La rareté compte la commande, pas la consommation

Quand le capital, l'architecture et la valeur se découplent, l'allocation des ressources suit le marché qui émet encore un prix. L'IA d'entreprise en est le cas courant.

La thèse de la déployabilité soutenait que le problème difficile de l'IA d'entreprise n'est pas le modèle mais son insertion institutionnelle. Ce texte énonce le mécanisme que le travail précédent contournait, au niveau où il est réutilisable plutôt qu'au niveau où il est spectaculaire.

Quand trois marchés qui devraient s'aligner se disjoignent, un marché du capital qui met un prix sur la capacité future, un marché de l'architecture qui met un prix sur ce qui se construit, et un marché de la valeur qui mettrait un prix sur les résultats réalisés s'il le pouvait, l'allocation des ressources est gouvernée par celui des trois qui émet encore un signal de prix.

En IA d'entreprise, à la mi-2026, ce signal provient de façon disproportionnée du marché du capital, parce que le marché de la valeur ne dispose pas encore d'un instrument de mesure standardisé. Le GPU, le token, l'agent, la rareté, le ROI client, sont le mécanisme et ses symptômes. La thèse, énoncée généralement, est que cela se produit partout où la création de valeur est différée ou difficile à mesurer et où le capital ne l'est pas, ce qui explique qu'elle devrait se lire hors de l'IA.

Deux termes doivent être fixés avant qu'ils ne dérivent.

- Premièrement, par rente ce texte entend le coût d'exploitation encouru au-delà de ce que la tâche exige pour parvenir à la même décision utile. Le surplus, et seulement le surplus, pas l'agent, pas la complexité qu'un problème exige réellement.
- Deuxièmement, disproportionnée est délibéré. Le marché de l'architecture répond à plusieurs signaux à la fois, le talent, la régulation, les standards ouverts, l'économie du cloud et le prix du compute. Le capital est simplement le plus fort d'entre eux aujourd'hui, pas le seul.

La formule du titre est une conséquence, et une compression. La rareté est surdéterminée, reflétant à la fois les anticipations, la demande réelle, les plans de construction, les contraintes industrielles et la consommation future attendue ; la formule isole la seule composante que la lecture haussière prend pour le tout, la

commande passée contre un plan, et n'exige qu'une chose, qu'on ne la confonde pas avec la charge servie. Le domaine de validité est étroit : le déploiement en entreprise sous facturation à l'usage, sur les chiffres réalisés de la mi-2026, plus fort là où l'achat est une décision de comité et où le signataire est isolé du compteur, plus faible là où un propriétaire unique met un prix sur sa propre trésorerie. Ce n'est pas un pronostic de bulle. Au moins un laboratoire de frontière projette un résultat opérationnel positif au deuxième trimestre 2026 sur des marges brutes proches de 77%.

I. Allocation du capital : qui fixe la demi-vie de l'actif

Les hyperscalers dépensent à un rythme qu'aucun revenu à court terme ne justifie, parce que le marché récompense le positionnement plutôt que la rentabilité. Les cinq plus grands sont en trajectoire vers environ 725 milliards de dollars de dépenses d'investissement en 2026, part d'un déploiement qu'Allianz Research chiffre près de 2 100 milliards sur 2026 à 2028, à une intensité capex de 34% du chiffre d'affaires, le double du pic de 15% de la construction internet de la fin des années 1990, et maintenue là quoi qu'il arrive, ce qui est le signe révélateur. Quand le marché punit celui qui cligne le premier, construire en avance sur la demande est une discipline défensive, et personne n'a besoin de comploter pour écouler du compute pour que le surinvestissement se produise.

Un acteur fixe l'horloge qui transforme le surinvestissement en urgence, et aujourd'hui c'est surtout Nvidia. Intel vendait des processeurs ; Nvidia fixe désormais la demi-vie économique du data center, chaque architecture qu'il livre, Hopper puis Blackwell puis la génération Rubin présentée à GTC 2026, comprimant la vie économique de la précédente. La position est contestée plutôt que détenue : AMD, Broadcom, le TPU de Google, le Trainium d'Amazon et les puces spécialisées en inférence de Groq et Cerebras pressent tous sur le rythme qu'il impose. Le point structurel survit à la concurrence. Celui qui fixe l'horloge de dépréciation force le propriétaire à remplir sa capacité avant qu'elle n'expire, et le round-tripping que le discours attribue vaguement au secteur vit ici, dans les flux de capital où le cash d'un investisseur revient sous forme de son propre revenu, bien plus que dans le revenu d'exploitation des laboratoires.

II. Économie de l'infrastructure : ce que mesure la pénurie

Deux faits cohabitent dans le même marché. La mémoire à haute bande passante et le packaging sont pratiquement en rupture, les délais de livraison des GPU se comptent en trimestres, et l'énergie a remplacé le silicium comme contrainte limitante, les desks d'Apollo et de JPMorgan décrivant une chaîne d'approvisionnement tendue jusqu'en

2026. Face à cela, la revue 2026 de Cast AI, portant sur environ vingt-trois mille clusters Kubernetes d'entreprise, trouve une utilisation GPU moyenne proche de 5%.

Le 5% doit être vu avec circonspection. Il mesure la capacité Kubernetes provisionnée en entreprise, pas le compute global, et pas l'inférence de production des laboratoires, qui tourne à plein ; un analyste de Constellation Research a justement mis en garde sur le fait que l'échantillon penche vers le test et les parcs mal optimisés. Lu à sa portée, il prouve une chose : une large population de clusters d'entreprise, provisionnés pendant la ruée, reste oisive tandis que le fournisseur la facture, qu'elle tourne ou non.

Le paradoxe apparent se dissout dès qu'on demande contre quoi la pénurie est mesurée. La mémoire est en rupture contre des commandes, et les commandes sont passées contre des objectifs de construction. Le contre-argument le plus fort mérite tout son poids : la fibre à la fin des années 1990 et le cloud à la fin des années 2000 ont tourné à des taux d'utilisation initiaux dérisoires et ont précédé leurs applications de plusieurs années, de sorte que la surcapacité pourrait être le décalage normal d'une transition dont la demande est latente. La disanalogie est l'horloge de dépréciation de la couche précédente. La fibre était bon marché à garder oisive ; la capacité GPU ne l'est pas, amortie en 18 à 36 mois. La lecture du décalage et la thèse divergent sur la durée pendant laquelle l'écart peut persister avant d'être passé en perte, et c'est pourquoi l'écart est le chiffre à surveiller : David Cahn, de Sequoia, situe la distance annuelle entre la dépense et le revenu de l'écosystème près de 600 milliards de dollars, et en creusement.

L'intuition de prix de l'acheteur mérite une note plus prudente que celle qu'écrirait un cynique. Il est tentant de dire que l'entreprise est dupée par le produit grand public subventionné, l'offre gratuite où environ 95% des utilisateurs d'un assistant ne paient rien, jusqu'à croire l'intelligence bon marché. C'est exagéré. L'acheteur de 2026 est souvent la même personne qui lit une facture AWS à la recherche d'anomalies depuis dix ans. L'écart qui compte n'est pas la culture du prix, que l'acheteur mûr possède, mais l'absence d'une métrique de qualité, que personne ne possède : un DSI peut lire la facture de tokens au centime près et ne détenir aucun instrument qui lui dise si l'architecture qui la génère était dimensionnée au problème.

III. Architecture logicielle : les cinq sources du poids

Une requête de chatbot est une inférence ; un workflow agentique qui planifie, appelle des outils, vérifie et s'auto-corrige déclenche 10 à 20 appels pour accomplir une tâche, un vrai changement d'échelle, et non le facteur mille qui circule sans source. L'orchestration multi-agents n'est pas, en 2026, un artifice de facturation. C'est la réponse opérante aux limites de raisonnement d'un modèle monolithique. La distinction n'est pas l'agent contre l'absence d'agent. C'est la consommation contre la consommation productive.

Le poids a cinq sources, et une seule est la cible. Il peut être une propriété du problème : l'identité, la sécurité, l'observabilité et l'orchestration sont difficiles, et un déploiement de qualité entreprise est plus lourd qu'une démonstration. Il peut être une propriété de la régulation : sous l'AI Act ou l'Espace européen des données de santé, une architecture qui journalise, isole, documente et garde un humain dans la boucle est plus lourde qu'une architecture qui ne le fait pas, et ce poids est une optimalité juridique, rationnelle sur un axe qui n'est pas l'axe économique. Il peut être l'optionnalité, et d'un genre particulier que la vue financière manque : un déploiement surdimensionné peut être rationnel non pour le retour du cas d'usage courant mais comme ticket d'entrée, le prix de la stabilisation des compétences internes, de la restructuration de la donnée de l'entreprise et du rodage d'une capacité dont l'organisation aura besoin ensuite. Ce surplus est des frais de scolarité, pas de la rente, et seul l'acheteur sait lequel il a acheté. Il peut être l'immaturité, un agent empilé sur un agent par une équipe sous délai. Et il peut être la rente, le surplus tel que défini. Quatre des cinq sont légitimes. L'argument porte sur la cinquième, et toute sa difficulté est que la cinquième est indiscernable des quatre autres sans un baseline que presque personne n'exécute, l'architecture moins chère placée à côté pour voir si elle aurait suffi. La thèse ne prétend pas que les architectures d'entreprise sont trop lourdes. Elle prétend que rien dans les instruments de l'acheteur ne sépare le poids justifié du poids qui est rente.

Une seule illustration fixe le point sans porter l'argument. Un modèle de toxicité calibré associant une molécule à quatorze endpoints, du genre construit sous le nom de ToxTwin sur le terrain propre de l'Institut, est un calcul qui prédit, rapporte une confiance et s'arrête, à une fraction de ce qu'un agent générique dépenserait pour parvenir au même résultat. C'est un cas où le baseline est évident. Dans la plupart des situations d'entreprise il ne l'est pas, et c'est cela, non l'existence de l'option frugale, qui est le problème.

IV. Déploiement en entreprise : deux incitations, une direction

L'architecture atteint le client par le déployeur, qui n'est pas un acteur mais deux fonctions d'utilité qui se trouvent alignées. Le bras conseil de l'hyperscaler hérite de l'impératif du propriétaire, le taux d'occupation. L'intégrateur de systèmes, Accenture, Capgemini, Deloitte et les autres, tourne sur une fonction différente, les heures facturables, la maintenance, le support et le change management, dont aucun ne récompense l'architecture plus légère, et c'est souvent l'intégrateur, plus proche du parc du client et payé à la complexité de l'intégration, qui est le plus déterminant des deux. Aucun n'exige de mauvaise foi. Même là où la complexité est entièrement légitime, aucun ne fait face à une incitation économique à rechercher la simplicité, et le biais survit aux

bonnes intentions de tous parce qu'il est une propriété du champ d'incitations, pas du caractère de quiconque.

Confrontons le récit du déploiement au relevé. Un déploiement est vendu comme un gain de productivité, un chiffre de 40% sur telle équipe. L'initiative NANDA du MIT, dans The GenAI Divide: State of AI in Business 2025, trouve environ 95% des pilotes d'IA générative sans impact mesurable sur le compte de résultat et environ 5% atteignant une accélération rapide des revenus ; le Widening AI Value Gap du Boston Consulting Group de septembre 2025 situe 60% des entreprises à aucune valeur matérielle ; l'enquête de McKinsey de fin 2025 trouve un impact sur l'EBIT à l'échelle de l'entreprise chez moins de quatre adoptants sur dix. Le récit survit au relevé parce que le coût qu'il omet est institutionnel : l'évaluation, l'observabilité, la gouvernance et la revue humaine qu'un système confiant mais faux impose. Une architecture peut être bon marché en tokens et ruineuse en organisation, et un chiffre de ROI client qui ne compte ni l'un ni l'autre ne mesure pas un retour.

V. La règle manquante, et pourquoi elle résiste à sa construction

Rien de ce qui précède ne conclut une vente si l'acheteur détient un instrument qui met un prix sur le poids face à la valeur. Aujourd'hui, le plus souvent, il ne l'a pas. George Akerlof, dans « The Market for Lemons » (1970), a montré que lorsque les acheteurs ne peuvent observer la qualité, le marché sélectionne ce qui est observable, et la qualité invisible ne peut être payée ; ici l'inobservable est la ligne entre consommation productive et rente, l'observable est la démonstration et le benchmark, et le marché sélectionne ce que l'offre est récompensée à montrer, c'est-à-dire le poids.

L'objection évidente est que si la règle avait tant de valeur, le marché l'aurait déjà construite, donc son absence serait temporaire ou illusoire. La vérité est que la règle est réellement difficile à construire, pour quatre raisons qui se composent. La causalité est diffuse : la valeur d'une bonne décision se répand sur les équipes et les trimestres et se rattache rarement à une ligne. La temporalité est longue : le retour sur la réingénierie d'un processus émerge des années après la facture de compute. L'attribution est quasi impossible : quand un humain valide la sortie du modèle, la valeur de la décision ne peut être proprement partagée entre le modèle et l'expert qui l'a corrigé. Et les externalités sont organisationnelles : les gains atterrissent sur le moral, la rétention, la vitesse et l'optionnalité, dont aucun n'est comptabilisé par la finance. Un marché ne manque pas de construire un instrument par paresse. Il en manque parce que la quantité résiste à la mesure, et la règle doit être forgée contre cette résistance plutôt qu'adoptée sur étagère.

Elle se forge malgré tout, et c'est pourquoi la défaillance est une phase et non un destin. Le FinOps est né pour discipliner la dépense cloud, le LLMOps et l'observabilité des

modèles pour instrumenter l'inférence, le benchmarking et le process mining pour attacher des chiffres aux résultats, et du côté de l'offre les modèles à poids ouverts et les petits modèles, la distillation et l'inférence sur site permettent à une entreprise d'échapper entièrement au tuyau facturé au compteur. La thèse est donc datée : la pression est réelle maintenant, tant que l'instrument est immature, et elle reflue à mesure que l'instrument se standardise. Le parallèle historique demande la même retenue. L'ERP tournait sur le lock-in, le cloud sur les économies d'échelle, le CRM sur de faibles effets de réseau, l'IA d'entreprise sur la facturation variable ; ce ne sont pas les mêmes marchés, et ce qu'ils partagent est seulement qu'à l'achat la qualité économique ne pouvait être mesurée, si bien que chacun a été acheté sur la foi avant d'être rationalisé plus tard, la condition que Solow a nommée en 1987, l'âge de l'ordinateur visible partout sauf dans les statistiques de productivité. Une chose est nouvelle : sous une licence, le coût de la période non mesurée était fixe, et sous la facturation à l'usage il ne l'est pas. Une vague qu'on peut rationaliser plus tard est une vague dont le compteur ne tourne pas pendant qu'on attend. Celui-ci tourne.

L'instrument, une fois nommé, n'est pas un instrument mais une famille, et pas une famille neutre. Appelons-le coût par décision utile : un ratio dont le dénominateur compte les décisions dont la valeur aval dépasse le coût de leur vérification, dont le numérateur somme le coût total sur la durée de vie de leur production, le calcul plus la supervision humaine plus la gouvernance plus la vérification plus l'exploitation, portant un terme d'option pour l'adaptation future qu'une architecture plus lourde achète réellement. Le dénominateur est spécifique au domaine et ne peut être autrement, une décision utile étant correcte et défendable pour un assistant juridique, rapide et sûre pour un système de triage, assez précise pour agir pour une prévision. Et voici sa limite la plus tranchante : celui qui définit la décision utile contrôle le ratio, de sorte que si l'intégrateur capture cette définition, la métrique devient marketing plutôt que discipline, et l'asymétrie se déplace simplement du token vers le mot utile. La règle est réelle et vaut d'être construite. Le combat autour de son dénominateur est le prochain endroit où la rente tentera de se cacher.

VI. La contribution, sa portée, et sa preuve

La version la plus profonde de la thèse est un découplage de la performance technique et de la performance économique. Il fut un temps où la performance technique se traduisait en valeur économique assez directement pour qu'optimiser la première serve la seconde ; cette traduction s'est amincie, et entre les deux le marché interpose désormais la consommation, puis la valeur attendue, puis un retour hypothétique, optimisant le premier maillon parce que c'est le seul qu'il sait lire. Les trois marchés sont emmêlés là où ils devraient être distincts, et le marché de la valeur n'émet aucun prix, donc le marché de l'architecture est piloté de façon disproportionnée par le marché du

capital. Cela n'est pas propre à l'IA. Partout où un marché du capital met un prix sur la capacité future, un marché de l'architecture sur ce qui se construit, et un marché de la valeur ne peut pas encore mettre un prix sur les résultats, le second est de plus en plus gouverné par le premier, un motif visible en pointillé dans les plateformes de biotechnologie, l'infrastructure énergétique, le quantique et l'achat de défense. L'IA est le cas où le compteur tourne le plus vite, ce qui explique qu'elle montre le motif la première et le pire.

Une tension doit être résolue plutôt que lissée, parce que deux défaillances sont en jeu et une seule est transitoire. L'incapacité de l'acheteur à mesurer est transitoire : elle se referme à mesure que la règle se construit. L'incapacité du vendeur à garantir un résultat probabiliste peut être permanente : un système non déterministe résiste à la garantie contractuelle qu'exige la tarification au résultat, si bien que la tarification au token pourrait ne pas disparaître à maturité, et l'acheteur pourrait internaliser le risque d'architecture pour de bon. Ce sont des choses différentes. La défaillance de marché que décrit ce texte, l'acheteur qui signe sans règle, est une phase. La structure de tarification qui laisse le risque d'architecture à l'acheteur pourrait être un trait permanent d'une technologie probabiliste. Un marché mûr n'est donc pas un marché où le vendeur porte le risque, mais un marché où l'acheteur le porte en connaissance de cause, avec un instrument pour le mettre en prix.

Un mot sur ce qu'est cet argument. Il propose un mécanisme et l'illustre ; il ne le mesure pas. Akerlof, Solow, Cast AI, MIT, Sequoia et le reste illustrent une chaîne qui n'a pas été testée comme chaîne. La variable indépendante est énoncée clairement, l'acheteur possède-t-il une métrique de qualité, et les variables dépendantes suivent, la complexité architecturale, le coût d'exploitation, le retour réalisé ; la thèse est que la première gouverne le reste tant que la métrique est absente, et le statut honnête de cette thèse est une hypothèse assortie d'un test spécifié, pas un résultat. Le pas suivant n'est pas un argument de plus mais un protocole qui mesure le coût par décision utile sur des déploiements appariés avec et sans la métrique et lit la différence. En attendant, ceci est une théorie avec de bonnes illustrations, chose plus petite qu'une preuve et plus grande qu'un essai.

Pour le décideur, la conséquence n'est ni l'adoption sur la foi ni le retrait dans la panique, la même erreur tournée dans deux directions opposées. C'est de construire la règle plus vite que le marché ne la standardise, et de refuser les deux substitutions sur lesquelles tourne la phase actuelle : la pénurie contre un plan lue comme une pénurie contre un besoin, et la dépense engagée lue comme un retour réalisé. Les organisations qui auront mis un prix sur leurs déploiements en décisions utiles seront encore debout quand le compteur aura fini de tourner. Les organisations qui les auront mis en prix en conviction découvriront à quoi servait la facture.

Un ROI client affirmé et jamais mesuré n'est pas un retour. C'est un récit avec une facture jointe.