Intelligence on the Device: Pricing Privacy, Latency and Hardware Reach in AI M&A

A Transaction Framework for Hybrid Inference Economics, Installed-Base Access and Defensible Edge Value

Authored by Chennakeshav Adya
Independent Researcher
September 2026

Abstract

Artificial-intelligence inference is moving across cloud, private infrastructure, personal computers, phones, vehicles, industrial systems and embedded devices. Apple states that Core ML can run predictions and fine-tuning on a person's device and can use CPU, GPU and Neural Engine resources while reducing network dependence. Google provides AI Edge tooling and documents mobile deployment of Gemma models. MLCommons publishes mobile inference benchmarks with defined scenarios, quality targets, latency metrics and power measurement. Qualcomm describes on-device and hybrid AI as a way to address privacy, latency and infrastructure-cost constraints, while acknowledging workload, power and memory challenges.

These capabilities create a strategic M&A thesis. A buyer may seek an application with privileged access to an installed device base, a model company with efficient edge technology, an orchestration layer that routes work across device and cloud, or a semiconductor and software stack that expands addressable workloads. The buyer can also overpay for a demonstration that reaches few active devices, uses unsupported model rights, performs poorly outside the laboratory or transfers cost from cloud invoices into engineering, battery, memory, distribution and support.

This paper develops a device-to-value framework for investors, boards and transaction teams. It maps workload placement, supported hardware, accepted outcome, privacy and latency evidence, complete cost, customer price, retention and strategic reach. It distinguishes on-device, edge-server, private-cloud and public-cloud routes; constructs a hardware-reach denominator; sets out hybrid-routing, model-lifecycle, cyber-security, data-rights, regulatory, accounting and valuation requirements; and translates evidence into diligence, deal terms and a 180-day value-creation programme.

A hypothetical enterprise acquisition case demonstrates the method. Every device count, activation rate, task volume, latency result, price, cost, retention rate, synergy and valuation input in that case is a management assumption created solely to show the calculation. The case is not a forecast, market quotation or representation of a named company. Six figures present the placement continuum, reach funnel, hybrid router, contribution bridge, valuation matrix and integration roadmap. Seven tables provide the workload taxonomy, reach ledger, privacy-latency proof map, hypothetical operating case, sensitivity matrix, diligence checklist and valuation bridge.

The central conclusion is that on-device capability has enterprise value when it expands accepted customer outcomes and durable contribution across a verified supported estate. Privacy, latency and offline operation are commercial propositions that require evidence and pricing. Hardware reach is an economic denominator, not a marketing count. M&A underwriting should therefore follow the complete path from model and device compatibility through activation, accepted outcome, customer behaviour, cost and cash.

AI, privacy, intellectual property, consumer, sector, cyber-security, export, tax, accounting, valuation, financing and investment decisions require current advice from qualified professionals in the relevant jurisdictions. Hardware, operating systems, models, provider terms and regulation can change. This paper provides general information for professional audiences and does not provide legal, regulatory, tax, accounting, valuation, credit, technical or investment advice.

JEL Classification: G24, L11, L21, L63, L86, O32

Keywords: on-device AI, edge inference, hybrid AI, privacy, latency, hardware reach, installed base, M&A, valuation, AI applications

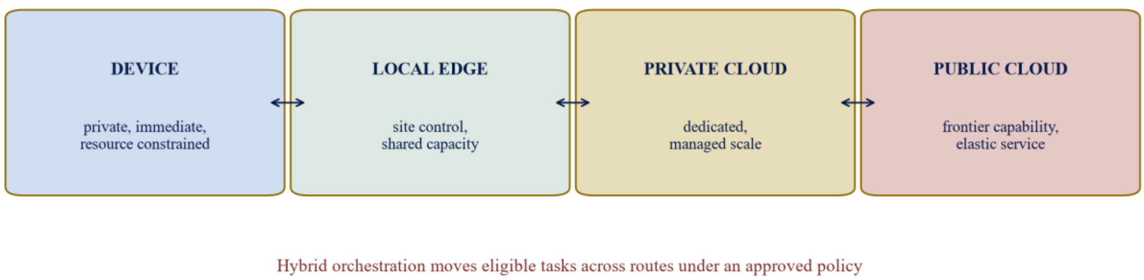
1. Define the on-device transaction thesis

An on-device AI acquisition should solve a specific strategic problem. The buyer may need lower cloud exposure, faster interaction, offline resilience, local data processing, access to a device ecosystem, differentiated hardware utilisation or a distribution channel that places intelligence close to the user. Each thesis produces a different diligence burden and valuation logic.

The seller's claim should be expressed as a measurable chain: compatible device, installed software, activated feature, eligible task, accepted outcome, customer value, realised price and retained cash contribution. A break in the chain limits value. A model can run on a device without users activating it. An activated feature can fail quality or consume excessive battery. A valuable outcome can remain trapped inside a fixed price.

Apple's Core ML documentation states that on-device execution can remove the need for a network connection, support privacy and responsiveness, and use CPU, GPU and Neural Engine resources.[1] Google's AI Edge materials position on-device models around low latency and local data.[2] These are platform capabilities. A transaction team still needs company-specific evidence across the supported estate.

Figure 1. The workload-placement continuum



Author framework. Placement follows task capability, consequence, connectivity, data and economics.

2. Classify the workload before selecting the route

The relevant question is not whether the product is on-device. It is which tasks belong on which route. A bounded classification, local transcription or sensor-detection task can fit a device envelope. A large multi-source investigation can require cloud models, search and extended context. A consequential action can require central controls or authorised review.

Task classification should record input type, model requirement, memory, energy, latency, connectivity, data sensitivity, tool access, consequence and quality threshold. The product can then define an approved route and fallback. The same application may use all four locations in Figure 1.

Table 1. Workload-placement taxonomy

Workload class	Primary value	Typical constraint	Candidate placement	Transaction evidence
Routine personalisation	private adaptation and immediacy	device compatibility and model size	device	activation, acceptance and battery impact
Interactive assistance	low response time and continuity	context, thermal envelope and quality	device or hybrid	tail latency, fallback and user retention
Sensitive enterprise task	local data control	policy, audit and tool permissions	device, local edge or private cloud	data flow, contract and control evidence
Complex synthesis	frontier quality and large context	network, cloud cost and data transfer	hybrid or cloud	quality-response curve and complete cost
Industrial perception	continuous local decision support	reliability, safety and hardware lifecycle	device or site edge	validated performance under operating conditions
Consequential execution	controlled action with side effects	authority, verification and rollback	hybrid with mandatory control	approved workflow, logging and incident evidence

Company-specific evaluation determines which route meets the approved outcome standard.

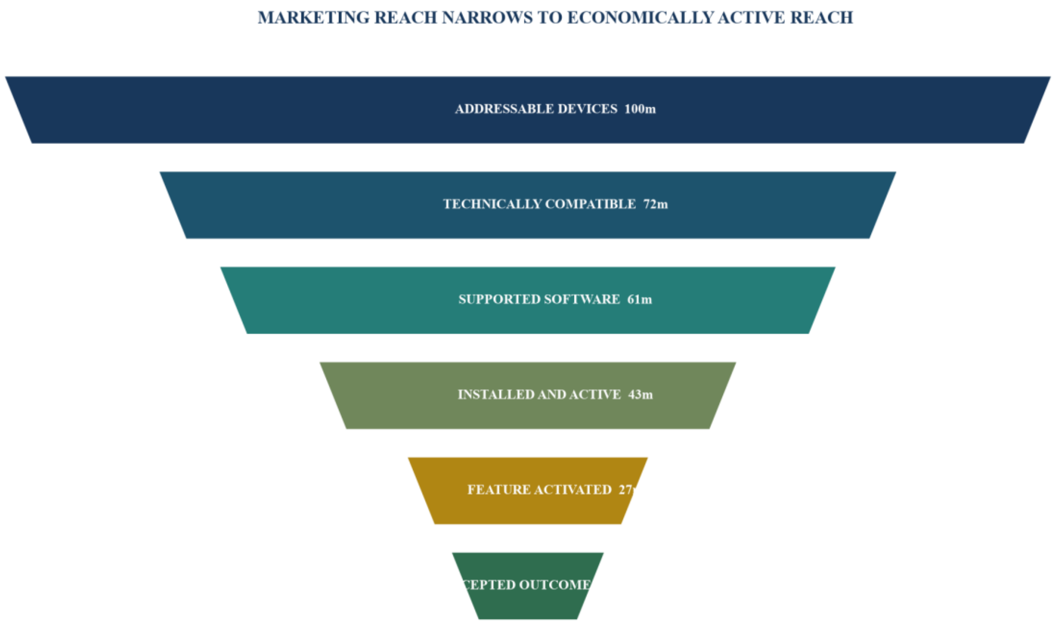
The taxonomy should be versioned. A route that works on a current flagship processor may fail on an older device or after an operating-system change. A seller should state the supported minimum hardware and software for every material feature.

3. Build the hardware-reach denominator

Installed base is frequently presented as strategic reach. The economic denominator is narrower. It begins with devices technically compatible with the model and runtime. It then removes unsupported operating systems, insufficient memory or storage, unavailable accelerators, excluded regions, uninstalled applications, disabled permissions, inactive accounts and users who do not activate the feature.

The remaining active supported estate should be measured by cohort. Devices age, batteries deteriorate, operating systems fragment and vendors change access. A buyer should understand the replacement cycle and the seller's ability to maintain performance without abandoning customers.

Figure 2. Hardware reach to accepted-outcome funnel



Every count is a hypothetical management assumption for method illustration.

Table 2. Hardware-reach ledger

Reach layer	Definition	Evidence	Valuation use
Addressable hardware	devices in the claimed category or ecosystem	platform and market records	upper market boundary only
Compatible hardware	devices meeting processor, memory, storage and accelerator requirements	compatibility test and device matrix	technical reach
Supported estate	compatible devices with supported operating system, region and application version	release and support records	serviceable reach
Installed active estate	supported devices with active application or service use	product telemetry	current distribution access
Activated feature	active devices whose users enable and use the AI feature	consent and product event	adoption denominator
Accepted-outcome users	users receiving the defined result at approved quality and service	evaluation and workflow evidence	economic reach denominator
Paying retained users	accepted-outcome users linked to realised price and renewal	billing, collection and cohort data	contribution and valuation base

The buyer should reconcile each layer to a system of record and a stated measurement date.

4. Turn privacy into an evidenced product claim

Local processing can reduce transmission of eligible raw inputs. It does not prove that all data stays on device. Telemetry, crash logs, retrieval, model updates, cloud fallback, support, advertising, analytics and account systems can still move data. The buyer should map the actual path for every material task and fallback.

The privacy claim should specify which inputs remain local, which derived data leaves the device, how long data is retained, whether a provider can use it for model improvement, where processing occurs and how deletion works. User controls and contract language should match implementation.

Apple's Private Cloud Compute materials describe a security architecture for cloud processing used when a request needs larger models, including verifiable software and privacy protections.[8] The existence of a privacy-oriented cloud design reinforces the need to examine hybrid paths rather than assume a binary device or cloud architecture.

Table 3. Privacy and latency proof map

Claim	Technical proof	Commercial proof	Failure mode
Data remains local	packet, log, code and configuration review	customer terms and privacy notice	hidden telemetry or cloud fallback
Offline availability	disconnected functional test across supported devices	observed use and retention in relevant cohorts	partial feature failure without network
Faster response	percentile end-to-end measurement under representative load	task completion and user preference	benchmark gain without workflow gain
Lower delivery cost	device and cloud cost ledger including engineering and support	contribution by cohort	cost shifted into support or hardware burden
Stronger trust	consent, control and incident evidence	willingness to adopt, renew or pay	marketing claim without behavioural response
Wider reach	supported active estate and distribution rights	activation, accepted outcomes and collection	headline installed base with low eligibility

Claims enter valuation only when technical, contractual and customer evidence agree.

5. Price latency and offline resilience

Latency is an end-to-end service measure. Model execution time is only one component. Model loading, prompt construction, retrieval, tool use, thermal throttling, memory pressure, operating-system scheduling and user-interface rendering can influence the result. The product should measure median and tail latency by device class and task.

MLCommons defines mobile inference scenarios and reports percentile latency, throughput and quality targets.[3] Standardised results can inform hardware comparison, while product diligence requires the actual application, model and device estate. A benchmark result on an available flagship system does not establish performance on the seller's full installed base.

Offline resilience can carry value in travel, field work, industrial operations and constrained networks. The product should state which functions work offline, how long they remain available, how updates and credentials behave, and how state synchronises when connectivity returns. Customer evidence should show that resilience influences adoption, retention or price.

6. Reconstruct complete device economics

On-device execution can reduce metered cloud inference for eligible tasks. It introduces another cost stack: model conversion and compression, device-specific optimisation, application size, download bandwidth, storage, evaluation, operating-system certification, telemetry, support, cyber-security, model update, rollback and a wider compatibility matrix.

Energy and battery use affect the customer experience and can create support or adoption cost. Memory pressure can limit concurrent functions. Thermal throttling can change sustained performance. Hardware vendors may include accelerators in the device price, while application companies still incur engineering and distribution cost.

The economic ledger should compare routes at the accepted-outcome level. A cloud route may cost more per request and support a wider capability. A device route may have low variable provider cost and high fixed engineering cost. The breakeven depends on eligible volume, reach, quality, lifecycle and customer price.

7. Govern model size, compression and quality

Device deployment often uses smaller models, quantisation, pruning, distillation or specialised architectures. These techniques can reduce memory, energy and latency. They can also change quality, robustness and behaviour. Every material variant requires evaluation on representative tasks and devices.

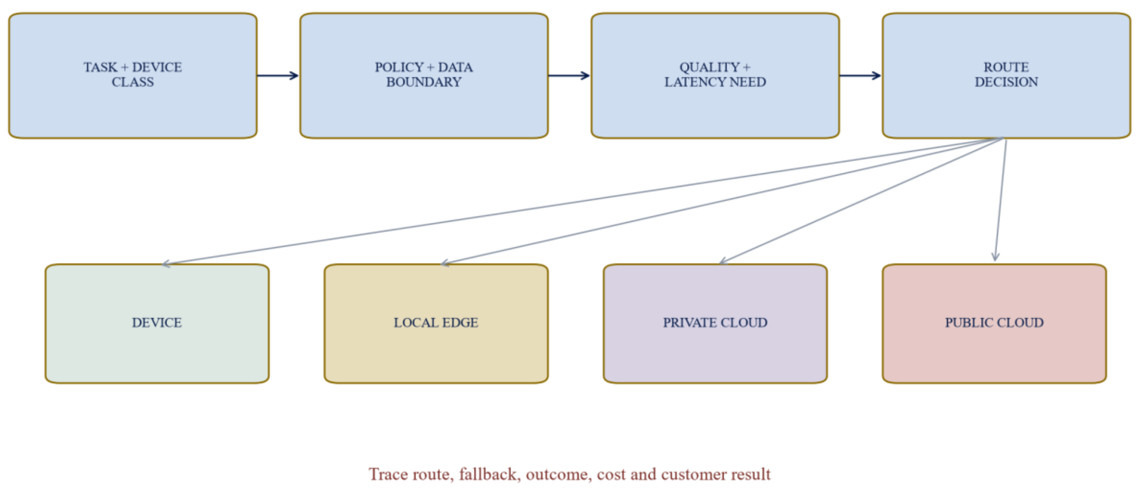
Apple publishes example model information with device-specific inference times and model variants.[4] Google documents mobile deployment paths for Gemma through its edge tooling.[5] These sources show that performance depends on model, precision, device and software stack. The target company's measured results remain the transaction source of truth.

The model register should record source, licence, training and fine-tuning rights, architecture, precision, supported devices, runtime, quality threshold, release date, rollback and end-of-support. A transaction should identify whether a critical optimisation depends on a founder, a non-transferable tool or confidential platform access.

8. Use hybrid routing as the control plane

Hybrid routing decides whether to execute locally, at site edge, in private cloud or in public cloud. It can preserve local processing for sensitive or interactive tasks and escalate complex work to stronger models. The router should evaluate device capability, task class, consequence, connectivity, data policy, quality, latency, cost and service availability.

Figure 3. Hybrid inference router and control points



Author framework. Every route remains subject to the approved task, data, quality and service envelope.

Fallback economics belong in the ledger. A device route can fail because the model is unavailable, the hardware is unsupported, resources are constrained or quality falls below threshold. The cloud fallback may protect the outcome while increasing cost and changing the data path. The product should disclose that path to the customer where required.

9. Underwrite distribution and platform control

On-device value often depends on a platform owner, semiconductor vendor, operating system, app store, original-equipment manufacturer or enterprise device manager. The target may have contractual access, a technical integration, a preferred distribution position or merely the ability to publish an application. These positions have different durability.

The buyer should examine approval rights, APIs, entitlements, store rules, revenue share, ranking, pre-installation, default status, update controls, security requirements and termination. A distribution agreement can expire. An operating-system owner can replicate a feature. A hardware roadmap can deprecate the target's optimisation.

Strategic reach also depends on enterprise deployment. Mobile-device management, security review, procurement, model update and data policy can determine adoption. Consumer download counts should not be treated as enterprise distribution evidence.

10. Secure model, software and data rights

The transaction needs transferable rights to every material model, weight, dataset, runtime, compiler, library and optimisation. Open-source labels do not remove licence obligations. Model licences can restrict use, redistribution, hosted service, branding or scale. Third-party code can impose notice, source or patent conditions.

Training and evaluation data require provenance, permission and retention records. Device telemetry can contain personal or commercially sensitive information. The buyer should understand whether customer contracts permit the intended integration, model improvement and cross-product use.

Employee and contractor assignments should cover inventions, code, models, data pipelines and documentation. Joint development and university arrangements can create background and foreground rights. A critical model optimisation that cannot transfer can weaken the acquisition thesis.

11. Treat cyber-security as part of edge economics

Distributing a model to devices changes the attack surface. Adversaries can inspect application packages, extract weights, manipulate inputs, tamper with runtime state or exploit update channels. The company should assess model confidentiality, code signing, secure boot, hardware-backed keys, attestation, sandboxing, update integrity and rollback.

Device compromise can produce false outputs or unsafe actions. Consequential workflows need validation, permissions, rate limits and confirmation. Telemetry should detect abnormal behaviour without undermining the privacy proposition.

Security cost belongs in the product model. A larger estate requires vulnerability response, supported-version management and incident communication. Unsupported devices can become both a security risk and a customer-retention problem. NIST's Secure Software Development Framework provides relevant development guidance, while its AI Risk Management Framework and Generative AI Profile support broader governance.[31][32][33]

12. Map regulatory and export constraints

Privacy, AI, consumer, cyber-security, product-safety, employment and sector rules can apply according to use and jurisdiction. Local processing can reduce some transfers. It does not remove transparency, lawful-basis, security, accuracy, discrimination, human-oversight or record obligations.

The European Union AI Act uses risk-based obligations and includes rules relevant to providers, deployers and general-purpose AI under the applicable timetable.[34] The General Data Protection Regulation and supervisory guidance remain relevant to personal-data processing.[35] Transaction teams should obtain current advice for the actual product and geography.

Export controls and sanctions can affect advanced hardware, software, encryption and technical support. Hardware reach should exclude jurisdictions or customers that cannot be served lawfully. A buyer should not value theoretical reach that is contractually or legally unavailable.

13. Define the M&A rationale and counterfactual

The acquisition case should identify what the buyer gains faster or more reliably than through partnership or internal development. Possible assets include edge-optimised models, compilers, runtime software, device telemetry, distribution rights, an active installed base, enterprise contracts, specialist engineers or hardware relationships.

The counterfactual should estimate build cost, time, platform access, opportunity cost and failure risk. A buyer with a large installed base may value the target's model and runtime. A model company may value the buyer's distribution. A semiconductor company may value workloads that increase accelerator utilisation. The synergy model should follow the relevant party's assets rather than assume every buyer can realise the same value.

Competition review can examine market definition, foreclosure, data, ecosystems, interoperability and innovation. The United States Department of Justice and Federal Trade Commission publish Merger Guidelines; the European Commission and United Kingdom Competition and Markets Authority publish merger-review guidance.[36][37][38] Current legal advice is required for the actual transaction.

14. Build the acquisition operating model

The buyer should decide which components remain independent and which integrate. Model development, device optimisation, application experience, distribution, cloud fallback, customer contracts and governance can have different owners. Integration that centralises every decision can slow a product whose advantage is rapid device-specific iteration.

Architecture and commercial governance should meet. A product team can select the route. Finance should reconcile cost and customer price. Security and privacy teams should approve the data path. Sales should avoid promising unsupported devices or unlimited expensive fallback. The board should see one reach and contribution ledger.

Retention of specialist engineering talent can be important. The diligence team should map critical knowledge, documentation, succession, incentives and work authorisation. A transaction should avoid attributing technology value to an individual whose knowledge has not been institutionalised.

15. Hypothetical acquisition case

Consider a hypothetical acquirer evaluating an enterprise productivity application with on-device and hybrid AI. Every number in this section is a management assumption created solely to demonstrate the method. It does not describe a named company or market forecast.

The seller cites 80 million addressable devices. Technical and support review identifies 52 million compatible devices and 41 million supported devices. The application is active on 14 million supported devices. Four million users activate the AI feature, and 2.8 million produce at least one accepted outcome in the month. Of those users, 1.2 million are linked to paying enterprise or premium accounts.

The product processes 48 million eligible tasks. Fifty-eight per cent execute on device, 12 per cent at local or private edge, and 30 per cent in public cloud or cloud fallback. Monthly recognised revenue allocated to the AI-enabled offer is AED 8.4 million. Complete direct cost is AED 3.36 million, producing AED 5.04 million of contribution, or 60.0 per cent.

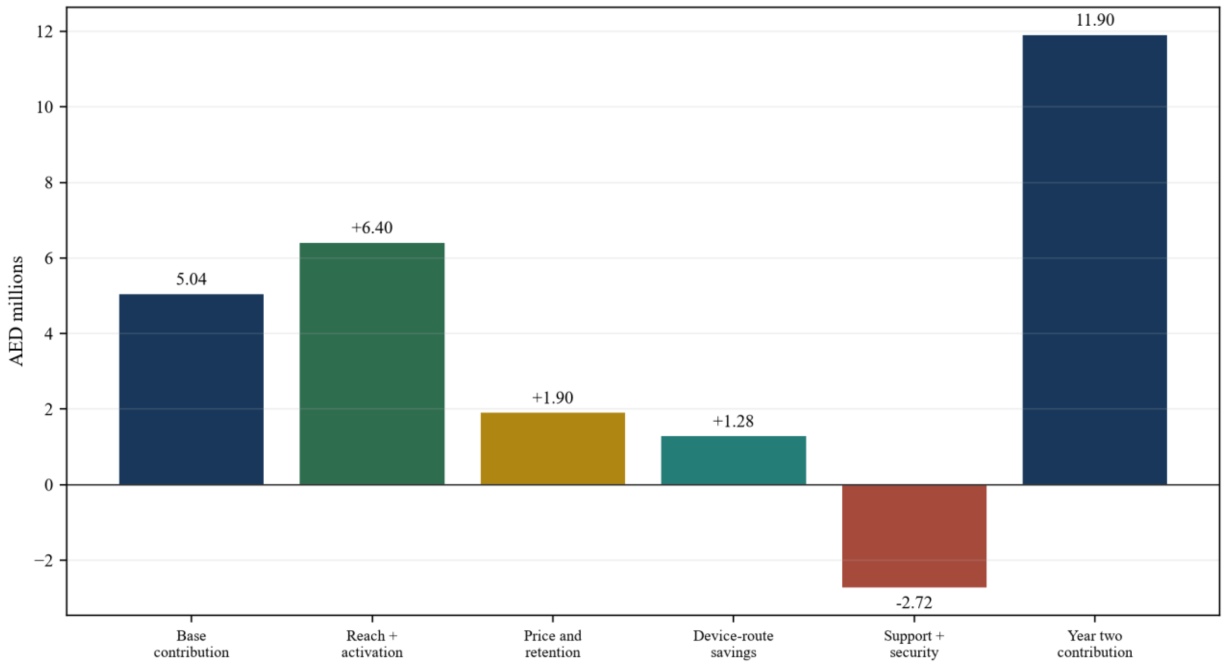
Table 4. Hypothetical on-device AI operating case

Metric	Base month	Year-two case	Evidence gate
Supported active devices	14.0m	22.0m	compatible hardware, software support and active telemetry
Accepted-outcome users	2.8m	6.2m	task evaluation and user event
Paying accepted-outcome users	1.2m	3.0m	billing, collection and account linkage
Eligible monthly tasks	48.0m	118.0m	classified task telemetry
Device and local-edge share	70%	78%	route trace and fallback reconciliation
Recognised monthly revenue	AED 8.40m	AED 18.60m	contracts, billing and revenue policy
Complete direct cost	AED 3.36m	AED 6.70m	device, cloud, engineering, control and support ledger
Contribution margin	60.0%	64.0%	consistent complete-cost policy
Monthly cloud fallback rate	14%	8%	failed-route and escalation records

Every value is an illustrative management assumption and requires company-specific evidence.

The year-two case assumes wider supported reach, higher activation, more paying users, improved device execution and a lower fallback rate. It also funds model updates, security and support. Each change requires operating evidence. A buyer should keep unsupported reach and untested routing savings outside the base case.

Figure 4. Hypothetical monthly contribution bridge



All values are illustrative management assumptions in AED millions.

16. Stress reach, routing and customer price

The downside should begin with reach. A compatibility problem can reduce eligible devices. An operating-system change can increase support cost. A quality gap can increase cloud fallback. Privacy concerns can reduce activation. A platform change can weaken distribution. These effects can occur together.

Table 5. Hypothetical transaction sensitivities

Scenario	Paying reach	Device route share	Realised price	Complete direct cost	Contribution margin	Interpretation
Base	index 100	70%	index 100	AED 3.36m	60.0%	illustrative current case
Strong privacy adoption	index 118	74%	index 106	AED 3.74m	64.0%	customer response supports premium value
Hardware fragmentation	index 78	55%	index 98	AED 3.62m	50.5%	reach falls and cloud fallback rises
Platform access loss	index 64	63%	index 95	AED 3.20m	45.1%	distribution and activation weaken
Quality regression	index 85	48%	index 94	AED 3.88m	39.8%	retries, fallback and support rise
Controlled hybrid routing	index 108	79%	index 101	AED 3.08m	66.2%	requires verified quality and estate coverage
Combined downside	index 58	42%	index 88	AED 4.10m	28.0%	liquidity and valuation protection required

Every value is an illustrative management assumption for method demonstration.

Sensitivity should flow into cash, integration funding and transaction structure. A buyer should model update and support cost through the hardware lifecycle rather than use one steady-state percentage. Device replacement can expand capability while leaving a tail of customers on older hardware.

17. Value supported reach and retained contribution

An on-device AI valuation should begin with the supported economic estate rather than the installed-base headline. The opening denominator is the number of devices that satisfy the required processor, memory, storage, operating-system, security and application conditions. The next gates are active use, feature activation, accepted outcomes, paying accounts and retained contribution. Each gate should have a reproducible source and a period definition.

The income approach can model cash flows from paying use, renewal, expansion and verified cost avoidance. Revenue should reflect the contract and revenue-recognition policy. Direct cost should include device testing, model optimisation, software updates, cloud fallback, telemetry, evaluation, security, support and platform fees. Working capital, capital expenditure, tax and integration spending then convert operating contribution into cash. IFRS 13 and the International Valuation Standards provide relevant fair-value and valuation principles; IFRS 3, IAS 36 and IAS 38 govern important accounting questions after a business combination.[40][41][42][43][45]

The market approach requires careful normalisation. A revenue multiple from a cloud-software company can misstate value when the target bears device-qualification cost, hardware fragmentation and long support tails. A semiconductor multiple can misstate value when the target controls neither chip economics nor distribution. Comparable transactions should be adjusted for supported reach, paying penetration, route mix, margin policy, customer concentration, recurring revenue, intellectual-property control and the maturity of the operating evidence.

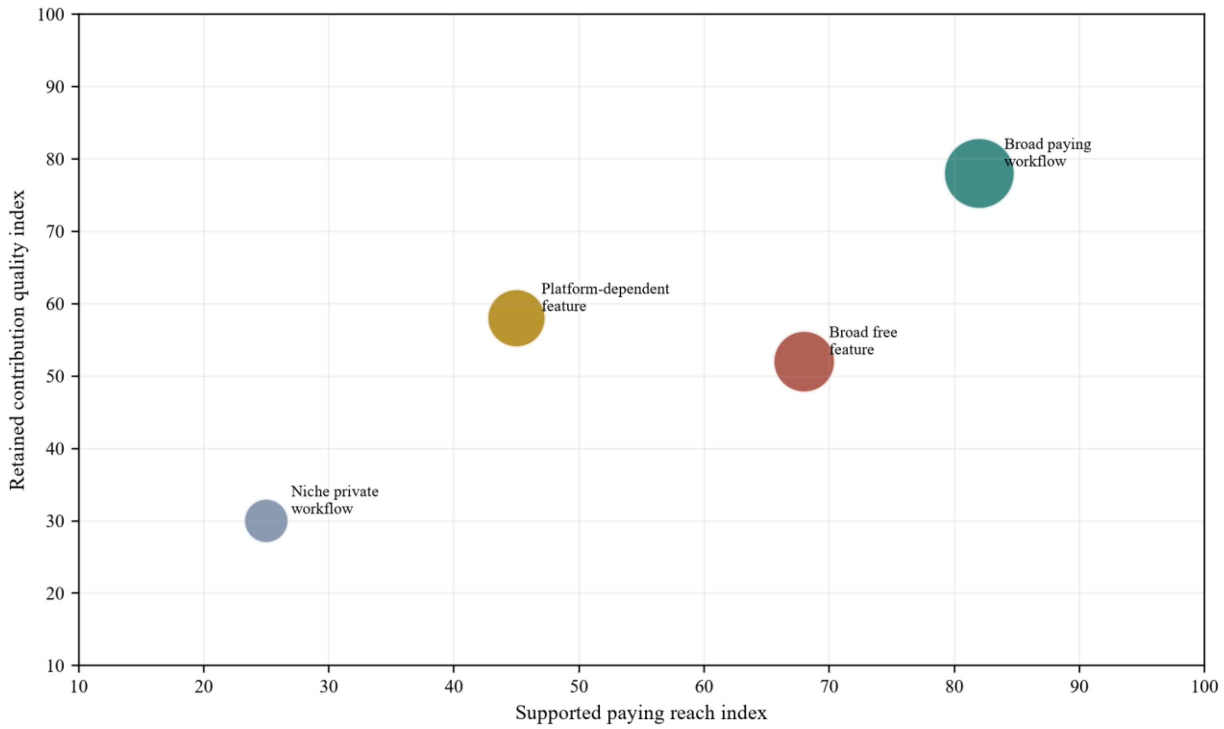
A cost approach can help test the replacement cost of models, runtimes, compilers, data, integrations and specialist teams. It rarely captures distribution, customer contracts or time-to-market on its own. The analysis should also recognise obsolescence. A technically impressive model can lose economic value if a platform owner supplies a substitute, if hardware generations change, or if the target lacks the rights to maintain and commercialise the stack.

Table 6. Valuation bridge from installed base to transaction value

Layer	Required evidence	Value treatment	Common overstatement
Installed base	platform and device records	contextual only	treating every shipped device as reachable
Compatible estate	minimum hardware and software test	scenario denominator	ignoring memory, thermal and OS constraints
Supported active estate	release matrix, telemetry and support policy	operating denominator	counting obsolete or unsupported devices
Activated use	consented feature events	adoption evidence	equating availability with use
Accepted outcomes	task evaluation and user acceptance	product-value evidence	counting calls or tokens as customer value
Paying accepted use	contract, billing and account linkage	revenue driver	attributing free use to paid demand
Retained contribution	complete-cost cohort ledger	cash-flow foundation	excluding device and control costs
Buyer-specific synergy	executable integration plan	separate probability-weighted case	paying the seller for the buyer's own assets

The bridge is a diligence structure. Company-specific values require verified records and an approved valuation method.

Figure 5. Illustrative reach and value matrix



The matrix is a transaction-screening framework. It does not represent observed market data.

18. Separate buyer synergies from target stand-alone value

Synergy analysis should name the resource, owner, action, timing, cost and dependency. A handset manufacturer may contribute distribution and hardware integration. A software platform may contribute accounts, developer access and a billing rail. A model provider may contribute research capacity. A semiconductor company may contribute optimisation tools and accelerator access. The target should receive stand-alone value for resources it controls. Buyer-owned distribution or procurement power belongs in a separate synergy case.

Revenue synergies require a route to customer. The model should identify eligible accounts, sales motion, activation, realised price, renewal and cannibalisation. Cost synergies require a complete baseline and a credible implementation plan. Cloud-cost reduction should deduct the added cost of device engineering, estate testing, observability and support. Headcount savings should account for retention, severance, knowledge transfer and the need to maintain the acquired capability.

Double-counting is a recurring risk. Higher revenue, higher margin and a higher multiple may all reflect the same improvement. The valuation committee should show the causal bridge and use one treatment. It should also separate synergies available to several plausible buyers from synergies unique to the selected acquirer. Competitive tension can affect price, while the investment case still needs an executable cash-flow basis.

19. Build a reproducible diligence room

The diligence room should connect product claims to records. A buyer should be able to select a device cohort, reproduce eligibility, trace a task through routing and quality controls, link the accepted outcome to the account, reconcile the invoice, and identify the direct cost. Sampling should cover important device families, operating-system versions, geographies, customer segments, workloads and failure states.

Technical evidence includes model cards, evaluation sets, release histories, runtime dependencies, device matrices, power and thermal tests, fallback logs, update mechanisms, security architecture and incident records. Commercial evidence includes contracts, price schedules, entitlements, usage cohorts, renewals, support tickets, channel terms and platform dependencies. Financial evidence includes revenue policy, cloud invoices, device-lab spend, engineering allocation, warranty or support provisions, capitalised development and cash collection.

Legal diligence should cover ownership, open-source obligations, model and data licences, employee and contractor assignments, privacy notices, consent, processor terms, export controls, customer restrictions and change-of-control clauses. Security diligence should test the update chain, key management, local data protection, model extraction, malicious input handling and fleet response. NIST's AI Risk Management Framework, Cybersecurity Framework and Secure Software Development Framework can support a structured review.[29][30][31][33]

Table 7. Minimum on-device AI transaction diligence checklist

Workstream	Minimum evidence	Decision output
Hardware reach	device matrix, benchmark results, active telemetry	supported paying denominator
Product quality	task taxonomy, evaluation set, accepted-outcome rate	route and product envelope
Economics	contract-to-cash trace, complete direct-cost ledger	cohort contribution and cash
Privacy	data map, consent, local/cloud path, deletion tests	claim boundary and remediation
Security	threat model, signing, update, incident and fleet controls	risk register and funding plan
Technology rights	code, model, data and dependency provenance	ownership and licence schedule
Platform exposure	distribution, OS, store and hardware terms	concentration and continuity plan
Regulation	product, AI, privacy, export and sector analysis	jurisdiction conditions
Organisation	critical roles, documentation and retention	integration and retention plan
Valuation	stand-alone case, sensitivities and synergies	price and protection range

The checklist requires tailoring to the product, transaction and jurisdictions.

20. Allocate risk through transaction terms

Price structure can reflect evidence quality. Consideration may combine cash at close, deferred payments, escrow, holdbacks, earn-outs or contingent value tied to measurable outcomes. Metrics should be within the operating team's influence, consistently defined and auditable. Installed-base or inference-volume metrics can reward uneconomic activity. Paying accepted use, retained contribution, renewal and supported estate quality are usually closer to durable value, subject to the actual deal.

Representations and warranties can address intellectual-property ownership, licences, privacy, security, compliance, contracts and financial information. Covenants can require maintenance of documentation, access, security response or regulatory cooperation between signing and closing. Specific indemnities may address identified issues. Warranty and indemnity insurance can alter recourse but does not replace diligence. Legal advisers should design the package for the transaction and jurisdictions.

Integration funding should sit beside purchase price. The buyer may need device laboratories, model optimisation, security remediation, contract migration, platform certification, retention and cloud capacity. A lower purchase price does not create value if the implementation budget is absent. The board should approve acquisition consideration, required investment and downside liquidity as one capital decision.

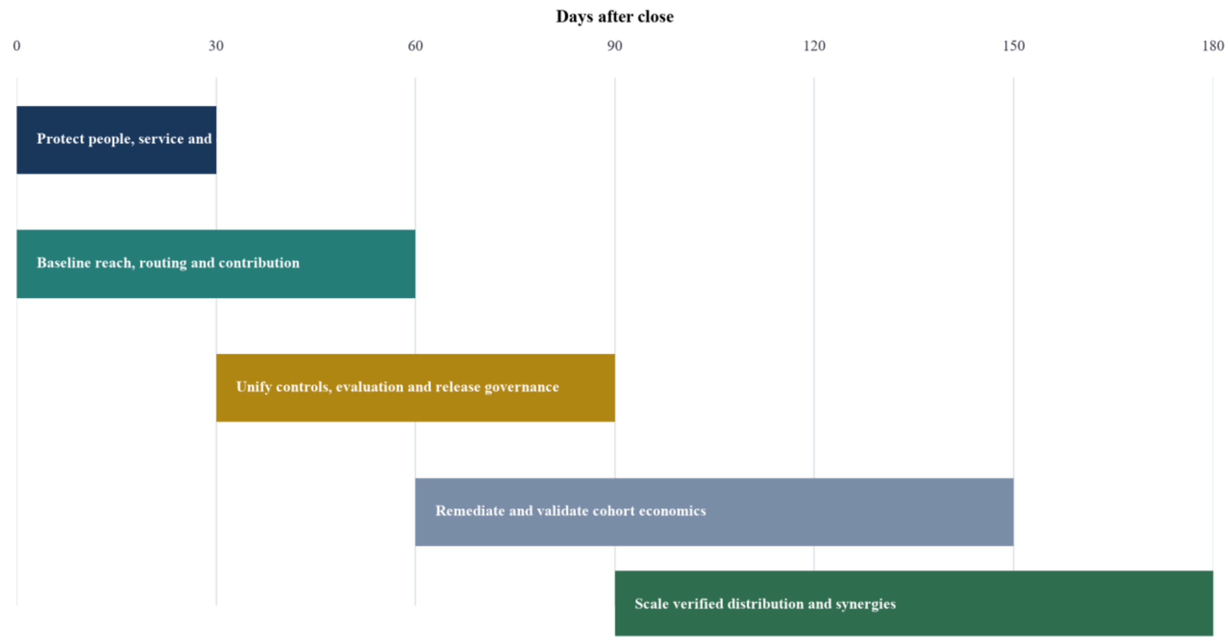
21. Run a 180-day evidence-led integration

The first thirty days should preserve service, people and evidence. The buyer confirms ownership, protects critical staff, freezes unsupported product promises, establishes the reach and contribution baselines, and names accountable leaders for product, finance, privacy, security and platform relationships. Material incidents and contractual deadlines receive immediate attention.

Days 31 to 90 establish common measurement and controls. The team reconciles the device matrix, route telemetry, evaluation method, complete-cost policy and customer entitlements. It remediates critical security or privacy gaps, aligns release governance and tests the investment case on selected cohorts. Commercial teams receive an approved claims and pricing framework.

Days 91 to 180 scale only verified improvements. The team expands supported hardware where quality and contribution pass the gate, renegotiates material platform or supplier terms where feasible, and launches cross-sell or premium offers with controlled measurement. The investment committee receives the stand-alone case, realised synergies, remediation spend and revised downside.

Figure 6. 180-day on-device AI integration sequence



The sequence is a general operating framework and requires transaction-specific adaptation.

22. Make the transaction decision explicit

The final decision memorandum should state the transaction thesis in one sentence, identify the evidence that supports it, and show the conditions that can invalidate it. It should present stand-alone value, buyer-specific synergies, integration cost, downside liquidity and the proposed risk allocation. Open findings need an owner, deadline and treatment in price, terms or closing conditions.

A proceed decision may be appropriate when the target controls important technology or distribution, supported paying reach is reproducible, accepted outcomes are strong, complete contribution is durable, rights are clear, and integration is executable. A conditional decision may require remediation, a staged investment, commercial partnership, licence, asset purchase or contingent consideration. A decline decision can be rational when reach depends on unsupported devices, quality requires persistent expensive fallback, rights are uncertain, platform concentration is uncontrolled, or the seller's valuation relies on buyer-owned synergies.

Governance continues after approval. Monthly operating review should follow supported reach, activation, accepted outcomes, route mix, fallback, complete cost, revenue, retention, incidents and integration spend. Quarterly investment review should reconcile realised cash and synergies to the approved case. Changes in hardware, operating systems, models, regulation or platform terms should trigger a refreshed view.

The board pack should preserve the relationship among technical, customer and financial evidence. A hardware-reach schedule should reconcile the opening and closing supported estate, additions, removals, software changes and inactive devices. The product schedule should reconcile activated users, task attempts, accepted outcomes, failures, fallbacks and customer support. The commercial schedule should reconcile entitlements, realised price, recognised revenue, invoicing and cash. The cost schedule should reconcile device execution, cloud execution, engineering, testing, security, platform, support and incident costs. This common data model prevents separate teams from presenting incompatible success measures.

Thresholds require explicit approval. Management can define minimum quality, maximum latency, privacy restrictions, energy or battery limits, support cost, fallback rate and contribution for each task class. The router applies those policies in production. Exceptions should identify the approving owner and the period of validity. A high-value regulated workflow can justify a different cost and control envelope from a low-value consumer feature. The investment case should reflect the actual mix rather than one portfolio average.

Evidence quality should affect confidence and capital release. Directly reproduced contract-to-cash and device-to-outcome records support the base case. Samples, management estimates and incomplete cohorts belong in a probability-weighted or downside case. A signed customer contract does not establish economic reach when the customer's deployed hardware cannot run the product. A successful laboratory demonstration does not establish retained contribution when production support and fallback are unknown. The decision memorandum should state these boundaries plainly.

The financing plan should match the risk. Stable recurring contracts with verified contribution may support acquisition debt or recurring-revenue facilities, subject to lender requirements. Volatile usage, concentrated platforms, rapid hardware change or material remediation can require more equity, delayed consideration or milestone funding. Debt capacity should use downside cash after integration and support costs. Covenants can monitor liquidity, customer concentration, supported estate, contribution or other agreed measures where definitions are auditable.

Exit planning begins at entry. A future strategic buyer may value the same technology differently because its distribution, hardware, data and platform position differ. A financial buyer will require an executable stand-alone case and management depth. Public-market investors may focus on recurring revenue, margin policy, concentration, security and capitalised development. The current acquisition model should therefore preserve separable evidence for stand-alone performance and each realised synergy. That record reduces dependence on a narrative at exit.

The board should also specify conditions for changing course. A platform restriction, critical vulnerability, unexpected regulation, adverse customer response or persistent quality regression can require route redesign, pricing change, narrower hardware support, additional capital or product withdrawal. Pre-agreed triggers accelerate action. They also prevent sunk acquisition cost from overriding current evidence.

An independent review can test whether the model remains reproducible. The reviewer selects cohorts, repeats reach and outcome calculations, traces costs and challenges synergy ownership. Findings feed the risk register and valuation. The review is especially useful before a major deferred payment, refinancing, impairment test or expansion into a new hardware family or regulated market.

Conclusion

On-device AI can create transaction value through privacy, responsiveness, offline resilience, distribution and reduced dependence on public-cloud inference. Those benefits become investable when the buyer can demonstrate supported hardware reach, accepted customer outcomes, clear rights and retained contribution after complete device and control costs.

The decisive unit is the paying accepted outcome delivered across a supported estate. Installed devices, model parameters and inference counts are supporting measures. A disciplined transaction process classifies workloads, verifies the route, reconstructs complete economics, tests rights and platform dependencies, separates stand-alone value from buyer synergies, and funds the operating plan.

Hybrid architecture is often a portfolio of routes rather than a binary design. The buyer therefore underwrites a control system: which tasks run where, under what privacy and quality constraints, at what complete cost, and with what fallback. That system should remain measurable as hardware, models and customer expectations change.

Frequently asked questions

Q: What does on-device AI mean in this paper? A: It means that some or all model inference occurs on the user's or enterprise's device. The economic analysis also covers local-edge and hybrid routes because many products move tasks among device, private infrastructure and public cloud.

Q: Why is installed base an unreliable valuation metric? A: Shipped devices can lack the processor, memory, software, security state, application version or support required for the product. Valuation should use the supported active estate and then test activation, accepted outcomes, paying use and retained contribution.

Q: Does on-device inference remove cloud cost? A: Device execution can reduce selected public-cloud inference. The complete ledger still includes model optimisation, device testing, update infrastructure, observability, security, support, platform fees and cloud fallback.

Q: How should privacy value be measured? A: The company should document the data path and test claims against actual product behaviour. Commercial evidence can include activation, win rate, retention, realised price and loss avoidance in privacy-sensitive cohorts.

Q: Which workloads belong on the device? A: Suitable workloads depend on model size, hardware, latency, power, quality, data sensitivity, offline need and control requirements. A workload taxonomy and reproducible benchmark should precede routing decisions.

Q: How should an acquirer treat hardware fragmentation? A: The acquirer should maintain a device and software matrix, measure active supported reach, budget testing and support by cohort, and model the long tail of older hardware in margins and integration funding.

Q: How does on-device architecture affect M&A valuation? A: It changes the supported reach, direct-cost structure, platform dependency, intellectual-property review, integration plan and risk allocation. Valuation should follow durable cash flow and separate buyer-owned synergies from target stand-alone value.

Q: Which Matchpoint practice is relevant to this work? A: This research connects to Matchpoint Partners' strategy and execution and M&A advisory work, including technology diligence, operating-model design, valuation, transaction structuring, integration planning and value creation.

References

[1] Apple, Core ML, https://developer.apple.com/documentation/coreml

[2] Google, Google AI Edge, https://ai.google.dev/edge

[3] MLCommons, Mobile Working Group, https://mlcommons.org/working-groups/benchmarks/mobile/

[4] Apple, Core ML Models, https://developer.apple.com/machine-learning/models/

[5] Google, Gemma on Mobile and Web, https://ai.google.dev/gemma/docs/integrations/mobile

[6] Apple, Core AI, https://developer.apple.com/documentation/coreml

[7] Apple, Foundation Models Framework, https://developer.apple.com/documentation/foundationmodels

[8] Apple, Private Cloud Compute Security Guide, https://security.apple.com/documentation/private-cloud-compute/

[9] Apple, Privacy, https://www.apple.com/privacy/

[10] Google, LiteRT, https://ai.google.dev/edge/litert

[11] Google, Google AI Edge Gallery, https://github.com/google-ai-edge/gallery

[12] Google, Gemini Nano, https://developer.android.com/ai/aicore

[13] Google, Built-in AI APIs in Chrome, https://developer.chrome.com/docs/ai/built-in

[14] Qualcomm, The On-Device AI Opportunity, https://www.qualcomm.com/news/onq/2024/02/the-on-device-ai-opportunity

[15] Qualcomm, The Future of AI Is Hybrid, https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/hybrid-ai-is-the-future-of-ai.pdf

[16] Qualcomm, Enabling On-Device Generative AI, https://www.qualcomm.com/news/onq/2023/05/enabling-on-device-generative-ai

[17] Qualcomm, AI Hub, https://aihub.qualcomm.com/

[18] Qualcomm, Annual Reports, https://investor.qualcomm.com/financial-information/annual-reports

[19] Arm Holdings, Annual Reports, https://www.arm.com/company/investors/financial-results

[20] Arm, Artificial Intelligence, https://www.arm.com/markets/artificial-intelligence

[21] Intel, OpenVINO Toolkit, https://docs.openvino.ai/

[22] Microsoft, ONNX Runtime Mobile, https://onnxruntime.ai/docs/tutorials/mobile/

[23] NVIDIA, Jetson Documentation, https://docs.nvidia.com/jetson/

[24] MLCommons, MLPerf Inference, https://mlcommons.org/benchmarks/inference-datacenter/

[25] MLCommons, MLPerf Inference Policies and Results, https://mlcommons.org/benchmarks/inference/

[26] International Energy Agency, Energy and AI, https://www.iea.org/reports/energy-and-ai

[27] Organisation for Economic Co-operation and Development, Measuring the Environmental Impacts of AI Compute and Applications, https://www.oecd.org/en/publications/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_7babf571-en.html

[28] National Institute of Standards and Technology, Privacy Framework, https://www.nist.gov/privacy-framework

[29] National Institute of Standards and Technology, Cybersecurity Framework 2.0, https://www.nist.gov/cyberframework

[30] National Institute of Standards and Technology, AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

[31] National Institute of Standards and Technology, Generative AI Profile, https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

[32] National Institute of Standards and Technology, Adversarial Machine Learning Taxonomy and Terminology, https://csrc.nist.gov/pubs/ai/100/2/e2025/final

[33] National Institute of Standards and Technology, Secure Software Development Framework, https://csrc.nist.gov/Projects/ssdf

[34] European Commission, AI Act, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[35] European Union, General Data Protection Regulation, https://eur-lex.europa.eu/eli/reg/2016/679/oj

[36] United States Department of Justice and Federal Trade Commission, Merger Guidelines, https://www.justice.gov/atr/merger-guidelines

[37] European Commission, Merger Control, https://competition-policy.ec.europa.eu/mergers/merger-control_en

[38] United Kingdom Competition and Markets Authority, Merger Assessment Guidelines, https://www.gov.uk/government/publications/merger-assessment-guidelines

[39] United States Federal Trade Commission, Premerger Notification Program, https://www.ftc.gov/enforcement/premerger-notification-program

[40] IFRS Foundation, IFRS 3 Business Combinations, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-3-business-combinations/

[41] IFRS Foundation, IFRS 13 Fair Value Measurement, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/

[42] IFRS Foundation, IAS 36 Impairment of Assets, https://www.ifrs.org/issued-standards/list-of-standards/ias-36-impairment-of-assets/

[43] IFRS Foundation, IAS 38 Intangible Assets, https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/

[44] IFRS Foundation, IFRS 15 Revenue from Contracts with Customers, https://www.ifrs.org/issued-standards/list-of-standards/ifrs-15-revenue-from-contracts-with-customers/

[45] International Valuation Standards Council, International Valuation Standards, https://www.ivsc.org/standards/

[46] Apple, Annual Reports, https://investor.apple.com/sec-filings/default.aspx

[47] Alphabet, Annual Reports, https://abc.xyz/investor/

[48] Qualcomm, SEC Filings, https://investor.qualcomm.com/financial-information/sec-filings

[49] Arm Holdings, SEC Filings, https://investors.arm.com/financials/sec-filings

[50] International Organization for Standardization, ISO/IEC 27001 Information Security Management Systems, https://www.iso.org/isoiec-27001-information-security.html

About the Author

Chennakeshav (CK) is a corporate finance and investment banking executive with 25+ years of global experience in deal origination, structuring and execution across M&A, growth capital and corporate strategy. He has led value-creation mandates for founders, corporates and funds — bridging the boardroom view to hands-on execution and close.

His career spans Morgan Stanley, HSBC, Lloyds Banking Group, EWEC, ADQ portfolio companies and Emirates Growth Fund, across TMT, real estate, fintech, deeptech, cleantech, infrastructure and energy. He has partnered with C-suite leaders, private equity and venture funds, sovereign wealth funds and family offices to finance complex fund raises and scale-up ventures, and has led M&A due diligence, post-merger integration and business-transformation initiatives to create value.

At Matchpoint Partners he is Managing Partner, leading the firm's corporate finance, M&A and capital-raising practice. He holds an MBA from London Business School, an engineering degree from VTU and a Master of Laws (LLM, in progress) from UCL London.

An active start-up mentor, CK mentors at Techstars, DIFC FinTech Hive, Startup Grind, Founder Institute and IN5, serves as Entrepreneur Mentor in Residence (EMiR) at London Business School, and judges the Entrepreneurship World Cup.

https://www.linkedin.com/in/ckadya/

https://www.matchpoint-partners.com/team/ck-adya.html

This paper is part of a continuing series on the structure of private and alternative markets. The views expressed are the author's own. The paper is for information only, describes market structure in general terms, and does not constitute investment, legal, tax or regulatory advice or a recommendation in respect of any security, vehicle or counterparty.