2.5.1 Title, researchers and institution

קול קורא 34/2026 · משרד האנרגיה והתשתיות · יחידת המדען הראשי · מסלול אקדמיה

JouleFlex: A Model-Based Decision and Control Framework for Maximizing Useful AI Work under Dynamic Power Limits

JouleFlex: מודלים ובקרה חכמה למקסום תפוקת AI תחת מגבלות הספק משתנות

Principal Investigator (מנהל הפרויקט): Alexander Apartsin, Senior Lecturer, Department of Computer Science · **Institution:** Holon Institute of Technology (HIT)

2.5.2 Abstract / תקציר

English (300/300 words)

In July 2026 Israel's Electricity Authority froze new data-center connection requests after the queue reached 27,000 megawatts. Under electricity limits, data centers switch GPUs (AI processors) off or cap their power, deciding in watts (the momentary draw), not joules (the energy consumed). A capped GPU computes slower and pays its fixed power floor longer, so energy per unit of useful work can rise even as the meter shows fewer watts: less AI per kilowatt-hour. A similar challenge exists in aviation, where it is solved: fuel flow rises far faster than speed, yet slow flight burns fuel longer, so airlines cruise at the speed minimizing fuel per kilometer.

The PI's recent study [1] measured the time-power tradeoff for AI workloads (5,500 measurements, 20 models, four GPUs). Energy per task is U-shaped in the power cap. On the large GPUs its minimum, the Joule Point (JP), sits at 43 to 46 percent of rated power; holding a card there cut energy per task 29 to 31 percent while running 1.2 times slower. Even under a strict deadline, capping saved 28 percent of energy while keeping 95 percent of workload value. A fleet simulation on the measured curves (only scheduling simulated) served equal work with 18 to 45 percent less energy.

These insights change how AI data centers are planned, managed, and priced, raising decisions no current tool supports: how to design a flexible-connection-ready fleet whose cards run near their JP; how to measure and predict the JP for new cards and jobs; and how to price service to motivate customers toward efficiency, not per-hour billing rewarding the energy-worst point. Each is a modeling problem. This project delivers the decision-support toolkit answering all three, released with an open measurement dataset. Run near its JP, a fleet frees dispatchable power headroom for Israel's flexible-connection scheme.

עברית (300/300 מילים)

ביולי 2026 הקפיאה רשות החשמל בקשות חיבור חדשות של מרכזי נתונים לאחר שהתור הגיע ל-27,000 מגה-ואט. כאשר אספקת החשמל מוגבלת, מרכזי נתונים מכבים מעבדי GPU (מעבדי בינה מלאכותית) או קובעים להם תקרת הספק; ההחלטות מתקבלות בוואטים, ההספק הרגעי, ולא בג׳ולים, האנרגיה הנצרכת. מעבד GPU תחת תקרת הספק מחשב לאט יותר וצורך את הספק הבסיס הקבוע שלו זמן ממושך יותר; לכן האנרגיה ליחידת חישוב מועיל עלולה לעלות אף שהמונה מציג פחות ואטים: פחות בינה מלאכותית לכל קילוואט-שעה. אתגר דומה קיים בתעופה, ושם הוא נפתר: קצב צריכת הדלק עולה מהר בהרבה מן המהירות, אך טיסה איטית שורפת דלק זמן רב יותר; לכן חברות התעופה משייטות במהירות הממזערת את צריכת הדלק לקילומטר.

מחקר שערך לאחרונה החוקר הראשי [1] מדד את יחסי הגומלין בין זמן להספק בעומסי בינה מלאכותית (5,500 מדידות, 20 מודלים, ארבעה מעבדי GPU). האנרגיה למשימה מתארת עקומת U כתלות בתקרת ההספק. בכרטיסים הגדולים נקודת המינימום שלה, נקודת הג׳ול (Joule Point; JP), מצויה ב-43 עד 46 אחוזים מההספק הנקוב; החזקת כרטיס בנקודה זו הפחיתה את האנרגיה למשימה ב-29 עד 31 אחוזים בעוד זמן הריצה התארך פי 1.2. אף תחת מועד קשיח, הגבלת ההספק חסכה 28 אחוזים מהאנרגיה תוך שמירת 95 אחוזים מערך העבודה. סימולציית צי על העקומות שנמדדו (רק התזמון מדומה) סיפקה כמות עבודה שווה בצריכת אנרגיה נמוכה ב-18 עד 45 אחוזים.

תובנות אלה משנות את האופן שבו מתוכננים, מנוהלים ומתומחרים מרכזי נתונים לבינה מלאכותית, ומעלות החלטות שאף כלי קיים אינו תומך בהן: כיצד לתכנן צי מוכן-לחיבור-גמיש שכרטיסיו פועלים סמוך ל-JP; כיצד למדוד ולחזות את ה-JP עבור עבודות וכרטיסים חדשים; וכיצד לתמחר את השירות באופן המתמרץ לקוחות להתייעל, במקום החיוב השעתי המתגמל את נקודת ההפעלה הגרועה ביותר אנרגטית. כל אחת מהן היא בעיית מידול. הפרויקט יספק את ערכת התמיכה בהחלטות העונה על שלושתן, שתשוחרר עם מסד מדידות פתוח. צי הפועל סמוך ל-JP משחרר עתודת הספק זמינה לניצול עבור הסדר החיבור הגמיש של ישראל.

2.5.3 Glossary

Term	Meaning
Power	How fast electricity is drawn at a given moment; watts (W) or megawatts (MW). The grid connection is sized by peak power.
Energy	The total electricity consumed: power × time; joules or kilowatt-hours. The bill and the carbon footprint are set by energy.
Power cap	A software limit on a GPU's power draw. Set with one command; takes effect in ~0.2 seconds. The control knob of this project.
TDP	A GPU's full rated power. Running at TDP is today's default.
Joule Point (JP)	The power cap at which energy per task is lowest for a given card; measured at 43–46% of TDP on large GPUs.
Operating point	The pair (GPU, power cap) a job runs at.

Power envelope	A limit on a facility's total power draw that can change over time, set by a budget, a price or carbon signal, or a grid instruction.
Power shaping	Keeping a fleet inside a power envelope by capping jobs, like traffic shaping in networks.
ELF	Our released dataset: dense power-cap sweeps of 20 AI models on 4 GPUs, ~5,500 measured rows.
SLO	Service level objective: the deadline or latency a job must meet.
Firm / flexible capacity	Grid connection capacity guaranteed at all times, versus capacity given on condition that the facility reduces power when instructed.
Card multiplier	The extra cards needed to keep total throughput when capping. It equals the latency ratio (~1.2×).

2.5.4 Objectives of the program

2.5.4.1 The objective

The objective is to construct **JouleFlex**: a toolkit of models, software tools and a simulator for the energy management of GPU-rich AI data centers, built on the JP concept, in three layers. **Models:** the power-response law validated on current hardware and whole servers; a job model calibrating the energy-optimal cap per workload; a transfer model for unmeasured GPUs; and a joint job-and-cap assignment method. **Tools:** a characterization kit qualifying a new GPU within the validated hardware domain in hours, and a fleet control engine holding a facility inside its power envelope while meeting deadlines. **Simulation:** a digital twin replaying measured curves, evaluating any policy or connection scenario before deployment. All layers are released openly with the calibrating corpus (ELF-2, the extended measurement dataset), serving two users: the operator who runs it, and the Ministry that plans, regulates and sizes connections with its numbers. The twin answers the fleet-design question directly: how many cards, of which type, run near the JP behind a given connection.

2.5.4.2 The research question

Can AI data centers operate as controllable electrical loads, each GPU near its energy-optimal cap and the facility inside a changing power limit, so that energy per unit of AI work falls and the power capacity required to serve it shrinks, while admitted work keeps its service promises, under Israeli grid constraints?

Six sub-questions structure the work plan:

1. **Does the law generalize?** Do the power law and JP, measured on 20 models and 4 GPUs, hold on H100-class and, as access allows, B200-class hardware, production serving software, and shared GPUs?
2. **When is one cap per card enough?** Where does the near-constant per-card JP hold, and where must a job model replace it?
3. **Where is the true optimum at the wall?** What is the JP at node and server-input (wall) power, which the grid meters, not just the board?
4. **Can we skip the measurement?** Can an unmeasured GPU's JP be predicted from published characteristics plus at most one short test?
5. **Can allocation and power be optimized together, in motion?** Can job placement and per-card caps be chosen jointly under a time-varying power limit with every deadline

met? The measured result covers fixed caps and a fixed placement rule; the joint problem is open.

6. **How much flexibility can a facility sell, and who has the incentive?** How much power can be shed on instruction, how fast, for how long, at what cost, and which pricing rules make offering flexibility profitable?

2.5.5 The proposed research

2.5.5.1 Scientific and technological background

The Israeli challenge is specific and pressing. AI data centers are among the world's fastest-growing electricity consumers, approaching 945 TWh by 2030 [3], and Israel's grid is an isolated island system with no cross-border interconnections, so new demand must be met from domestic capacity; in July 2026 the Electricity Authority halted new data-center connection requests for 140 days, after the queue reached about 27,000 MW, roughly three times average national consumption [2]. The binding limit is grid capacity in megawatts: the national need is more AI work from every connected megawatt.

Today's operating decisions think in watts, not joules. Under stress a data center switches GPUs off, discarding service, or caps power energy-blind: a capped GPU runs slower and pays its fixed power floor longer, so capping too deep lowers the dashboard watts while raising the joules on the bill, and unconstrained the default is full rated power, where the last increment of speed carries a steep surcharge. A further choice goes unweighed: for a fixed workload, whether to concentrate it on fewer cards at full power or spread it across more at partial power, where our measurements show the second is sometimes more efficient.

Our discovery. The applicant's study [1] mapped, over 20 AI models and four GPU types (~5,500 instrumented measurements, released as the ELF dataset), how energy per task depends on the power cap: board power follows a simple physical law in the computing rate θ,

$$P(\theta) = P_0 + a\,\theta^{\wedge}\beta, \quad \text{median } R^2 = 0.99,$$

a fixed floor P_0 plus a superlinear term. Energy per task, $E = P/\theta$, is therefore U-shaped in the cap; its minimum, the **JP**, sits at 43 to 46 percent of rated power on the large GPUs, cutting energy per task by 29 to 31 percent at an exactly priced cost: tasks run about 1.2 times slower, and holding throughput takes that factor more cards. The cap is one software command, settling in about 0.2 seconds, fast enough to follow grid signals; behind a fixed feed the measured A100 curve implies about 50 percent more AI tasks per megawatt [1].

2.5.5.2 Review of existing knowledge

Four research lines meet here. *Energy-proportional computing* [4] identified the fixed, work-independent part of server power as the central inefficiency; our floor P_0 measures it for modern GPUs, and the JP is where paying it stops being worthwhile. *Speed-scaling theory* proved that energy is minimized at a sub-maximal speed [5,6,7]; the ELF dataset gives the measured form for GPU inference, where the law holds per operating point and breaks when pooled across batch sizes. *GPU energy systems*, notably Zeus [8] and the ML.ENERGY benchmarks [9,10], search for a good setting per job at runtime; the per-card-constant JP replaces that search with a one-time characterization. *Production context*: cluster studies [11] and public traces [12] supply realistic job arrivals; energy-aware GPU schedulers [13,14] leave each device's power draw at its default; methodology reviews [15] motivate our fixed protocol. Recent work power-caps GPU inference servers for energy savings [20] but reports a single recommended cap, not the optimum's measured location across many models and GPUs, and one analysis finds capping inert for memory-bound decode [23], a regime ELF's loaded-serving sweeps bound. Grid-interactive AI data centers are now demonstrated as grid assets [21], following a facility power target with priority-preserving curtailment; JouleFlex instead places jobs at the measured per-(card,job) energy optimum and quantifies the firm-plus-flexible capacity in MW, ramp and recovery.

2.5.5.3 Knowledge gaps

- **The law's reach is unmapped.** It is established for single jobs at fixed settings on three inference GPUs; its persistence on H100 and, as access allows, B200-class hardware, production serving software and shared cards is unknown, and placing the JP on a new GPU still requires a full sweep [1].
- **The job-dependence structure is unmapped.** The per-card constant holds in the measured loaded regime; outside it the optimum moves, and no model predicts a job's operating point from its features.
- **All published numbers are board numbers.** The grid meters the wall; host overhead (CPU, memory, cooling, power-supply losses) shifts the true optimum upward by an unmeasured amount, so no megawatt claim is yet grounded.
- **Allocation and power control are not optimized together in the cited work.** The measured 18 to 45 percent fleet saving used a fixed budget, a fixed placement rule and caps chosen afterward; the joint problem, coupling discrete scheduling to continuous control through the shared power budget, is unsolved.
- **Flexibility is unquantified.** How much power an AI facility can shed, how fast, for how long, at what cost has never been measured or standardized; Israel is rationing connections without those numbers.
- **The incentive problem is identified but unsolved.** Per-hour billing makes the renter's cheapest choice the energy-worst point, and a per-joule surcharge is too weak [1]; which

structural instruments work is open.

2.5.5.4 Description of the proposed research

JouleFlex builds one measured foundation and three tools on it: first work out the energy-optimal way to run each job on each card, then control power to achieve it. **WP1 and WP2 measure and predict**: they extend the dataset to current hardware, production software, and whole-server power, then turn those measurements into models that place any card's or job's efficient point without a full sweep. **WP3 controls**: a real-time engine that decides which job runs on which card and at what power cap, tested in a digital twin before any deployment. **WP4 and WP5 connect to the grid**: they define and measure the flexibility a controlled facility can offer, and study the pricing and connection rules that make efficient operation worthwhile. Everything is released openly.

2.5.5.5 Methods (work packages)

WP	Objective	Approach and output
WP1	Extend the energy-versus-power measurements to current GPUs, production software, and whole-server power.	Re-run the study's measurement sweep on new GPUs (H100/L40S class) and production serving engines (vLLM [16], TensorRT-LLM), reading power at three rented tiers: the GPU board (NVML), the host CPU domain (RAPL), and the server's power-supply input, the wall tier (BMC, read out-of-band via IPMI/Redfish). Accuracy is taken from vendor specifications and cross-checked across tiers; the JP is where the slope of P(θ) is zero, so a constant meter bias cancels in its location: the JP and per-card savings are calibration-robust, while absolute board-to-wall and megawatt figures carry that caveat. *Output:* the ELF-2 open dataset (board and server-input).
WP2	Predict any new card's or job's efficient operating point without a full measurement sweep.	A job model predicts a job's best power cap from its features; a transfer model predicts a new card's curve from its published specifications plus one short test. Both report error bars and a stated range of validity, and a short proof bounds the small penalty of using one cap per card; the transfer model is a calibrated-prior predictor anchored by one test per card, its error bars reflecting the small card count. *Output:* the characterization kit (models and one-test protocol).
WP3	Decide, in real time, which job runs on which card and at what power cap, to do the most useful work within a changing power limit.	A controller that makes both choices together, since moving a job changes the power left for the others, tested in a digital twin that replays the measured curves against exact baselines. When power is too tight it defers or rejects low-priority work and reports the cost. The joint problem is NP-hard; the per-card capping step is solved exactly. *Output:* the open control engine and digital twin.
WP4	Define what flexibility a controlled facility can offer the grid, and measure it.	Specify the contractable quantities, a committed power envelope, a flexibility offer, and an auditable delivery record, with ramp and recovery metrics; demonstrate them on a rented metered server and at fleet scale in the twin, under an Israeli firm-plus-flexible scenario. *Output:* the flexibility products and the Israel demonstration.

WP5	Find pricing and connection rules that reward efficient operation, and release everything openly.	An incentive model showing why per-hour billing rewards the energy-worst point and what changes fix it; a valuation of the flexible-connection option from public Noga and Electricity Authority data; each rule tested in the WP3 twin. *Output:* the policy report and the open release of the dataset, control engine, twin, and interfaces.

Success metrics are pre-registered per work package; the stage table in 2.5.7.1 states each threshold.

2.5.5.6 Novelty and originality

The measured law and the JP are established preliminary results [1]; the project's novelty is the frontier built on them:

- **AI load as a grid resource, grounded in the measured optimum.** Unlike grid-following curtailment to a power target [21], JouleFlex runs each card at its measured energy-optimal cap and makes the resulting flexibility contractable and audited, quantified in MW, ramp and recovery, the form Israel's connection challenge needs.
- **From per-job search to one-time measurement, with an exact price.** In the loaded regime, one per-card-model measurement replaces earlier per-job runtime search at under one percent penalty [1], and the extra cards to hold throughput equal the latency ratio, so every recommendation carries its exact capital cost.
- **Energy first, power as the means.** Inference power-capping saves energy at a chosen cap [20] and can be inert for memory-bound decode [23]; JouleFlex instead sets power to the computed energy optimum per (card, job) in the loaded regime where the cap binds, having measured the cost of universal full-power operation at about a third of inference energy.
- **From the board to the wall.** Published GPU energy results are board-power results; measuring the node's JP at server-input power moves the claim toward what the grid actually meters.
- **An open dataset with no equivalent.** WattGPU [22] predicts inference power at a single operating point and ML.ENERGY [9,10] reports one point; open power-cap sweeps cover training or HPC. The intersection, a dense power-cap sweep across many inference models and GPUs at the granularity to locate the energy-optimal point, does not exist. ELF fills it, and the project extends an asset that already exists rather than one it must first build.
- **The joint allocation-and-capping problem, to the measured optimum.** Placing jobs and setting every cap at the per-(card,job) energy optimum, coupled through the shared power budget, defines a distinct scheduling-control problem that grid-following

controllers [21] do not address; the JouleFlex engine is a first solution driven by the measured curves.

2.5.5.7 Fit to the Ministry's R&D goals

The roles divide cleanly: operators run the hardware, the Ministry sets the rules. Just as a regulator sets fuel-economy standards without driving the cars, JouleFlex gives the Ministry the measured basis for connection terms, planning coefficients and efficiency standards, without operating any facility itself. The proposal addresses the Ministry's 2026 R&D policy [17] under call 34/2026 [18], in particular high-priority clauses of Chapter 2:

Clause	How the project addresses it
16.2 · טכנולוגיות תכנון וניהול הספקת אנרגיה לחוות שרתים	The core deliverable: power shaping manages a server farm's energy supply in real time; the flexibility framework tells planners how much a farm needs.
6.2 · טכנולוגיות לניהול אנרגיה מקומי וניהול ביקושים	A facility following a power envelope on instruction, at ~0.2-second actuation, is a demand-response resource of grid-relevant size.
8 · התייעלות באנרגיה וביזור מערכות אנרגיה (priority 2026)	29–31 percent less energy per AI task at the card, 18–45 percent less per served job at the fleet, in software on installed hardware.
16.1 · שימוש בבינה מלאכותית ככלי עזר לפיתוח ומחקר בכל הנושאים לעיל (priority 2026)	AI compute is the subject, learned models and control the instruments, an energy-system tool the deliverable.
15.3 · פיתוח מתודולוגיה לבניית תחזיות ומודלים ופיתוח כלי עזר בקבלת החלטות	Flexibility products, compliance metrics and the regulator-facing report are decision-support instruments for connection policy.

2.5.5.8 Preliminary results

The preliminary results are the applicant's study *The Joule Point* [1] and its ELF dataset: about 5,500 instrumented measurements over 20 models and four GPUs establish the response law (median $R^2 = 0.99$), locate the JP at 43 to 46 percent of TDP with a 29 to 31 percent energy saving per task at 1.2 times longer runs, show one static cap per card serving all twenty workloads at under one percent mean penalty, and demonstrate an 18 to 45 percent fleet-level saving at equal-or-better deadline compliance in a measured-curve simulation, with about 50 percent more work per megawatt behind a fixed GPU power budget. Cap actuation settles in about 191 milliseconds. Under the per-instance-hour billing model analyzed in [1], the renter's cost optimum is the energy-worst operating point. ELF spans four GPU types; the fine sweeps behind the law and the JP cover three (A100, A10G, L4), with the T4 swept coarsely for coverage, and $R^2 = 0.99$ is the median per fixed-setting sweep while pooled per-card fits across workloads are 0.88 to 0.92 (Fig. A1). The benefit persists under tight service: in the strict-deadline experiment of [1], capping saved 28 percent of energy while keeping 95 percent of

workload value. The harness, simulator and analysis pipeline ran end to end; the project scales them out rather than building from zero.

2.5.6 Application: expected benefit and impact

2.5.6.1 What the project delivers, concretely

The project delivers five artifacts: the ELF-2 open dataset, the characterization kit, the control engine, the digital twin, and the grid interfaces and policy report. Appendix Table A1 states what each is and who uses it. Each artifact is released openly on delivery, every release crediting the Ministry's R&D program under call 34/2026 [18].

2.5.6.2 Benefit and impact

For the Israeli grid and regulator (the primary benefit). Under the July 2026 freeze [2], Israel faces a binary choice: deny AI-load growth, or overbuild firm generation to serve it. JouleFlex supplies a third path. The binding constraint is connection capacity in megawatts, and the JP converts efficiency into capacity: on the measured A100 curve, a fixed-feed calculation implies about 50 percent more AI work per megawatt of GPU power [1]. Every 1,000 MW of the 27,000 MW queue connected under JP operation could carry the AI work of up to 1,500 MW uncapped, a board-level figure WP1 and WP4 verify at server input and facility level before any coefficient is recommended. WP5 puts a shekel-per-megawatt value on the flexible-connection option from public Noga and Electricity Authority data, entering measured numbers into the Authority's open consultation [19] while the rules are written. With flexibility measured, committed and audited (WP4), the regulator gains a third option and the frozen queue stops being all-or-nothing. The same efficiency avoids about 0.3 TWh per TWh of AI workload, roughly 150,000 tonnes of CO_2 at 0.5 kg per kilowatt-hour.

Training and human capital (an explicit impact objective). Israel has little research capacity at the AI-energy interface, and operators and the regulator now both need it. It builds on three levels: one postdoctoral researcher trained across a stack no single Israeli group combines (GPU energy measurement with NVML/RAPL/BMC, ML-systems control, and energy-market analysis), employable by Israeli operators, Noga or the Authority; the open ELF-2 dataset, twin and control engine as teaching material at HIT, reusable by any institution; and the Hebrew M8 and M21 briefings plus a public release that carry the methods and numbers to Ministry, Authority and Noga staff and to operators. A trained specialist, an open national dataset and a better-equipped regulator outlast the grant.

For operators. On installed hardware, software control delivers 29 to 31 percent lower energy per task (measured) and 18 to 45 percent lower per served job (fleet simulation).

2.5.7 Work plan

2.5.7.1 Stages, deliverables, duration and success metrics

Months	Stage	Deliverable	Success metric
M1–M6	WP1: harness; new-generation and production-engine (vLLM) sweeps; board and node instrumentation	D1: ELF-2 v1	Law fit $R^2 \geq 0.95$ per sweep, below the measured 0.99 median [1], or deviation documented; JP located (or absence documented) per platform at board and node power
M5–M8	WP1: shared GPUs, LLM phases, fine-tuning; sweeps where the constant should break	MS1: regime map	Server-input (BMC) wall tier validated; validity domain documented, including the pre-registered JP shift (percentage points of TDP) at which the per-card constant breaks; an M8 technical brief to the chief scientist unit
M7–M12	WP2: job model; transfer model; one-test characterization protocol	MS2: models result	Job-model placement within 5 percentage points of TDP where the constant breaks; transfer-model placement within 5 percentage points of TDP on held-out cards; protocol cost ≤ 8 hours and ≤ ₪1,500 per card model
M9–M16	WP3: twin; joint controller under fixed, price, carbon and curtailment envelopes	MS3: controller result	Energy per served job and SLO compliance versus uncapped, static-cap and cap-only baselines; a pre-registered floor of at least 10 percent below cap-only at equal compliance (the incremental gain over an already-capped fleet on held-out traces, not the saving over full power); zero exceedance beyond the WP1 noise band in committed windows, with overshoot and transient violations (under 1 percent of epochs) reported
M14–M20	WP4: flexibility products and metrics; demonstration on rented BMC-metered servers; Israel scenario	MS4: flexibility demonstration	Sheddable power as a fraction of the committed envelope (pre-registered floor: at least 15 percent within the ramp deadline at bounded SLO loss), measured on a BMC-metered server, projected to fleet scale in the twin (projection error versus the metered runs reported); ramp within 60 seconds and recovery within 5 minutes, within a grid flexibility product's response window [19]; delivery-record compliance and curtailment service cost
M16–M21	WP5: incentive model; connection valuation on published Israeli data; twin evaluation	MS5: policy report draft	Alignment condition per billing structure, confirmed in the twin; option value per MW and break-even curtailment frequency, with sensitivity to every price input
M21–M24	WP5: release, documentation,	D2: open release; D3:	Public repositories with reproducing pipelines; metered validation; final scientific and policy

	validation, publication	final report	reports

Two interim briefings are named deliverables: the M8 technical brief and the M21 policy draft, in Hebrew, circulable to the Authority and Noga [19]. Each milestone is tested against pre-stated criteria; a bounding result is itself a deliverable. Minimum-success path: H100/L40S validation, one serving engine, BMC wall measurement, the job and transfer models, the joint controller, and one Israeli flexibility scenario; B200-class cards, a second engine, fine-tuning, shared-GPU regimes and the formal guarantee are extension objectives as access allows.

2.5.7.2 Budget by stage

Stage	Share	Amount (₪)	Principal costs
M1–M8 (WP1; WP2 begins)	30%	148,314	Postdoctoral researcher; GPU rentals for cap sweeps
M9–M16 (WP2 ends; WP3; WP4 begins)	33%	163,145	Postdoctoral researcher; card-model rentals for transfer validation; twin and controller work
M17–M21 (WP4 ends; WP5)	25%	123,595	Postdoctoral researcher; BMC-metered rentals; flexibility campaign
M22–M24 (WP5: release and validation)	12%	59,326	Postdoctoral researcher; metered validation runs; documentation; dissemination and publication
Total		**494,380**	Full itemization in the מפרט תקציבי

2.5.7.3 Risks and mitigation

Risk	Impact	Mitigation
Production serving software (continuous batching, shifting phases) blurs the per-card constant	Static cap loses near-optimality for LLMs	Measured first (WP1, M1); the job model predicts the optimum where the constant breaks
Power-cap control unavailable on some rented platforms	Measurement campaign narrowed	AWS EC2 gives passthrough and root; too-high cap floors are exposed via the graphics clock (L4 [1]); BMC rentals cover the wall tier
Postdoc recruitment slips, or the single hire is lost mid-project	Schedule and delivery at risk	PI runs the WP1 bench directly (the ELF harness is his); the minimum-success path is one-FTE executable; recruitment via HIT and national channels with a fallback
Out-of-band PSU telemetry (BMC/IPMI/Redfish) blocked on a rented server	Wall-tier measurement narrowed	Named BMC-capable bare-metal vendors with dated quotes (GPU Mart and equivalents); fallback to an in-line metered PDU at the wall

2.5.8 The applicant

Project manager and principal investigator: Alexander Apartsin, Senior Lecturer, Computer Science, HIT; single-PI, responsible for all five work packages, committing about one day per week (roughly 20 percent) across the 24 months and drawing no salary from the grant (§2.6).

Fit of expertise to the proposed work. The scientific foundation is the PI's own study [1]: the ELF dataset, law, JP and fleet result. The work plan maps onto that record across all five packages, and the harness, models and simulation are built and validated. Beyond the study, the PI brings three decades in industry and academia: capacity and cost engineering, including system-sizing that saved a large operator about $10M; research and patents in computationally efficient computing (CalexNet); and graduate teaching at HIT, Afeka, Bar-Ilan and Tel-Aviv University.

Research personnel. One postdoctoral researcher at 90% position, both years, carries the measurement, modeling and control work under the PI's supervision, reported within a month of signature (נספח טז' §2.5). Core profile required at hiring: GPU systems and energy instrumentation (NVML/RAPL/BMC), machine-learning systems, and optimization or control. The grid-flexibility and techno-economic analysis of WP4 and WP5 is public-data modeling on Noga and Electricity Authority figures, developed on the project under the PI's direction, and needs no power-systems specialist. Recruitment runs through HIT and national postdoctoral channels; if the month-1 start slips the PI runs the WP1 bench directly (the ELF harness is his), so the minimum-success path holds.

Advisor. A senior expert with long experience in energy systems, energy efficiency and energy-systems management will advise the facility-level energy measurement and grid-flexibility work (WP4) and the techno-economic analysis (WP5), adding an energy-systems perspective to the team's computing and ML expertise. The advisor will be named before signature.

Resources. The project owns no hardware: board sweeps on AWS EC2 (root), the CPU-domain tier on AWS bare metal, and the wall tier and WP4 demonstration on hourly BMC-metered dedicated servers. HIT provides laboratory space, computing and grant administration.

2.5.9 Funding

No other funding exists or is applied for; per נספח טז' §1.5.2 and §2.10, no complementary funding will be received and no student receives parallel state funding.

2.5.10 Bibliography

1. Apartsin, A. and Aperstein, Y. (2026). *The Joule Point: an Energy-Optimal Operating Point for AI Inference*. Preprint, September 2026. DOI: 10.13140/RG.2.2.20385.36964 (CC BY 4.0), with the released ELF dataset.
2. JNS (22 July 2026). *Israel halts new data-center grid requests for 140 days amid surge in demand*. Press report on the Electricity Authority's temporary order: pending applications of approximately 27,000 MW, about three times Israel's average electricity consumption. https://www.jns.org/news/israel-news/israel-halts-new-data-center-grid-requests-for-140-days-amid-surge-in-demand (the Authority's order is published on gov.il).
3. International Energy Agency (2025). *Energy and AI: Energy Demand from AI*. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
4. Barroso, L. A. and Hölzle, U. (2007). *The Case for Energy-Proportional Computing*. IEEE Computer 40(12), pp. 33–37.
5. Yao, F., Demers, A., and Shenker, S. (1995). *A Scheduling Model for Reduced CPU Energy*. Proceedings of FOCS 1995, pp. 374–382.
6. Miyoshi, A., Lefurgy, C., Van Hensbergen, E., Rajamony, R., and Rajkumar, R. (2002). *Critical Power Slope: Understanding the Runtime Effects of Frequency Scaling*. Proceedings of ICS 2002, pp. 35–44.
7. Gandhi, A., Harchol-Balter, M., Das, R., and Lefurgy, C. (2009). *Optimal Power Allocation in Server Farms*. Proceedings of ACM SIGMETRICS 2009, pp. 157–168.
8. You, J., Chung, J.-W., and Chowdhury, M. (2022). *Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training*. arXiv: 2208.06102. https://arxiv.org/abs/2208.06102
9. Chung, J.-W., Ma, J. J., Wu, R., et al. (2025). *The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization*. arXiv: 2505.06371. https://arxiv.org/abs/2505.06371
10. Chung, J.-W., Wu, R., Ma, J. J., et al. (2026). *Where Do the Joules Go? Diagnosing Inference Energy Consumption*. arXiv: 2601.22076. https://arxiv.org/abs/2601.22076
11. Hu, Q., Sun, P., Yan, S., Wen, Y., et al. (2021). *Characterization and Prediction of Deep Learning Workloads in Large-Scale GPU Datacenters*. arXiv: 2109.01313. https://arxiv.org/abs/2109.01313
12. Alibaba Group (2020). *cluster-trace-gpu-v2020: production GPU cluster traces from Alibaba PAI*. https://github.com/alibaba/clusterdata
13. Lipe, E., Karia, N., Espenshade, C., Stein, C., Tantawi, A., and Tardieu, O. (2026). *Energy Efficient Scheduling of AI/ML Workloads on Multi-Instance GPUs with Dynamic Repartitioning*. arXiv: 2606.25082. https://arxiv.org/abs/2606.25082
14. Lu, T. and Reda, S. (2026). *Agentic CPU-GPU Scheduling for Heterogeneous AI Workloads*. arXiv: 2607.22242. https://arxiv.org/abs/2607.22242
15. Rodriguez, C., Degioanni, L., Kameni, L., Vidal, R., et al. (2024). *Evaluating the Energy Consumption of Machine Learning: Systematic Literature Review and Experiments*. arXiv: 2408.15128. https://arxiv.org/abs/2408.15128
16. Kwon, W., Li, Z., Zhuang, S., et al. (2023). *Efficient Memory Management for Large Language Model Serving with PagedAttention*. Proceedings of SOSP 2023. arXiv: 2309.06180. https://arxiv.org/abs/2309.06180
17. Ministry of Energy and Infrastructure (2026). *מדיניות התמיכה במחקר ופיתוח לשנת 2026*. Chapter 1 clauses 6, 8, 15, 16; Chapter 2 priority list.

18. Ministry of Energy and Infrastructure (2026). *קול קורא 34/2026 למימון מחקרים בתחום האנרגיה ומדעי האדמה והים.*

19. Calcalist (June 2026). Report on the Electricity Authority's public hearing (שימוע): a dedicated connection track for transmission-grid consumers of 50 MVA and above, including a system-operator right to curtail consumption at peak and payments for holding connection capacity. Under public consultation; cited as regulatory direction, not settled regulation. https://www.calcalist.co.il/calcalistech/article/rylkyxemgx (the hearing documents are published on gov.il).

20. Ma, Y., Srinivasan, S., and Wang, X. (2025). *Power Capping of GPU Servers for Machine Learning Inference Optimization.* Proceedings of ICPP 2025 (54th International Conference on Parallel Processing). DOI: 10.1145/3754598.3754670.

21. Colangelo, P., Coskun, A. K., Megrue, J., et al. (2025). *AI data centres as grid-interactive assets.* Nature Energy. DOI: 10.1038/s41560-025-01927-1; field demonstration in Phoenix, AZ, arXiv: 2507.00909.

22. Fadel Argerich, M., Fürst, J., and Patiño-Martínez, M. (2026). *WattGPU: Predicting Inference Power and Latency on Unseen GPUs and LLMs.* arXiv: 2607.02391. https://arxiv.org/abs/2607.02391

23. Ma, B., Afzal, A., Eitzinger, J., and Wellein, G. (2026). *The Illusion of Power Capping in LLM Decode: A Phase-Aware Energy Characterisation.* arXiv: 2605.11999. https://arxiv.org/abs/2605.11999

Appendix A: Figures (within the 5-page allowance of 2.2§ נספח א׳1)

JouleFlex · קול קורא 34/2026 · Appendix to the research proposal, permitted under נספח א׳1 §2.2 (graphs, tables and diagrams, up to 5 A4 pages, outside the 10-page limit). All figures are measured results from the applicant's study [1].

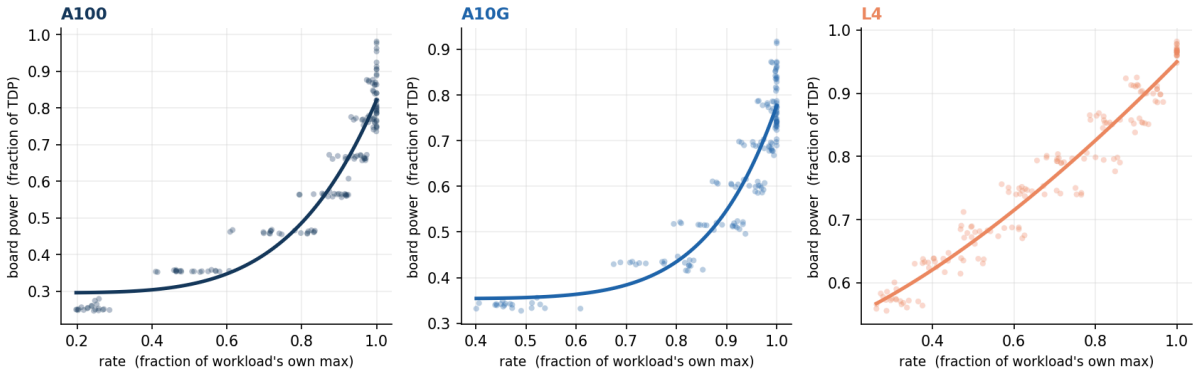

Figure A1. The response law is a card property. One panel per GPU (A100, A10G, L4); each dot is a measured operating point for one of the 20 workloads in the loaded regime, rate normalized to the workload's own maximum and power to the card's TDP. All workloads follow a single fitted law per card, $P(\theta) = P_0 + a \cdot \theta^{\beta}$ (pooled per-card R^2 of 0.92, 0.88 and higher per-sweep), showing the exponent travels with the card rather than the model. This is the empirical basis for the one-time per-card characterization that WP2 turns into a transfer model.

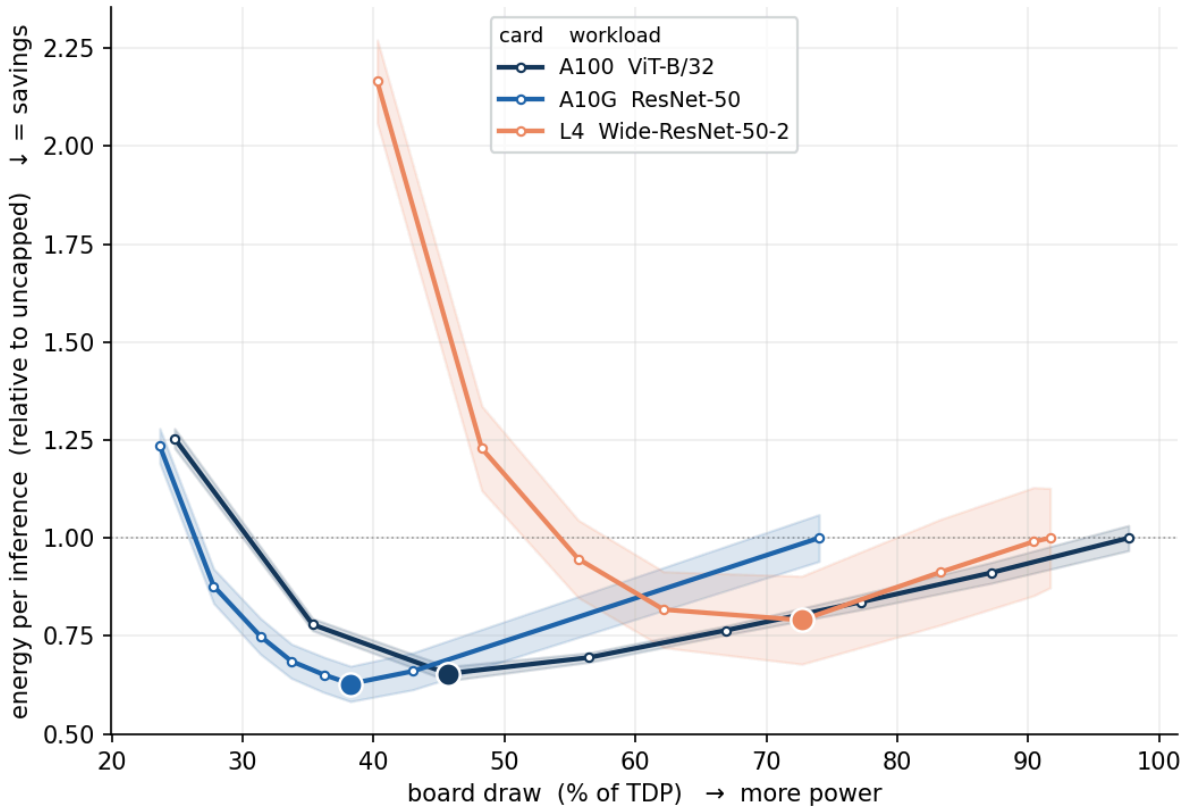

Figure A2. Every card has an energy U-shape. Energy per inference, normalized to each card's own uncapped value, against board draw as a fraction of TDP; the dot marks the JP, the shaded band one standard deviation over repetitions. On the large GPUs the minimum sits at 43–46 percent of TDP and capping to it cuts energy per inference by 29–31 percent; on the L4 the interior minimum lies below the power cap's reachable floor and is exposed by the graphics clock instead, mapping which actuator applies to which card class.

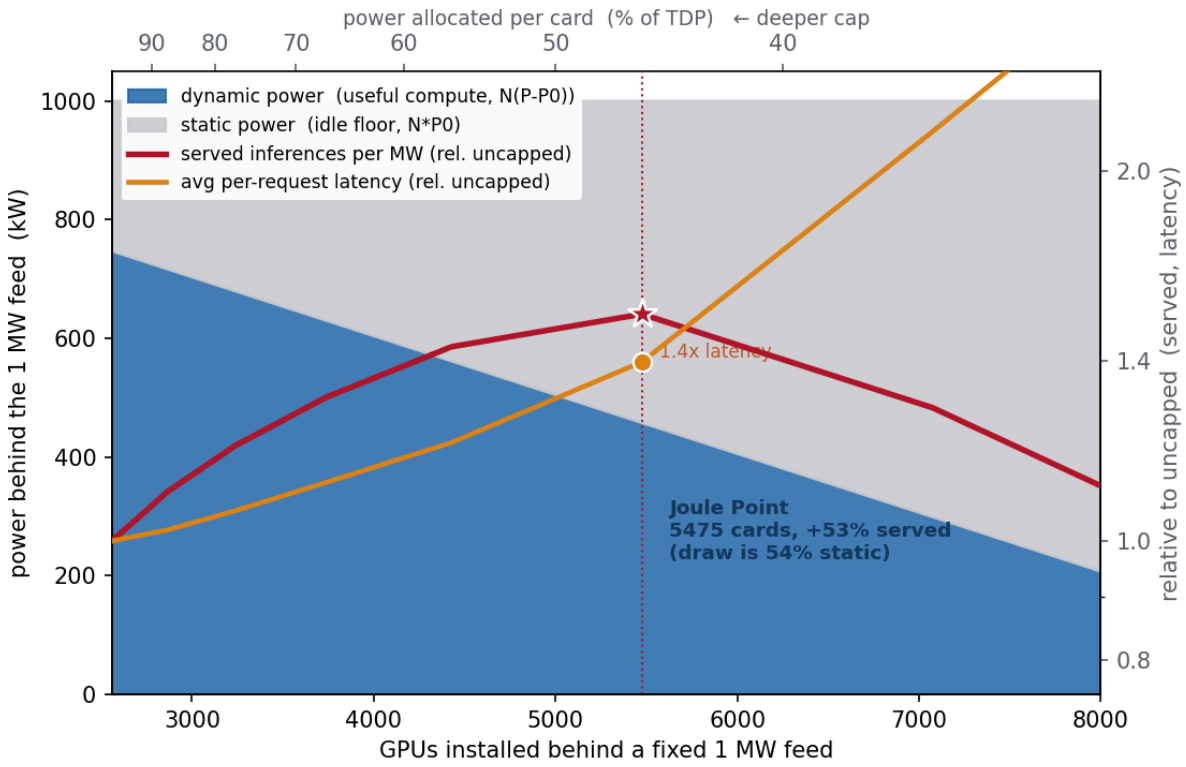

Figure A3. When power, not capital, binds, the JP is the interior optimum. Behind a fixed 1 MW feed (A100, ViT-B/32 at batch 64), installing more GPUs forces each to a deeper cap, shifting the megawatt from useful dynamic compute (blue) to the static idle floor (grey). Served inferences per megawatt (red) peak at the JP, near 5,500 installed cards, about 50 percent above the uncapped configuration. This is the capacity lever the Israel demonstration (WP4) operationalizes: under a rationed connection, efficiency becomes throughput. The same measured curves, allocated exactly across a fleet, dominate uncapped operation at every power budget from 100 down to 40 percent of total TDP; the preliminary measured-curve fleet simulation result appears in section 2.5.5.8.

Table A1. The five deliverables: what each is, who uses it

The project delivers five named artifacts, each with a defined user and use:

Deliverable	What it is	Who uses it, for what
ELF-2	An open dataset: power, throughput and latency at every operating point, board to wall, on current GPUs and production serving software	Researchers and vendors, as ground truth; every other deliverable is built and validated on it
The characterization kit	Software plus protocol: a one-test procedure and the two fitted models placing the JP of a new or unmeasured card within the validated hardware domain, in hours	A GPU engineer or operator qualifying a new card model or a new serving stack before deployment, at known cost
The control engine	An open-source control engine that assigns jobs and sets caps to keep a fleet inside its envelope, evaluated in the twin and demonstrated on a metered rented server before any deployment	An AI data-center manager, to cut energy per job and honor a committed envelope; its report doubles as the compliance record
The digital twin	A simulator replaying the measured curves of a described fleet and job mix, evaluating any policy or envelope scenario exactly before hardware is touched	An operator testing policies before deployment; a planner or regulator asking what a proposed facility could deliver and shed under a given connection
The grid interface and policy report	A blueprint: open formats for the envelope commitment, the flexibility offer and the delivery record, plus a regulator-facing report quantifying firm-plus-flexible connections against the national queue	The Electricity Authority and Noga, as the technical basis for offering conditional connections and auditing compliance

Table A1. The chain of use is direct: ELF-2 calibrates the models, the models feed the controller, the twin proves the controller before deployment, and the grid interface turns the controlled facility into something the regulator can contract with.