CAPACITAS

THE TECHNOLOGY EDGE

CLOUD CAPACITY PLANNING FOR CTOS

A pragmatic approach to long-range cloud efficiency, risk visibility, and forecasting

Dr Manzoor Mohammed
Co-founder & Chief Innovation Officer, Capacitas

Contributors
Anderson Quach, CTO, Qualtrics
Steve Jang, Chief Architect, Qualtrics

AUTHORS

Dr. Manzoor Mohammed
Co-Founder & CINO - Capacitas

CONTRIBUTIONS FROM

Anderson Quach
CTO - Qualtrics

Steve Jang
Chief Architect - Qualtrics

BUSINESS IMPERATIVE

Modern CEOs expect technology to be a growth engine: platforms that scale efficiently, perform reliably, **shorten time-to-market**, and enable continuous innovation while remaining secure and cost-effective. Over time, they also expect this to translate into improved margins and EBITDA.

For CTOs, meeting these expectations is increasingly complex. Reliability, security, performance, and cost discipline—once secondary concerns—are now core responsibilities, often pushed deep into engineering teams.

At the same time, cloud has become the second-largest line item on many technology budgets, behind people, and its behaviour is becoming harder to predict as organisations scale.

In early growth stages, it is often rational for CTOs to prioritise speed and customer outcomes, accepting higher cloud costs as the price of momentum. As products mature, customers become more demanding, and enterprises introduce security, regulatory, and contractual expectations, **this approach becomes harder to sustain.**

Long-range planning, forecasting, and architectural discipline begin to matter.
This tension is now amplified by AI and other compute-intensive workloads. Investment in new tooling and platforms is essential, but CFOs increasingly expect this to be absorbed within tightly constrained budgets.

WHY THIS PAPER EXISTS

This paper is for CTOs and senior technology leaders responsible for critical platforms—teams that value reliability, delivery speed, and engineering ownership, and who are under increasing pressure to explain how cloud costs will behave as the business scales.

Most organisations face a consistent set of business imperatives:

- Increased reliability and improved delivery velocity are non-negotiable
- Forecasts are expected, despite complex and evolving systems
- Cloud spend is now discussed alongside margins, growth, and investor expectations

For many CTOs, forecasting cloud costs is not always a priority—and that is understandable. Some are focused on building products at pace; others are increasingly expected to contribute at the executive level.
Either way, rising cloud costs—particularly with AI and compute-intensive workloads—make understanding cost behaviour essential.

Taking a proactive, analytical approach benefits CTOs in two ways:

Personal "why":

- Maintain a credible voice at the executive table
- Anticipate trends and act before they become problems
- Avoid reactive cost-cutting or loss of control to finance

Strategic "why":

- Enable long-range planning (LRP) and engineering alignment
- Anticipate service and operational risks ahead of impact
- Link engineering activity directly to business demand
- Drive architectural simplification and more efficient delivery

This paper does not propose a heavy optimisation programme or restrict engineering autonomy. Instead, it introduces a lightweight modelling approach that enables organisations to:

- Make scaling behaviour visible over longer time horizons
- Separate temporary cost noise from structural growth drivers
- Identify where demand and capacity are tightly coupled—and where they are not

EXECUTIVE SUMMARY

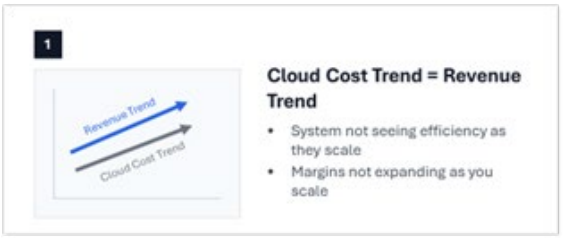
Most technology leaders already expect their platforms to become more efficient as they scale. The challenge is rarely motivation—it is visibility.

As organisations grow, cloud cost patterns are shaped by multiple overlapping forces:

- Business demand
- Feature expansion
- Architectural decisions
- Technology migrations
- Short term operational activity

Some parts of a platform scale efficiently with revenue. Others scale faster or slower.
In complex environments, these behaviours are often masked by short-term fluctuations, making long-range forecasting and capacity planning difficult without introducing heavy process.

This complexity increases further in multi-cloud environments.



A PRACTICAL APPROACH TO CAPACITY PLANNING

This paper outlines a pragmatic, engineering-led approach to understanding how cloud capacity scales relative to demand.

Rather than assuming cloud spend must track revenue, the model focuses on identifying the drivers of capacity usage and their long-term behaviour.

By modelling platforms at a domain level and correlating demand signals with infrastructure usage over multiple time horizons, teams can distinguish:

- Operational noise
- Structural growth patterns
- The objective is not cost reduction for its own sake. The value lies in:
- Validating where architecture already scales efficiently
- Identifying where demand and capacity are misaligned
- Anticipating where cost or reliability risks are likely to emerge

This enables more confident forecasting, better-timed investment decisions, and a clearer narrative for leadership and investors.

WHY CLOUD COSTS ARE ASSUMED TO SCALE WITH REVENUE

Business leaders and investors expect margins to improve with scale: revenue should grow faster than cost.

Cloud platforms are no exception.

For CTOs, this expectation typically arises during:

- Annual budgeting and multi-year forecasting
- Negotiation of long-term cloud commitments (e.g. AWS EDPs)

Cloud providers often support these discussions with forecast models based on historical spend or assumed linear growth.

These models are simple and defensible—but they do not explain what is actually driving that growth.

As a result, organisations can meet short-term budget targets while accumulating longer-term structural risk.

DOMAIN MODELLING: MAKING COMPLEXITY MANAGEABLE

Modern cloud environments are inherently complex:

- Dozens or hundreds of teams
- Thousands of services
- Multiple persistence layers
- Diverse growth patterns

Modelling every service individually is not practical.

Instead, systems can be grouped into domains based on shared characteristics:

- Similar demand drivers
- Comparable scaling behaviour
- Common architectural patterns

This approach allows leaders to reason about growth at the right level of abstraction. Domains are not organisational units—they are functional groupings whose capacity usage responds similarly to demand.

Understanding how each domain scales is critical to predicting long-term cost behaviour.

UNDERSTANDING COST GROWTH PATTERNS

Cloud cost growth typically falls into four categories:

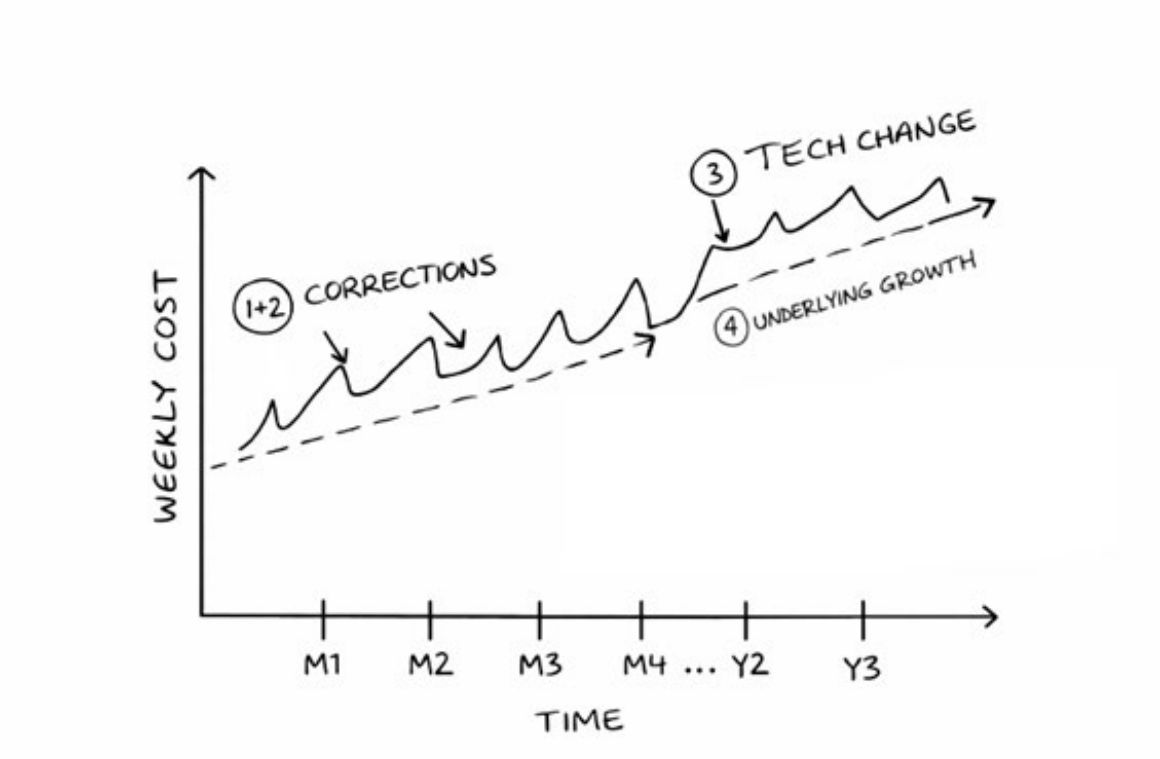
Temporary

1. Housekeeping (idle resources, orphaned assets)
2. Short-term project environments

Permanent

3. Technology changes or migrations
4. Underlying demand growth

These effects often overlap, making it difficult to isolate true long-term drivers without structured modelling.

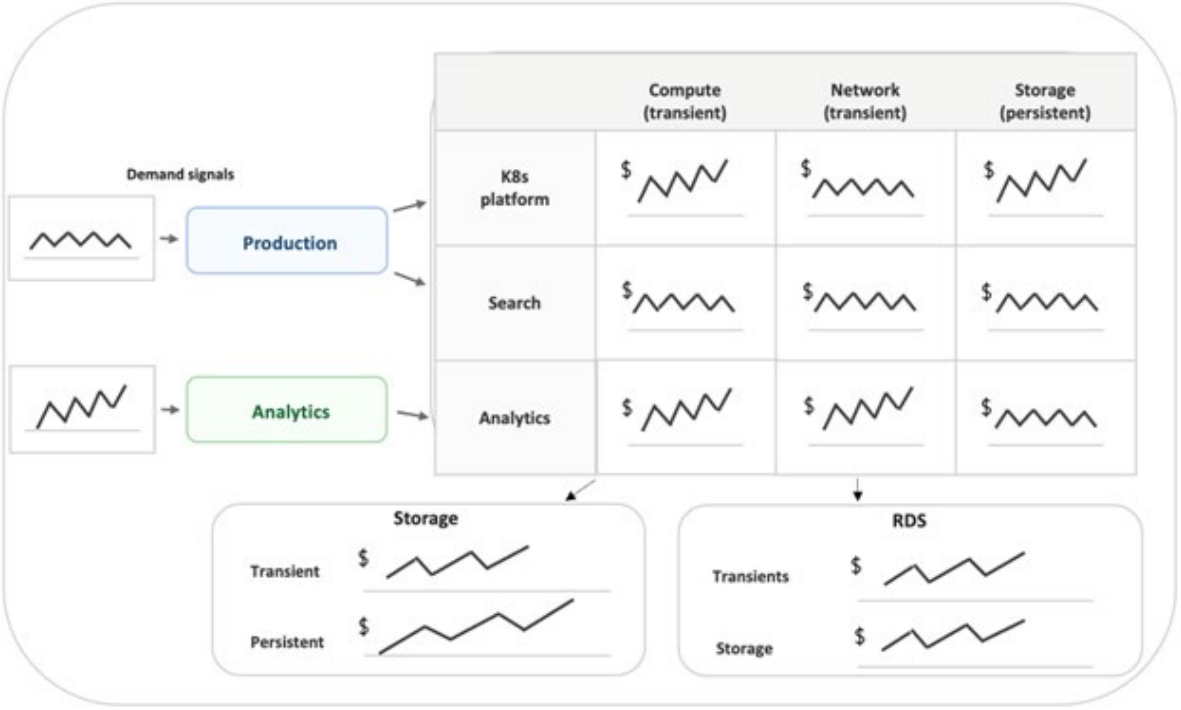


PATTERN MATCHING AND ANALYSIS

By correlating demand signals (e.g. transactions, users, traffic) with capacity usage across multiple time horizons, organisations can identify:

- Which resources track business demand
- Which follow project or migration patterns
- Where efficiency is improving or deteriorating

This creates a clearer, evidence-based capacity model grounded in real behaviour.



BENEFITS FOR CTOs AND LEADERSHIP

Organisations adopting this approach typically achieve:

- Earlier, more effective architectural investment
- Reduced end-of-year cost pressures
- Better identification of risk accumulation
- Increased confidence in forecasting
- Stronger narratives for boards and investors

Importantly, this is achieved without adding heavy process for engineering teams.

THE GROWING IMPACT OF AI

The rapid adoption of generative AI introduces new scaling dynamics.

In many organisations, AI-related cloud spend is growing faster than overall revenue. Technology leaders must be able to explain both why this is happening and how it delivers business value.

Three key drivers typically emerge:

1. Product enhancement through AI

AI features are increasingly embedded into products. In most cases, inference, not training, becomes the dominant cost driver. These costs scale with usage, making long-term patterns visible—but not always optimised.

In practice, incorrect scaling assumptions can significantly distort cost trajectories.

2. Faster feature development

AI accelerates development velocity, increasing the rate at which new features are released. While beneficial, this can lead to cloud consumption growing ahead of revenue in the short term.

3. AI-driven operational efficiency

AI can reduce operational costs (e.g. customer support automation) while increasing cloud spend. In these cases, rising cloud cost is not inherently negative—but must be understood in the context of total cost reduction.

AUTOMATION AND THE ROLE OF AI

The analysis outlined in this paper can initially be performed manually.

However, sustaining this approach requires automation.

Agentic AI can analyse infrastructure behaviour, detect anomalies, and identify scaling patterns at scale—but these systems still require a structured framework.

This model provides that foundation.

CLOSING THOUGHTS

This approach is not about minimising spend at all costs.

It is about making long-term behaviour visible so that technology leaders can act early, invest wisely, and ensure that their platforms scale efficiently alongside the business.

When domain-level behaviour is understood and owned by engineering leaders, cloud economics improves not through constraint—but through clarity.

ACKNOWLEDGEMENTS

Mark Gillett, Managing Director and Head of Value Creation at Silver Lake, for his continued support, guidance, and insight into capacity planning in product-led technology companies.

Thank you also to the following for their invaluable feedback:
Matt Kane, Michael Todd (Silver Lake), Nik Sathe (Blackhawk Network), Ramana Thuma (Expedia Group), Brendan Farrell (BMC Helix).

CAPACITAS

THE TECHNOLOGY EDGE

www.capacitas.co.uk
sales@capacitas.co.uk

Second Floor
8-10 Hatton Garden
London
EC1N 8AH