Clause AI-4

SRAM Thermal Integrity Bound for Fused Attention Kernels

Version 1.0

Ryan Fields

UncleBroFields@proton.me

Abstract

This clause establishes a formal thermal integrity bound for fused attention kernels—principally FlashAttention, FlashAttention-2, and FlashAttention-3—executing on GPU on-chip Static Random-Access Memory (SRAM). The FlashAttention family of kernel optimizations eliminates off-chip memory round-trips by tiling all attention computation within SRAM, achieving near-continuous duty cycles on the memory arrays of NVIDIA A100 (TSMC 7nm) and H100 (TSMC 4nm) accelerators. While this has enabled unprecedented throughput for long-context Large Language Model (LLM) training and inference, it introduces an under-characterized thermal risk: sustained SRAM power density can drive localized junction temperatures beyond the assumptions embedded in standard JEDEC qualification envelopes, accelerating Bias Temperature Instability (BTI), degrading Static Noise Margin (SNM), and increasing vulnerability to Silent Data Corruption (SDC).

No existing safety standard—including ISO 26262, IEC 61508, DO-254, NIST AI RMF 1.0, or the EU AI Act—addresses hardware-level thermal fault tolerance for AI inference workloads. This clause fills that regulatory gap by specifying four interlocking operational controls: a thermal envelope with headroom, mandatory ECC monitoring on attention-score SRAM, pre-deployment silicon-level screening, and duty-cycle limiting for safety-critical inference. Every claim is grounded in peer-reviewed literature and verified vendor specifications; the complete fault path is presented as convergent evidence across adjacent domains rather than as a single experimentally verified chain.

1. Instability Surface — Fused Attention and Sustained SRAM Thermal Load

1.1 The Memory-Wall Bypass and Its Thermal Consequence

Standard ("vanilla") self-attention materializes the full $N \times N$ score matrix to High Bandwidth Memory (HBM), creating natural thermal pauses during data-transfer latencies. FlashAttention disrupts this equilibrium. By keeping Query (Q), Key (K), and Value (V) blocks resident in on-chip SRAM and fusing matrix multiplication, masking, softmax, and dropout into a single kernel, the workload shifts from memory-bound to compute-bound. The intermediate score and probability matrices never touch HBM—they exist transiently in SRAM and registers.

The thermal implication is direct: the idle periods that previously allowed micro-scale silicon cooling are architecturally eliminated. The SRAM banks operate not as a cache but as a high-frequency scratchpad at maximum electrical load.

1.2 FlashAttention Tiling Mechanics

FlashAttention (Dao et al., NeurIPS 2022 [1]) partitions Q, K, V into blocks sized to fit within shared memory—approximately 192 KB per Streaming Multiprocessor (SM) on A100 (108 SMs, approximately 20 MB total SRAM). Block sizes follow $B_c = \lceil M/4d \rceil$, yielding IO

complexity $\Theta(N^2 d^2 M^{-1})$, proven optimal within a specified SRAM-size range. All intermediate computation—QK^T, softmax normalization, output accumulation—executes entirely within SRAM.

FlashAttention-2 (Dao, 2023 [2]) reduces SRAM read/write traffic further by switching from split-K to split-Q warp partitioning, eliminating inter-warp shared-memory synchronization. This achieves 50–73% of theoretical peak FLOPs on A100 (forward pass) and approximately 225 TFLOPs/s in end-to-end GPT-class training. Crucially, it introduces parallelism across the sequence-length dimension, ensuring all SMs are fully occupied—distributing thermal load globally across the die rather than concentrating it in a subset of units.

1.3 The Hopper Thermal Apex — FlashAttention-3

FlashAttention-3 (Shah et al., 2024 [3]) exploits the NVIDIA H100's Tensor Memory Accelerator (TMA) and warp specialization to create a circular shared-memory buffer pipeline: producer warps continuously load tiles from HBM via TMA while consumer warps execute Warpgroup Matrix Multiply-Accumulate (WGMMA) tensor-core operations reading directly from SRAM. This architecture sustains near-continuous SRAM read/write activity, reaching 740–840 TFLOPs/s on H100 SXM5 (75–85% utilization in BF16). The H100 provides approximately 228 KB shared memory per SM across 132 SMs.

The result is a "double-stress" regime: while tensor cores read from one SRAM bank via WGMMA, the TMA simultaneously writes to another bank to prepare the next tile. The SRAM arrays are subject to concurrent high-current read and write operations with no architecturally scheduled relaxation period.

1.4 The Thermal Measurement Gap

No published thermal measurements of GPU SRAM banks during FlashAttention execution exist in the public literature. NVIDIA does not expose SM-level or SRAM-level thermal sensors publicly. Sub-component thermal instrumentation requires specialized infrared imaging on exposed dies, which has not been performed for fused attention workloads on A100 or H100.

The closest available work is Patel et al. [27], which provides GPU-level power traces during LLM workloads on A100 via `nvidia-smi`/DCGM at 100 ms intervals—but reports only whole-GPU power draw with no SRAM-specific breakdown. Per-kernel power profiling is not supported in Nsight Compute or Nsight Systems; workarounds involve NVML polling (`nvmlDeviceGetPowerUsage()`) at 10–100 ms intervals correlated with CUDA events, providing device-level but not per-kernel resolution.

This measurement gap is itself a finding: the industry lacks the instrumentation to characterize the very thermal risk this clause addresses, which strengthens rather than weakens the case for precautionary operational bounds.

2. SRAM Reliability Under Thermal Stress — The Physics of Failure

2.1 Static Noise Margin Degradation

The stability of an SRAM cell is quantified by its Static Noise Margin (SNM). SNM degrades with temperature through two mechanisms: immediate parametric shifts (threshold voltage reduction, mobility degradation from phonon scattering) and accelerated Bias Temperature Instability (BTI).

Vattikonda et al. [9] demonstrated via Monte Carlo simulation in 45 nm PDSOI that both Negative BTI (NBTI) and Positive BTI (PBTI) degrade read stability and increase cell failure probability over time. Grasser et al. [10] confirmed that NBTI-induced SNM degradation affects cell metastability.

FinFET SRAM is more vulnerable to BTI than planar CMOS. Ndiaye et al. [11] showed strong supply-voltage dependence of BTI degradation in 14 nm FinFET cells. Li et al. [12] demonstrated that self-heating exacerbates NBTI-driven SNM degradation in next-generation nanosheet field-effect transistors—directly relevant to the thermal profile of fused attention workloads.

Zhang et al. [13] provided a comprehensive methodology combining BTI, Hot Carrier Injection (HCI), Gate Time-Dependent Dielectric Breakdown (GTDDB), and Random Telegraph Noise (RTN) for FinFET SRAM reliability, showing failure probability increases with both temperature and duty cycle.

2.2 Arrhenius-Scaled BTI Acceleration

BTI degradation follows the Arrhenius relationship:

$$\Delta V_{\text{th}} \;\propto\; t^{n} \cdot \exp\!\left(\frac{-E_a}{kT}\right), \tag{1}$$

where $E_a \approx 0.5$–$0.7\,\text{eV}$ for NBTI in high-κ/metal-gate FinFETs and $n \approx 0.16$–0.25. The acceleration factor from 85 °C to 120 °C is approximately 4–6× for aging-related cell failure rate. A device operating at sustained 110 °C ages significantly faster than one at 85 °C, effectively consuming the silicon's reliability budget at an accelerated rate and increasing the likelihood of stuck bits or timing violations over the device's lifetime.

2.3 Soft Error Rate — A Non-Monotonic Relationship

The relationship between temperature and Soft Error Rate (SER) in FinFET SRAM is more complex than often claimed. Bagatin et al. [14] found the direction of temperature dependence is vendor- and design-specific.

In FinFET SRAM, higher temperature initially *reduces* SER through increased critical charge (Q_{crit}) from threshold voltage reduction—but long-term BTI aging degrades Q_{crit}, so aged SER worsens. A 2024 experimental study on 28 nm embedded SRAM [15] measured a 39.8% Single Event Upset (SEU) cross-section increase from 23 °C to 109 °C with approximately 4.8% Q_{crit} decrease.

Any operational clause must account for this non-monotonic behavior: short-term thermal excursions may not immediately increase SER, but sustained elevated temperatures degrade aging-dependent margins that make future soft errors more likely.

2.4 Quantitative BER Gap

No publicly available quantitative Bit Error Rate (BER) versus temperature curve exists for 7 nm, 5 nm, or 4 nm SRAM. TSMC's IEDM 2020 paper [16] demonstrates reliability attributes but does not publish temperature-dependent BER data. This data is foundry-proprietary. The clause therefore specifies bounds parametrically—operators supply vendor-specific constants under NDA—rather than hardcoding values that cannot be publicly verified.

3. GPU Thermal Specifications and the Instrumentation Gap

3.1 Published Thermal and Power Specifications

Table 1: NVIDIA Data-Center GPU Thermal and Power Specifications

GPU Variant	TDP	Process Node	Die Size	L2 (SRAM)
H100 SXM5	700 W	TSMC 4nm (4N)	814 mm^2	50 MB
H100 PCIe	350 W	TSMC 4nm (4N)	814 mm^2	50 MB
A100 SXM4 (80 GB)	400 W	TSMC 7nm (N7)	826 mm^2	40 MB
A100 PCIe (80 GB)	300 W	TSMC 7nm (N7)	826 mm^2	40 MB

Sources: NVIDIA H100 Tensor Core GPU Datasheet, 2022 [20]; H100 PCIe Product Brief PB-11133-001_v02 [21]; A100 Tensor Core GPU Datasheet [22]; A100 80 GB PCIe Product Brief PB-10577-001_v03 [23].

The H100 SXM5 nearly doubles the power density (W/mm^2) compared to the A100 SXM4 despite similar die areas. This increase in heat flux places extreme demands on thermal interface materials and cooling solutions.

3.2 Junction Temperature — NDA-Restricted

Maximum junction temperature (T_{j_max}) is *not* publicly published for H100 or A100 data-center GPUs. The H100 PCIe Product Brief [21] provides thermal qualification temperatures of GPU $T_{AVG} = 87\,°C$ and HBM $T_{HBM} = 95\,°C$, with hardware slowdown and 50% clock reduction at $T_{LIMIT} - 2\,°C$ (relative to the slowdown threshold), and hardware shutdown at $T_{LIMIT} - 5\,°C$. The absolute threshold temperatures are NDA-restricted.

Any clause citing a specific T_{j_max} for these GPUs cannot reference a public NVIDIA source. This clause therefore defines thermal bounds *relative* to vendor-supplied T_{j_max} values provided under procurement agreements.

3.3 SRAM Thermal Constants — Not Publicly Available

SRAM thermal resistance (R_θ) and thermal capacitance (C_{th}) are not publicly available for any NVIDIA GPU SRAM banks. NVIDIA publishes thermal design guides with junction-to-case (R_{jc}) and junction-to-board (R_{jb}) values only for embedded products (Jetson Orin). The clause specifies these as vendor-supplied parameters.

3.4 The Overshoot Mechanism

FlashAttention kernels ramp from idle to maximum power in microseconds ($< 10\,\mu s$). The thermal management system—sensors, fan/pump controllers, driver-level throttling—operates on a loop of tens to hundreds of milliseconds. Hardware-level clock gating is faster but reactive. The thermal time constant ($\tau = R_\theta \times C_{th}$) of on-die features is in the millisecond range.

This latency mismatch allows **thermal overshoot**: local SRAM temperature can spike above the reported average before active cooling or throttling mechanisms engage. The overshoot magnitude depends on the vendor-specific R_θ and C_{th} of the SRAM banks—parameters that are not public but are known to the GPU vendor and system integrator.

4. The Fault Path — Convergent Evidence Across Adjacent Domains

The proposed causal chain—sustained SRAM thermal stress → bit-level corruption → model output degradation—is individually supported at each segment by peer-reviewed literature but has never been experimentally demonstrated end-to-end in any single published study. This section maps each segment to verified sources.

4.1 Segment A — Temperature Drives SDC at Fleet Scale

Wang et al. [4] studied Silent Data Corruption (SDC) across over one million processors at Alibaba Cloud (SOSP 2023). Their Observation 10 states that the occurrence frequency of SDCs demonstrates *exponential growth in response to increasing temperatures*. They developed "Farron," a system using temperature control to mitigate less reproducible SDCs.

Dixit et al. at Meta [5] documented SDC across hundreds of CPUs over 18+ months at a rate of approximately 1 in 1,000 devices—orders of magnitude higher than the 1 in a million rate assumed by traditional soft error models. Temperature, voltage noise, and data patterns were identified as key triggers.

Hochschild et al. at Google [6] coined the term "mercurial cores"— cores producing errors only under specific, reproducible conditions attributed to near-limit CMOS feature sizes.

Important caveat: These studies focus on CPUs, not GPU SRAM specifically. However, the physical mechanisms (BTI, HCI, manufacturing-edge variability) are common to all advanced-node CMOS, and the temperature–SDC correlation is a material property, not an architecture-specific artifact.

4.2 Segment B — Hardware Faults Corrupt Neural Network Outputs

Li et al. [8] (SC 2017, NVIDIA co-authored) was the first study of soft-error propagation in DNN systems, finding that $0 \rightarrow 1$ bit flips in high-order exponent bits are most likely to cause SDC in model outputs.

A 2025 study (arXiv:2601.19912) performed the first instruction-level fault injection on LLM inference (GPT-2, Qwen, Llama), finding substantial Detected Unrecoverable Error (DUE) rates from single bit-flip injection—though using simulated rather than thermally-induced faults.

4.3 Segment C — Real SDC Corrupts LLM Training

Ma et al. [7] (ACL 2025) is the first study using real unhealthy nodes (not simulated faults) to demonstrate SDC impact on LLMs: models converge to different optima, and some nodes cause *full corruption of model weights* during fine-tuning.

Meta's Llama 3 training report documented 6 unplanned job interruptions from SDC during 54 days on 16,384 H100 GPUs (466 total interruptions, 78% from hardware). Neither study isolates the specific physical mechanism, but both confirm that hardware-level faults—of the type accelerated by thermal stress—produce measurable model corruption in production.

The OCP whitepaper "Silent Data Corruption in AI" [24] (led by Nishant George, NVIDIA, with contributions from Meta and others, 2024) provides the most comprehensive industry treatment, enumerating hardware fault types and AI impact while acknowledging the gap between hardware fault metrics and AI correctness metrics.

4.4 The Softmax Amplification Mechanism

The attention mechanism's softmax function,

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{2}$$

is highly non-linear. A single bit flip in the exponent field of a BF16 or FP16 attention score can transform a value from, for example, -5.0 to $+5.0$ or larger. This artificially inflated score dominates the softmax denominator, collapsing the probability distribution onto a single, potentially irrelevant token. In autoregressive generation, this corrupted attention vector retrieves the wrong Value (V), and the error propagates forward through subsequent tokens.

4.5 Epistemological Status of the Fault Path

Each segment—temperature accelerates hardware faults, hardware faults corrupt DNN outputs, real SDC corrupts LLM weights—is individually supported by peer-reviewed research from independent groups (Alibaba, Meta, Google, NVIDIA, academic institutions). The convergence of evidence across these adjacent domains constitutes a *well-grounded risk inference*. This clause presents the fault path as such, not as a proven end-to-end causal demonstration.

This epistemological honesty is a feature, not a limitation: it reflects the current state of knowledge accurately while establishing precautionary bounds appropriate to the risk. It also identifies the end-to-end experimental demonstration as an urgent research priority.

5. The Regulatory Vacuum

5.1 Standards Surveyed

The following standards and frameworks were examined for provisions addressing hardware-level thermal fault tolerance for AI inference workloads:

ISO 26262:2018 (Road Vehicles—Functional Safety): Addresses hardware fault metrics (SPFM, LFM, PMHF) and includes semiconductor guidance in Part 11, but contains no AI-specific or GPU-thermal-specific provisions. NVIDIA has pursued ISO 26262 assessment only for automotive SoCs (DRIVE Orin, assessed by TUV SUD for ASIL D systematic / ASIL B random).

DO-254 (Design Assurance Guidance for Airborne Electronic Hardware): Covers FPGA/ASIC design assurance but does not address GPUs or AI inference.

IEC 61508:2010 (Functional Safety of E/E/PE Systems): Provides generic functional safety methods (SIL, HFT, SFF) applicable to any electrical/electronic system but contains no AI-accelerator provisions.

NIST AI RMF 1.0 (AI 100-1, January 2023 [26]): Operates at the organizational/process level and explicitly does not address hardware.

EU AI Act (Regulation EU 2024/1689): Requires "technical robustness" for high-risk AI systems but does not extend to silicon-level thermal specifications.

ISO/PAS 8800:2024, ISO/IEC TR 5469:2024, ISO/IEC 42001:2023: Emerging AI safety standards focused on algorithmic/model-level safety, not hardware thermal behavior.

5.2 The Finding

No cloud provider, GPU vendor, or standards body has proposed formal thermal safety bounds for AI/ML accelerator workloads. The ACM Computing Surveys paper "GPU Devices for Safety-Critical Systems" [25] (2022) confirms that applicable safety standards have "none or limited consideration of GPU devices."

No "silicon-level audit" requirements formally exist in any regulated industry—finance, defense, or medical AI—for GPU-accelerated inference. FDA guidance on AI/ML-enabled medical devices focuses on software lifecycle, not hardware thermal reliability. Financial regulators focus on data governance and auditability, not silicon verification.

MLPerf Power (MLCommons, 2024 [28]) standardizes energy-efficiency measurement but defines no thermal safety thresholds.

This regulatory vacuum is the structural justification for this clause.

6. Governing JEDEC Standards — Corrected Reference Chain

6.1 JESD47 — Stress-Test-Driven Qualification of Integrated Circuits

Current revision JESD47L [17]. This is the parent standard defining the baseline qualification test suite, including High-Temperature Operating Life (HTOL, per JESD22-A108), which directly stresses the BTI, HCI, TDDB, and electromigration mechanisms that degrade SRAM. JESD47 includes explicit SRAM/NVM data retention bake requirements: blocks cycled to $\leq 10\%$ of maximum specified cycles must retain data for 100 hours at 125 °C; blocks at 100% of maximum cycles must retain for 10 hours at 125 °C.

If a workload drives the chip to temperatures approaching the HTOL stress condition (125 °C) during normal operation, the device is effectively being stress-tested continuously, consuming its reliability budget at an accelerated rate.

6.2 JESD94 — Application-Specific Qualification

Revision JESD94A (2008) [18]. Provides the framework for adapting test conditions when application use deviates from JESD47 assumptions. JESD47 explicitly references JESD94 for this purpose. GPU SRAM operating at sustained 90–120 °C with near-continuous FlashAttention duty cycles represents exactly the kind of deviation requiring JESD94-based application-specific qualification.

6.3 JESD89B — Soft Error Rate Measurement and Reporting

JESD89B [19] standardizes SER measurement methodology for alpha, thermal neutron, and high-energy neutron induced soft errors. Relevant for SER characterization at elevated temperatures but does not itself address temperature-dependent qualification.

6.4 Supporting Documents

JEP122 (Failure Mechanisms and Models): Provides activation energies and acceleration factors for BTI, HCI, TDDB, and electromigration—the quantitative basis for reliability budget calculations.

JEP148 (Reliability Qualification Based on Physics of Failure): Provides the conceptual framework for physics-based qualification rather than purely statistical approaches.

6.5 Note on JESD22-A104

JESD22-A104 (Temperature Cycling) tests package-level thermo-mechanical failures—solder joint fatigue, delamination, coefficient of thermal expansion mismatch. It is relevant for packaging reliability under repeated thermal cycling but is *not* the correct primary reference for SRAM cell-level thermal reliability. It should be cited as a supplementary concern (FlashAttention's rapid power transients do induce thermal cycling) but not as the governing standard.

7. Clause Specification — Operational Controls

7.1 Control 1 — Thermal Envelope with Headroom

The workload scheduler shall ensure that sustained SRAM junction temperature (T_{j_SRAM}) does not exceed the manufacturer's rated $T_{j\_\max}$ minus a safety margin of $\Delta T_{\mathrm{margin}}$ (recommended: 5 °C) for any continuous duration exceeding the thermal time constant of the die (τ, approximately 100 ms for on-die SRAM features).

Formally: for all $t \in [0, T_{\mathrm{run}}]$, if the moving-average SRAM junction temperature over any window of length τ exceeds $(T_{j\_\max} - \Delta T_{\mathrm{margin}})$, the runtime shall initiate the fail-safe action defined in Control 4.

Rationale: The latency mismatch between FlashAttention power ramp (microseconds) and thermal management response (tens to hundreds of milliseconds) permits transient overshoot. This bound ensures that overshoot events cannot persist beyond the thermal time constant without intervention.

Parameters: $T_{j\_\max}$ and τ are vendor-supplied under procurement NDA. $\Delta T_{\mathrm{margin}} = 5\,°\mathrm{C}$ is recommended based on the JESD47 HTOL stress margin (125 °C qualification versus 105–110 °C rated operation).

7.2 Control 2 — ECC Monitoring Mandate

All SRAM arrays utilized for attention-score computation (the matrices $S = QK^T$ and $P = \mathrm{softmax}(S)$) shall have Single-Error Correction Double-Error Detection (SECDED) ECC enabled and actively monitored. Any Double-Bit Error (DBE) or uncorrectable error shall trigger an immediate kernel halt and device drain from the inference pool.

Correctable single-bit errors shall be logged to an append-only audit ledger with SHA-256 hash of the offending request, timestamp, SM identifier, and SRAM bank address. If the correctable error rate exceeds a configurable threshold (recommended: 1 correctable error per 10^9 SRAM access cycles) over any 60-second window, the device shall be flagged for silicon-level screening per Control 3.

Rationale: The non-monotonic SER behavior in FinFET SRAM [14, 15] means that elevated temperatures degrade aging-dependent noise margins, increasing future vulnerability to both single-event upsets and multi-cell upsets. Active ECC monitoring is the only operational mechanism that can catch thermally-accelerated soft errors before they propagate through the softmax amplification path.

7.3 Control 3 — Silicon-Level Screening (The "Ripple" Mandate)

Prior to deployment in production inference, all accelerator units shall undergo a silicon-level screening audit using a high-stress workload: FlashAttention-3 (or the highest-available fused attention kernel) at maximum supported context length, sustained for a minimum of 4 hours at rated TDP.

During screening, the unit shall execute a deterministic reference workload with known-correct outputs. Any device exhibiting calculation variance—defined as any divergence from the reference output beyond the expected floating-point non-determinism bounds for the given precision format—shall be quarantined and excluded from safety-critical deployment.

Rationale: Meta's fleet-scale study [5] identified SDC-susceptible devices at a rate of approximately 1 in 1,000—orders of magnitude higher than soft error models predicted. Google's "mercurial cores" [6] confirmed that these defects are reproducible under specific workload conditions. FlashAttention's combination of high thermal stress, high switching activity, and complex data patterns represents a maximally stressing workload for identifying such latent defects.

Devices passing screening shall be recorded in a hardware provenance ledger with screening date, firmware version, thermal telemetry summary, and reference-output hash.

7.4 Control 4 — Duty-Cycle Limiting and Fail-Safe Action

For safety-critical inference (corresponding to ASIL-B or higher under ISO 26262, or SIL 2 or higher under IEC 61508), the kernel shall enforce one of the following:

(a) Inject thermal relaxation micro-operations ("cooling micro-ops") at configurable intervals to maintain a maximum effective SRAM duty cycle of $D_{\max}$ (recommended: 90%), unless real-time thermal telemetry confirms $T_{j\_\mathrm{SRAM}} < (T_{j\_\max} - 2\,\Delta T_{\mathrm{margin}})$; or

(b) Limit the occupancy of high-power instructions (WGMMA on Hopper, HMMA on Ampere) to $D_{\max}$ percent of available warp slots.

Fail-safe action upon thermal bound breach (Control 1 violation):

1. Flush the key-value cache (`flush_kv_cache()`) and switch to a safe-attention kernel (unfused, standard memory-bound implementation).
2. Record the breach event to the append-only audit ledger: SHA-256 hash of the offending request, timestamp, thermal telemetry snapshot, kernel identifier.
3. Require two successive monitoring windows (each of length τ) with $T_{j\_\mathrm{SRAM}} < (T_{j\_\max} - 2\,\Delta T_{\mathrm{margin}})$ before re-enabling fused attention.

Rationale: FlashAttention architecturally eliminates the thermal relaxation periods present in memory-bound attention. For workloads where model output correctness has safety, financial, or clinical consequences, re-introducing controlled idle time is a necessary engineering trade: throughput is reduced by at most 10% while the silicon operates within its qualified reliability envelope. The fallback to unfused attention ensures continued service availability—at reduced throughput—during thermal recovery.

8. Capabilities Unlocked

Implementation of this clause enables the following:

Chip-level safety ledger. Deterministic recording of thermal overshoot events, ECC errors, and screening results creates an auditable hardware provenance trail for root-cause analysis and regulatory compliance.

Run-time reliability bound. Prevents undetected token corruption at 100k+ sequence lengths by ensuring SRAM operates within its qualified thermal envelope during the sustained duty cycles unique to fused attention.

Certified fused-attention service level. Cloud providers can offer "thermally clause-backed attention" as a differentiated tier for regulated users in finance, defense, and medical AI—converting a latent risk into a marketable assurance.

Latency predictability. Bounded thermal oscillations reduce throttling jitter across multi-tenant GPU pods, improving tail-latency SLAs for inference serving.

Procurement qualification. Hardware buyers can incorporate this clause into GPU procurement specifications, requiring vendors to supply the $T_{j\_\max}$, R_θ, and C_{th} parameters necessary for clause enforcement—creating market pressure for thermal transparency.

Regulatory readiness. As GPU safety standards mature (the ACM Computing Surveys finding [25] of "none or limited consideration" will not persist indefinitely), organizations with clause-compliant infrastructure will be positioned ahead of mandatory requirements.

9. Audit Hook — Operational Parameters (Example Configuration)

The following values are illustrative. Deployers shall substitute vendor-supplied and site-specific parameters.

Table 2: Example Operational Parameters for Clause AI-4

Parameter	Example Value	Source
$T_{j\_\max}$	Per vendor NDA	GPU vendor datasheet (NDA)
ΔT_{margin}	5 °C	This clause (recommended)
τ (thermal time constant)	100 ms	Vendor-supplied or estimated
$D_{\max}$ (max duty cycle)	90%	This clause (ASIL-B+)
ECC correctable-error threshold	1 per 10^9 cycles / 60 s	This clause (recommended)
Screening duration	4 hours at rated TDP	This clause (minimum)
Recovery criterion	2 windows $< (T_{j\_\max} - 2\Delta T)$	This clause
Audit ledger format	Append-only, SHA-256	This clause

10. Conclusion

The transition to fused, tile-based attention mechanisms has been instrumental in scaling AI capabilities. It has also fundamentally altered the reliability profile of the underlying silicon. By bypassing the memory wall, these kernels expose nanometer-scale SRAM cells to sustained, high-intensity thermal stresses that were previously rare in standard computing workloads and that are not addressed by any existing safety standard.

The evidence assembled in this clause—from FlashAttention's documented near-continuous SRAM duty cycles, through Arrhenius-scaled BTI degradation in FinFET SRAM, to fleet-scale temperature–SDC correlations at Alibaba, Meta, and Google, to real SDC-induced LLM weight corruption—constitutes convergent evidence from independent research groups across

adjacent domains. The individual links are peer-reviewed and empirically grounded. The end-to-end chain has not been experimentally demonstrated in a single study; this clause identifies that gap as both a limitation of the current literature and an urgent research priority.

What is not in doubt is the regulatory vacuum. No standard, no framework, and no vendor specification currently governs the thermal integrity of SRAM under fused attention workloads. This clause is the first public specification to address that gap. Its four operational controls—thermal envelope, ECC monitoring, silicon screening, and duty-cycle limiting—are designed to be implementable with existing hardware instrumentation and enforceable within existing procurement and SLA frameworks.

The pursuit of speed must not come at the cost of truth. For organizations deploying AI in regulated, safety-critical, or high-stakes environments, the implementation of this clause is not merely precautionary—it is a necessary engineering standard for the infrastructure upon which their systems depend.

References

[1] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," *Proc. NeurIPS*, 2022. arXiv:2205.14135.

[2] T. Dao, "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning," 2023. arXiv:2307.08691.

[3] J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao, "FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision," 2024. arXiv:2407.08608.

[4] Y. Wang, Z. Zhang, et al., "Understanding Silent Data Corruptions in a Large Production CPU Population," *Proc. SOSP*, 2023. ACM DOI:10.1145/3600006.3613149.

[5] H. D. Dixit, S. Pendharkar, M. Beadon, C. Mason, T. Chakravarthy, B. Muthiah, and S. Sanber, "Silent Data Corruptions at Scale," 2021. arXiv:2102.11245.

[6] P. H. Hochschild, P. Turner, and J. C. Mogul, "Cores that don't count," *Proc. HotOS*, 2021. ACM DOI:10.1145/3458336.3465297.

[7] Z. Ma, Y. Pei, et al., "Understanding Silent Data Corruption in LLM Training," *Proc. ACL*, 2025. arXiv:2502.12340.

[8] G. Li, S. K. S. Hari, M. Sullivan, T. Tsai, K. Pattabiraman, J. Emer, and S. W. Keckler, "Understanding Error Propagation in Deep Learning Neural Network Accelerators and Applications," *Proc. SC*, 2017.

[9] R. Vattikonda, W. Wang, and Y. Cao, "Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability," *Microelectronics Reliability*, 2009.

[10] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, "Impact of NBTI/PBTI on SRAM Stability Degradation," *IEEE Electron Device Letters*, 2011.

[11] C. Ndiaye, F. Cacho, and V. Huard, "NBTI-Related Variability Impact on 14-nm Node FinFET SRAM Performance and Static Power," *IEEE Transactions on Electron Devices*, 2018.

[12] Y. Li et al., "Interaction of Negative Bias Instability and Self-Heating Effect on Threshold Voltage and SRAM Stability of Nanosheet Field-Effect Transistors," *Micromachines*, 2024. DOI:10.3390/mi15030420.

[13] Z. Zhang et al., “SRAM Stability Analysis and Performance–Reliability Tradeoff for Different Cache Configurations,” *IEEE TVLSI*, 2020.

[14] M. Bagatin et al., “Temperature dependence of neutron-induced soft errors in SRAMs,” *Microelectronics Reliability*, 2012.

[15] “Experimental Study of the Impact of Temperature on Atmospheric Neutron-Induced Single Event Upsets in 28 nm Embedded SRAM of SiP,” *Electronics*, 2024. DOI:10.3390/electronics13112012.

[16] C. C. Liu et al., “A Reliability Enhanced 5nm CMOS Technology Featuring 5th Generation FinFET with Fully-Developed EUV and High Mobility Channel for Mobile SoC and High Performance Computing Application,” *Proc. IEDM*, 2020. DOI:10.1109/IEDM13553.2020.9372009.

[17] JEDEC Solid State Technology Association, “Stress-Test-Driven Qualification of Integrated Circuits,” JESD47 (current revision JESD47L).

[18] JEDEC Solid State Technology Association, “Application-Specific Qualification Using Knowledge-Based Test Methodology,” JESD94A, 2008.

[19] JEDEC Solid State Technology Association, “Measurement and Reporting of Alpha Particle and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices,” JESD89B.

[20] NVIDIA Corporation, “H100 Tensor Core GPU Datasheet,” 2022.

[21] NVIDIA Corporation, “H100 PCIe GPU Product Brief,” PB-11133-001_v02, November 2022.

[22] NVIDIA Corporation, “A100 Tensor Core GPU Datasheet.”

[23] NVIDIA Corporation, “A100 80 GB PCIe GPU Product Brief,” PB-10577-001_v03, March 2022.

[24] N. George et al., “Silent Data Corruption in AI,” Open Compute Project Whitepaper, 2024.

[25] “GPU Devices for Safety-Critical Systems: A Survey,” *ACM Computing Surveys*, 2022. DOI:10.1145/3549526.

[26] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, January 2023.

[27] P. Patel et al., “Characterizing Power Management Opportunities for LLMs in the Cloud,” *Proc. ASPLOS*, 2024.

[28] MLCommons, “MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI,” 2024. arXiv:2410.12032.

Intellectual Property (IP) Declaration

The methods, logic structures, and operational specifications contained in this work are the sole property of Ryan Fields.

Public License (Non-Commercial)

- **Academic Use:** Researchers may share and use this framework for non-commercial academic purposes, provided full attribution is given to Ryan Fields.
- **No Derivatives:** No modifications or adaptations of the clause specifications, operational parameters, or audit protocols are permitted without express written consent.

Commercial Prohibition

Commercial use of this framework is strictly prohibited. This includes, but is not limited to:

- Integration into cloud provider service-level agreements (SLAs) for GPU-accelerated inference.
- Use within proprietary hardware procurement or qualification pipelines.
- Incorporation into commercial AI safety, governance, or compliance software.
- Use by semiconductor vendors for thermal qualification marketing claims.