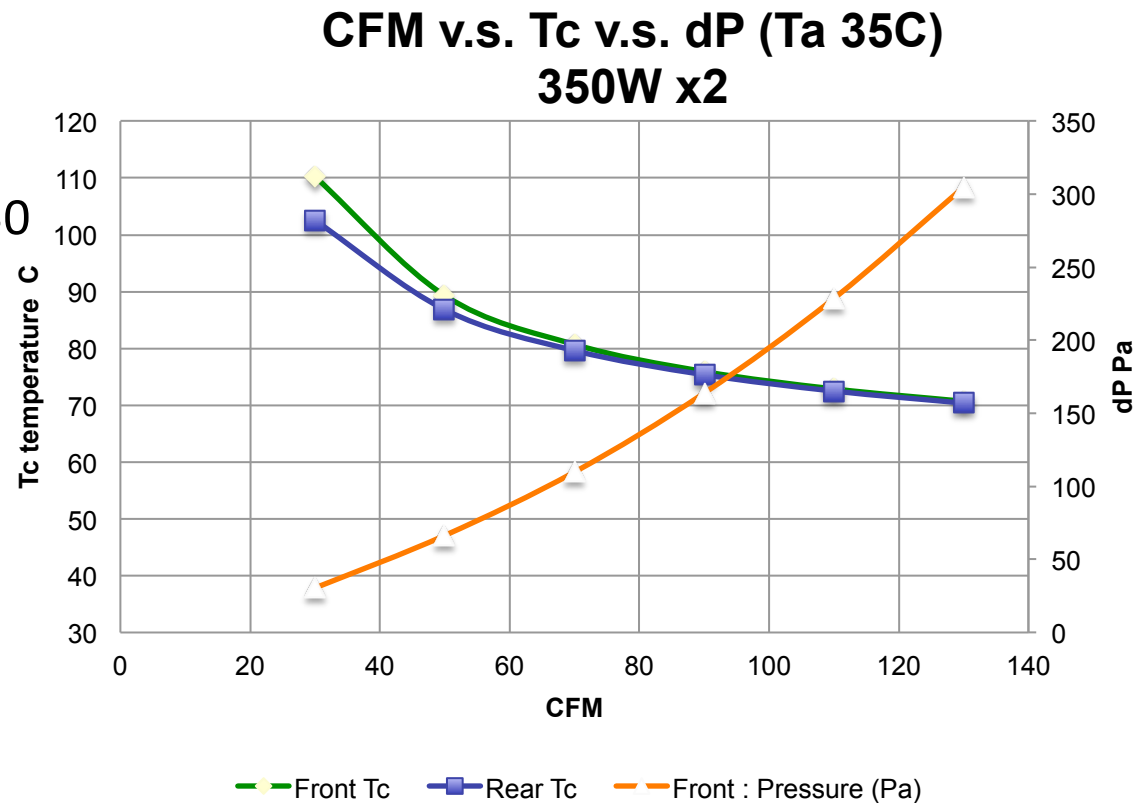


Chip temps of 68mm(H) & 105mm(H)



Result of 35C inlet

- The Step-Up special layout and deliberate leakage design, this enables front and rear module to have close temperature after 50 CFM

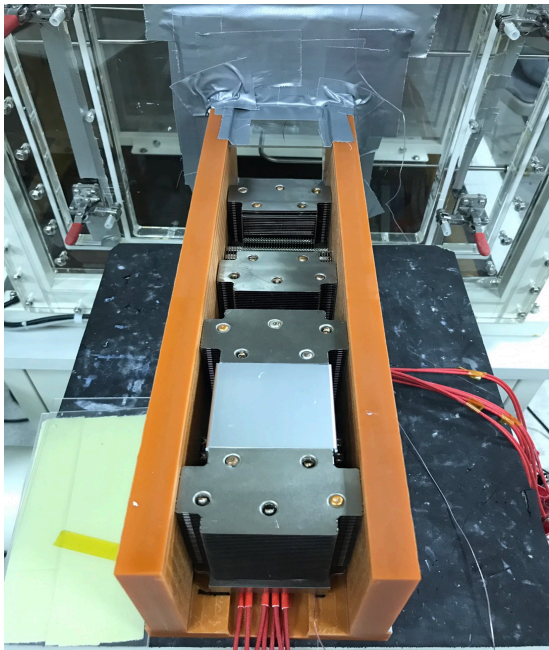


Step-Up design

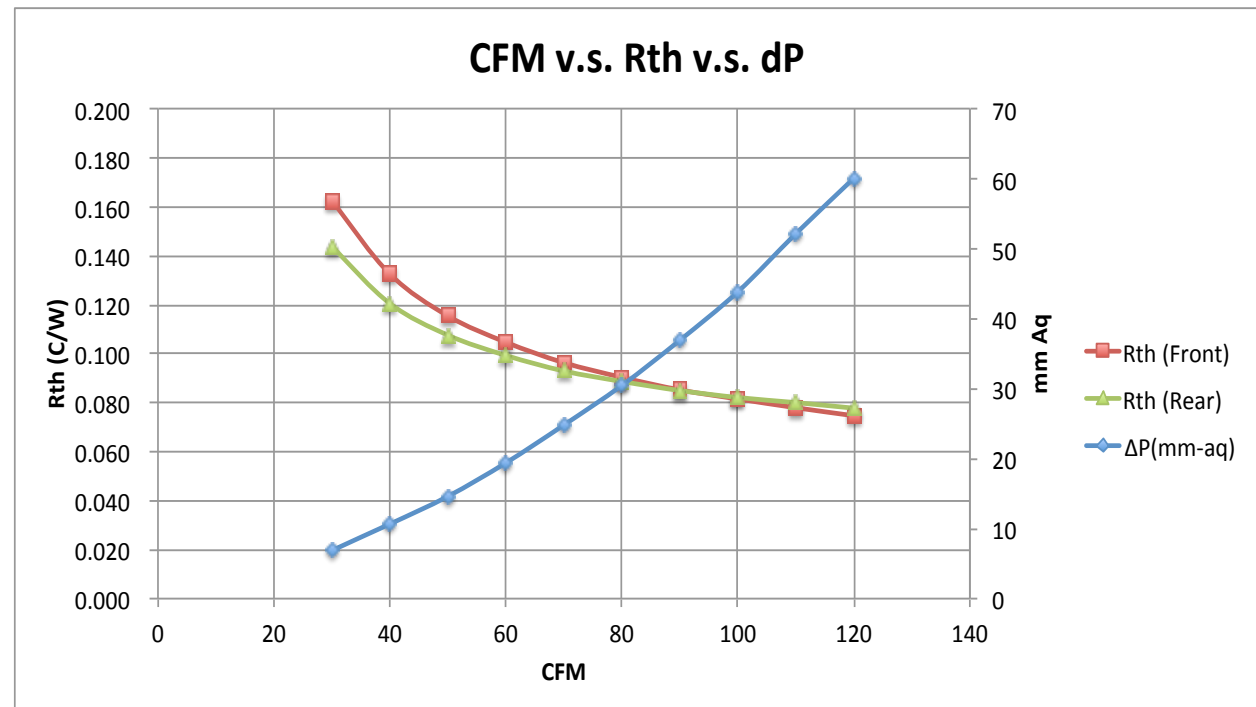
CPU1 temperature = CPU2 temperature!!

Experimental Result

- The Step-Up design with deliberate leakage design that enables front and rear module to have close temperature after 50 CFM
- Test results are in agreement with the simulation results!

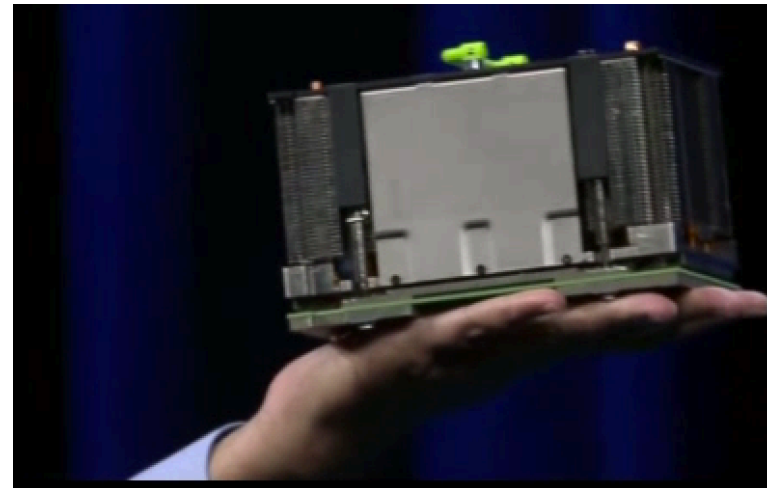
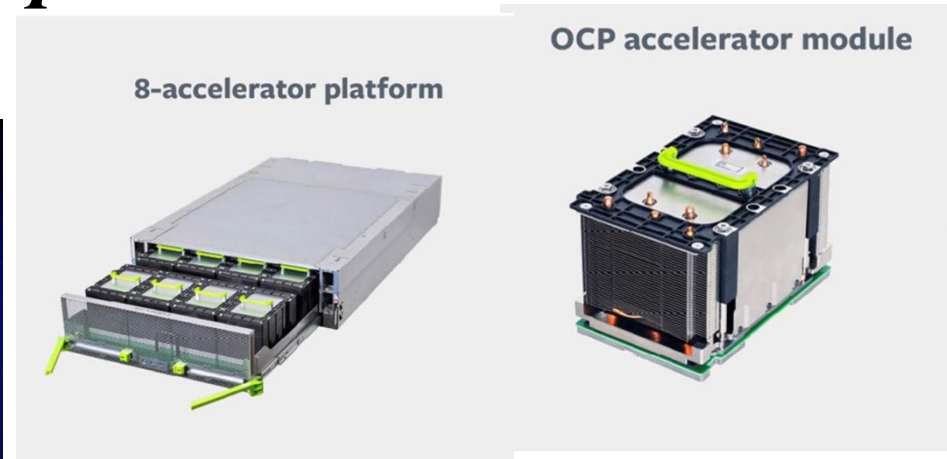


Step-Up design



CPU1 temperature=CPU2 temperature!!

Auras U shape module become OCP AI chip standard solution up to 400W!



8-accelerator platform



8-accelerator baseboard



OCP accelerator module



OAM (OCP Accelerator Module) Partners



AMD

XILINX

habana

GRAPHCORE

Lenovo

NVIDIA

Google

Microsoft

intel

Tencent 腾讯

QCT

Baidu 百度

facebook

Alibaba Group
阿里巴巴集团

“

AMD is excited to join with Facebook, Microsoft, and the industry to contribute the open accelerator module (OAM) specification to the Open Compute Project. The entire ecosystem now has an open path toward increasing adoption of machine learning through the development of accelerator modules and systems that support the connectivity needs of high-performance computing workloads.

- David Wang, Senior Vice President Engineering, Radeon Technologies Group, AMD



Auras Auras Auras Auras Auras Auras Auras Auras Auras Auras Auras

Thank you for your attention!