International Journal of Engineering Research and Reviews ISSN 2348-697X (Online)
Vol. 14, Issue 2, pp: (10-16), Month: April - June 2026, Available at: www.researchpublish.com

Desk Power and Concurrent Human-Speed Workflows: A Dual-Metric Framework for AI Infrastructure Decision-Making

Eng. Nawaf F. Al-Mutairi

Institute of Industrial Studies (IIS), PAAET, Kuwait

Department of General Studies

DOI: https://doi.org/10.5281/zenodo.20408317

Published Date: 27-May-2026

Abstract: **Organizations deploying AI inference systems for knowledge-work automation currently lack a shared, intuitive vocabulary for capacity planning and procurement. Existing metrics — tokens per second, FLOPS, benchmark scores — carry precise technical meaning but fail to communicate operational impact to the managers and executives who make infrastructure investment decisions. This paper introduces a dual-metric framework that addresses this gap. The first metric, Desk Power (DP), quantifies the economic replacement capacity of an AI server in terms of equivalent human office workers, drawing explicit analogy to James Watt's mechanical horsepower. The second metric, Concurrent Human-Speed Workflows (CHW), quantifies the number of simultaneous users or workflow streams a given system can serve at or above a specified responsiveness threshold. Together, DP and CHW answer the two questions that drive real procurement decisions: how many employees can this system replace (DP), and how many employees can it serve at once (CHW). A formal mathematical framework is developed for both metrics, incorporating token generation rate, system availability, parallelism, quality correction, and response-speed thresholds. Worked examples are provided for systems ranging from local consumer-GPU edge nodes to enterprise multi-GPU clusters. The proposed framework offers a vendor-neutral, audience-appropriate language for AI capacity planning, workforce transition analysis, and regulatory impact assessment.**

Keywords: **AI capacity measurement, large language models, office automation, workforce displacement, inference throughput, concurrent users, server sizing, decision-making framework.**

1. INTRODUCTION

The deployment of large language model (LLM) systems capable of performing complex knowledge-work tasks — drafting correspondence, editing spreadsheets, synthesizing reports, scheduling, and data entry — has created an urgent practical problem: the people who decide whether and how to deploy such systems cannot meaningfully interpret the metrics by which those systems are characterized. Tokens per second, parameter counts, and benchmark scores are technically rigorous but organizationally opaque. A procurement officer approving a six-figure server purchase, a workforce planner modeling a departmental restructure, or a regulator assessing labor-market impact has no framework within which these numbers become actionable.

This paper argues that two distinct questions drive AI infrastructure decisions at the organizational level, and that these questions require two distinct metrics. The first question is strategic and economic: given the cost of deploying a specific AI system, how many human employees does it effectively replace? This is the return-on-investment question, addressed by Desk Power (DP). The second question is operational and technical: given a specific workforce size, how many employees can a given system serve simultaneously at an acceptable response speed? This is the server-sizing question, addressed by Concurrent Human-Speed Workflows (CHW).

Neither question is currently answerable using existing standardized metrics. This paper fills that gap by developing a formal, unified framework grounded in empirically calibrated human baseline parameters. The analogy to James Watt's mechanical horsepower — which solved an identical communication problem during the first industrial transition — provides both methodological guidance and historical validation for the approach.

The paper is structured as follows. Section 2 reviews the horsepower precedent. Section 3 establishes the human baseline. Section 4 develops the Desk Power metric. Section 5 develops the CHW metric and introduces response-speed thresholds. Section 6 demonstrates how the two metrics function as a complementary pair for the CFO and the IT manager respectively. Section 7 provides worked examples and a comparative table. Section 8 discusses limitations. Section 9 concludes with recommendations for standardization.

2. HISTORICAL PRECEDENT: MECHANICAL HORSEPOWER

James Watt introduced the concept of horsepower circa 1782 to provide mill and factory owners with a frame of reference immediately relatable to their existing operations [1]. By observing a mill horse's work rate and translating steam engine output into an equivalent number of horses, Watt created a single number that made the purchasing decision trivial: if a 10-horsepower engine costs less to operate than 10 horses, the decision is self-evident.

Three methodological choices in Watt's design are directly instructive here. First, he identified the correct audience (the buyer, not the engineer) and designed the metric for their decision context. Second, he deliberately calibrated the baseline conservatively — overstating a horse's actual output by approximately 50% — so that engines would reliably outperform their rated capacity in practice [2]. Third, the unit explicitly referenced the labor category it replaced, embedding economic meaning directly into the metric's name.

The present framework adopts all three principles. DP and CHW are designed for the buyer's decision context, not the engineer's benchmarking context. The human baseline is calibrated conservatively. And both units explicitly reference the human workforce being characterized.

3. ESTABLISHING THE HUMAN BASELINE

3.1 Effective Token Production Rate

In the LLM context, productive output is naturally measured in tokens. Research consistently places average office typing speed at 40–60 words per minute (WPM) under optimal conditions [3]. At the standard approximation of 1.33 tokens per word, this yields a peak token production rate of approximately 0.9–1.3 tok/s. However, empirical time-motion studies of knowledge workers indicate that active writing and data-entry occupies only 15–25% of total working time [4]. Applying a utilization factor of 0.20 to a midpoint rate of 1.0 tok/s yields a sustained effective rate of approximately 0.15–0.20 tok/s.

Following Watt's conservative calibration principle, this paper adopts Th = 0.15 tok/s as the human effective token baseline. This ensures that DP ratings reliably understate AI capability relative to the human benchmark.

3.2 Availability and Parallelism

A standard office worker is available for approximately 8 hours per day, 5 days per week, adjusted for leave and interruptions, yielding a normalized daily availability factor of Ah = 0.283 (8 hours × 0.85 uptime factor, expressed as a fraction of 24 hours). An office worker handles one primary task at a time (Ph = 1). These parameters are summarized in Table 1.

Table 1: Human Baseline Parameters

Parameter	Symbol	Value	Basis
Effective token rate	Th	0.15 tok/s	Peak rate × utilization factor (conservative)
Daily availability	Ah	0.283	8 hr × 0.85 uptime ÷ 24 hr
Task accuracy	Qh	0.95	Typical knowledge-worker error rate
Parallelism	Ph	1	One primary task at a time

4. METRIC 1: DESK POWER (DP)

4.1 Definition and Formula

Desk Power (DP) is defined as the ratio of effective AI knowledge-work throughput to the effective throughput of the human office worker baseline, capturing the economic replacement capacity of an AI system operating continuously.

$$\mathbf{DP = (T_{ai} \times A_{ai} \times P \times Q) \div (T\mathit{h} \times A\mathit{h})}$$

Where:

- T_{ai} = AI token generation rate (tok/s, measured during active decode)
- A_{ai} = AI system availability (1.0 for 24/7 continuous operation)
- P = Number of simultaneous independent sessions
- Q = Quality correction factor $\in (0, 1]$; see Section 4.2
- $T\mathit{h}$ = Human effective token baseline = 0.15 tok/s
- $A\mathit{h}$ = Human daily availability = 0.283

DP is a dimensionless scalar expressing how many equivalent office employees the system replaces in continuous operation. A system rated at 500 DP sustains the equivalent productive throughput of 500 office workers around the clock.

4.2 Quality Correction Factor (Q)

Raw throughput does not capture output quality. Q is defined as the fraction of AI outputs meeting a predefined acceptance standard without human correction on a representative task benchmark. Indicative values are given in Table 2. For general planning, composite values of Q = 0.76 (7B class), Q = 0.85 (20B class), and Q = 0.94 (70B+ class) are recommended pending task-specific calibration.

Table 2: Indicative Quality Factors (Q) by Task and Model Class

Task Category	7B Class	20B Class	70B+ Class
Email drafting & correspondence	0.80	0.88	0.95
Spreadsheet editing (simple)	0.75	0.85	0.92
Spreadsheet editing (complex)	0.55	0.72	0.88
Report summarisation	0.85	0.90	0.96
Data entry & formatting	0.88	0.93	0.97
Meeting minutes & transcription	0.82	0.89	0.95
Composite office average	0.76	0.85	0.94

4.3 Sub-units

Following SI convention: milliDesk (mdP = 0.001 DP) for constrained edge devices; kiloDesk (kDP = 1,000 DP) for workstation-class servers; megaDesk (MDP = 10^6 DP) for hyperscale inference platforms.

5. METRIC 2: CONCURRENT HUMAN-SPEED WORKFLOWS (CHW)

5.1 Motivation

DP measures replacement capacity — it answers a strategic question about workforce equivalence. But it does not answer the question that an IT manager faces when sizing a server for a live organization: how many employees can this machine serve simultaneously at a speed they will actually find useful?

A server generating 50 tok/s serves one user at 50 tok/s, or ten users at 5 tok/s each, or 333 users at 0.15 tok/s each. All three scenarios have the same DP rating, but they represent fundamentally different operational realities. The Concurrent Human-Speed Workflows metric makes this distinction explicit and actionable.

5.2 Definition and Formula

CHW is defined as the maximum number of simultaneous independent workflow sessions a system can sustain at or above a specified per-session token delivery threshold T*:

$$\mathbf{CHW(T^*) = \lfloor T_{ai} \div T^* \rfloor}$$

Where T* is the chosen responsiveness threshold (tok/s per session) and $\lfloor \cdot \rfloor$ denotes the floor function. T* is not a single universal value but a design parameter chosen to reflect the intended use case:

Table 3: Standard CHW Responsiveness Thresholds

Threshold Name	T* (tok/s)	User Experience	Recommended Use Case
Human Parity (CHW-HP)	0.15	AI keeps pace with a human typist	Bulk async automation, document pipelines
Readable (CHW-R)	10	User can read output as it streams	Interactive chat, email drafting
Instant (CHW-I)	30	Response feels immediate	Real-time decision support, live assistants

Reporting CHW at all three thresholds provides a complete picture of a system's operational envelope. A procurement specification might read: "The system shall deliver CHW-HP ≥ 200, CHW-R ≥ 20, CHW-I ≥ 5" — three numbers that fully specify the capacity requirements for a 200-person organization with varied usage patterns.

5.3 CHW as a Server Sizing Rule

The CHW framework directly converts organizational headcount into a minimum server specification. For an organization of N employees with a target responsiveness threshold T*:

$$\mathbf{T_{ai} \geq N \times T^* \Rightarrow \text{Required tok/s}}$$

This is the server sizing equation: multiply the number of users by the required per-user token rate to obtain the minimum aggregate throughput the server must deliver. This single equation replaces the current practice of ad hoc benchmarking and vendor-specific capacity claims.

6. THE DUAL-METRIC FRAMEWORK: CFO AND IT MANAGER

The central contribution of this paper is not either metric individually, but their combination as a complementary pair serving two distinct organizational decision-makers.

Table 4: The Dual-Metric Framework

Metric	Question Answered	Primary Audience	Decision Type
Desk Power (DP)	How many employees does this replace?	CFO, Finance, HR	ROI, headcount planning, capex justification
CHW (T* = 0.15)	How many employees can it serve (async)?	IT Manager, Operations	Bulk workflow server sizing
CHW (T* = 10)	How many employees can it serve (interactive)?	IT Manager, Dept. Head	Interactive tool deployment
CHW (T* = 30)	How many employees get instant response?	IT Manager, UX Lead	Real-time assistant sizing

These metrics are complementary, not redundant. A CFO evaluating ROI needs DP. An IT manager configuring user access quotas needs CHW. A workforce planner modeling phased automation needs both: DP to understand the ceiling of substitution, CHW to understand the rate at which the transition can be operationalized given real server capacity.

This dual structure mirrors the relationship between mechanical horsepower (economic comparison) and steam pressure ratings (operational safety). Both were necessary, and neither made the other obsolete. DP and CHW serve the same complementary roles for AI infrastructure.

7. WORKED EXAMPLES

7.1 Local Edge Node: Consumer GPU Server

Configuration: NVIDIA RTX 4070 Super (12 GB VRAM), 20B parameter model in MXFP4 quantization via llama.cpp, approximate throughput 50 tok/s, continuous 24/7 operation, single inference stream, Q = 0.85.

DP = (50 × 1.0 × 1 × 0.85) ÷ (0.15 × 0.283) ≈ 1,002 DP ≈ 1.0 kDP

CHW-HP = ⌊ 50 ÷ 0.15 ⌋ = 333 | CHW-R = ⌊ 50 ÷ 10 ⌋ = 5 | CHW-I = ⌊ 50 ÷ 30 ⌋ = 1

Interpretation: This single local server carries the replacement capacity of 1,000 office workers (DP), can serve an entire 333-person department at human-parity throughput (CHW-HP), serve 5 users simultaneously at readable streaming speed (CHW-R), or serve 1 user at instant-response speed (CHW-I). For a 5-person team requiring interactive use, this hardware is adequate. For a 50-person department requiring interactive use, it is insufficient.

7.2 High-Performance Workstation: Apple M3 Ultra

Configuration: Apple M3 Ultra Mac Studio (March 2025), up to 512GB unified memory, 819 GB/s memory bandwidth, 70B parameter model Q4 quantization via llama.cpp, ~20 tok/s (community benchmark: 17–18 tok/s on Llama 3.3 70B Q4), continuous operation, single session, Q = 0.94.

DP = (20 × 1.0 × 1 × 0.94) ÷ (0.15 × 0.283) ≈ 443 DP

CHW-HP = 133 | CHW-R = 2 | CHW-I = 0

Interpretation: The M3 Ultra's DP rating appears modest at 443 DP compared to systems running smaller models, but this comparison requires important context. The M3 Ultra's defining capability is its capacity to load and run models exceeding 600 billion parameters on a single desktop node — a capability no other system in this comparison table can match. Its 819 GB/s bandwidth means that on a 20B MXFP4 model it would deliver approximately 75–80 tok/s (~1,500 DP), but its strategic value lies in enabling cloud-quality reasoning models to run entirely locally. For interactive use with a 70B model, it serves 2 simultaneous users at readable speed. This illustrates a key principle of the DP/CHW framework: the same hardware carries very different ratings depending on the model deployed, and the choice of model reflects a deliberate quality-versus-throughput tradeoff.

7.3 Enterprise Cluster: 8× NVIDIA A100

Configuration: 8× A100 80GB, 200 tok/s per session, 8 parallel inference streams, A = 1.0, Q = 0.94.

DP = (200 × 1.0 × 8 × 0.94) ÷ (0.15 × 0.283) ≈ 35,400 DP ≈ 35 kDP

CHW-HP = 10,667 | CHW-R = 160 | CHW-I = 53

Interpretation: Enterprise-scale capacity. Can serve 160 users interactively or over 10,000 users in async workflow mode.

Table 5: Comparative DP and CHW Ratings — Representative Systems

System	tok/s	DP	CHW-HP	CHW-R	CHW-I
Human office worker (baseline)	0.15	1 DP	1	—	—
RTX 4070 Super — 7B Q4	100	~1,790 DP	667	10	3
RTX 4070 Super — 20B MXFP4	50	~1,002 DP	333	5	1
AMD Ryzen AI Max+ 395 128GB — 20B MXFP4 †	46	~919 DP	306	4	1
NVIDIA DGX Spark (GB10) — 20B MXFP4 ‡	47	~940 DP	313	4	1
NVIDIA DGX Spark (GB10) — 120B MXFP4 ‡	38	~841 DP	253	3	1
Apple M3 Ultra 512GB — 70B Q4 ¶	20	~443 DP	133	2	0
2× RTX 4090 — 70B Q4	120	~2,660 DP	800	12	4
8× A100 enterprise node	1,600§	~35,400 DP	10,667	160	53
Hyperscale API (1000 sessions)	—	>10 MDP	>10M	>100K	>30K

† AMD Ryzen AI Max+ 395 128GB: 256 GB/s unified memory bandwidth. Benchmarked via ROCm + llama.cpp on Framework Desktop 128GB (Ubuntu 24.04). Source: AMD internal testing, Nov 2025; GMKTec Evo-X2 community benchmarks, Oct 2025.

‡ NVIDIA DGX Spark GB10: 273 GB/s unified LPDDR5X, 128 GB. Benchmarked via CUDA llama.cpp / Ollama with TensorRT-LLM optimizations post-CES 2026 update. 120B figure sourced from AMD comparative testing (Dec 2025) and LMSYS benchmark study (Oct 2025). Key differentiator: DGX Spark is the only single desktop node capable of running 120B+ parameter models locally.

¶ Apple M3 Ultra (Mac Studio, March 2025): 819 GB/s unified memory bandwidth, up to 512GB. Benchmark: ~17–18 tok/s on Llama 3.3 70B Q4 (community benchmarks, May 2025). Low DP reflects large model size, not hardware weakness — the M3 Ultra's unique capability is running models with 600B+ parameters that no other single desktop node can load. On smaller models (20B MXFP4) bandwidth yields ~75–80 tok/s.

§ Aggregate across 8 parallel streams at 200 tok/s each.

8. DISCUSSION

8.1 Why CHW Is the More Operationally Critical Metric

While DP provides the economic framing, CHW addresses the more immediately actionable organizational question. An IT manager does not need to know that a server replaces 1,000 employees — they need to know whether it can serve their 50-person team at acceptable response speeds. CHW answers this directly and produces a simple pass/fail evaluation against a headcount requirement.

Furthermore, CHW exposes a critical non-linearity in AI server capacity that DP conceals. A server's DP rating is fixed for a given configuration, but its CHW rating varies dramatically depending on the responsiveness threshold chosen. A procurement decision made on DP alone may result in a system that technically has enormous replacement capacity but cannot serve more than a handful of interactive users at once — a mismatch between strategic and operational requirements that CHW makes immediately visible.

8.2 Implications for Procurement Specifications

The dual-metric framework suggests a natural structure for AI server procurement specifications: state a minimum DP requirement (driven by ROI analysis) and minimum CHW requirements at relevant thresholds (driven by organizational headcount and use-case mix). This structure forces vendors to optimize for both economic efficiency and user-serving capacity simultaneously, rather than optimizing for benchmark scores that may not reflect either.

8.3 Limitations

Several limitations must be acknowledged. Both metrics reduce multi-dimensional AI capability to scalars, abstracting away task specificity, context dependency, and model behavior on edge cases. The quality factor Q requires empirical calibration per deployment and cannot currently be predicted from model architecture alone. The framework addresses text-based knowledge work; extension to multimodal tasks requires additional baseline studies. Finally, as model architectures evolve and inference hardware improves, the absolute values of DP and CHW ratings will shift — the metrics are meaningful only in context of a stated date and configuration.

8.4 Standardization Pathway

For these metrics to achieve their potential as an industry-standard vocabulary, standardization through a recognized technical body is necessary. IEEE, ISO, or NIST would be appropriate venues. Standardization should focus on: (1) ratification of the human baseline parameters (Th, Ah); (2) development of a standardized office-task benchmark suite for Q calibration; and (3) formal definition of CHW threshold tiers (HP, R, I). This paper is intended as a foundational contribution toward that standardization effort.

9. CONCLUSION

This paper has introduced a dual-metric framework for AI infrastructure decision-making consisting of Desk Power (DP) and Concurrent Human-Speed Workflows (CHW). DP, calibrated against an empirically grounded human office worker baseline, provides the economic framing needed for ROI analysis and headcount planning. CHW, parameterized by a responsiveness threshold, provides the operational capacity metric needed for server sizing and user-serving capacity assessment.

The two metrics are complementary by design: DP answers the CFO's strategic question, CHW answers the IT manager's operational question. Neither metric alone is sufficient; together they provide a complete and actionable language for AI infrastructure decision-making that is currently absent from the field.

The historical analogy to mechanical horsepower — Watt's solution to an identical communication problem at the dawn of the first industrial transition — validates both the need for such metrics and the methodological approach taken here. As AI systems increasingly constitute critical organizational infrastructure, the development of shared, intuitive, and formally grounded measurement standards becomes not merely convenient but essential.

REFERENCES

[1] R. Hills, *Power from Steam: A History of the Stationary Steam Engine.* Cambridge University Press, Cambridge, UK, 1989.

[2] D. Cardwell, *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age.* Cornell University Press, Ithaca, NY, 1971.

[3] A. Dhakal, K. Feit, S. Feit, and P. Kristensson, 'Observations on typing from 136 million keystrokes,' in *Proc. ACM CHI Conference on Human Factors in Computing Systems,* Glasgow, UK, pp. 1–12, 2018.

[4] M. González and G. Mark, 'Constant, constant, multi-tasking craziness: Managing multiple working spheres,' in *Proc. ACM CHI Conference on Human Factors in Computing Systems,* Vienna, Austria, pp. 113–120, 2004.

[5] T. Brown *et al.,* 'Language models are few-shot learners,' in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[6] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W. W. Norton & Company, New York, NY, 2014.

[7] M. Chen *et al.,* 'Evaluating large language models trained on code,' arXiv preprint arXiv:2107.03374, 2021.

[8] A. Nuvolari and B. Verspagen, 'Numerical data for the development of British steam technology, 1700–1800,' *Technology and Culture,* vol. 50, no. 1, pp. 1–31, 2009.