

THE AI INSTITUTE / OPERATING INTELLIGENCE

EXECUTIVE AGENDA · 27 SEP / 2026



Find the Bottleneck Before Buying More AI

Next week's infrastructure events should sharpen a buying decision: add computing capacity, fix the data path or improve how the existing system runs.

Boards, C-suite and senior management

Research period: September 2026 – September 2026

First published: 27 September 2026

Current to 27 September 2026

PUBLICATION RECORD

An Institute thesis and decision framework

Allocate the next AI infrastructure budget to the measured constraint on useful work, not the most impressive performance claim.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- 01
- Locate the constraint before pricing more capacity.
- 02
- Compare added compute with data and serving improvements.
- 03
- Test cost and quality on the actual workload and languages.
- 04
- Use next week’s events to gather evidence, not to rush a commitment.

Research method

Institute synthesis of official event programmes and three research preprints. Event dates and agenda topics were checked; advertised performance claims are not adopted as findings. Research methods and limitations inform the buying questions. The proposed Monday–Friday agenda is Institute analysis, not a tested intervention.

First published 27 September 2026 · Updated 27 September 2026 · Research period September 2026 – September 2026 · Version 1.0 · Suggested citation: The AI Institute, *Find the Bottleneck Before Buying More AI* (2026).

CONTENTS

Inside this paper

01	More capacity is not always the next purchase The buying decision
02	Find where the work waits Monday 28 September
03	Make each claim answer the same question Tuesday to Thursday
04	Efficiency depends on what the system is doing Evidence to watch
05	Use local constraints, not a global average Across markets
06	Buy, test or defer Friday 2 October
M	Method and limitations
R	References

How to use this brief

Start with the decision and use the framework to challenge the current plan. Figures retain their original populations and definitions. Institute analysis is labelled, and limitations are recorded before the references.

More capacity is not always the next purchase

Before approving another AI infrastructure commitment, ask where useful work is waiting. Is the model short of computing capacity, is it waiting for data, or is the business spending too much time correcting its output? Those conditions require different investments.

Next week offers a concentrated opportunity to challenge suppliers. NetApp INSIGHT runs from 29 September to 1 October in Las Vegas. CoreWeave's Fully Connected main programme runs on 30 September and 1 October in San Francisco, following arrival activities on 29 September. Their published agendas emphasise storage, model serving and infrastructure choices. These are scheduled commercial discussions, not independently verified performance results.¹²

02

MONDAY 28 SEPTEMBER

Find where the work waits

Ask the technology and operations owners to trace one important workload from request to usable result. Record elapsed time, waiting for data, computing time, correction work and total cost. Use an ordinary busy period as well as a quiet one. Keep failed requests in the record.

Then identify the dominant constraint. If the system is waiting for permission to retrieve documents, more processors may do little. If demand exceeds available computing capacity while data arrives promptly, additional capacity may be justified. These are diagnostic examples, not claims about how most businesses operate.

Finance should set the comparison: the current system, an improvement to its data or serving setup, and added capacity. Hold the task and quality requirement constant. A cheaper answer that creates more repair work is not necessarily a cheaper service.

Make each claim answer the same question

At NetApp INSIGHT, the NVIDIA programme includes sessions on shared model memory and access to enterprise data. CoreWeave's agenda covers capacity choices, storage and operational monitoring. Ask which part of your measured delay each offer would remove, what integration it requires and which charges remain outside the quoted price.¹²

Bring the same workload to competing discussions. Request the hardware, model, workload mix, concurrency and quality conditions behind any improvement claim. A faster internal operation is not the same as a faster customer journey. Ask whether the result includes data movement, failed attempts and human correction.

The AI Conference's San Francisco programme, with workshops on 29 September and conference sessions on 30 September–1 October, provides a parallel place to question evaluation and infrastructure assumptions. Use it to seek counter-evidence, not another collection of product demonstrations.³

Efficiency depends on what the system is doing

Two revised research preprints give a reason to test alternatives to buying more capacity. TIDE studies adapting a model-serving method to changing workloads, including selected non-English datasets. NearOracleKVSelection studies reducing the model memory retained during generation. Both report results under specified experimental conditions; neither establishes savings for your organisation.⁵⁶

A new speech-recognition preprint supplies a caution. In experiments on Whisper models, some compression methods worsened error differences between groups, while distillation sometimes narrowed them. Its estimate of correction time uses an assumed time per error, not observed employee productivity. Lower computing requirements therefore need a separate quality and repair-cost check.⁷

The practical direction is promising: better use of existing systems may release capacity. What remains unproven is the net benefit for your mix of tasks, languages, traffic and service commitments. Do not convert a research speedup into a budget saving without that bridge.

Use local constraints, not a global average

A US vendor event is not evidence of worldwide readiness. Connectivity, available cloud regions, power reliability, data permissions and language coverage can change the preferred investment. A central team should compare those constraints explicitly for each operating market rather than impose one infrastructure answer.

Zurich's AI Policy Summit on 29–30 September and AI+X Summit on 1 October offer a different institutional perspective within the Zurich AI Festival. Conference discussion is not enacted law or regulatory clearance. Use policy specialists to identify the applicable obligations; use operating evidence to assess performance. ⁴

For teams in Africa, the Middle East, Latin America, Asia-Pacific, Europe and the United Kingdom, the missing evidence is often organisation-specific. The studies cited here do not support a regional ranking of cost or capability.

Buy, test or defer

Close the week with a short decision record: the measured constraint, the alternatives considered, the remaining uncertainty and the next commitment. Buy where a comparable test supports the case. Fund a bounded trial where the opportunity is credible but unproven. Defer where the proposed purchase does not address the constraint.

The Monday and Friday dates are Institute planning suggestions, not external deadlines. The objective is a better allocation of capital, not a purchase by the end of conference week.

EXHIBIT / A CONSTRAINT-LED BUYING DECISION

BASE	TEST	COST	BUY
Current work Locate waiting, errors and cost. Institute proposed agenda	Alternatives Compare fixes with added capacity. Institute proposed agenda	Whole service Include integration and repair. Institute proposed agenda	Evidence first Commit, trial or defer. Institute proposed agenda

RESEARCH RECORD

Method and limitations

Method

Institute synthesis of official event programmes and three research preprints. Event dates and agenda topics were checked; advertised performance claims are not adopted as findings. Research methods and limitations inform the buying questions. The proposed Monday–Friday agenda is Institute analysis, not a tested intervention.

Limitations

No representative enterprise adoption or return-on-investment estimate is claimed. Conference programmes can change. Research results concern selected models, datasets and hardware; workshop acceptance is author-reported where noted. There is no cross-market commercial comparison. No partnership, endorsement or guest participation is implied.



Sources and limitations

Source facts link to the original publication. Institute analysis reconciles definitions and turns the material into a management proposition. Frameworks are decision aids, not claims of universal causality. Read every important statistic with its population, date, definition and caveat.

Research cutoff: 27 September 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01 NVIDIA at NetApp INSIGHT

Official event programme; 29 September–1 October 2026, Las Vegas. Publication date unspecified; checked 27 September. Agenda and vendor claims, not independent outcome evidence.

02 CoreWeave: Fully Connected 2026 agenda

Official programme; arrivals 29 September, main programme 30 September–1 October, San Francisco. Checked 27 September; schedule may change.

03 The AI Conference: 2026 agenda

Official organiser schedule: workshops 29 September, conference 30 September–1 October, San Francisco. Checked 27 September.

04 Zurich AI Festival: programme

Festival 28 September–3 October 2026; Policy Summit 29–30 September, AI+X Summit 1 October. Event information only, not a legal authority.

05 Park et al.: TIDE

Revised preprint: v1 5 February, v2 24 September 2026; listed 25 September. SC26 acceptance author-reported. Methods: <https://arxiv.org/html/2602.05145v2>

06 Gao et al.: NearOracleKVSelection

Revised preprint: v1 9 February, v2 23 September 2026; listed 25 September. Selected models and benchmarks; no peer-reviewed publication verified.

07 Ginjala et al.: Temporal Taxation Compounds Under Post-Training Compression of Whisper Models

New v1 submitted 23 September 2026; listed 25 September. IMPACT-SPEECH acceptance author-reported. Whisper-only experiments and assumed correction time: <https://arxiv.org/html/2609.28739v1>

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Research · Readiness · Capability

theai.institute