

L201 LEAD · PRACTITIONER

IT Economics: Cost, Value and AI Spend

Cost transparency, unit economics and value realization across cloud, SaaS, data cloud platforms, data centers and AI.

FORMAT

Knowledge pack · 13 slides

FOR

IT and finance leaders

LEVEL

Practitioner



THE PROBLEM

Technology cost is easy to total and hard to explain

IT spend is usually reported by ledger account. Saying what a **business capability** costs, who consumes it and what would change if it stopped requires an allocation model.

Allocation

Spend traced to products, services and the people who consume them.

Unit cost

Cost expressed per transaction, user or outcome.

Value

Benefits evidenced in terms finance recognizes.

COST MODEL

Trace every dollar from ledger to consumer

TBM Taxonomy 5.0.1 (TBM Council, July 2025) names these layers and adds an AI Models sub-tower.

Cost pools

General-ledger spend by type: staffing, outside services, cloud services, hardware, software and SaaS.

Resource towers

What the spend buys: compute, storage, network, data, applications, end user, security.

Solutions

Business-facing products and services; applications now sit in the towers.

Consumers

Business functions, value streams and external customers that use the solutions.

WHY IT MATTERS

Allocation shows each consumer what its demand costs, so the people who create the demand can change it.

FINOPS

FinOps now reaches well beyond public cloud

The FinOps Framework now gives guidance for five technology categories: public cloud, SaaS, data cloud platforms, data center and AI.

Inform

Ingest billing data, ideally in FOCUS format, then allocate, forecast and benchmark unit costs.

Optimize

Identify where rates, usage and workload placement can improve efficiency and value.

Operate

Act on the chosen optimizations, with engineering, finance and business accountable.

Principles

Two of six:
“Business value drives technology decisions” and
“Everyone takes ownership for their technology usage.”

LEVERS

Each cost lever has a different owner and risk

LEVER	WHAT IT CHANGES	TYPICAL OWNER	MAIN RISK
Rate	Price per unit, through commitments and negotiated discounts	FinOps team, procurement	Unused commitments
Usage	Units consumed, through rightsizing, scheduling and removing idle resources	Engineering teams	Performance regressions
Architecture	How cost grows with volume	Architects	Migration effort
Portfolio	Number of systems, through retirement and consolidation	Portfolio governance	Missed dependencies
Demand	Consumer behavior, through showback and policy	Business partners	Friction with the business

UNIT ECONOMICS

Unit costs connect spend to volume and revenue

Prefer a unit the business already manages, so cost can be read alongside volume and revenue.

Cost per transaction

Run cost of a product divided by the transactions it processes.

Cost per active user

Platform or SaaS cost divided by active users, by business unit or role.

Cost to serve

Technology cost per customer, account or segment.

Cost per inference

Total inference cost divided by the number of inference requests.

Run cost to revenue

Technology run cost as a share of the revenue a product earns.

Cost per environment

Non-production cost per environment, and its share of the total.

GENERATIVE AI

Generative AI has a different cost structure

Spend follows consumption at a fine grain, so small design choices multiply across millions of calls.

Tokens

Output tokens, including reasoning tokens, typically cost several times more than input.

Context length

Instructions, tool definitions and retrieved text are billed on every call; caching only discounts them.

Model choice

Per-token prices vary more than tenfold and tokenizers count text differently; compare cost per task.

Agent loops

Each call, including retries, re-sends the growing history, so cost outpaces step count.

GPU capacity

Provisioned capacity is billed whether used or not; on-demand GPUs can be scarce.

Supporting costs

Embeddings, vector storage, evaluation, logging and human review.

CONTROLS

Control AI spend without slowing adoption

- ✓ **Route by task**
Send each task to the least expensive model that meets the quality bar.
- ✓ **Bound agents**
Limit steps, tokens and time for every agent run.
- ✓ **Measure per outcome**
Track cost per resolved task, not only per call.
- ✓ **Re-test models**
Try cheaper models again as prices and capabilities change.
- ✓ **Cache**
Cache stable prompt prefixes; reads can cost a tenth of the input rate, but writes can cost more.
- ✓ **Batch**
Send non-interactive work to batch APIs, typically billed at half the standard rate.
- ✓ **Tag**
Attribute every AI charge to a product, team and use case.
- ✓ **Set budgets and alerts**
Per product and environment, each with an owner and an agreed action at the limit.

VALUE

Report value in terms finance accepts

VALUE TYPE	EVIDENCE REQUIRED	COMMON MISTAKE
Run-rate cost removed	Invoice stopped or asset decommissioned, net of any replacement run cost	Counting a retirement before the bill stops
Cost avoided	Spend in an approved budget or forecast that will no longer occur	Constructing the baseline after the fact
Capacity released	Hours redeployed to named work	Reporting hours saved as cash savings
Risk reduced	Exposure removed, such as unsupported systems retired	Presenting avoided breaches as savings
Revenue enabled	An attribution method agreed with finance in advance	Taking full credit for shared outcomes

ACCOUNTING

Agree the accounting treatment before comparing options

Options with the same total cost can report differently, depending on what is capitalized and what is expensed.

SPEND	USUAL TREATMENT	WHAT TO CHECK
Cloud, SaaS and AI usage fees	Operating expense, as consumed or over the subscription term	Replacing depreciated assets with fees lowers EBITDA
SaaS implementation, US GAAP	Qualifying costs capitalized, then expensed over the hosting term in the same line as the fees (ASU 2018-15)	The term includes renewals reasonably certain to be exercised
SaaS configuration, IFRS	Usually expensed as the supplier does the work, or over the access period (IFRIC agenda decision, 2021)	Code the customer controls may be an intangible asset
Internal-use software, US GAAP	From annual periods beginning after December 15, 2027 (ASU 2025-06): capitalized once funded and probable to complete	Significant development uncertainty delays capitalization

PITFALLS

Common errors in technology business cases

✗ Counting license fees only

✓ Include infrastructure, support, integration and labor

✗ Double counting across programs

✓ One register of benefits, owned by finance

✗ Ignoring parallel running

✓ Budget for old and new systems running together

✗ Missing contract dates

✓ Plan around notice periods, renewals and true-ups

✗ Pricing AI from the pilot

✓ Model cost at production volume and concurrency

✗ Benefits with no owner

✓ Name an owner and review each benefit quarterly

GETTING STARTED

Build cost transparency in three stages

STAGE 1

Visibility

- ◆ Map the ledger to cost pools and towers
- ◆ Tag cloud and AI spend; map SaaS contracts to products
- ◆ Identify the 20 most expensive applications

STAGE 2

Accountability

- ◆ Allocate cost to products and consumers
- ◆ Publish showback every month
- ◆ Introduce unit costs for three products

STAGE 3

Decisions

- ◆ Budget and forecast by product
- ◆ Require run cost in every investment case
- ◆ Track benefits quarterly against the baseline

SUMMARY

Key takeaways

- 01 Trace spend from ledger to consumer, so each cost has an accountable owner.
- 02 FinOps now spans public cloud, SaaS, data cloud platforms, data centers and AI.
- 03 Tokens, context length, model choice and agent loops drive AI cost; design for them.
- 04 Use unit costs the business already manages.
- 05 Report only value that finance can verify, with an owner for every benefit.

NEXT COURSE · U101

Technology Portfolio Intelligence

Understand · Foundation