

Running orbital AI data centers hot: chip leakage matters as much as temperature tolerance

Jiawei Jing

jiawei.jing@mkhr.co.jp

Makuhari Development Corporation <https://orcid.org/0009-0000-8365-367X>

Research Article

Keywords: orbital AI data center, space radiator, leakage power, junction temperature, electronics cooling, mass optimization

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11473415/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Running orbital AI data centers hot: chip leakage matters as much as temperature tolerance

Jiawei Jing*

Makuhari Development Corporation, Chiba, Japan

*Corresponding author. E-mail address: jiawei.jing@mkhr.co.jp. ORCID: 0009-0000-8365-367X

Abstract

Mass estimates for orbital AI data centers place radiators at 350–385 K to shrink them through the fourth-power law. Running that hot is usually treated as a question of chip and high-bandwidth memory (HBM) temperature ratings. This study shows that chip leakage, whose power rises with temperature, matters as much. A closed on-orbit mass model links junction, coolant loop, radiator, solar array and storage, treating chip and HBM limits as design variables and constraining leakage with vendor power-estimation tools, transistor measurements and GPU measurements. The literature radiating temperatures require chip and HBM ratings of about 117–174 °C, and at present leakage most leakage shapes leave such fixed-clock designs without a stable operating point. Raising the chip and HBM limits from 85 and 95 °C to 150 °C would save 23.5% of on-orbit mass for a chip without leakage, but present leakage cancels about nine-tenths of this, leaving 2.6% with the nominal radiator. The tolerable leakage, measured as leakage-driven power growth per kelvin, increases almost linearly with the radiator's share of the power-dependent mass; present GPUs (0.42–0.51% K⁻¹ at 85 °C) lie below the break-even value (0.80% K⁻¹) but above the 0.33% K⁻¹ that a 5% saving requires. Leakage can be bypassed: reducing both chip- and HBM-to-coolant temperature differences from 40 to 25 K lets the coolant run hotter without heating the chips, saving 8.0%, more than higher ratings save at present leakage. Chips for orbital AI data centers therefore need low leakage as much as high temperature ratings.

Keywords: orbital AI data center; space radiator; leakage power; junction temperature; electronics cooling; mass optimization

Nomenclature

Symbol	Description	Unit
A	radiator planform area	m^2
b	leakage activation temperature	K
F	radiator view factor not blocked by the spacecraft's own structure	—
f_L	leakage fraction of chip (module) power at T_{ref} , full load, nominal clock	—
$g(T)$	leakage temperature function, $T^2 \exp(-b/T)$	—
M	on-orbit mass per kW_{ref}	$\text{kg kW}_{\text{ref}}^{-1}$
M_{hp}	heat-pump mass	$\text{kg kW}_{\text{ref}}^{-1}$
$m_{\text{IT}}, m_{\text{fix}}$	compute-hardware and fixed mass	$\text{kg kW}_{\text{ref}}^{-1}$
m_{bat}	energy-storage mass per kW of bus power	kg kW^{-1}
$m_{\text{comp}}, m_{\text{evap}}$	compressor and evaporator specific mass, per kW of compressor electric power (kW_e) and per kW of heat (kW_{th})	$\text{kg kW}_e^{-1}, \text{kg kW}_{\text{th}}^{-1}$
n_{chip}	number of chips per kW_{ref}	—
P	chip power normalized to its value at T_{ref}	—
P_{tot}	total bus power	kW
Q	heat rejected by the radiator	kW
q_{env}	absorbed environmental heat flux, both faces	W m^{-2}
s	leakage-driven power growth, $(1/P) dP/dT_j$, quoted at 85 °C	$\% \text{ K}^{-1}$
s_0, s_5	break-even value and 5% threshold of s for higher ratings	$\% \text{ K}^{-1}$
SP_{pv}	solar-array specific power	W kg^{-1}
T_j, T_{hbm}	junction and HBM temperature	°C
T_c	local coolant temperature	°C
$T_{\text{in}}, T_{\text{out}}$	coolant temperature at the loop inlet (radiator outlet) and loop outlet (radiator inlet)	°C
T_{ref}	reference junction temperature of P and f_L , 87.5 °C	°C
T_{cond}	heat-pump condensing temperature	K
$T_{\text{cap}}, T_{\text{hbm, cap}}$	chip and HBM temperature limits	°C
T_{eff}	effective radiating temperature	K
ΔT_{jc}	junction-to-coolant temperature difference at the power drawn at T_{ref}	K
ΔT_{hbm}	HBM-to-coolant temperature difference (to the coolant outlet)	K
ΔT_{loop}	coolant temperature rise along the loop	K
W	heat-pump compressor power	kW
ε	radiator emissivity	—
$\eta_{\text{fin}}, \eta_{\text{PMAD}}$	fin efficiency; power management and distribution efficiency	—
σ	Stefan–Boltzmann constant	$\text{W m}^{-2} \text{ K}^{-4}$
σ_{rad}	radiator areal mass per planform area	kg m^{-2}
ϕ_{rad}	radiator mass share: radiator part of the power-dependent mass	—

Symbol	Description	Unit
ϕ_{eff}	radiator share: the radiator's share of the power-dependent mass, weighted by the area-temperature sensitivity	—
—	leakage share: part of chip power that is leakage at the actual junction temperature	%
kW_{ref}	unit of compute, equal to that delivered by chips drawing 1 kW at a junction temperature of 80 °C and nominal clock	—

Temperatures in Eqs. (3)–(5) are absolute (K). Chip, HBM and coolant temperatures are given in °C, radiating temperatures in K, and temperature differences in K.

1. Introduction

Proposals to place AI computing in orbit multiplied in 2024–2026. Starcloud's white paper argues for training large models in orbit [1], Google's Suncatcher study tested tensor processing units (TPUs) under proton irradiation and designed a satellite constellation for them [2], Turyshev estimated the specific mass of compute satellites including solar arrays, storage and radiators with a lumped analytical model [3, 4], and Gaalema et al. integrated solar cells, compute boards and radiators into a single panel [5]. All of these designs share one physical constraint. Every watt drawn by the chips ends up as heat that only a radiator can reject to deep space, while the solar array and storage that supply that watt must also be launched. Radiators and power systems therefore dominate on-orbit mass: about 29 kg kW^{−1} in Turyshev's baseline [3], and about 78% of infrastructure mass when radiators are scaled from International Space Station (ISS) hardware [6].

The most direct way to shrink the radiator is to raise its temperature. Because radiated flux grows with the fourth power of temperature, the literature often places the radiator at 350–385 K. Turyshev anchored his estimate at 350 K and noted that 385 K would reduce the area to about two-thirds [3], and the Starcloud white paper proposes heat pumps to raise the radiating temperature [1]; at the chip end, van Berkel quotes graphics processing units (GPUs) designed to operate at 370 K (97 °C) [7]. Other estimates go the opposite way and scale the radiator directly from the ISS area constant [6], although ISS radiators serve crew-cabin and battery temperature levels. Both groups treat the radiating temperature as an input. Turyshev wrote out the temperatures from junction to cold plate, coolant and radiator, but noted only qualitatively that the fourth-power benefit competes with junction limits, leakage and throttling thresholds [3].

System-level optimization of the radiating (heat-rejection) temperature has precedents in spacecraft thermal engineering. Thermophysical models of 100 kW-class heat rejection systems for space nuclear power have been developed and used to optimize radiator design [8, 9]; radiator-condensers for heat-pump thermal control have been studied under lunar conditions [10]; heat-pump and radiator combinations for a lunar base [11] and for high-power spacecraft [12] have yielded optimum radiating temperatures once the power-system mass penalty is included; a space heat-pump prototype has been characterized with measured performance and compressor mass [13]; and radiating temperatures of 370–845 K have been examined for high-temperature SiC electronics [14]. In these studies the heat source either has a fixed temperature or dissipates power independent of temperature. Computing chips do not. Their leakage current, and with it their power,

risers steeply with temperature, so a hotter chip produces more heat, and when the feedback is strong enough an integrated circuit runs away thermally [15].

Terrestrial data centers already face this coupling. Zapater et al. proposed leakage-aware cooling management for servers [16], and Price et al. measured and fitted the temperature dependence of GPU power [17]. Coolant supply (inlet) temperatures of 45 °C are being introduced to reduce chiller energy [18, 19]. On the ground the objective is energy, and the optimum coolant temperature results from the trade between chiller work and leakage. In orbit the objective is mass, radiation is the only heat sink, and the chip temperature limit itself may rise with future high-temperature chips qualified for space.

Running an orbital AI data center hot therefore involves two opposing effects. A hotter radiator is smaller, but it requires hotter coolant, hotter coolant requires chips and high-bandwidth memory (HBM) that tolerate higher temperatures, and hotter chips leak more. The extra leakage power must be supplied by a larger solar array and rejected by the radiator, and in the limit the loop becomes unstable. If temperature tolerance were the only barrier, the problem would disappear once high-temperature chips became available. If leakage matters as well, a higher rating may buy little. How much of the saving from running hot leakage takes away, and what a thermal designer can do about it, has not been quantified.

This study treats the chip and HBM temperature limits as design variables, scanned from 85 to 150 °C, and includes chip leakage, HBM limits, loop temperature rise and pumping power, radiator environmental load and fin efficiency, solar array and storage, and real heat-pump cycles in a single closed mass model. The temperature dependence of leakage is constrained by vendor power-estimation tools and transistor measurements, and the leakage of present GPUs by data-center GPU measurements. The analysis asks, in turn, where present chips place the design point, what the literature radiating temperatures demand of the chips, how much of the saving from higher ratings leakage takes away, how low leakage must be for higher ratings to pay off, and how the coolant can be made hotter without making the chips hotter. The answer is that leakage matters as much as temperature tolerance: it cancels most of the saving from higher ratings, and the most effective way to run hot is to bypass it.

2. Model

2.1. System and mass model

The modeled system is the thermal and electrical system of a liquid-cooled compute satellite, one of the satellites that make up an orbital AI data center (Fig. 1). Chips are mounted on cold plates in series along a single-phase water loop that flows from the cold plates directly into a deployable radiator, without an intermediate heat exchanger; the series string represents one branch of a modular loop, and a megawatt-class system consists of many such branches in parallel. A solar array supplies power through the power management and distribution (PMAD) system, and batteries cover the seasonal eclipses of a dawn–dusk Sun-synchronous orbit. Mass is expressed per kW_{ref} , a unit of compute equal to that delivered by chips drawing 1 kW at a junction temperature of 80 °C and nominal clock. The on-orbit mass is

Symbol	Description	Nominal	Range
$\epsilon; F; \eta_{\text{fin}}$	emissivity; view factor; fin efficiency	0.9; 0.85; 0.85	0.8–0.92; 0.75–0.95; 0.61–0.95
σ_{rad}	radiator areal mass, kg m^{-2} (planform)	10	2–17.5
q_{env}	environmental load, W m^{-2}	127.3	64–255
$\Delta T_{\text{jc}}; \Delta T_{\text{hbm}}$	temperature differences, K	40; 40	15–50; 15–50
T_{cap}	chip limit, $^{\circ}\text{C}$	85	85–150 (design variable)
$T_{\text{hbm, cap}}$	HBM limit, $^{\circ}\text{C}$	95	95–150; $\max(T_{\text{cap}}, 95^{\circ}\text{C})$ unless stated
f_{L}	leakage fraction at 87.5°C	0.15	0–0.35 (scanned)
b	leakage activation temperature, K	3315	2584–4515
SP_{pv}	solar-array specific power, W kg^{-1}	100	50–300
$\eta_{\text{PMAD}}; m_{\text{bat}}$	PMAD efficiency; storage mass, kg kW^{-1}	0.92; 2	0.88–0.96; 0–6
$m_{\text{IT}}; m_{\text{fix}}$	compute-hardware and fixed mass, $\text{kg kW}_{\text{ref}}^{-1}$	7; 6	4–10.5; 3–12
$m_{\text{comp}}; m_{\text{evap}}$	compressor, $\text{kg kW}_{\text{e}}^{-1}$; evaporator, $\text{kg kW}_{\text{th}}^{-1}$	4.6; 1.5	0.92–9.2; 0.5–2.72

2.2. Temperature stack and limits

The junction temperature of each chip solves the fixed-point equation

$$T_j = T_c + \Delta T_{\text{jc}} P(T_j), \quad (2)$$

where T_c is the coolant temperature at that chip’s position along the loop and P is chip power normalized to its value at a junction temperature of 87.5°C , so that ΔT_{jc} is the junction-to-coolant difference at the power drawn at 87.5°C . Because P rises with T_j through leakage (Section 2.3), Eq. (2) is a feedback loop. The loop power is obtained by solving Eq. (2) for each chip along the loop and summing; the chip at the loop outlet is the hottest chip, and “junction temperature” below refers to it. Other cold-plate arrangements, in which all chips see the outlet, mid-loop or supply temperature, are examined in Section 4.5 and Supplementary Section S9. The nominal ΔT_{jc} of 40 K rests on two independent sources: the estimate of an 85–90 $^{\circ}\text{C}$ junction at 45 $^{\circ}\text{C}$ coolant for a new-generation GPU by Gaalema et al. [5], and direct cold-plate measurements on H100 extrapolated to full power [21]. The loop temperature rise is optimized together with the coolant temperature. The loop pressure drop is 200 kPa at a 10 K rise and scales with the square of the flow, and the pump efficiency is 0.5, so pumping power is 1% of the rejected heat at a 10 K rise and falls with the cube of the rise (Supplementary Section S1).

The present limits are taken from published specifications: a chip limit T_{cap} of 85°C , the software throttling onset of data-center GPUs (83°C for V100 [22], 85°C for A100 [23]), and an HBM limit $T_{\text{hbm, cap}}$ of 95°C , the H100 HBM specification [24]. The HBM temperature is the coolant outlet temperature plus ΔT_{hbm} . The junction-to-coolant difference ΔT_{jc} , also called the chip-to-coolant difference, and the HBM-to-coolant difference ΔT_{hbm} , both 40 K nominally, together form the temperature stack between the chips, the HBM and the coolant; “a stack of X K” means that both differences equal X K. Reducing the stack is the lever

examined in Section 4.5. The thresholds are sensor readings and are treated here as limits on the junction and HBM temperatures. A chip or HBM with a higher temperature rating is represented by a higher limit, and “higher ratings” below means raising T_{cap} and $T_{\text{hbm, cap}}$. The scan raises the chip limit to 150 °C, beyond the 125 °C junction rating of automotive-grade processors [25], and lets the HBM limit rise with it, $T_{\text{hbm, cap}} = \max(T_{\text{cap}}, 95 \text{ °C})$. Because HBM, like other dynamic random-access memory (DRAM), has no standard operating mode beyond about 105 °C and its refresh power grows steeply with temperature, the HBM part of this scan is an upper bound; results with the HBM limit held at 95 or 105 °C are reported alongside.

2.3. Leakage

Chip power is split into a dynamic part and a leakage part that grows with temperature:

$$P(T_j) = (1 - f_L) + f_L \frac{g(T_j)}{g(T_{\text{ref}})}, \quad g(T) = T^2 \exp(-b/T), \quad (3)$$

with $T_{\text{ref}} = 87.5 \text{ °C}$, a full-power junction temperature within the 85–90 °C range used for ΔT_{jc} in Section 2.2 (normalization points in Supplementary Section S1.3). Two numbers describe the leakage. The leakage fraction f_L is the part of module power at 87.5 °C that varies with temperature, defined at module level (die plus on-package HBM, as reported by board power telemetry), at full load and nominal clock; chip power in this paper is this module power, and gate leakage and other static power that hardly varies with temperature belong to $1 - f_L$. The activation temperature b sets how steeply leakage grows: leakage doubles every 22 K near 85 °C at the median value. The functional form follows the measurements of Price et al. on GPUs [17]. For the mass model, power is rescaled so that 1 kW_{ref} corresponds to 1 kW at a junction temperature of 80 °C. Together, f_L and b set the leakage-driven power growth s , the rise of chip power per kelvin,

$$s = \frac{1}{P} \frac{dP}{dT_j} = \frac{f_L g(T_j)/g(T_{\text{ref}})}{P(T_j)} \left(\frac{2}{T_j} + \frac{b}{T_j^2} \right), \quad (4)$$

a quantity that can be measured on a chip and is quoted here for the hottest chip of the present design, at 85 °C. Its break-even value s_0 and 5% threshold s_5 are the values of s at which higher ratings stop saving mass and save 5%. The leakage share is the part of chip power that is leakage at the actual junction temperature; in words, s is the leakage share multiplied by the growth rate of leakage, $\ln 2$ divided by its doubling interval. With the median shape at 85 °C, $f_L = 0.15$ gives a leakage share of 14.0% and $s = 0.44\% \text{ K}^{-1}$, about 3 W K^{-1} for a 700 W chip; each $0.1\% \text{ K}^{-1}$ of s corresponds to about 3 percentage points of leakage share. Three reference junction temperatures thus appear: 87.5 °C (T_{ref}) normalizes P and defines f_L and ΔT_{jc} ; 80 °C defines kW_{ref}; and 85 °C, the present chip limit, is where s and the leakage share are quoted.

How steeply leakage grows. Evidence on b comes from vendor power-estimation tools and transistor measurements (Fig. 2a). The TI AM62x and AM62D tools give leakage–temperature curves for several device libraries and process corners in TSMC N16FFC up to 125–150 °C [26, 27], and the AMD UltraScale+ tool gives the static current of the core supply of 16 nm FinFET parts up to 125 °C [28]. High-temperature measurements on FinFET, stacked-nanowire and nanosheet transistors [29–31] are converted into the temperature dependence of off-state current from the measured threshold-voltage temperature coefficient and subthreshold swing. All curves were fitted to Eq. (3) over 50–150 °C (50–125 °C for the AMD tool) with equal weight per source group, giving 10th, 50th and 90th percentiles of 2584, 3315 and 4515 K, for which

leakage doubles every 27, 22 and 17 K near 85 °C, or apparent activation energies of 0.28–0.45 eV at 85 °C; these are called the flat, median and steep shapes below. The vendor tools alone give 3042–3963 K, while the transistor data spread more widely because their threshold voltage is a parameter. The only GPU fits available, from the V100 telemetry below, give about 2600–3000 K, inside the evidence range but too uncertain (95% profile interval 1925–3750 K) to select a shape. Details are given in Supplementary Section S2.

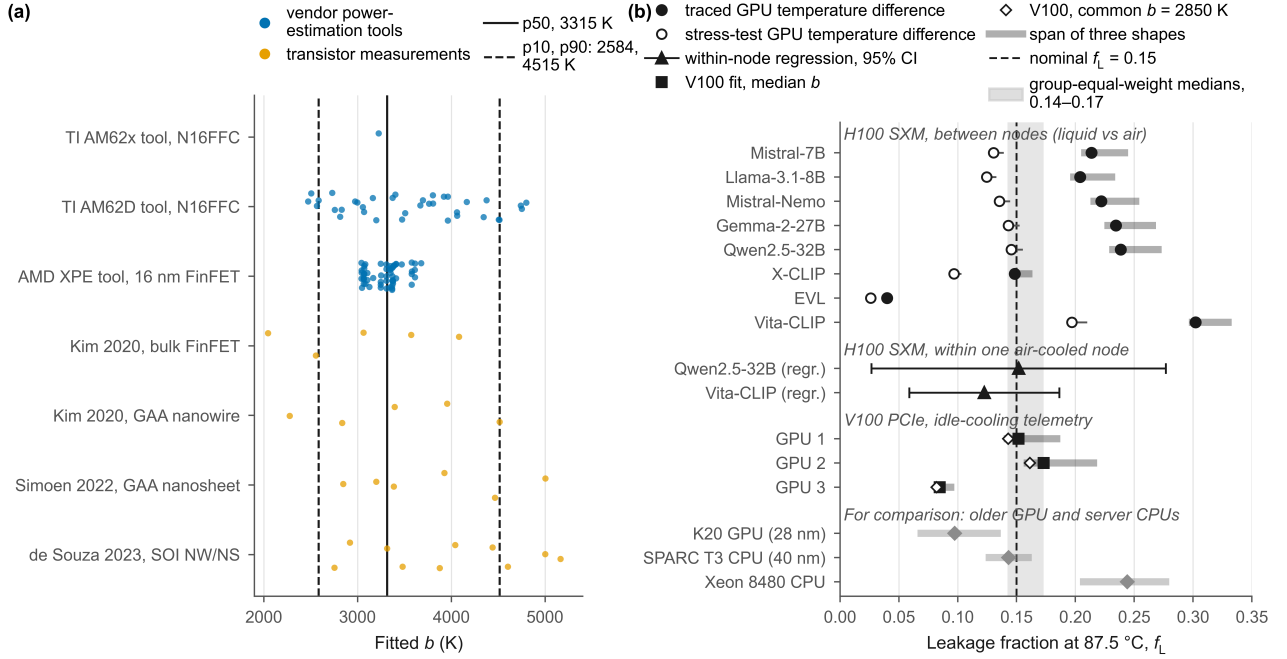


Fig. 2. Leakage evidence. (a) Leakage activation temperature b from fits of leakage–temperature curves to Eq. (3) over 50–150 °C (50–125 °C for the AMD tool), grouped by source (vendor power-estimation tools [26–28]; transistor measurements [29–31]); lines mark the group-weighted 10th, 50th and 90th percentiles. (b) Leakage fraction f_L estimated for H100 [21] from node-to-node differences (filled: node-to-node GPU temperature differences traced from the published curves; open: those from the reported stress-test ranges) and from within-node regressions, and for three V100 GPUs [32, 33]; older-GPU [17] and server-CPU [16, 39] values are shown for comparison. Bars span the three leakage shapes; the shaded band marks the group-weighted medians and the dashed line the nominal value.

How much leakage present GPUs have. Leakage is estimated from data-center GPUs in three ways (Fig. 2b). For H100, liquid-cooled and air-cooled nodes running the same workloads differ in GPU power and temperature [21]; with a 700 W full-load denominator, the eight workloads give node-to-node estimates, and regressions across the eight GPUs within the air-cooled node give two more. For V100, public data-center telemetry [32, 33] contains periods in which an idle GPU cooled down at a fixed clock after a job; fitting power against temperature for the three GPUs that met a predefined selection rule gives estimates with their own b . Weighting the H100 and V100 groups equally, the median f_L is 0.14–0.17, depending on how the H100 temperature differences are taken; the within-node H100 regressions give 0.12 and 0.15, and a common fit to the V100 GPUs gives 0.13 at $b = 2850$ K. The nominal f_L is 0.15, and results are reported over 0–0.35. By Eq. (4), leakage of this size gives present GPUs $s = 0.42$ – 0.51% K^{−1} at 85 °C (group medians; the interquartile ranges of the individual estimates cover 0.33–0.68% K^{−1}). The sample is small and its confounders act in both directions (Supplementary Section S3).

2.4. Radiator and power closure

The radiator planform area is obtained by integrating along the fluid temperature drop,

$$A = \frac{Q}{T_{\text{out}} - T_{\text{in}}} \int_{T_{\text{in}}}^{T_{\text{out}}} \frac{dT}{q(T)}, \quad q(T) = 2\varepsilon F \eta_{\text{fin}} \sigma T^4 - q_{\text{env}}, \quad (5)$$

where Q is the rejected heat, the coolant enters the radiator at the loop outlet temperature T_{out} and returns at the loop inlet temperature T_{in} , and σ is the Stefan–Boltzmann constant. The environmental load q_{env} is computed for a dawn–dusk Sun-synchronous orbit at 650 km with the radiator plane containing the Sun and nadir vectors; it comprises Earth infrared emission (107.9 W m^{-2}), albedo (15.3 W m^{-2}) and direct sunlight equal to 2% of the solar constant on one face, representing pointing error (4.1 W m^{-2}), 127.3 W m^{-2} over both faces. Exchange with the Earth enters only through q_{env} ; F is the part of each face’s view not blocked by the spacecraft’s own arrays and panels. The bus power, including chips, pump and other loads, ends up as heat in the radiator, whereas PMAD losses are rejected by the power system; radiator rejection, solar-array output and chip power are closed iteratively. For comparison with the literature, the effective radiating temperature T_{eff} is defined as the temperature at which an isothermal surface with the same ε , F and q_{env} but $\eta_{\text{fin}} = 1$ achieves the same net flux.

2.5. Radiator share of the power-dependent mass

Running hot trades a smaller radiator against more leakage power, so its value depends on how much of the mass that scales with power is radiator. Only the radiator, solar array and storage scale with bus power; the radiator mass share ϕ_{rad} and the radiator share ϕ_{eff} are defined at the present optimum as

$$\phi_{\text{rad}} = \frac{\sigma_{\text{rad}} A}{\sigma_{\text{rad}} A + P_{\text{tot}} / (S P_{\text{pv}} \eta_{\text{PMAD}}) + m_{\text{bat}} P_{\text{tot}}}, \quad \phi_{\text{eff}} = \phi_{\text{rad}} \frac{|\partial \ln A / \partial T_{\text{out}}|}{0.01554 \text{ K}^{-1}}. \quad (6)$$

For the nominal radiator ϕ_{rad} is 0.570, whereas the radiator is 40% of the on-orbit mass (Fig. 3b), because compute hardware and fixed mass do not scale with power. The radiator share ϕ_{eff} weights ϕ_{rad} by how strongly the radiator area responds to the coolant outlet temperature at a fixed loop temperature rise; the normalizing value of 0.01554 K^{-1} is that of the nominal radiator, which shrinks by 1.55% per kelvin of coolant outlet temperature. ϕ_{eff} can exceed unity. A heavier radiator, a lower fin efficiency, a larger environmental load, a lighter solar array or less storage all raise ϕ_{eff} . Three reference radiators span the plausible range (Table 2): the advanced radiator at 5 kg m^{-2} , the nominal radiator at 10 kg m^{-2} , and the heavy radiator with the 17.5 kg m^{-2} areal mass of the ISS photovoltaic radiator and the fin efficiency of 0.819 at which the model reproduces that radiator’s rated flux (Section 3).

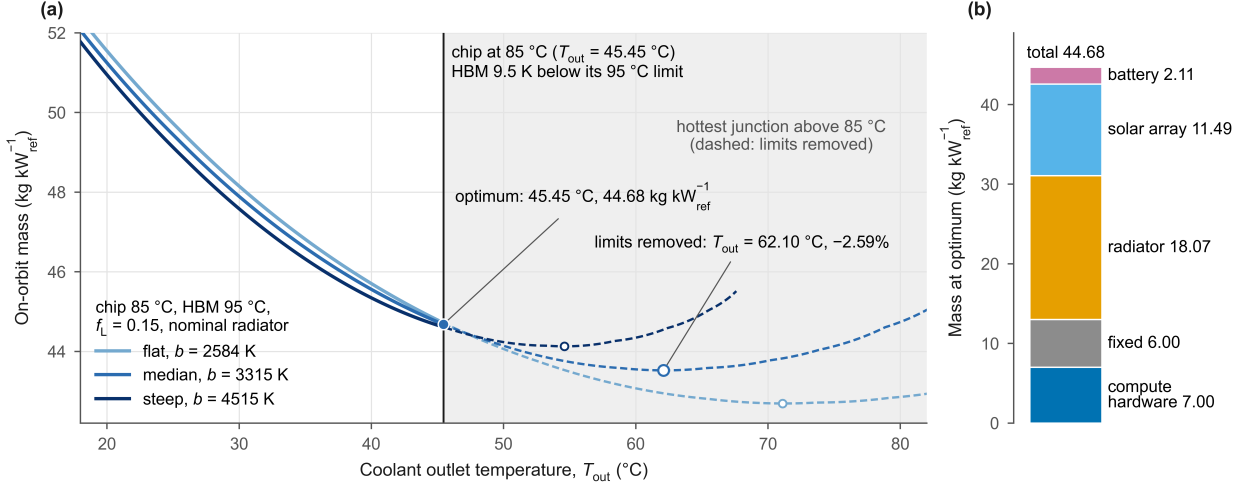


Fig. 3. Design point for present chips. (a) On-orbit mass versus coolant outlet temperature (chip limit 85 °C, HBM limit 95 °C, $f_L = 0.15$, nominal radiator, three leakage shapes); the vertical line marks the coolant outlet temperature (45.45 °C) at which the hottest junction reaches the 85 °C chip limit, and dashed curves show the continuation without limits. (b) Mass breakdown at the optimum.

Table 2. Reference radiators ($f_L = 0.15$, median shape, present limits).

Case	σ_{rad} (kg m ⁻²)	η_{fin}	ϕ_{rad}	ϕ_{eff}	M (kg kW ⁻¹ _{ref})
Advanced	5	0.85	0.403	0.406	35.60
Nominal	10	0.85	0.570	0.570	44.68
Heavy	17.5	0.819	0.709	0.715	59.70

2.6. Literature radiating temperatures and stability

A radiating temperature quoted in the literature is converted into chip requirements in two ways. The perfect-radiator mapping sets $\varepsilon = F = \eta_{fin} = 1$, neglects environmental load and assumes an isothermal radiator; because the radiator cannot be hotter than the coolant, this gives a lower bound on the coolant and junction temperatures that is independent of radiator parameters. The closed mapping finds, with the full model and no leakage, the coolant outlet temperature at which the mass-optimal design has an effective radiating temperature equal to the literature value.

Leakage can also make a design unstable. At chip level, the fixed point of Eq. (2) disappears when the loop gain $\Delta T_{jc} dP/dT_j$ reaches unity [15]. In terms of Eq. (4) this gain is $\Delta T_{jc} Ps$; for present GPUs at 85 °C it is about 0.17, and it approaches unity as both P and s grow with temperature. The classic criterion assumes a fixed ambient [15]. In orbit the coolant temperature itself rises with heat load through the thermal resistance of a radiator of fixed size, which adds a system-level term to the loop gain; the critical b , the largest b that still gives a stable operating point, is evaluated at chip and system level with the same loop integration as the mass model. Coolant-temperature control, such as a radiator bypass, can remove the system-level term only by holding radiator margin, whereas no loop control removes the chip-level criterion. A real chip that violates either criterion would throttle rather than run away, so instability here means that the design cannot hold a fixed nominal clock.

2.7. Heat pumps, clock scaling and lifetime

In heat-pump designs the chip loop feeds an evaporator and the condenser is an isothermal radiator at T_{cond} . The coefficient of performance (COP) is computed for a single-stage vapor-compression cycle with CoolProp properties [34, 35], using at each operating point the best of a set of practical working fluids, with component efficiencies matched to the space heat-pump prototype of van Gerner et al. [13]; results are reported with a constant fin efficiency and with one that declines with temperature at fixed fin geometry (Supplementary Section S5). Clock scaling is analyzed as an additional design variable: dynamic power falls with the square of the supply voltage, which follows the measured voltage–frequency curve of A100 down to its floor [36], the voltage dependence of leakage is taken from the same vendor tool as its temperature dependence [28], and the number of chips grows with the inverse clock ratio (Supplementary Section S6). Lifetime is examined with two replacement models (Supplementary Section S7).

3. Model verification

The model is checked in two ways (Table 3). Consistency checks confirm that it reproduces published calculations from the same inputs. With Turyshev’s inputs [3], it reproduces his emitted flux, megawatt-class radiator area, specific mass and area–temperature elasticity to within 0.2%; with 10 kg m^{−2} per planform area it gives his 12.5 kg kW^{−1}, which confirms the areal-mass convention. Both emission values of the Starcloud white paper [1] are also reproduced; its environmental-load expression uses the solar absorptance for the Earth infrared term, and the corrected value is used here (Supplementary Section S4).

Table 3. Consistency checks and plausibility anchors.

Check	This model	Source
<i>Consistency</i>		
Turyshev emitted flux at 350 K, W m ^{−2} [3]	650.9	651.0
Turyshev radiator specific mass, kg kW ^{−1}	12.48	12.5
Turyshev area–temperature elasticity	−5.20	−5.2
Starcloud two-sided blackbody emission, W m ^{−2} [1]	837.5	838
<i>Plausibility anchors</i>		
NASA Glenn Research Center (GRC) radiator heat rejection, kW [37]: $\eta_{\text{fin}} = 0.85 / 0.61$	8.79 / 6.31	6.28
ISS photovoltaic radiator flux, W m ^{−2} [20]: nominal / heavy radiator parameters	347 / 330	330
Space heat-pump prototype COP (model convention) [13]	2.53	2.60

The hardware anchors constrain the radiator parameters; no orbital measurement exists for the leakage coupling itself. The full-scale radiator test at GRC [37] is overpredicted by 40% with the nominal fin efficiency of 0.85 and matched with 0.61 at its test temperature of about 400 K; for the same geometry this corresponds to about 0.75 at the mean coolant temperature of the present designs, about 313 K (40 °C). The ISS photovoltaic radiator [20] is rated at about 330 W per square meter of planform. Assuming an ammonia loop cooling from 291 to 276 K and the present orbital environment, the nominal radiator parameters give 347 W m^{−2}, and a fin efficiency of 0.819 reproduces the rating; this value is an effective factor that also absorbs differences in orbit, coating and rating conditions (0.897 with a 34% higher environmental load). The

nominal fin efficiency is therefore on the optimistic side, and its effect is examined over 0.61–0.85 (Section 4.4). At the prototype operating point, the heat-pump cycle with the measured overall efficiency and evaporator pressure drop gives a COP 3% below the measurement.

4. Results and discussion

4.1. Present chips: a design point pinned by the temperature limits

For present data-center GPUs, the mass optimum lies where the hottest chip reaches its 85 °C limit (Fig. 3a; $f_L = 0.15$, median shape, nominal radiator, HBM limit 95 °C). The coolant outlet temperature is 45.45 °C, the HBM runs at 85.45 °C, the effective radiating temperature is 301.0 K, and the on-orbit mass is 44.68 kg kW_{ref}⁻¹. The leakage share of the hottest chip is already 14.0%. The optimum is a corner solution: a hotter coolant would shrink the radiator, but the hottest junction has reached its 85 °C limit. This holds for all three reference radiators and leakage shapes, and in 37 of the 40 one-at-a-time sensitivity cases that affect the passive design. The chip limit rather than the HBM limit binds because at 85 °C the hottest chip dissipates slightly less than its power at T_{ref} ($P < 1$) and therefore sits about 0.45 K below the HBM when both differences are 40 K; a larger HBM-to-coolant difference, a smaller junction-to-coolant difference or an HBM limit of 85 °C moves the binding to the HBM (Supplementary Section S9). On the feasible side the mass curve is flat: a coolant outlet temperature 10 K below the optimum adds 4.1% mass.

The corner suggests that the temperature limits hold the design back. Whether raising them pays off depends on how much more the chips leak when they run hotter.

4.2. What the literature radiating temperatures demand of the chips

Fig. 4a shows what radiating temperatures of 350 and 385 K require of the chips. In the perfect-radiator mapping, 350 K requires coolant of at least 76.9 °C everywhere in the loop and, with $\Delta T_{jc} = 40$ K, a junction temperature of at least 116.9 °C; 385 K requires a junction temperature of at least 151.9 °C. In the closed mapping, which includes emissivity, view factor, fin efficiency, environmental load and loop temperature rise, the required coolant outlet temperatures are 96.4 and 133.9 °C and the required junction temperatures 136.4 and 173.9 °C. Radiating at 350–385 K therefore requires chip ratings of about 117–174 °C, and HBM ratings about as high, comparable to or above the 125 °C junction rating of automotive-grade processors [25] and 32–91 K above present throttling onsets.

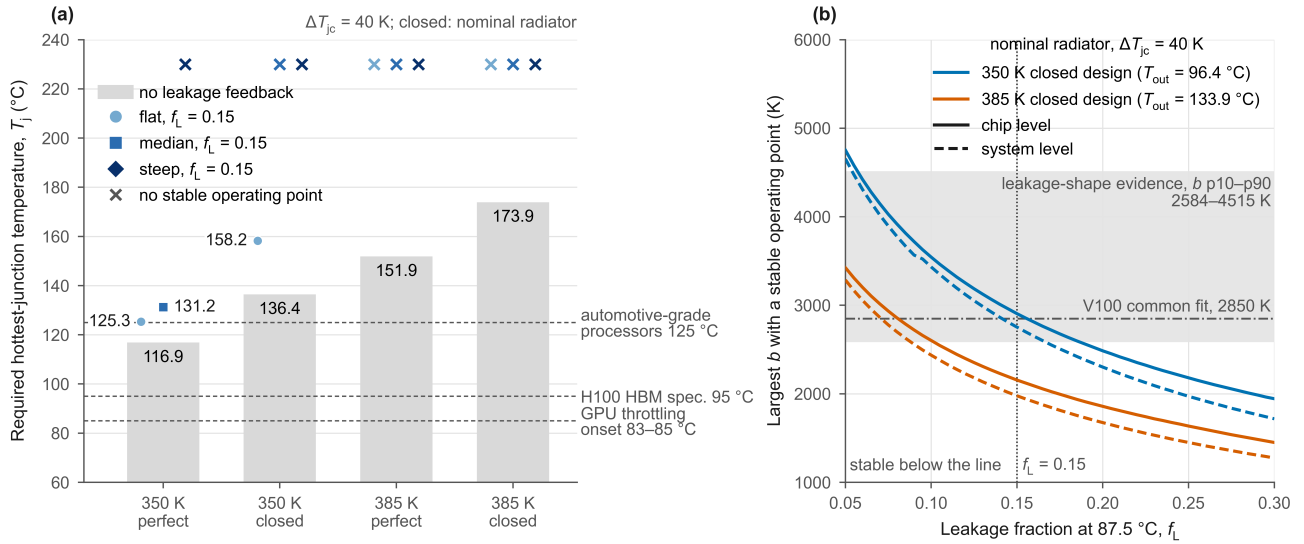


Fig. 4. What the literature radiating temperatures demand. (a) Junction temperature of the hottest chip required for literature radiating temperatures of 350 and 385 K ($\Delta T_{jc} = 40$ K): bars without leakage, markers for $f_L = 0.15$ with the three leakage shapes, crosses where leakage leaves no stable operating point at nominal clock; reference lines mark present throttling onsets, the H100 HBM specification and the 125 °C rating of automotive-grade processors. (b) Critical b (largest b with a stable operating point) of the closed 350 and 385 K designs versus f_L , at chip level (solid) and system level (dashed); legend values are the coolant outlet temperatures of the two designs; the band marks the evidence range of b and the line the V100 fit.

Higher ratings alone are not enough, because leakage at these temperatures destabilizes the loop. At $f_L = 0.15$ and $\Delta T_{jc} = 40$ K, the closed design for an effective radiating temperature of 350 K needs a junction temperature of 158.2 °C with the flat leakage shape and has no stable operating point at nominal clock with the median or steep shape. In the perfect-radiator mapping, which gives the lowest possible junction temperatures, 350 K is stable for the flat and median shapes (junction temperatures of 125.3 and 131.2 °C) and 385 K for none. The closed 350 K design is stable only if b is below 2753 K at system level and 2905 K at chip level, so the median shape is unstable even at chip level, which no coolant control can relax. The 2850 K fitted to the V100 telemetry lies between the two critical values: a chip with this shape would hold its fixed point on a coolant of fixed temperature but not on a radiator of fixed size (Fig. 4b). For 385 K the critical b is 1978 K, below every fitted value. A smaller temperature stack restores stability: a 350 K design with the median shape needs ΔT_{jc} below 31.9 K at system level, and a 385 K design below 11.4 K, the same lever that Section 4.5 shows to save mass at present ratings. Without such a stack, the literature radiating temperatures imply permanently reduced clocks.

4.3. How much leakage takes away

In Fig. 5 the chip limit is raised from 85 °C and the HBM limit from 95 °C, both to 150 °C, with $T_{hbm, cap} = \max(T_{cap}, 95 \text{ °C})$; this case is called “raising the limits to 150 °C” below. For a chip without leakage, it would reduce on-orbit mass by 23.5% with the nominal radiator. With present leakage ($f_L = 0.15$), the saving falls to 2.59% (1.05–4.55% across leakage shapes), or $1.16 \text{ kg kW}_{ref}^{-1}$. Leakage thus cancels about nine-tenths of the saving, and the same holds for every reference radiator: it removes 96% of the saving with the advanced radiator and 83% with the heavy one (Table 4). Across the three leakage shapes leakage cancels 81–96% of the saving with the nominal radiator and 72–99% over all three reference radiators, so the conclusion does not rest on the median shape.

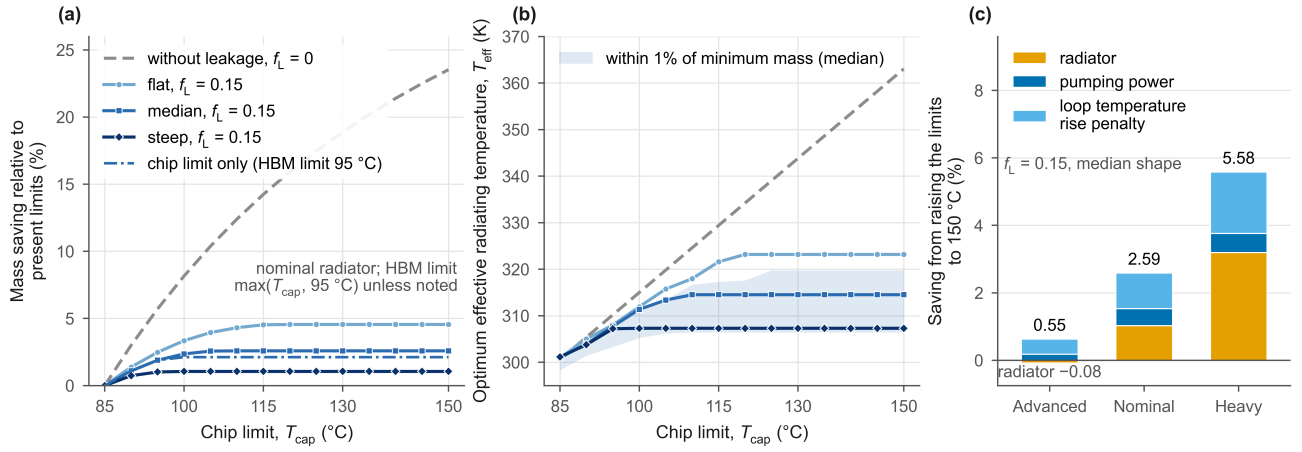


Fig. 5. What leakage takes away (nominal radiator). (a) Mass saving relative to the present limits and (b) effective radiating temperature at the optimum versus chip limit, with the HBM limit at $\max(T_{cap}, 95\text{ °C})$: dashed lines for a chip without leakage, solid lines for $f_L = 0.15$ with three leakage shapes, dash-dotted line for the chip limit raised alone with the HBM limit held at 95 °C; the band in (b) marks designs within 1% of the minimum mass. (c) Saving from raising the limits to 150 °C split into three parts for the three reference radiators: radiator, the difference of the pump-free optimum masses; pumping power and loop temperature rise penalty, the changes in the mass added by pumping power and by the pump-limited loop temperature rise.

The mechanism is visible along the scan. As the limits rise, the leakage share of the hottest chip grows from 14.0% to 23.7%. The extra leakage power must be supplied by a larger solar array and rejected by the radiator, which cancels most of the radiator shrinkage. The saving therefore saturates early: 90% of it is reached at a chip limit of about 100 °C, and from a chip limit of 110 °C the optimum leaves the limits and stays at an effective radiating temperature of 314.5 K, far below the literature values, with designs within 1% of the minimum mass spanning 306–320 K. At this optimum the junction temperature is about 107 °C, so the 2.6% saving depends on the leakage shape only up to that temperature, inside the range of the vendor tools; extrapolation beyond it enters only the stability criteria of Section 4.2.

Leakage removes almost all of the radiator shrinkage itself. Of the 2.59% that remains, only 1.03 percentage points come from a smaller radiator, and with the advanced radiator the radiator even grows slightly (Fig. 5c). The other 1.56 percentage points come from the pumped single-phase loop: at the present limits, pumping power holds the loop temperature rise at 10 K, whereas the higher limits allow a 15 K rise, which cuts pumping power from 10.6 to 3.4 W kW_{ref}^{-1} (0.51 percentage points) and reduces the mass penalty of the loop temperature rise (1.05 percentage points). A two-phase loop has neither penalty, and there the saving is only 1.03%; with cold plates in parallel it is 1.08%.

The same holds for the HBM and for the chip separately. Raising the chip limit alone, with the HBM limit held at its 95 °C specification, saves 2.12% before the HBM binds at a coolant outlet temperature of 55 °C. Holding the HBM limit at 105 °C changes nothing, because leakage stops the optimum at an HBM temperature of about 102 °C. The leakage fraction itself matters strongly: with the nominal radiator the saving is 6.07% at $f_L = 0.10$ and 0.38% at 0.22.

4.4. How low leakage must be

How much leakage can be tolerated depends on how heavy the radiator is. Raising the coolant temperatures by 1 K at a fixed loop temperature rise shrinks the nominal radiator by about 1.55%, which saves

$\phi_{\text{eff}} \times 1.55\%$ of the power-dependent mass; the same kelvin raises chip power by s , and every extra watt needs its share of solar array, storage and radiator. To first order, higher ratings therefore pay off while s is below about $(1.55 \phi_{\text{eff}})\% \text{ K}^{-1}$, so a heavier radiator tolerates more leakage. The full model gives a smaller break-even value, because the saving is integrated over the whole rise of the limits and leakage steepens as the chip warms. With the nominal radiator, higher ratings stop saving mass at $s_0 = 0.80\% \text{ K}^{-1}$ (0.76–0.83% K^{-1} across shapes), a leakage share of about a quarter of chip power at 85 °C with the median shape; the corresponding leakage fraction ranges from 0.22 to 0.31 with the shape. Across 126 radiator, power-system and environment settings, the break-even value follows

$$s_0 \approx (0.964 \phi_{\text{eff}} + 0.220)\% \text{ K}^{-1} \quad (7)$$

for the median shape, and the 5% threshold follows $s_5 \approx (0.803 \phi_{\text{eff}} - 0.140)\% \text{ K}^{-1}$ (Fig. 6). Written in s , the break-even values of the three leakage shapes fall on one line ($R^2 = 0.965$ pooled, 0.976 for the median shape); written in f_L , they scatter ($R^2 = 0.821$). The leakage that limits higher ratings is therefore the leakage-driven power growth, not the leakage fraction alone.

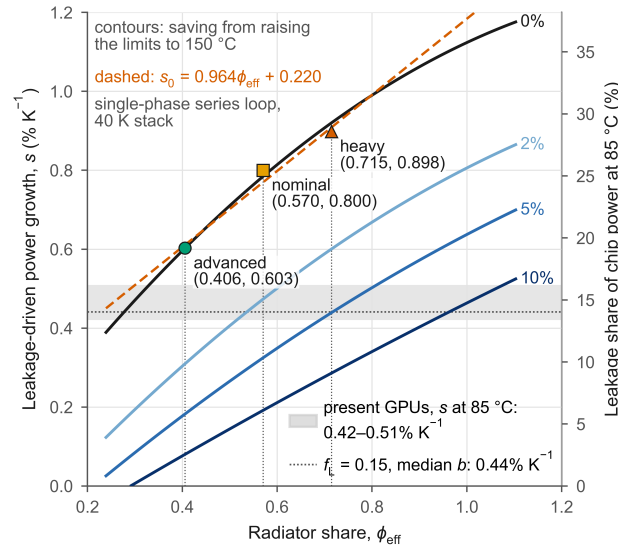


Fig. 6. How low leakage must be. Contours of the mass saving from raising the limits to 150 °C over the radiator share ϕ_{eff} and the leakage-driven power growth s (median shape; fits to the values found by root finding for 126 settings); the dashed line is Eq. (7), markers show the three reference radiators, and the shaded band marks present GPUs (s at 85 °C). The right axis gives the corresponding leakage share of chip power at 85 °C for the median shape. Valid for the single-phase series loop with a 40 K stack.

Table 4. What leakage takes away and the alternative of a smaller temperature stack (median shape; ranges across the three shapes in brackets; $f_L = 0.15$ unless stated). For the heavy radiator the median and flat shapes give the same s_0 .

Case	Raising the limits to 150 °C, saving at f_L = 0 / 0.10 / 0.15 / 0.22 (%)	s_0 (% K ⁻¹)	s_5 (% K ⁻¹)	Both differences 40 → 25 K, saving (%)	f_L above which the stack lever is larger
Advanced	14.85 / 2.18 / 0.55 / 0.00	0.603 (0.577–0.647)	0.173 (0.151–0.188)	5.04	0.058 (0.040–0.076)
Nominal	23.53 / 6.07 / 2.59 / 0.38	0.800 (0.763–0.825)	0.329 (0.301–0.349)	7.97	0.081 (0.055–0.107)
Heavy	32.23 / 11.09 / 5.58 / 1.53	0.898 (0.898–0.958)	0.462 (0.432–0.484)	10.97	0.100 (0.068–0.133)

Present GPUs, at $s = 0.42\text{--}0.51\%$ K⁻¹, lie below the break-even value of every reference radiator, so higher ratings do save some mass. They lie well above the 5% threshold with the advanced and nominal radiators, and near it with the heavy one. With the nominal radiator, the 5% threshold is 0.33% K⁻¹, a leakage share of about 10% at 85 °C against 14% today; without a design margin, the leakage share would have to fall by a quarter to a third, and by more than half with the advanced radiator.

Leakage can be lowered at the transistor level by raising threshold voltages or, in nanosheet transistors, through a steeper subthreshold slope, but both also steepen its temperature dependence (larger b). A leakage fraction lowered this way must therefore meet a lower value, whereas the 5% threshold in s changes little with the shape (Table 4), which is why s is the quantity to specify. With the nominal radiator, meeting the 5% threshold requires raising the threshold voltage of the leaking transistors by about 12 mV, and with the advanced radiator by about 37 mV (Supplementary Section S3.6); the cost is speed, which must be recovered through supply voltage or clock (Section 4.6). Gating blocks that are idle at full load lowers leakage without changing its shape, but at full load its reach is limited.

Three choices in the nominal design work against higher ratings. A 10 K design margin on both limits raises the saving with the nominal radiator to 5.71%, an end-of-life environmental load 34% higher to 3.49%, and a fin efficiency of 0.61, as in the GRC test, to 5.47%. Each raises the radiator share ($\phi_{\text{eff}} = 0.636$ with the margin) and moves the break-even value up with it ($s_0 = 0.88\%$ K⁻¹), as Eq. (7) anticipates. Conversely, a fin efficiency that falls with temperature at fixed fin geometry, which the constant value ignores, lowers the saving with the nominal radiator to 1.79% (heavy radiator 3.83%). The break-even relation holds within the sampled range of 0.24–1.10 in ϕ_{eff} and for the single-phase, series-loop design with a 40 K stack. With a two-phase loop, parallel cold plates or a 15 K stack, the break-even value shifts by –21% to +18%, and outside the sampled range the linear relation overestimates it by up to about 40% (Supplementary Section S9).

4.5. Bypassing leakage: a smaller temperature stack

A higher rating makes the coolant hotter by making the chips hotter, and leakage takes most of the saving. A smaller temperature stack makes the coolant hotter while the chips stay at the same temperature, so leakage does not grow. Reducing both ΔT_{jc} and ΔT_{hbm} from 40 to 25 K raises the coolant outlet temperature by 15 K and reduces mass by 7.97% with the nominal radiator, 5.04% with the advanced radiator and 10.97% with the heavy one (Table 4). Because the chips do not get hotter, the saving hardly depends on leakage: across $f_L = 0.10\text{--}0.22$ and all three shapes, the saving with the nominal radiator varies only between 7.70% and 8.06%, whereas the saving from higher ratings collapses as leakage grows (Fig. 7). Within the model this lever is the

same as raising both limits by 15 K for a chip whose leakage does not grow. Each kelvin is worth 0.24–0.28 $\text{kg kW}_{\text{ref}}^{-1}$ with the nominal radiator, which sets the mass allowance for better cold plates and HBM cooling; the cold-plate mass and pressure drop needed to achieve it are not modeled.

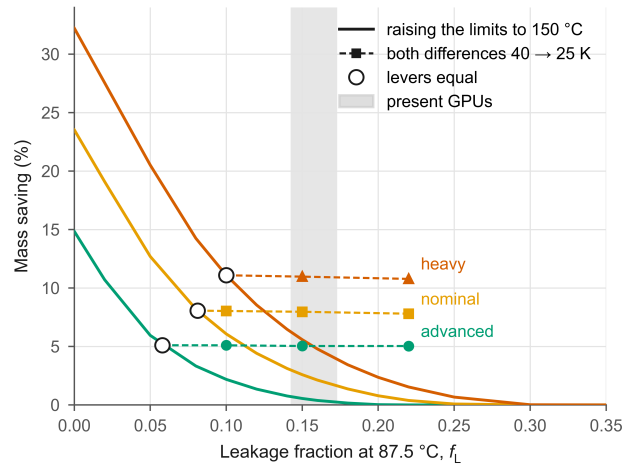


Fig. 7. Bypassing leakage. Mass saving from raising the limits to 150 °C (solid lines) and from reducing both chip- and HBM-to-coolant temperature differences from 40 to 25 K (dashed lines with markers, computed at the crossing and at $f_L = 0.10$, 0.15 and 0.22) versus the leakage fraction, for the three reference radiators (median shape); open circles mark where the two levers are equal, and the shaded band marks present GPUs.

Most of this lever is on the chip side, because the chip limit binds at present. Reducing ΔT_{jc} alone to 25 K saves 6.71%, and holding ΔT_{hbm} at 30 or 35 K costs almost nothing (7.97% and 7.91%), because the HBM keeps about 9.5 K of margin to its 95 °C specification. If the cold plate itself contributes only 5–7 K, as estimated for distributed vapor chambers on an orbital compute panel [5], most of the 15 K must come from the die, the thermal interface and the lid, so the stack is a package specification, much like leakage. The mass allowance for it is generous: 15 K is worth about 3.6 $\text{kg kW}_{\text{ref}}^{-1}$ with the nominal radiator, whereas doubling the pumping power of the present design to drive a higher-performance cold plate would cost about 0.3 $\text{kg kW}_{\text{ref}}^{-1}$.

At present leakage, bypassing leakage is worth more than higher ratings (Fig. 7). The stack lever is the larger one whenever f_L exceeds 0.055–0.107 with the nominal radiator and 0.068–0.133 with the heavy one, so it is larger for all three reference radiators at $f_L = 0.15$; at a lower leakage fraction of 0.10 with the flat shape, higher ratings would be the larger lever. The ordering holds in all four cold-plate arrangements examined, where the stack lever gives 6.57–8.08% and higher ratings 0.51–2.59% with the nominal radiator. The two levers are partly substitutes: after the stack is reduced to 25 K, raising the limits to 150 °C adds 1.75%.

4.6. Clock scaling and heat pumps: other routes around leakage

Lowering the clock reduces chip power, leakage included, but trades thermal and power-system mass for compute hardware, and it leaves the leakage result unchanged. With the clock as a design variable, the junction temperature is about 72 °C and the HBM limit binds; raising the limits to 150 °C then saves 1.71%, and the stack lever still gives 3.55%; only with the heavy radiator and a lower leakage fraction of 0.10 do higher ratings overtake the stack lever (8.08% against 6.64%), as Eq. (7) anticipates. Clock scaling itself is a large lever. The optimum lowers the clock to the knee where the supply voltage stops falling (0.72 of nominal on the A100 curve), which reduces mass by 25.1% with the nominal radiator, to 33.45 $\text{kg kW}_{\text{ref}}^{-1}$

(20.9–28.6% across model variants), but requires about 39% more chips; it stops paying off in cost once the chip price per kW_{ref} exceeds about 63 times the launch price per kilogram (Supplementary Section S6).

A heat pump is another way to make the radiator hotter without heating the chips, but it pays for the bypass with compressor power, which, like leakage, must be supplied by the array and rejected by the radiator. With the nominal radiator it does not pay: forcing the effective radiating temperature to 350 K, a condensing temperature of about 365 K with the nominal fin efficiency, adds 4.7% mass (8.9% with a fin efficiency that declines with temperature), 385 K adds 35.7% at a pressure ratio of 10 or cannot be reached once the fin efficiency declines, and even the best heat-pump design is 3.4–5.0% heavier than the passive one. Only with the heavy radiator does it save mass, 4.9%, or 1.0% with the declining fin efficiency. For the compressor itself these penalties are conservative, because its nominal specific mass of 4.6 kg kW_e^{-1} is 4.6 times that of the prototype [13].

4.7. Implications for mass estimates and chip specifications

Hot-radiator estimates [3, 7] carry an unstated set of chip assumptions: chip and HBM ratings of about 117–174 °C, leakage low enough for the saving to survive and for the loop to remain stable, and no lifetime loss. Published concepts can be placed on the same map. The mean radiating temperature of 293 K (20 °C) in the Starcloud white paper [1] corresponds to a coolant outlet temperature of 36.2 °C and, at $f_L = 0.15$, a junction temperature of 74.1 °C, compatible with present chips. The reported figures for SpaceX’s AI1, an average compute payload of 120 kW and up to 110 m² of deployable liquid radiators [38], imply, if the 110 m² is planform area radiating from both faces and the 120 kW is the rejected heat, a coolant outlet temperature of 90.4 °C and, at $f_L = 0.15$, a stable operating point only for the flat leakage shape (junction temperature 146.8 °C) (Supplementary Section S8).

Lifetime is a separate condition whose weight depends on how failures are replaced; its activation energies of 0.5–0.9 eV describe wear-out mechanisms distinct from leakage. If chips must be replaced in batches as their thermally activated lifetime shortens, raising the limits to 110 °C or more turns the 2.6% saving into a mass increase of 15–46% for the median leakage shape; if failures are covered by spares at a constant base rate of 2% per year, the net saving lies between –2.3% and +0.7% (Supplementary Section S7). Qualification at the new limit is therefore part of what a higher rating means.

For chips intended for orbital AI data centers, leakage, stated as the leakage-driven power growth s at 85 °C, should rank with the junction temperature rating as a first-order specification: s is about 0.44% K^{-1} today, or 3 W K^{-1} for a 700 W chip, against a 5% threshold of about 0.33% K^{-1} (2.3 W K^{-1}) with the nominal radiator. HBM ratings must rise with the chip’s. For satellite designers, the radiating temperature should follow from these specifications and the radiator share ϕ_{eff} rather than be set a priori, and for a given chip and clock the temperature stack between chips, HBM and coolant is the first quantity to reduce.

4.8. Limitations

The evidence on leakage is indirect and the sample is small. The H100 node-to-node estimates rise with workload power (correlation 0.99 across the eight workloads), whereas a purely thermal leakage effect, once divided by the temperature difference, would give the same fraction for every workload; this points to a load-proportional difference between the two nodes. The within-node regressions, which are free of node-to-node

differences, give 0.12 and 0.15, and the V100 telemetry covers three GPUs. The measurements are made at 40–60 °C and extrapolated with the leakage shape to T_{ref} for f_L and to 85 °C for s , which changes them by a factor of up to about 1.8 between the flat and steep shapes. If the supply voltage is lowered as temperature rises at a fixed clock, as on a server central processing unit (CPU) [39], total power understates leakage, there by a factor of 2.5–3.5; for the GPUs the voltages were not recorded, so the sign of this bias is unknown. The ordering of the two levers holds unless f_L is below 0.055–0.107 with the nominal radiator (Table 4). Measurements on 4–5 nm-class GPUs with a fixed clock, controlled coolant temperature and recorded voltage would narrow this input; Eq. (4) defines the quantity to measure.

The shape evidence comes from 16 nm FinFET vendor tools and transistor measurements, whereas present GPUs use 4–5 nm-class processes whose threshold-voltage mix and gate-leakage share are not published. The HBM part of f_L is scaled with the junction temperature and the leakage shape of logic transistors. Timing margins at high temperature and the refresh power of hot HBM are not modeled; both would reduce the saving from higher ratings further. The HBM-to-coolant difference has no first-hand measurement under liquid cooling. The break-even relation and the absolute savings depend on the loop design, which here is a single-phase water loop with chips in series; other arrangements give smaller savings (Section 4.5).

The model assumes steady full-load operation. It does not include transients, survival heating during launch, early operations and eclipse shutdowns, power-cycling fatigue of the package, loop redundancy, or coating degradation beyond the end-of-life load of Section 4.4. Items that add mass in proportion to radiator area, such as deployment hardware, manifolds and micrometeoroid shielding, raise σ_{rad} and move a design along the radiator-share axis of Fig. 6; items that add temperature differences, such as an intermediate heat exchanger used to keep an external water loop from freezing, lengthen the stack and cost about 0.25 kg $\text{kW}_{\text{ref}}^{-1}$ per kelvin. Cold-plate mass and compressor electronics are not modeled, and the two-phase variant ignores microgravity effects on flow boiling.

5. Conclusions

The orbital AI data center literature places radiators at 350–385 K to shrink them through the fourth-power law. Treating the chip and HBM limits as design variables and including leakage in a closed mass model shows that leakage matters as much as temperature tolerance. Three conclusions follow.

1. The literature radiating temperatures require chip and HBM ratings of about 117–174 °C, and at present leakage a fixed-clock design for an effective radiating temperature of 350 K is stable only for the flat leakage shape unless the junction-to-coolant difference falls below about 32 K.
2. Present leakage cancels about nine-tenths of the saving from higher ratings. Raising the chip and HBM limits from 85 and 95 °C to 150 °C would save 23.5% of on-orbit mass for a chip without leakage but saves 2.6% at present leakage with the nominal radiator, and 0.6–5.6% across the three reference radiators. Higher ratings pay off only if the leakage-driven power growth s is below a break-even value, about $(0.96 \phi_{\text{eff}} + 0.22)\% \text{ K}^{-1}$, that rises with the radiator share ϕ_{eff} . Present GPUs, at 0.42–0.51% K^{-1} , lie below the break-even value of 0.80% K^{-1} for the nominal radiator but above the 5% threshold of 0.33% K^{-1} , so their leakage share would have to fall by a quarter to a third.

3. Leakage can be bypassed. Reducing both chip- and HBM-to-coolant temperature differences from 40 to 25 K makes the coolant hotter without heating the chips and saves 8.0% with the nominal radiator, nearly independent of leakage and more than higher ratings save at present leakage; each kelvin is worth about $0.25 \text{ kg kW}_{\text{ref}}^{-1}$. Lowering the clock saves more mass but needs about 39% more chips, and under it the stack lever still exceeds higher ratings.

Chips for orbital AI data centers therefore need low leakage as much as high temperature ratings, and for a given chip and clock, shrinking the temperature stack saves more mass than a higher rating.

CRedit authorship contribution statement

Jiawei Jing: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The model code, the scripts that extract and fit the leakage evidence, and the result files underlying all figures and tables are available at <https://doi.org/10.5281/zenodo.23576307> (version 1.0.0) [40]. The V100 telemetry is available from the MIT Supercloud Datacenter Challenge dataset [33] under its own license and is not redistributed; the repository contains the scripts that retrieve and process it.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used Claude (Anthropic) to assist with code development, literature retrieval, drafting and language editing. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- [1] E. Feilden, A. Oltean, P. Johnston, Why we should train AI in space, White Paper v1.03, Lumen Orbit (now Starcloud), September 2024. <https://starcloudinc.github.io/wp.pdf> (accessed 23 October 2026)
- [2] B. Agüera y Arcas, T. Beals, M. Biggs, J.V. Bloom, T. Fischbacher, K. Gromov, U. Köster, R. Pravahan, J. Manyika, Toward a future space-based, highly scalable AI infrastructure system design, Joule (2026) 102678. <https://doi.org/10.1016/j.joule.2026.102678>

- [3] S.G. Turyshev, Orbital data centers: spacecraft constraints and economic viability, arXiv:2604.27197 (2026). <https://arxiv.org/abs/2604.27197>
- [4] S.G. Turyshev, Operations, maintenance, and industrial scaling of MW-class orbital data centers, arXiv:2608.27499 (2026). <https://arxiv.org/abs/2608.27499>
- [5] S. Gaalema, S. Indyk, C. Staley, Reduced-mass orbital AI inference via integrated solar, compute, and radiator panels, arXiv:2604.07760 (2026). <https://arxiv.org/abs/2604.07760>
- [6] R. Ohs, G.F. Stock, A. Schmidt, J.A. Fraire, J. Ott, H. Hermanns, Dark clouds rising in low-Earth orbit: on environmental limits to massive orbital AI, in: Proc. ACM SIGCOMM 2026 Conf. (companion volume; 4th ACM Workshop on LEO Networking and Communication, LEO-NET 2026), Denver, CO, USA, 2026, pp. 1960–1965. <https://doi.org/10.1145/3789240.3827598>
- [7] K. van Berkel, The cost and network limits of space-based AI compute, arXiv:2607.14172 (2026). <https://arxiv.org/abs/2607.14172>
- [8] Q. Guo, J. Fu, K. Zang, Y. Yang, C. Wang, W. Tian, S. Qiu, G. Su, Design and optimization analysis of a 100 kW-level heat rejection system for space nuclear power systems, Appl. Therm. Eng. 300 (2026) 131536. <https://doi.org/10.1016/j.applthermaleng.2026.131536>
- [9] J. Fu, Y. Fei, C. Wang, G.H. Su, W. Tian, S. Qiu, Modeling and analysis of heat rejection subsystems for space power systems: steady-state and transient thermal management, Appl. Therm. Eng. 278 (2025) 127245. <https://doi.org/10.1016/j.applthermaleng.2025.127245>
- [10] Z. Lu, Z. Zhou, J. Luo, R. Jia, N. Hua, R. Xu, L. Yang, W. Zhou, Q. Eri, Study on the conjugate radiation heat transfer performance of radiator/condenser for heat pump thermal control system under lunar conditions, Appl. Therm. Eng. 298 (2026) 131071. <https://doi.org/10.1016/j.applthermaleng.2026.131071>
- [11] K.R. Sridhar, M. Gottmann, Thermal control systems for low-temperature heat rejection on a lunar base, NASA Contractor Report NASA-CR-190063, University of Arizona, February 1992 (grant NAG5-1572). <https://ntrs.nasa.gov/citations/19920011027>
- [12] W.E. Lear, S.A. Sherif, Weight optimization of active thermal management using a novel heat pump, final report, grant NAG3-2630, University of Florida, Department of Mechanical Engineering, submitted to NASA Glenn Research Center, 2004. <https://ntrs.nasa.gov/citations/20040075537>
- [13] H.J. van Gerner, G. van Donk, A. Pauw, J. van Es, S. Lapensée, A heat pump for space applications, ICES-2015-35, in: 45th International Conference on Environmental Systems, Bellevue, WA, USA, 12–16 July 2015. <http://hdl.handle.net/2346/64305>
- [14] A.J. Juhasz, R.C. Tew, G.E. Schwarze, Impact of radiation hardness and operating temperatures of silicon carbide electronics on space power system mass, in: Space Technology and Applications International Forum (STAIF-99), Albuquerque, NM, USA, 31 January–4 February 1999, AIP Conf. Proc. 458 (1999) 610–615. <https://doi.org/10.1063/1.57627>
- [15] A. Vassighi, M. Sachdev, Thermal runaway in integrated circuits, IEEE Trans. Device Mater. Reliab. 6 (2) (2006) 300–305. <https://doi.org/10.1109/TDMR.2006.876577>
- [16] M. Zapater, O. Tuncer, J.L. Ayala, J.M. Moya, K. Vaidyanathan, K. Gross, A.K. Coskun, Leakage-aware cooling management for improving server energy efficiency, IEEE Trans. Parallel Distrib. Syst. 26 (10) (2015) 2764–2777. <https://doi.org/10.1109/TPDS.2014.2361519>
- [17] D.C. Price, M.A. Clark, B.R. Barsdell, R. Babich, L.J. Greenhill, Optimizing performance-per-watt on GPUs in high performance computing: temperature, frequency and voltage effects, Comput. Sci. Res. Dev. 31 (4) (2016) 185–193. <https://doi.org/10.1007/s00450-015-0300-5>
- [18] (a) K. Aubrey, Inside the NVIDIA Vera Rubin platform: six new chips, one AI supercomputer, NVIDIA Technical Blog, 5 January 2026 (updated 21 August 2026). <https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer/> (accessed 23 October 2026). (b) J. Parker, Hotter than a hot tub: the 45 °C breakthrough to

cool AI's biggest machines, NVIDIA Blog, 22 June 2026 (updated 8 July 2026). <https://blogs.nvidia.com/blog/liquid-cooling-ai-factories/> (accessed 23 October 2026).

- [19] Z.-X. Wang, W.-Q. Tao, Numerical study on an air-cooling and liquid-cooling combined server for high-temperature data center application, *Appl. Therm. Eng.* 280 (2025) 128111. <https://doi.org/10.1016/j.applthermaleng.2025.128111>
- [20] The Boeing Company, Active Thermal Control System (ATCS) overview, ISS fact sheet, Boeing IDS Business Support, St. Louis, MO, USA, 2006 (undated; PDF created 22 November 2006), 19 pp. https://www.nasa.gov/wp-content/uploads/2021/02/473486main_iss_atcs_overview.pdf (accessed 23 October 2026)
- [21] I. Latif, M.A. Shafique, H. Ullah, A.C. Newkirk, X. Yu, A. Munir, Cooling matters: benchmarking large language models and vision-language models on liquid-cooled versus air-cooled H100 GPU systems, *IEEE Trans. Cloud Comput.* 14 (3) (2026) 1645–1656. <https://doi.org/10.1109/TCC.2026.3700971>
- [22] P. Sinha, A. Guliani, R. Jain, B. Tran, M.D. Sinclair, S. Venkataraman, Not all GPUs are created equal: characterizing variability in large-scale, accelerator-rich systems, in: *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, Dallas, TX, USA, 13–18 November 2022, pp. 1–15. <https://doi.org/10.1109/SC41404.2022.00070>
- [23] (a) slow-zhang, Question about MIG config persistent, NVIDIA/k8s-device-plugin, GitHub issue #343, 3 November 2022. <https://github.com/NVIDIA/k8s-device-plugin/issues/343> (accessed 23 October 2026). (b) csibbitt, DCGM-Exporter crashes on startup when using MIG w/ A100, NVIDIA/dcgm-exporter, GitHub issue #443, 16 January 2025. <https://github.com/NVIDIA/dcgm-exporter/issues/443> (accessed 23 October 2026).
- [24] NVIDIA, NVIDIA H100 PCIe GPU product brief, PB-11133-001_v02, November 2022. https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf (accessed 23 October 2026)
- [25] (a) NVIDIA, Configuring thermal shutdown, NVIDIA DriveOS 7.0.3 Linux SDK Developer Guide (last updated 16 December 2025). https://developer.nvidia.com/docs/drive/drive-os/7.0.3/public/drive-os-linux-sdk/platform-customization/System_Software_Settings/TSENSEThermalSensor88.html (accessed 23 October 2026). (b) DAVE Embedded Systems, Template:Thermal management, DAVE Developer's Wiki. https://wiki.dave.eu/index.php/Template:Thermal_management (accessed 23 October 2026)
- [26] Texas Instruments, AM62x Power Estimation Tool, AM62X-PET-CALC, version 01.00.00.00, released 30 May 2022; application note SPRAD31, May 2022. <https://www.ti.com/tool/download/AM62X-PET-CALC/01.00.00.00> (accessed 23 October 2026)
- [27] Texas Instruments, AM62D Power Estimation Tool, AM62D-PET-CALC, version 1.1, released 21 February 2025; application note SPRADO7. <https://www.ti.com/tool/download/AM62D-PET-CALC/1.1> (accessed 23 October 2026)
- [28] AMD (Xilinx), Xilinx Power Estimator (XPE) for UltraScale+, version 2023.1.2, released 16 August 2023. <https://www.amd.com/en/products/adaptive-socs-and-fpgas/technologies/power-efficiency/power-estimator.html> (accessed 23 October 2026)
- [29] S. Kim, J. Kim, D. Jang, R. Ritzenthaler, B. Parvais, J. Mitard, H. Mertens, T. Chiarella, N. Horiguchi, J.W. Lee, Comparison of temperature dependent carrier transport in FinFET and gate-all-around nanowire FET, *Appl. Sci.* 10 (8) (2020) 2979. <https://doi.org/10.3390/app10082979>
- [30] E. Simoen, C.H.S. Coelho, V.C.P. da Silva, J.A. Martino, P.G.D. Agopian, A. Oliveira, B. Cretu, A. Veloso, Performance perspective of gate-all-around double nanosheet CMOS beyond high-speed logic applications, *J. Integr. Circuits Syst.* 17 (2) (2022) 1–9. <https://doi.org/10.29292/jics.v17i2.617>
- [31] M. de Souza, A. Cerdeira, M. Estrada, S. Barraud, M. Cassé, M. Vinet, O. Faynot, M.A. Pavanello, High temperature and width influence on the GIDL of nanowire and nanosheet SOI nMOSFETs, *IEEE J. Electron Devices Soc.* 11 (2023) 672–680. <https://doi.org/10.1109/JEDS.2023.3264876>
- [32] S. Samsi, M.L. Weiss, D. Bestor, B. Li, M. Jones, A. Reuther, D. Edelman, W. Arcand, C. Byun, J. Holodnack, M. Hubbell, J. Kepner, A. Klein, J. McDonald, A. Michaleas, P. Michaleas, L. Milechin, J. Mullen, C. Yee, B. Price, A. Prout, A. Rosa, A. Vanterpool, L. McEvoy, A. Cheng, D. Tiwari, V. Gadepally, The MIT Supercloud dataset, in: 2021

- IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 20–24 September 2021, pp. 1–8. <https://doi.org/10.1109/HPEC49654.2021.9622850>
- [33] [dataset] S. Samsi, M. Weiss, D. Bestor, et al., MIT Supercloud dataset, Datacenter Challenge labelled data, release 202201 (Version 1), Registry of Open Data on AWS, 2022. <https://registry.opendata.aws/dcc/> (accessed 23 October 2026). License CC BY-NC-ND 4.0. s3://mit-supercloud-dataset/datacenter-challenge/202201/
- [34] I.H. Bell, J. Wronski, S. Quoilin, V. Lemort, Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library CoolProp, *Ind. Eng. Chem. Res.* 53 (6) (2014) 2498–2508. <https://doi.org/10.1021/ie4033999>
- [35] [software] CoolProp, version 8.0.0, 2026. <https://github.com/CoolProp/CoolProp> (accessed 23 October 2026).
- [36] R. Schoonhoven, B. Veenboer, B. van Werkhoven, K.J. Batenburg, Going green: optimizing GPUs for energy efficiency through model-steered auto-tuning, in: 2022 IEEE/ACM International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), Dallas, TX, USA, November 2022, pp. 48–59. <https://doi.org/10.1109/PMBS56514.2022.00010>
- [37] D.L. Ellis, J. Calder, J. Siamidis, Summary of the manufacture, testing and model validation of a full-scale radiator for fission surface power applications, in: *Proc. Nuclear and Emerging Technologies for Space 2011 (NETS 2011)*, Albuquerque, NM, USA, 7–10 February 2011, paper 3181 (NASA report E-17664). <https://ntrs.nasa.gov/citations/20110012006>
- [38] L. James, Elon Musk’s first-gen orbital data center craft spans wider than a Boeing 747 and runs an interchangeable chip payload — AI1 satellite compute payload is 120 kW, peaks at 150 kW, *Tom’s Hardware*, 9 June 2026. <https://www.tomshardware.com/tech-industry/spacex-details-its-ai1-compute-satellite> (accessed 23 October 2026). Primary source: SpaceX (@SpaceX), technical update video on SpaceX’s capability to manufacture, launch, and operate AI satellites at scale, X post, 8 June 2026. <https://x.com/SpaceX/status/2064099405758906727> (accessed 23 October 2026)
- [39] J. Crop, H. Moore, S. Pasricha, Revisiting “Cooler is Better”: ITD-aware per-CPU thermal optimization for sustainable data center operation, in: *Proc. 16th ACM International Green and Sustainable Computing Conference (IGSC ’26)*, Canandaigua, USA, 2026, pp. 1–7. <https://doi.org/10.1145/3797248.3815399>; arXiv:2606.11163
- [40] [dataset] J. Jing, Closed mass model for liquid-cooled compute satellites with temperature-dependent chip leakage, version 1.0.0, Zenodo, 2026. <https://doi.org/10.5281/zenodo.23576307>
- [41] Vertiv, Keep it at 30 °C: a roadmap for the future chip coolant temperature, Vertiv blog, 9 April 2026. <https://www.vertiv.com/en-emea/insights/articles/blog-posts/a-roadmap-for-the-future-chip-coolant-temperature/> (accessed 23 October 2026)
- [42] V. Kandiah, S. Peverelle, M. Khairy, J. Pan, A. Manjunath, T.G. Rogers, T.M. Aamodt, N. Hardavellas, AccelWattch: a power modeling framework for modern GPUs, in: *Proc. 54th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO-54)*, 2021, pp. 738–753. <https://doi.org/10.1145/3466752.3480063>
- [43] U.S. Department of Energy, National Renewable Energy Laboratory, EnergyPlus reference data set Chillers.idf (ElectricEIRChiller, Trane RTWA 383 kW, COP 4.17) [dataset], EnergyPlus GitHub repository. <https://github.com/NatLabRockies/EnergyPlus/blob/develop/datasets/Chillers.idf> (accessed 23 October 2026).
- [44] JEDEC Solid State Technology Association, Failure mechanisms and models for semiconductor devices, JEDEC Publication JEP122H, September 2016.
- [45] P. Johnston, interviewed by C. Simms, AI hits a power wall. Starcloud launches data centers into orbit, *Inevitable podcast (MCJ)*, 13 January 2026. <https://mcj.vc/inevitable-podcast/starcloud> (accessed 23 October 2026).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [orbitalaidcleakagepreprintSI.pdf](#)