capacitas®
COST | SCALABILITY | PERFORMANCE

Cloud Capacity Planning for CTOs

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

Dr Manzoor Mohammed, Co-found & CINO, Capacitas

Contributions from Anderson Quach, CTO, Qualtrics, Domain modelling concept from Steve Jang, Chief Architect, Qualtrics.

capacitas®
COST | SCALABILITY | PERFORMANCE

Authors

Dr. Manzoor Mohammed
Co-Founder & CINO - Capacitas

Contributions from

Anderson Quach
CTO - Qualtrics

Steve Jang
Chief Architect - Qualtrics

Why this paper exists

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 improving margins and EBITDA.

For CTOs, meeting these expectations is increasingly complex. Reliability, security, performance, and cost discipline, once secondary concern, 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 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, that approach becomes harder to sustain. Long-range planning, forecasting, and architectural discipline start to matter.

This tension is now amplified by AI and other compute-intensive workloads. Investment in new tooling and platforms is essential, but CFOs expect it to be absorbed within existing tightly constrained budgets.

Who is this paper for?

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

Most organisations have the following business imperatives, increased reliability and improved delivery velocity are non-negotiable, forecasts are expected, and cloud spend is now discussed alongside business margins, growth, and investor expectations. For many CTOs, forecasting cloud costs is not top of mind and that's normal. Some are focused purely on building products, while others aspire to a seat at the executive table, shaping strategic decisions. Either way, rising cloud costs especially with AI and other compute-intensive workloads make understanding cost trends essential.

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

Personal "why":

- Maintain a voice at the executive table and influence business decisions.
- Be the hero who protects the business, spotting trends before they become problems and steering the rest of the business to adjust to maintain margins
- Avoid reactive firefighting, budget cuts, or loss of control when finance takes over.

Logical "why":

- Support long-range planning (LRP) and engineering delivery planning PPA (Private Pricing Agreement).
- Anticipate service and operational issues before they impact users.
- Understand what teams are working on, the value they are delivering and how costs tie to business demand.
- Effective modelling and cost control drive architectural simplification, accelerating the delivery of business change

This paper does not propose a new optimisation programme or challenge engineering autonomy. It describes a rapid lightweight modelling approach that organisations can use to:

- Make architectural scaling behaviour visible over longer horizons.
- Separate temporary cost noise from persistent growth drivers.
- Identify where demand and capacity are

The result: CTOs gain foresight, influence, and control, while helping their teams deliver efficiently and ensuring cloud costs grow more slowly than revenue, without compromising performance or resilience.

capacitas
COST | SCALABILITY | PERFORMANCE

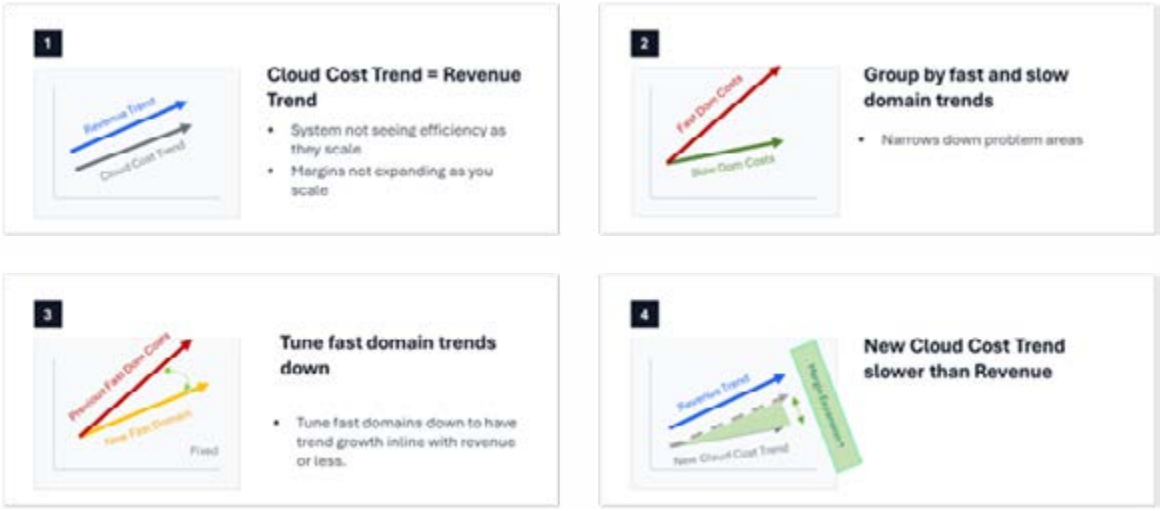
Executive summary

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

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

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

Some parts of a platform scale efficiently with revenue. Others scale much faster or much slower. In complex environments, these behaviors are often masked by short-term variance, making long-range forecasting and capacity planning difficult without introducing heavy process. This complexity increases with the use of multi-cloud environments.



This paper outlines a pragmatic, engineering-led approach to understanding how cloud capacity actually scales relative to demand. Rather than assuming cloud spend must track revenue, the model focuses on identifying drivers of capacity usage and their long-term gradients.

By modelling platforms at a domain level and correlating demand signals with infrastructure usage over multiple time horizons, teams can distinguish operational noise from structural growth.
The outcome is not cost reduction for its own sake. The value lies in:

- Validating where architecture already scales well
- Identifying where demand and capacity are misaligned
- Anticipating where cost or reliability risks are likely to emerge as the business grows

This approach enables more confident forecasting, better-timed architectural investment, and a clearer story for leadership and investors about how cloud economics improve with scale.

This paper recognizes that introducing new technologies, such as generative AI, can make forecasting more challenging. In most organisations, the majority of cloud spend remains tied to well-understood platform components. It is therefore practical to focus on understanding the proportion of spend in these known areas and to invest in improving forecasting accuracy there. Doing so provides greater flexibility to accommodate new technology initiatives without compromising reliability or budget discipline

capacitas®
COST | SCALABILITY | PERFORMANCE

Why cloud costs are assumed to scale with revenue

Business leaders and investors reasonably expect margins to improve with scale: revenue should grow faster than costs. Cloud platforms are no exception.

For CTOs, this expectation typically surfaces in two places:

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

Cloud providers often support these discussions with 3–5 year forecasts that extrapolate historical spend or assume linear growth with revenue. These models are useful but they don't ask what drives that growth or whether usage truly needs to follow the same gradient as the business.
Given time pressure, operational risk, and forecasting uncertainty, many technology leaders adopt these proportional models. They are simple, defensible, and usually accurate in the near term.
The limitation and risks appear over longer horizons.

The limits of procurement and tooling alone

Most organisations already apply a combination of:

- Procurement strategies (discounts, commitments)
- FinOps processes
- Engineering-led efficiency work
- Tooling for cost visibility and anomaly detection

These approaches are necessary and valuable. They help teams stay within budget and manage short-term variance.

What they typically do not do is explain:

- What fundamentally drives capacity usage
- Which parts of the platform scale with demand and which do not
- Whether long-term cost gradients are improving or deteriorating

As a result, organisations can meet near term budget goals while still accumulating structural cost or reliability risk.

capacitas®
COST | SCALABILITY | PERFORMANCE

Systems that do not scale efficiently carry hidden risk

In our experience working with large, scaling platforms, systems whose costs scale linearly or super-linearly with revenue often hide issues that only surface later: This is supported by Werner Vogels, Chief Architect at AWS with his 3rd law of frugal architecture. “Systems that last align with the business” Laws

- Excess capacity can delay the visibility of architectural constraints
- Demand-agnostic systems may consume resources without delivering proportional business value
- Cost growth frequently correlates with future reliability or performance risk

Understanding how systems scale is therefore not only a financial concern it is a platform health concern.

A modelling approach that simplifies without oversimplifying

Modern cloud environments are inherently complex: dozens or hundreds of teams, thousands of services, multiple persistence layers, and heterogeneous growth patterns.

To manage this complexity without overwhelming teams, the approach described here relies on two core ideas:

1. **Domain modelling:** grouping systems by architectural and scaling characteristics rather than organisational structure
2. **Pattern matching:** correlating demand signals with capacity usage over multiple time horizons

The goal is to simplify the system enough to see structural behaviour, without losing the signals that matter.

capacitas®
COST | SCALABILITY | PERFORMANCE

Domain modelling: reducing complexity

As organisations scale, cloud behaviour becomes harder to manage. The number of services grows, teams become more autonomous, architectures diversify, and demand signals diverge. Modelling every service or team quickly becomes impractical.

To make structural behaviour visible without overwhelming teams, platforms can be grouped into product and/or architectural domains: collections of systems that share similar demand drivers, scaling behaviour, and technical characteristics. This approach used in large scale environments such as Qualtrics and others allows leaders to reason about growth patterns at the right level of abstraction.

Domains are not organisational units. They are functional or architectural groupings whose capacity usage responds to demand in broadly similar ways. A single domain may span multiple teams; a single team may contribute to more than one domain.

Typical domains might include:

- Customer facing transaction or API workloads
- Data and storage growth platforms
- Internal platform and shared services
- Compute-intensive or AI workloads
- Non-production and project environments
- Products with similar growth patterns/characteristics

Each of these tends to scale differently with business demand, and understanding those differences is key to predicting long-term capacity behaviour.

This approach drawn from work by Steve Jang (Chief Architect at Qualtrics) becomes increasingly important as organisations scale into dozens or hundreds of teams.

When domain modelling becomes useful

Not every organisation needs domain modelling. In smaller or less complex environments, direct mapping between demand signals and infrastructure usage is sufficient.

- Domain modelling becomes valuable when:
- No single demand signal explains cloud growth
- Different parts of the platform scale at noticeably different rates
- Forecast accuracy declines as the organisation grows
- Architectural or migration activity obscures underlying trends
- Leaders cannot clearly explain why capacity is increasing

In these conditions, grouping systems by scaling behaviour rather than organisational structure reduces complexity while preserving meaningful signals.

The objective is not to simplify the system artificially, but to simplify it enough to observe how it truly behaves at scale.

Complexity Team Scale (number of Teams)	Complexity Product Range (number of products)	Complexity Architecture	Domain Modelling	Approach
Small (< 10)	Small (< 10)	Medium	No	Direct pattern matching of demand signal to infrastructure usage
Large (> 30)	Medium (>10 & < 20)	Medium	Yes	Group into domains by size followed by pattern matching
Large (> 30)	Large (>20)	Medium	Yes	Group into domains by product, capacity size followed by pattern matching
Large (> 30)	Small (<10)	High (microservices etc)	Yes	Group into domains by architecture & growth pattern (e.g. some architectural patterns increase faster than revenue/demand signal)

capacitas
COST | SCALABILITY | PERFORMANCE

Understanding cost increase patterns

Cloud cost growth generally falls into four patterns:

Temporary

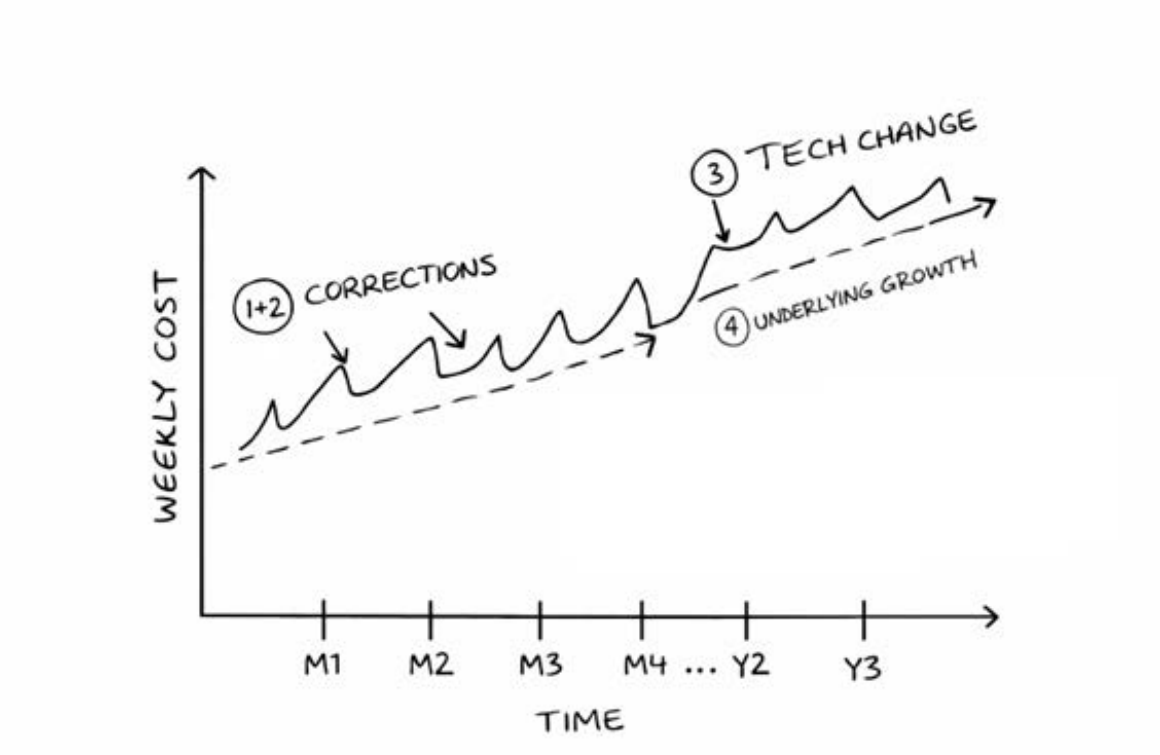
1. Housekeeping: unattached volumes, idle environments, hygiene issues
2. Short-term projects: environments spun up and later torn down

These create noise but do not change the underlying growth gradient.

Permanent

3. Technology changes or migrations: step changes caused by new platforms or capabilities
4. Underlying demand growth: slow, persistent increases driven by business or technical demand

The challenge is that temporary and permanent effects overlap, making it difficult to see the true long-term driver without modelling.
In addition, corrections may be difficult to see if housekeeping activities are carried out very quickly.



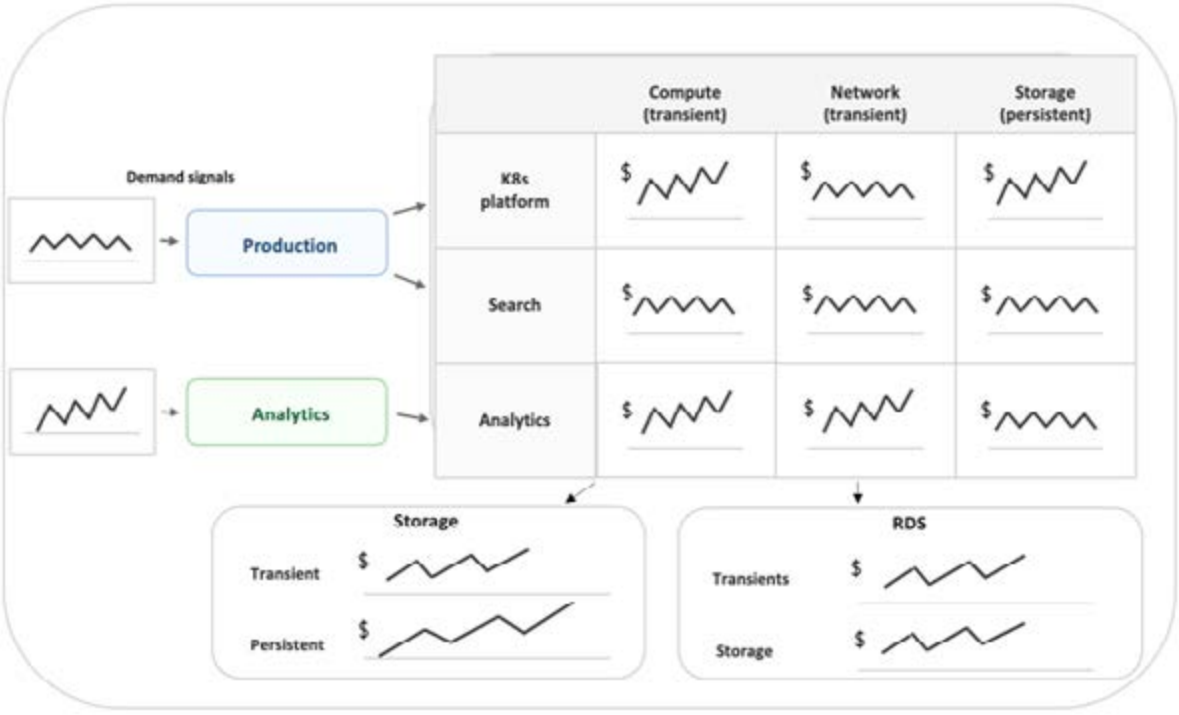
capacitas®
COST | SCALABILITY | PERFORMANCE

Pattern matching at Scale to understand drivers

By correlating demand signals (e.g. traffic, transactions, users) with undiscounted capacity usage across multiple time horizons e.g. hours, weeks, months teams can identify:

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

This end-to-end view helps avoid false conclusions, such as assuming efficiency gains when systems are actually becoming chattier or more coupled.
The result is a clearer capacity model grounded in observed behaviour.



Data and granularity

This model does not require perfect data. Useful inputs typically include:

- Revenue (monthly)
- Global demand signals (daily and monthly)
- Capacity usage by domain (production and non-production)
- Cost data split by transient vs persistent usage

Where direct signals are missing, proxies (such as network traffic) are often sufficient.

Benefits for CTOs and leadership

Organisations using this approach will see benefits such as:

- Earlier, better-timed architectural investment
- Reduced end-of-year cost fire drills
- Clearer identification of domains accumulating risk
- Improved confidence in multi-year forecasts
- Stronger narratives for boards and investors around margin expansion

Importantly, these benefits come without introducing heavy new process for engineering teams.

Choosing the right level of modelling

Not every organisation needs the same depth:

- Reactive: anomaly-driven, sufficient for slow-growth environments
- Trend-based: focuses on long-term gradients
- Bottom-up: links demand drivers to future capacity
- Dependency-aware: models interactions between domains (most complex)

The right approach depends on growth rate, architectural complexity, and business expectations.

From model to operational dashboards

The same model can support operational dashboards to monitor both team /domain and system health trends / patterns over the short, medium and long term that combine:

- Demand
- Cost (transient vs persistent)
- Usage
- Performance

Used carefully, these dashboards help senior leaders track platform health and efficiency without turning cost into a blunt performance metric.

Closing thoughts

The value of this approach is its speed and simplicity. The value also comes from making scaling behaviour visible on a regular basis, and embedding that visibility into how teams think, plan, and operate.

In mature organisations, understanding how capacity scales is not a centralised finance or architecture exercise. It becomes part of normal engineering ownership. Domains are most effective when they are owned by the leaders closest to the systems. Team leads and engineering directors will understand both the technical drivers and the product context behind demand.

This shifts the conversation to “How does my domain scale, and is that behaviour improving?”

When domain-level patterns are visible over time, teams can compare trends / behaviours rather than absolute spend. Peer comparison becomes constructive and curiosity led conversations. Some domains will scale faster than others, some of these will be explainable but others will be opportunities to reduce risk, increase understanding and managing cost. Over time, this builds a culture of cost ownership and curiosity grounded in engineering not financial pressure.

The growing importance of AI and impact on cost

The rapid adoption of generative AI introduces new scaling dynamics. In many organisations, AI-related cloud spend is growing faster than overall business revenue. Technology leaders need to explain the increase in cloud cost trendlines and how they are adding business valuable.
In practice, three broad forces are typically driving AI-related cost growth.

1.Product Enhancement using AI features

Many teams are embedding AI capabilities directly into their products. In most cases, the dominant cost driver is inference, not training. Unlike one-off training events, inference grows steadily with product usage, making long-term behaviour easier to understand. However, the slope of that trendline can be incorrect due to the sizing methodologies used by the team.

In recent work with large-scale AI platform, we observed teams scaling growth on the incorrect vector, they used GPU utilization rather the CPU utilization as driver for impacting performance. This resulted in a cost trajectory approximately twice what demand justified. The issue was not technical inefficiency, but an incorrect scaling assumption. As AI workloads become more prominent, identifying the correct scaling vector by understanding the relationship between demand and usage will be key to avoiding these scaling model errors.

2.Faster, more autonomous feature development

Advances in generative AI are accelerating how quickly new product capabilities can be built and deployed. In some organisations, this is increasing the rate of feature introduction significantly. While this improves product velocity, it can also cause cloud consumption to grow ahead of revenue realisation, shifting the cost curve upward in the short term.

It is still early to determine how persistent this effect will be. Historically, however, more features tend to translate into higher steady-state infrastructure consumption, and organisations should monitor whether this growth remains aligned with long-term business value.

3.AI-driven Operational Efficiency.

Many organisations are also deploying AI to improve internal efficiency particularly in areas such as customer support and operational automation. A common example is call centre transformation, where AI reduces people-related costs but increases cloud consumption.

In these cases, cloud cost growth is not inherently negative. The key is to track both sides of the equation. When cloud costs rise but total operating cost falls, the net effect is positive, but this relationship and trendlines must be visible to demonstrate real business value.

The role of automation and agentic AI

The analysis outline in this paper can be carried out initially manually relatively quickly, but doing this on a regular basis requires automation. Agentic AI can analyse infrastructure behaviour, detect anomalies, and identifying scaling patterns across large datasets.

However, these systems still require a framework. It's hoped by using the framework in this paper it will help organisations use AI.

Final reflection

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 maintain confidence that their platforms will scale efficiently as the business grows.

When domain-level scaling becomes understood and owned by engineering leaders, cloud economics improve 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 on the importance of capacity planning in product led tech companies. Mark set the challenge to produce a paper that tech teams can use to apply in their own environments a few years ago. I'm hoping I've gone some way to achieveing it.

Thank you to following for their invaluable feedback, Matt Kane, Operating Partner, Silver Lake, Michael Todd, Operating Partner, Silver Lake, Nik Sathe, CPO & CTO, Black Hawk Network, Ramana Thuma, CTO, Expedia Group, Brendan Farrel, Chief Architect, BMC Helix.

capacitas®

COST | SCALABILITY | PERFORMANCE

www.capacitas.co.uk

sales@capacitas.co.uk

Second Floor

8-10 Hatton Garden

London

EC1N 8AH