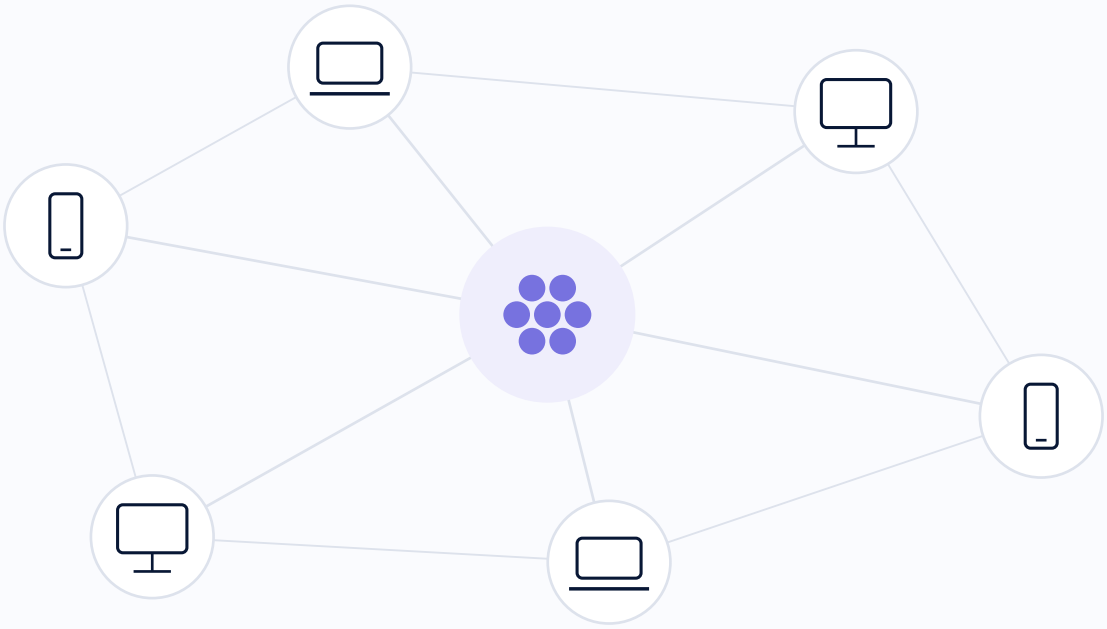
Quantx

RESEARCH PAPER / SYSTEMS 01

The distributed AI mesh.

Capacity, execution and economics.



A public analytical framework for useful AI computation across phones, Macs, and PCs.

Analytical working paper. Modeled scenarios, not production benchmarks.
Public methods edition. Not peer reviewed.

01 / READING GUIDE

A network is useful when its work is useful.

The opportunity is to reuse qualified device-time for workloads that meet the same quality, privacy, and timing requirements as their alternatives.

300,000

device-hours per day

Working scenario

12,500

average active workers

24-hour equivalent

46.96 M

accepted tasks per day

Conditional calculation

Scenario inputs: 1 M daily users; 50% participation; 60% conditional qualification; 1 effective hour; 20 seconds per attempt; 1.15× effort amplification. These are assumptions, not observations.

What the paper establishes

The equations show how participation, qualification, service time, and overhead determine useful capacity. They also show why aggregate throughput, individual response speed, and operator cost are separate questions.

What still needs measurement

No Quantx production benchmark is reported. Eligible device hours, sustained task quality, peak availability, privacy boundaries, and fully loaded cost must be measured before a product advantage is claimed.

How to read the evidence

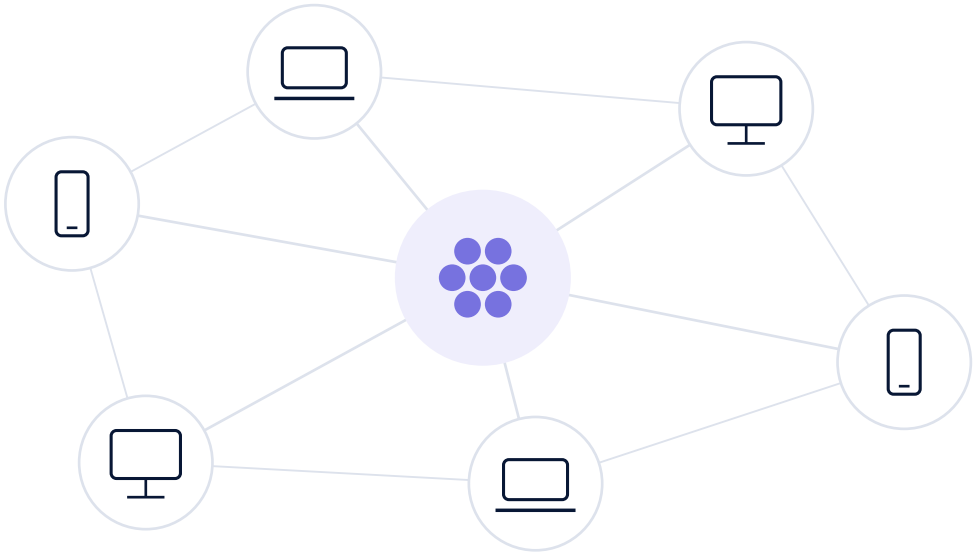
Evidence class	Meaning
Published evidence	External research or first-party disclosures, with citations.
Analytical result	A mathematical consequence of the inputs stated here.
Product hypothesis	A claim requiring Quantx-specific implementation and measurement.

Reading path: device roles p. 3; capacity model p. 4; sensitivity p. 5; timing p. 6; economics pp. 7-8; evaluation p. 9; sources p. 10.

02 / DEVICE ROLES

Different devices. Complementary capacity.

Phones, Macs, and PCs can serve different roles. A useful mesh respects those differences rather than treating every enrolled device as an equivalent worker.



Device role	Potential benefit	Constraint to measure
Phones	Local inputs, supported small models, short eligible jobs.	Battery, thermals, memory, OS background rules.
Macs and PCs	Larger working sets or more sustained execution where qualified.	Memory bandwidth, model support, wall power, interruptions.
Combined fleet	More independent task capacity and access to heterogeneous resources.	Compatibility, placement, availability, communication and validation.

Privacy is a placement constraint

A private task can use only resources permitted for its data. Spare public capacity does not authorize disclosure of private inputs. Device enrollment and job partitioning do not establish anonymity. This paper explains public models and evaluation criteria; it omits proprietary scheduling, verification, and security implementation.

03 / CAPACITY MODEL

Count qualified work, not enrolled devices.

Define a bounded workload j with a declared acceptance threshold, deadline, and service-level target. Only eligible productive time contributes to its capacity.

$$Q_j \geq Q_{0j} \quad \text{and} \quad \Pr(T_j \leq D_j) \geq \eta_j \tag{1}$$

Q is quality, T completion time, D a deadline, and η a target probability. The target is not an observed guarantee.

$$C_j(\tau) \approx (1 / A_j) \sum_i \int_0^{\tau} x_{ij}(t)\, v_{ij}(t)\, dt \tag{2}$$

x is 1 only when device i is qualified, available, permitted, and productive. v is its attempt rate. A converts attempt-equivalent work into accepted output.

$$H = N\, p\, q\, h \qquad C = 3{,}600\, H / (s\, A) \tag{3}$$

A planning simplification with one candidate device per user and a common workload. Use separate models for distinct device classes, then sum compatible accepted capacity.

Symbol	Definition and unit
N	Daily users; persons.
p and q	Participation, then conditional task qualification. These are sequential proportions.
h	Effective eligible execution hours per qualified device per day.
s	Seconds per reference attempt.
A	Total attempt-equivalent effort per accepted output; A ≥ 1.
H and C	Eligible device-hours and accepted tasks per day.

The integral does not assume independent device availability. Correlated outages, shared sleep schedules, and platform restrictions affect the observed indicators and uncertainty.

04 / WORKED SCENARIO

One million users. A conditional capacity case.

Daily-average feasibility can be substantial even when participation and task qualification are partial. Peaks and model suitability remain separate tests.

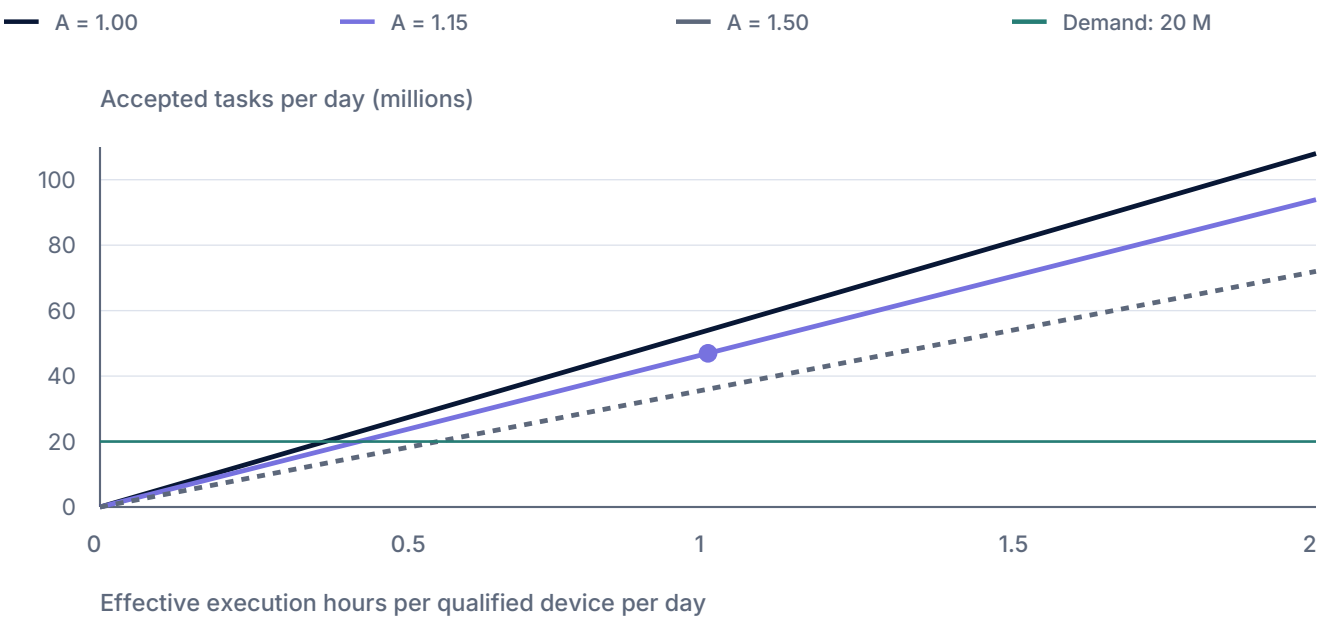


Figure 1. Capacity from Equation (3). N = 1 M, p = .50, q = .60, s = 20 seconds. Lines vary amplification A; h varies along the horizontal axis. The dot marks the working case. No device measurements are plotted.

Scenario	p / q / h	Device-hours	Accepted tasks/day
Restricted	.25 / .50 / .25	31,250	4.89 M
Working	.50 / .60 / 1.00	300,000	46.96 M
Favorable	.75 / .75 / 2.00	1,125,000	176.09 M

All rows use s = 20 seconds and A = 1.15. At 20 compatible tasks per daily user, demand is 20 M tasks/day. The working case has average utilization of about 42.6%; daily totals alone do not establish response-time performance.

Effective hours are real eligible compute time, not time plugged in. The scenario does not establish that a particular phone can contribute an hour or execute the chosen task in 20 seconds.

05 / TIMING AND PARALLELISM

Throughput is not single-answer speed.

Independent tasks can run in parallel. Sequential dependencies, setup, communication, and interruptions limit the speed of one task.

$$\rho = \lambda / \mu < 1 \quad (4)$$

For the workload and route under study, long-run offered rate λ must be below useful service rate μ for stationary operation to be viable. This is necessary, not a deadline guarantee.

$$\varepsilon = s / (s + b) \quad (5)$$

When b seconds of setup cannot overlap s seconds of useful processing, ε is the elapsed worker-time fraction used for that processing.

99.0%

setup efficiency

s = 20 s; b = 0.2 s

83.3%

setup efficiency

s = 20 s; b = 4 s

$$S(k) \le 1 / [(1 - f) + f / k + \omega(k)] \quad (6)$$

Generalized Amdahl bound [10]. f is the parallel fraction; k workers; ω is added communication and coordination time relative to one-worker execution.

Ten workers can still yield at most 4.17× speedup

For f = .90, k = 10, and ω = .05, the denominator is .24 and the bound is 4.17. No task-decomposition method is assumed. Adding devices does not show that a reasoning task has a large parallel fraction.

Measure timing percentiles, burst load, network effects, common interruptions, cold starts, and sustained thermals. High utilization can create long waits even when daily capacity exceeds daily demand.

06 / ECONOMIC MODEL

Lower purchased compute is an opportunity to test.

Compare accepted results at matched requirements. A low subscription price is not evidence of a lower cost to serve them.

$$c_{API} = 10^{-6} (n_{in} r_{in} + n_{out} r_{out}) \quad (7)$$

Input and billed output tokens multiply the applicable USD tariff per million tokens. Tools, search, caching, storage, and application costs require their own accounting.

Example request	Standard text tariff	Batch text tariff
1,000 input tokens	$0.10 / 1 M tokens	$0.05 / 1 M tokens
300 billed output tokens	$0.40 / 1 M tokens	$0.20 / 1 M tokens
Cost per 1 M requests	$220	$110

Gemini 2.5 Flash-Lite example; tariff checked 5 November 2026 [9]. Excludes other charges. Similar output size or schema does not establish equal model quality.

$$c_M = F / n + A s r_h / 3{,}600 + c_o + z c_{API} \quad (8)$$

F: shared fixed expense; n: accepted service tasks; r_h: contribution payment per worker-hour; c_o: other per-task expense; z: hosted fallback fraction. A s represents mesh worker effort amortized per accepted task, including failed attempts where incurred.

Cash break-even

A necessary operator-cost condition is $c_M < c_{API} + c_a$, where c_a is the applicable hosted application expense. Include coordination, communication, quality checks, retries, fallback, support, and any contributor compensation. Do not count the same attempt effort twice.

The hosted tariff is a purchasing reference, not a claim that the modeled mesh workload achieves the same quality or latency as that service.

07 / COST INTERPRETATION

Who pays changes. Resources still count.

The hosted buyer bill, Quantx operator expenditure, and whole-system resource cost are different boundaries.

127,778

worker-hours

20 M tasks; 20 s; A = 1.15

$4,400

hosted token-only bill

20 M standard example requests

Hypothetical worker payment	Daily payment	Remaining versus $4,400
$0.01 per device-hour	$1,278	$3,122 before other expenses
$0.05 per device-hour	$6,389	Already $1,989 above baseline

Rounded analytical values, not a proposed contributor policy. Batch purchasing would lower the token-only comparison to $2,200/day.

Operator expenditure

Unpaid contribution can reduce the operator’s cash expense. The operator still incurs services, support, fallback, reliability, billing, and any contracted payments.

Participant burden

Electricity, added awake time, connectivity, thermal effects, and hardware wear remain real even when no contributor payment is made.

Whole-system accounting

A payment transfers money between parties. Do not add a reimbursement and the same underlying electricity as two separate physical resource costs. Distinguish cash savings from reduced energy or environmental impact.

08 / VALIDATION AND CONCLUSION

Demonstrate the stronger claim separately.

A remote execution demo is a useful start. It is not yet a capacity, cost, privacy, or reliability benchmark.

Claim	Evidence required
Can execute remotely	Accepted multi-device completion.
Adds useful capacity	Matched quality and timing, measured over representative devices and days.
Cheaper to operate	Full cost per accepted result against an appropriate hosted alternative.

A reproducible evaluation agenda

Use a fixed task corpus and held-out labels or another suitable reference. Compare a capable single device, a desktop cohort, a heterogeneous cohort, and a suitable hosted service. Require the same acceptance threshold; returning the same schema is insufficient. [1-8]

Measure cold and warm runs separately, complete input/output work, sustained performance, eligible device-time, rejected and interrupted work, failures, retries, resource use, and participant-visible effects. Report latency distributions and fully loaded cost per accepted result.

Repeat across independent devices and days. Account for clustering and temporal correlation; many jobs on one device are not many independent device observations. Test correlated interruptions, shared bottlenecks, peak demand, and non-flagship hardware.

Conclusion

A heterogeneous mesh is plausible when qualified device-time, suitable parallel work, and common acceptance criteria coincide. The model identifies both the opportunity and its failure conditions: insufficient eligible hours, poor task quality, high amplification, communication cost, and demand mismatch.

Analytical study; no new Quantx measurements. This public edition omits implementation-specific methods and establishes neither patent novelty nor freedom to operate. Environmental accounting is addressed in the companion paper.

09 / SOURCES

Research foundations

Primary research and provider disclosures. Source links are active in the PDF. Dynamic provider information was checked for this edition on 5 November 2026.

[1] M. Satyanarayanan. The Emergence of Edge Computing. Computer, 2017.
Read source ↗

[2] W. Shi et al. Edge Computing: Vision and Challenges. IEEE Internet of Things Journal, 2016.
Read source ↗

[3] D. P. Anderson. BOINC: A System for Public-Resource Computing and Storage. Grid Computing, 2004.
Read source ↗

[4] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI, 2004.
Read source ↗

[5] P. Moritz et al. Ray: A Distributed Framework for Emerging AI Applications. OSDI, 2018.
Read source ↗

[6] A. Borzunov et al. Distributed Inference and Fine-tuning of Large Language Models Over the Internet. NeurIPS, 2023.
Read source ↗

[7] Y. Kang et al. Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. ASPLOS, 2017.
Read source ↗

[8] S. Laskaridis et al. MELTing Point: Mobile Evaluation of Language Transformers. 2024.
Read source ↗

[9] Google. Gemini Developer API pricing, Gemini 2.5 Flash-Lite. Text tariff checked 5 November 2026.
Read source ↗

[10] G. M. Amdahl. Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. AFIPS, 1967.
Read source ↗