Thermal, Power & TCO

The Economics of AI Infrastructure — Why a 700-Watt Chip Changes Everything About Building a Datacenter

Book: AI/ML Infrastructure from Silicon to Scale — V2 Edition | **Author:** Venkat Vinjam | **Cross-ref:**
Ch 4: Memory Hierarchy | Ch 7: GPU Kernels | Ch 12: Benchmarking | Ch 17: Telemetry

> *"Every watt you waste cooling hardware is a watt you didn't spend training a model. The most important equation in AI infrastructure isn't a matrix multiply — it's **Power × PUE × $/kWh × 8,760 hours**."*
>
> — Principal Infrastructure Architect's Cost Axiom

The H100 SXM5 has a Thermal Design Power (TDP) of 700W. A rack of 8 DGX H100 nodes draws 120+ kilowatts. A 10,000-GPU training cluster consumes power comparable to a small town. At this scale, thermal management is not a facilities concern — it is a **performance engineering concern**. Thermal throttling directly reduces MFU. Poor power allocation inflates TCO by millions of dollars per year.

LEARNING OBJECTIVES

FOUNDATIONAL

Understand why GPU heat is a performance problem. Learn what PUE means and why cooling choice affects throughput.

INTERMEDIATE

Apply thermal resistance formulas, DVFS mechanics, and power-capping tradeoffs. Build TCO models. Instrument clusters with DCGM.

PRINCIPAL LEVEL

Design a cooling strategy for 10,000+ GPUs from first principles. Model $/token as a function of power, throughput, and PUE. Make build/buy/rent decisions quantitatively.

5.1 Where Heat Comes From — Semiconductor Power Dissipation

5.1.1 The Three Sources of Power Dissipation

```
POWER DISSIPATION IN CMOS CIRCUITS:

P_total = P_dynamic + P_static + P_short_circuit

1. DYNAMIC POWER (60-75% of total):
   P_dynamic = alpha x C x V^2 x f
   Key: Power scales as V^2 -- reducing V by 10% cuts dynamic power by ~19%.
   This is the lever behind DVFS.

2. STATIC POWER / LEAKAGE (20-35%):
   P_static = V_dd x I_leakage
   At 4nm, leakage ~25% of total power even at zero workload.
   Idle GPUs draw 100-200W from leakage alone.

3. SHORT-CIRCUIT POWER (<5%):
   Brief PMOS+NMOS conduction during transitions. Not relevant to system engineers.
```

5.1.2 H100 Power Distribution

```
H100 SXM5 Power Distribution (700W TDP, estimated):

COMPONENT                       POWER     % of TDP
-----------------------------------------------------
Tensor Cores (132 SMs x 4 TC)   ~280W       40%
HBM3 Interface + Controllers     ~90W       13%
NVLink 4.0 PHY (18 links)        ~70W       10%
CUDA Cores + Special Function    ~70W       10%
L2 Cache + Memory Fabric         ~63W        9%
PCIe Gen5 Interface              ~35W        5%
NVSwitch + I/O Fabric            ~35W        5%
Misc (clock gen, power mgmt)     ~57W        8%
-----------------------------------------------------
TOTAL                            700W      100%

Hotspot: TC clusters nearest HBM stacks run 15-25 deg C above average die temp.
```

5.1.3 Thermal Throttling Hierarchy

Every GPU has a maximum junction temperature (T_j_max). For H100 SXM: **83°C operational**, 95°C absolute maximum. Throttling is not binary — it cascades:

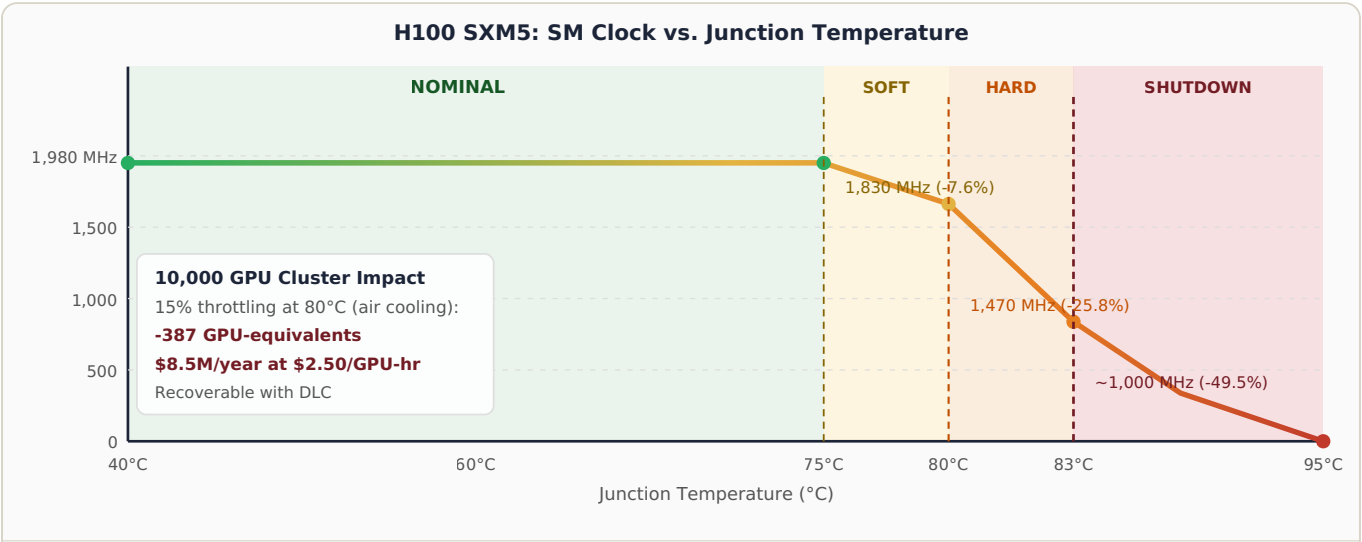


Figure 5.1 — H100 SM clock throttling cascade vs. junction temperature. Soft throttle onset at 75°C; hard throttle at 80°C; emergency shutdown at 83°C. At 10,000-GPU scale, 15% soft-throttling costs $8.5M/year in lost capacity.

> **⚠ The MFU Cost of Thermal Throttling**
>
> A 10,000-GPU cluster where 15% of GPUs soft-throttle at 25% clock reduction loses 387 GPU-equivalents — worth **$8.5M/year** at $2.50/GPU-hour. Entirely recoverable with proper cooling. Always check DCGM throttle counters before attributing MFU loss to software.

5.2 Cooling Technologies for AI Infrastructure

5.2.1 Thermal Resistance Chain

```
T_junction = T_ambient + (P_gpu x R_theta_total)

Air cooling  R_theta ~0.07 C/W: DeltaT = 700 x 0.07 = 49 C
  Required ambient <= 83 - 49 = 34 C  -- dangerously tight

DLC          R_theta ~0.03 C/W: DeltaT = 700 x 0.03 = 21 C
  Coolant inlet up to 62 C -- comfortable margin
```

5.2.2 Direct Liquid Cooling (DLC)

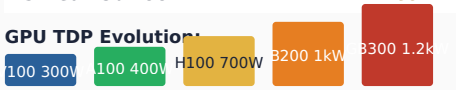
DLC brings coolant to GPU cold plates directly. With coolant inlet at 45°C: T_j = 45 + (700 × 0.02) = **59°C** — 24°C below soft-throttle. Supports TDP up to 1,500W+, enables 100–300 kW/rack density, allows waste heat reuse at 55–65°C outlet. CapEx premium: $50,000–$200,000+ per CDU.

5.2.3 Cooling Comparison

Cooling Technology Comparison Matrix

Metric	Air Cooling	Direct Liquid (DLC)	Single-Phase Imm.	Two-Phase Imm.
Max GPU TDP	500W marginal	1,500W	2,000W+	2,000W+
Rack Density	40 kW/rack	300 kW/rack	400 kW/rack	500 kW/rack
PUE Range	1.20 – 1.60	1.03 – 1.15	1.02 – 1.08	1.01 – 1.05
CapEx / kW	$1K–3K	$5K–20K	$15K–50K	$30K–80K
B200 (1,000W) Viable?	✗ No	✓ Yes (primary)	✓ Yes	✓ Yes
Maintenance Complexity	Low	Medium	High	Very High
5-Year Outlook	Declining	Dominant	Growing	Pilot Phase

GPU TDP Evolution:



← Air cooling insufficient above ~700W at rack density needed for AI
← DLC mandatory for B200/GB300 production; 8x DGX B200 = 240 kW/rack

Figure 5.2 — Cooling technology comparison matrix and GPU TDP trajectory. DLC is the clear choice for H100+ sustained workloads. B200 at 1,000W makes standard air cooling non-viable; rack density reaches 240 kW for a full 8-node DGX B200 configuration.

5.3 GPU Power Management — DVFS and Power States

5.3.1 DVFS — The 1ms Feedback Loop

```
DVFS FEEDBACK LOOP (every ~1ms on H100):
  1. Power sensor reads current draw
  2. Thermal sensor reads junction temp
  3. If current_power > limit OR T_j > threshold:
     Reduce V_core by ~5mV, f_core by ~13MHz  (--3% power per step)
  4. If headroom exists: increase f_core (boost)

Measured sustained SM clocks:
  Optimal DLC (25C coolant):  1,980 MHz  -- baseline
  Good DLC (35C coolant):     1,830 MHz  -- -7.6% MFU
  Air (well-ventilated):      1,650 MHz  -- -16.7% MFU
  Air (hot aisle, 50C inlet): 1,470 MHz  -- -25.8% MFU
  Hard throttle (83C T_j):   ~1,000 MHz  -- -49.5% MFU
```

> 💡 **Sustained Clock ≠ Advertised Boost Clock**
>
> H100 boost clock of 1,980 MHz is transient only. Under sustained 700W load, SM clocks stabilize at **1,600–1,800 MHz** depending on cooling — a 9–24% gap never captured in burst benchmarks. Design capacity plans from DCGM sustained measurements, not spec sheets.

5.3.2 P-States

```
P-State Hierarchy (H100):
  P0  Maximum Performance:  700W  | Full TDP, max clocks
  P1  High Performance:     500W  | ~80% max clocks
  P2  Balanced:             350W  | Moderate clocks
  P8  Low Power Idle:       125W  | Initialized, not computing
  P12 Deep Idle:             40W  | Maximum clock gating

P0 <-> P8 transition: 10-50ms latency
  Inference servers: lock at P0 to avoid TTFT spikes
  Fleet cost: 100-150W/GPU idle P0 vs P8 -- material at scale
```

5.3.3 Power Capping

```
H100 Performance-per-Watt (GEMM, approximate):
  500W --> 800 TFLOPS  = 1.60 TFLOPS/W  (+13.5% efficiency vs TDP)
  600W --> 910 TFLOPS  = 1.52 TFLOPS/W  (+7.8% vs TDP)
  700W --> 989 TFLOPS  = 1.41 TFLOPS/W  (baseline, worst efficiency)

Last 100W of TDP buys only ~9% more performance at 14.3% more power.
For $/token optimization: 500W may be more profitable than 700W TDP.
```

5.4 Datacenter Power Infrastructure

5.4.1 Power Delivery Chain

```
Utility (480V AC) --> UPS (96%) --> PDU (98%) --> PSU (94%) --> VRM (90%) --> GPU
Cumulative: 0.985 x 0.96 x 0.98 x 0.94 x 0.90 = 80.7% efficient

For every 1,000W at the GPU die: 1,239W from the utility grid.
239W of losses = heat that must also be cooled.
```

5.4.2 Power Usage Effectiveness (PUE)

```
PUE = Total Facility Power / IT Equipment Power

Example (10,000 H100 cluster):
  IT Power:    7,663 kW (GPUs + CPU + network + storage)
  Overhead:    1,027 kW (DLC + UPS losses + misc)
  PUE = 8,690 / 7,663 = 1.134

Benchmarks:
  1.01-1.05  World-class (Google Oregon with waste heat reuse)
  1.05-1.15  Excellent (modern hyperscaler with DLC)
  1.20-1.30  Good (modern air-cooled)
  1.40-1.60  Average enterprise datacenter
  >2.00      Legacy (half of IT power wasted)

Cost impact: every 0.10 PUE improvement saves ~$3.1M/year (10K GPU, $0.05/kWh)
```

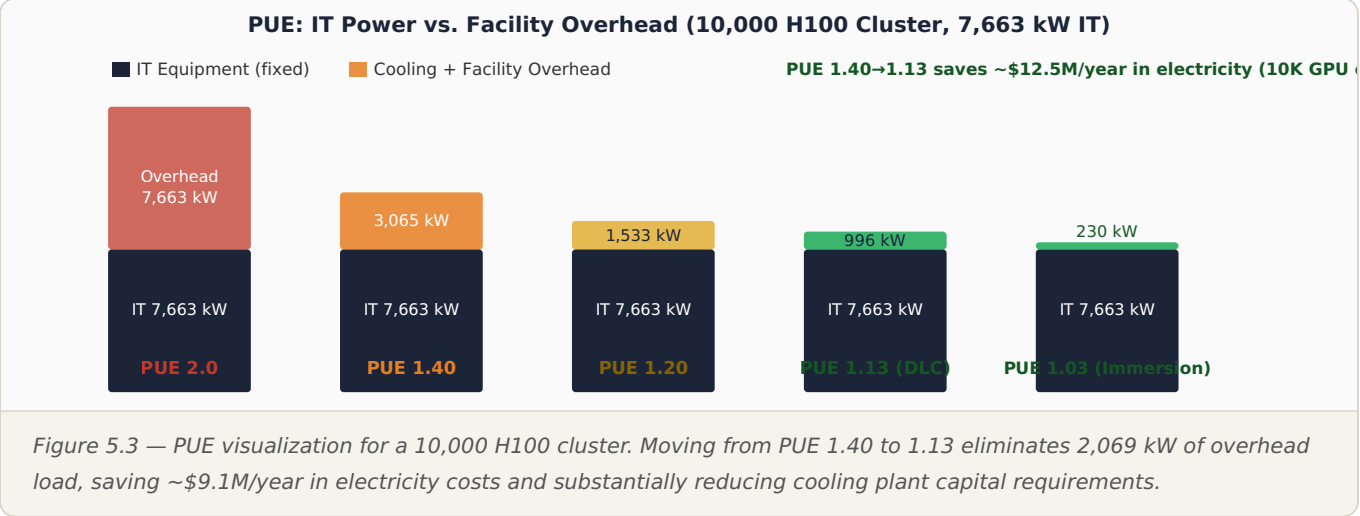


Figure 5.3 — PUE visualization for a 10,000 H100 cluster. Moving from PUE 1.40 to 1.13 eliminates 2,069 kW of overhead load, saving ~$9.1M/year in electricity costs and substantially reducing cooling plant capital requirements.

5.4.3 Rack Density and Power Infrastructure

Generation	GPU	TDP	DGX Node	8-Node Rack	Cooling Required
Pre-AI (2018)	V100 SHIPPED	300W	2.4 kW	19 kW	**Standard air**
First wave (2021)	A100 SHIPPED	400W	6.5 kW	52 kW	**High-density air**
Current (2023)	H100 SHIPPED	700W	10.2 kW	82 kW	**DLC recommended**
Next gen (2025)	B200 SHIPPED	1,000W	14.3 kW	115 kW	**DLC mandatory**
Future (2026+)	GB300 ROADMAP	~1,200W+	~18 kW	~144 kW	**DLC + specialized**

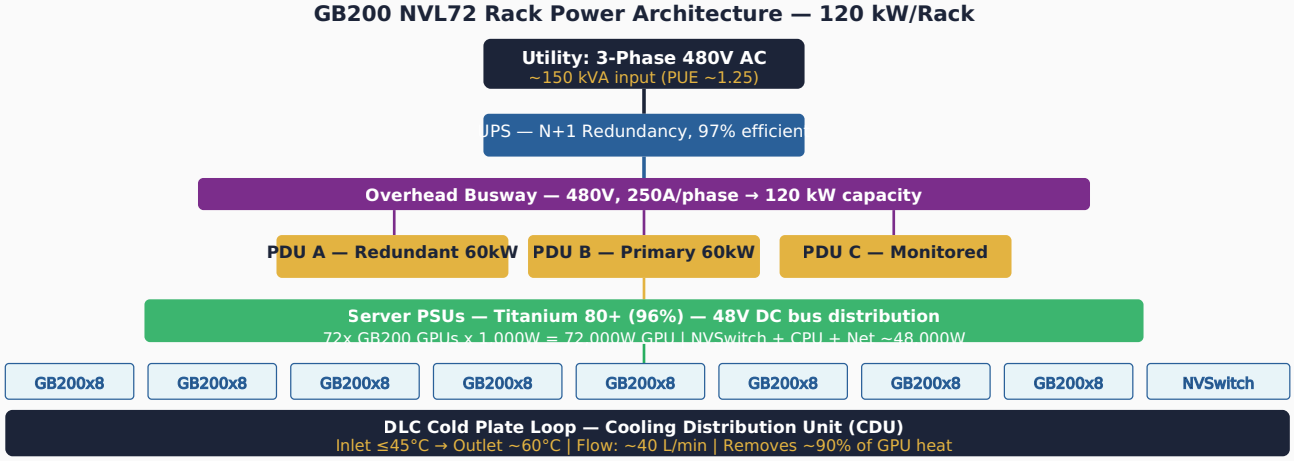


Figure 5.4 — GB200 NVL72 rack power architecture. 72 GB200 GPUs at 1,000W = 72 kW GPU power; total rack draw reaches 120 kW with NVSwitch, CPU, and networking. Requires 3-phase 480V circuits, overhead busway, and DLC infrastructure. Standard 30 kW enterprise circuits are completely inadequate.

5.5 TCO Modeling for AI Infrastructure

5.5.1 TCO Formula and 3-Year Breakdown

```
TCO = CapEx + (Power_kW x PUE x $/kWh x 8,760 x years) + OpEx_other

3-YEAR TCO -- 10,000 GPU H100 Cluster:

CapEx (3yr amortized):
  GPU hardware (~$30K avg):    $3.0B  -- dominates at 83% of TCO
  InfiniBand network:            $64M
  Storage:                       $20M
  DLC + DC buildout + power:     $95M
  Total CapEx:                ~$3.18B (amortized: ~$1.06B/year)

Annual OpEx:
  Power (7,663kW x 1.13 x 8,760 x $0.05):  $34.4M
  HW maintenance (3% CapEx):                $90.0M
  Staff (50 FTEs x $180K):                   $9.0M
  Software licenses + other:                $13.0M
  Total Annual OpEx:                       ~$146M

3-Year Total: $3.18B + ($146M x 3) = ~$3.62B

Cost/GPU-hour (80% util, 3yr):
  GPU-hours = 10,000 x 8,760 x 3 x 0.80 = 210M
  $/GPU-hr  = $3,620M / 210M = ~$17.24 (owned H100)

Compare: Cloud H100 on-demand = $3.50-4.50/GPU-hr
  Owned infrastructure only beats cloud at >65% sustained utilization
```

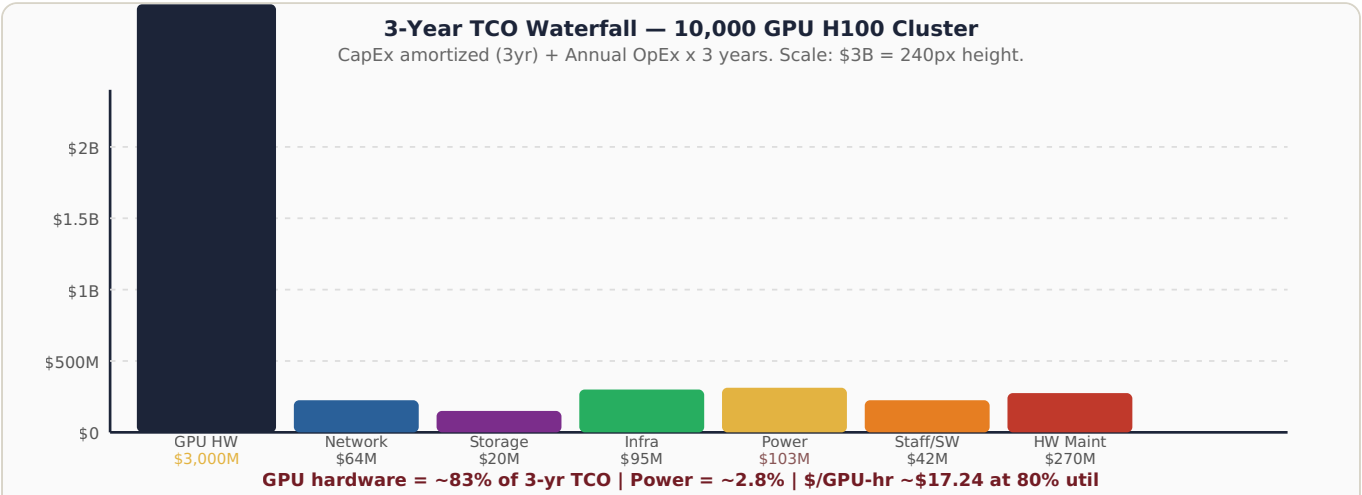


Figure 5.5 — 3-year TCO waterfall. GPU hardware overwhelmingly dominates at ~83% of total cost. Power, while $34M/year, is only ~2.8% of TCO at current H100 prices — but this ratio rises to 10–15% as GPU costs fall in future generations.

5.5.2 $/Token — The True Production Cost Metric

```
$/token = GPU_cost_per_hour / (tokens_per_second x 3,600)

Optimization impact (H100, LLaMA-3 70B, $0.05/kWh, PUE 1.13):

Configuration           Throughput    $/1M tokens  Saving
-----------------------------------------------------------
BF16 baseline           1,200 tok/s   $0.0092       --
FP8 quantization        2,000 tok/s   $0.0055      -40%
INT4 quantization       3,500 tok/s   $0.0032      -65%
+ Power cap 550W        3,500 tok/s   $0.0025      -73%
+ Flash Attention v3    4,200 tok/s   $0.0020      -78%

At 100B tokens/month, 10,000 GPUs:
  BF16:          $9.2B/month total power cost
  Fully opt.:    $2.0B/month total power cost
  Monthly saving: $7.2B -- entirely from software optimization
```

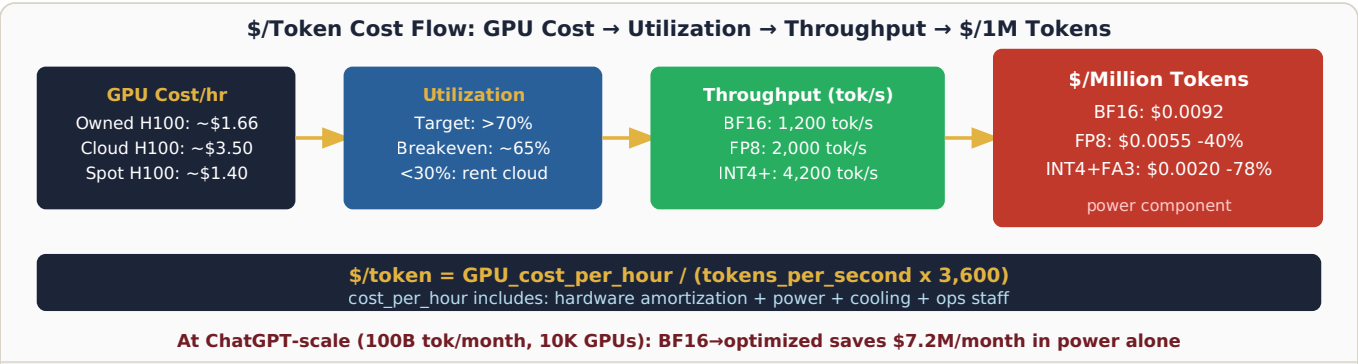


Figure 5.6 — $/token cost flow. The optimization stack (FP8 → INT4 → FlashAttention v3 → power cap) achieves a 78% cost reduction without changing hardware. At production scale, software optimization is the highest-ROI activity in AI infrastructure.

WORKED EXAMPLE 5.1

DLC vs. Air Cooling — True ROI Beyond Electricity Savings

Problem: Air (PUE 1.35) vs. DLC (PUE 1.10). DLC premium: $25M. At $0.05/kWh, when does DLC pay back — and what changes when MFU recovery is included?

```
Step 1: Annual electricity -- Air (PUE 1.35)
  Facility power = 7,663 kW x 1.35 = 10,345 kW
  Annual cost    = 10,345 x 8,760 x $0.05 = $4.53M/year

Step 2: Annual electricity -- DLC (PUE 1.10)
  Facility power = 7,663 kW x 1.10 = 8,429 kW
  Annual cost    = 8,429 x 8,760 x $0.05 = $3.69M/year

Step 3: Electricity savings = $4.53M - $3.69M = $0.84M/year
  Payback from electricity alone: $25M / $0.84M = 29.8 years  ← Not viable!

Step 4: Include MFU recovery (conservative: 15% GPUs throttle at 10% loss)
  Lost capacity = 1,500 GPUs x 10% = 150 GPU-equivalents
  At $2.50/GPU-hr: 150 x 8,760 x $2.50 = $3.29M/year recovered
  (Aggressive: 15% at 25.8% loss = $8.5M/year recovered)

Step 5: Combined payback
  Conservative: $25M / ($0.84M + $3.29M) = 6.1 years
  Moderate:     $25M / ($0.84M + $8.50M) = 2.7 years  GOOD ROI

Principal Lesson: DLC ROI is driven by performance recovery (MFU),
NOT electricity savings alone. Pitch DLC on recovered GPU-hours,
not on the power bill. The electricity saving is a rounding error.
```

5.5.3 Build vs. Buy vs. Rent

```
PROCUREMENT DECISION FRAMEWORK:

Step 1 -- Utilization:
  < 30%?   RENT cloud on-demand/spot
  30-60%?  HYBRID reserved cloud + spot
  > 60%?   BUY on-premise or colo

Step 2 -- Timeline:
  < 3 months: RENT (GPU lead times 6-18 months)
  > 6 months: BUY viable

Step 3 -- Breakeven:
  Cloud H100 (1yr reserved): ~$2.50-3.00/GPU-hr
  Owned H100 (80% util, 3yr): ~$1.50-2.50/GPU-hr
  Breakeven utilization: ~65-70%

Step 4 -- TCO risk factors:
  GPU cycle: new generation every ~2 years (vs 3yr depreciation)
  Power rate: $0.01/kWh change = ~$6.1M/year (10K cluster)
  Cooling retrofit: $1,000-5,000/kW for legacy facilities
  Ops: on-prem 40-80 FTEs vs cloud 10-20 FTEs
```

5.6 Thermal Monitoring — Profiling Cookbook

🔧 PROFILING COOKBOOK — THERMAL & POWER DIAGNOSTICS

```
WHAT TO MEASURE FIRST:
  DCGM field 203 (GPU_TEMP)        alert: >78 C
  DCGM field 204 (MEM_TEMP)        alert: >90 C
  DCGM field 155 (SM_CLOCK)        detects throttling in real time
  DCGM field 240 (THROTTLE_REASON) non-zero = active throttle event
  DCGM field 112 (POWER_DRAW)      compare to configured power limit

COMMANDS:
  # Quick check -- all GPUs
  nvidia-smi --query-gpu=index,temperature.gpu,power.draw,clocks.sm,
    clocks_throttle_reasons.hw_slowdown,clocks_throttle_reasons.sw_thermal_slowdown
    --format=csv,noheader

  # Find throttling GPUs immediately
  nvidia-smi --query-gpu=index,temperature.gpu,
    clocks_throttle_reasons.hw_slowdown,clocks_throttle_reasons.sw_thermal_slowdown,
    power.draw --format=csv,noheader |
    awk -F',' '$3=="Yes" || $4=="Yes" {print "THROTTLE:",$0}'

  # DCGM real-time monitor (production clusters)
  dcgmi dmon -e 203,204,155,156,112,240

  # Prometheus/Grafana exporter
  dcgm-exporter &
  # Exposes: DCGM_FI_DEV_GPU_TEMP, DCGM_FI_DEV_SM_CLOCK,
  #          DCGM_FI_DEV_POWER_USAGE, DCGM_FI_DEV_THROTTLE_REASON

  # Set power cap
  nvidia-smi --power-limit=600       # All GPUs
  nvidia-smi -i 3 --power-limit=550  # Single GPU index 3

DECISION TREE when GPU_TEMP > 78 C:
  Check inlet temp (IPMI):
    Inlet >35 C?  YES: Facility issue (CRAC/CDU failure, hot aisle containment)
                  NO:  Server issue (fan RPM, DLC flow rate, TIM degradation)
  Can resolve in <30 min?
    YES: Monitor at 5-min intervals
    NO:  IMMEDIATE: power-cap to 500W, drain workloads gracefully
         Training: trigger async checkpoint NOW
         Inference: remove from load balancer, reduce batch size

MFU drops 15-30% with no code change?
  Check SM_CLOCK via DCGM -- below expected?
  YES: check THROTTLE_REASONS (thermal or power limit)
  Check NVSwitch temp: >90 C causes fabric BW drop 30-50%
  Check NVLink error counters + board BMC temperature
```

5.7 Thermal Considerations in System Design

5.7.1 Thermal-Aware Job Scheduling

```
THERMAL-AWARE SCHEDULING (SLURM + DCGM):
  thermal_score(node) = 1.0 - (T_j - T_nominal) / (T_max - T_nominal)
  placement_score = 0.7 x utilization_score + 0.3 x thermal_score
  Avoid placement if T_j > 76 C
  Result: throttling events reduced ~40-60% across cluster
  (Meta MWFS, Google Borg, NVIDIA Base Command Manager)
```

5.7.2 Batch Size and Thermal Load

```
Decode phase thermal profile:
  Batch=1:    ~400-450W | T_j ~55-60 C | no throttling
  Batch=32:   ~550-600W | T_j ~68-72 C | mild risk
  Batch=128:  ~650-700W | T_j ~78-83 C | guaranteed throttle in air

Prefill is always compute-bound regardless of sequence length.
Power draws near TDP within 30-60 seconds of sustained prefill.

Thermal governor pattern:
  if T_j > 76 C: reduce max_batch_size by 20%
  Prevents throttle-thrash cycles, stabilizes production MFU.
```

5.8 The Green AI Imperative

```
ENERGY: Notable AI Training Runs
  GPT-3 (175B):    ~1K A100,  ~34 days,  ~1,060 MWh
  LLaMA-3 (405B): ~16K H100, ~56 days, ~50,200 MWh
  Frontier model: ~100K H100, ~90 days, ~504,000 MWh

504,000 MWh = annual electricity for ~47,000 US homes
CO2 (US grid 0.386 kg/kWh): frontier = 194,544 tonnes

Performance optimization IS sustainability optimization.
INT4 quantization: -65% $/token = -65% CO2/token (same metric, different units).
```

🌿 **Engineering Excellence = Sustainability**

Every throughput optimization halves the energy footprint at the same time it halves the cost. Performance engineering and sustainability engineering are the same discipline measured in different units. The engineer who achieves 2x throughput for the same hardware has also halved the carbon footprint.

⚡ POWER & THERMAL QUICK REFERENCE — KNOW THESE COLD

```
GPU TDP:
  H100 SXM5:  700W (SXM)  | 350W (PCIe)   [SHIPPED]
  H200 SXM:   700W (same TDP, 2x HBM BW)  [SHIPPED]
  B200 SXM:   1,000W -- DLC mandatory      [SHIPPED]
  MI300X AMD: 750W                         [SHIPPED]
  GB300:      ~1,200W+                     [ROADMAP]

Thermal Limits (H100):
  T_j operational max: 83 C | Soft throttle: ~75 C | Hard: ~80 C
  HBM max: ~95 C | HBM throttle onset: ~88 C

Cooling PUE:
  Air:          1.20-1.60 | Rear-door HX: 1.15-1.25
  DLC:          1.03-1.15 | Two-phase:    1.01-1.05

TCO Rules (10K H100, $0.05/kWh):
  Annual power cost:        ~$34M/yr
  GPU amortization (3yr):   ~$1.0B/yr
  Power % of TCO:           ~3% (rises 10-15% as GPU prices fall)
  Each 0.10 PUE improvement: ~$3.1M/yr savings
  $0.01/kWh rate change:     ~$6.1M/yr impact

Performance-per-Watt (LLaMA-3 70B FP8):
  A100: 2.75 tok/s/W | H100: 3.14 tok/s/W
  H200: 4.43 tok/s/W | B200: 6.80 tok/s/W

Power Capping Sweet Spots (H100):
  500W: 1.60 TFLOPS/W  +13.5% efficiency vs TDP
  600W: 1.52 TFLOPS/W   +7.8% efficiency vs TDP
  700W: 1.41 TFLOPS/W   baseline (worst efficiency)
```

Key Takeaways — Chapter 5

CORE CONCEPTS

- GPU power follows $P = \alpha C V^2 f$; DVFS controls V and f together — reducing V by 10% cuts dynamic power by ~19% (quadratic), not 10% (linear). This is the physics behind every power optimization.

- Thermal resistance governs sustainable performance: T_j = T_ambient + (P × R_theta). Air cooling at 700W leaves only a 1°C margin in a standard hot aisle — any ambient temperature spike triggers throttling.
- PUE = Total Facility Power / IT Power. Every 0.10 PUE improvement on a 10,000-GPU cluster saves ~$3.1M/year. DLC enables PUE 1.03–1.15 vs. air cooling's 1.20–1.60.
- H100 advertised boost clock (1,980 MHz) is transient. Sustained frequency under 700W load is 1,600–1,800 MHz — a 9–19% gap that burst benchmarks never capture. Always plan from sustained DCGM measurements.

COST & PRODUCTIVITY IMPACT

- 15% of GPUs throttling at 25% clock reduction on a 10,000-GPU cluster wastes 387 GPU-equivalents — $8.5M/year at $2.50/GPU-hr. DLC eliminates this entirely; the MFU recovery is the real DLC ROI, not the electricity bill.
- FP8 reduces $/token 40% vs. BF16; the full stack (INT4 + FlashAttention v3 + power cap) achieves 78% cost reduction — saving $7.2M/month per 10,000 GPUs at ChatGPT-scale with zero hardware changes.
- Power capping H100 at 500W vs. 700W improves TFLOPS/W by 13.5%. For cloud providers optimizing margin, lower absolute power may yield higher $/token efficiency despite lower raw throughput.

PRINCIPAL-LEVEL INSIGHT

- DLC payback from electricity savings alone takes ~30 years at $0.05/kWh. With MFU recovery included (conservative), payback drops to 2–6 years. Never pitch DLC on power savings — always pitch it on recovered GPU-hours and eliminated throttle events.
- The $/token formula is the correct optimization target, not raw TFLOPS. It integrates utilization, quantization, power state, and cooling into one metric that aligns hardware, software, and business objectives simultaneously.
- Owned vs. cloud breakeven (~65–70% utilization) is dominated by GPU amortization, not power. For frontier model companies at 90%+ utilization, owned infrastructure wins by 2–3× on $/GPU-hr. For product companies with bursty demand, cloud spot remains superior regardless of at-scale economics.

Review Questions

FOUNDATIONAL Build the mental model — no math required

1. Explain why a GPU running at 82°C junction temperature is a performance problem, not just a hardware reliability concern. What mechanism connects temperature directly to computational throughput?
2. A datacenter engineer says "our air-cooled facility has PUE 1.35, better than the industry average." Why does this statement, while accurate, miss the most critical constraint for a B200-class AI cluster?
3. Why does an H100 GPU still draw 100–150W at idle with no active workloads? Trace this back to the physics of CMOS transistors at 4nm process nodes.

INTERMEDIATE Apply and calculate — production context

1. H100 GPUs at batch=64 run at 700W and 82°C with occasional throttling (SM clock ~1,470 MHz). At batch=48 they run at 580W and 72°C with zero throttling but 15% lower throughput. Calculate which configuration has lower cost per token using chapter numbers at $0.05/kWh, PUE 1.13.
2. A 5,000-GPU H100 cluster has PUE 1.40 with air cooling. Calculate: (a) annual electricity cost at $0.06/kWh, (b) annual cost at PUE 1.10 with DLC, (c) years to recoup a $12M DLC installation from electricity savings alone, (d) GPU-equivalents lost if 20% of GPUs throttle to 1,470 MHz.
3. Using $T_j = T_{ambient} + (P \times R_{theta})$, with H100 SXM at R_{theta} 0.07°C/W, inlet air at 38°C, GPU at 700W: calculate T_j, identify the throttle zone, and state the expected SM clock and MFU impact from the chapter's measured data.

PRINCIPAL LEVEL Design, decide, and defend — architecture reviews

1. **System Design:** Architect a 20,000 B200 GPU cluster for frontier model training at $0.07/kWh. Design the complete thermal and power stack from first principles: rack layout, cooling technology, PDU architecture, substation sizing, and PUE target. Quantify the 5-year TCO delta between your optimal design and a "ship it fast" air-cooled alternative. Name the three largest risk factors and your mitigation for each.
2. **Tradeoff Analysis:** Your company owns 8,000 H100s on-prem in an air-cooled facility limited to 30 kW/rack at PUE 1.35. Evaluate: (a) upgrade to B200, (b) add more H100s, or (c) shift variable workloads to cloud spot. Build the quantitative decision framework: TCO per token for each option, cooling infrastructure requirements, timeline constraints, and the utilization rate at which each option dominates.
3. **Performance Incident:** During a 30-day LLaMA-4 pre-training run on 12,000 H100s, average MFU drops from 42% to 31% on Day 8. GPU temperatures are nominal (72–76°C). NVLink error counters are unchanged. SM clocks look stable. Power draw is 660W vs. expected 700W. Walk through your complete diagnostic process, name three likely root causes given these symptoms, and describe the specific DCGM metrics that would definitively distinguish between them.