

Die versteckte KI-Steuer

Nvidia, ASML und die Frage, wer im KI-Stack eigentlich das Nadelöhr hält — vier Ebenen, vier Margen, eine Rechnung.

Lesezeit 6 Min

Investoren-Brille

Handwerk statt Hype

1 • Was passiert ist

Nvidia hat für sein jüngstes Quartal einen Umsatz von rund **96 Milliarden US-Dollar** gemeldet — mehr als doppelt so viel wie im Vorjahr — bei einer Bruttomarge von etwa **75 Prozent**. Die Bruttomarge ist der Anteil vom Umsatz, der nach Abzug der reinen Herstellkosten übrig bleibt.

Parallel dazu erinnert ein Marktkommentar daran, dass ohne die Lithografie-Systeme von **ASML** gar kein moderner KI-Chip entstehen könnte. Lithografie heißt: Maschinen, die feinste Chip-Muster mit Licht auf Silizium belichten. Die modernsten Anlagen kosten teils **über 200 Millionen US-Dollar pro Stück**.

Beide Meldungen zusammen werfen eine einfache, aber wichtige Frage auf: Wer entlang der Kette vom Werkzeug bis zur fertigen KI-Antwort verdient eigentlich wie viel — und wie stabil ist das?

2 • Die Mechanik dahinter

Der KI-Stack besteht aus mehreren Schichten, die jede für sich Geld verdienen, bevor überhaupt eine Frage an ein Sprachmodell beantwortet wird:

Ebene	Wer	Was verdient wird
Werkzeug	ASML	Verkauft die EUV-Maschinen, ohne die keine Foundry moderne Chips fertigen kann. Kein Modell-Risiko, keine Cloud-Konkurrenz — reines Ausrüstergeschäft.
Fertigung	TSMC	Baut die Chips im Auftrag, mit laut Marktbeobachtung zweistelligen Margen. Trägt das Fertigungsrisiko, aber nicht das Design- oder Absatzrisiko.
Design	Nvidia	Verkauft die fertigen GPU-Systeme zu Preisen im hohen fünfstelligen bis niedrigen sechsstelligen Bereich pro Einheit — mit den höchsten Margen in der Kette.
Cloud	Hyperscaler	Microsoft, Amazon, Google verkaufen die Rechenzeit auf diesen Chips nochmals mit Aufschlag weiter, verpackt als Cloud-Dienst oder API-Zugang.

Jede Ebene addiert ihre Marge. Das Ergebnis ist eine Art **KI-Steuer**, die am Ende jedes Unternehmen zahlt, das Inferenz einkauft — also das tatsächliche Ausführen eines trainierten Modells, wenn es Antworten liefert. Gerechnet wird pro verarbeitetem Token, der kleinsten Texteinheit, die ein Sprachmodell verarbeitet.

3 • Der historische Befund

Solche Stack-Steuern sind kein neues Phänomen. In der PC-Ära hielten Intel und Microsoft über Jahrzehnte eine ähnliche Doppelmacht — bis Cloud und mobile Chips die Architektur aufbrachen. In der Smartphone-Ära verdiente Qualcomm über Lizenzgebühren an praktisch jedem verkauften Gerät, unabhängig vom Hersteller.

Das Muster ist immer gleich: Solange eine Ebene physisch oder patentrechtlich ein Nadelöhr bleibt, hält sie die Marge. Sobald Wettbewerb oder Effizienzsprünge die Engpässe auflösen, bröckelt sie.

Genau das deutet sich jetzt auf der Modell-Ebene an: Der chinesische Anbieter Zhipu hat mit **GLM-5.3-Flash** ein offenes Modell mit rund 320 Milliarden Parametern vorgestellt, das bei Coding-Aufgaben nah an etablierte Konkurrenz herankommen soll. Alibaba zieht mit dem effizienten, multimodalen **Qwen3.8-Flash** nach, das Kosten explizit drücken soll. Und dass das mysteriöse Benchmark-Modell „Ox Alpha“ sich als eben dieses GLM-5.3-Flash entpuppte, zeigt: Der Wettlauf um Modell-Vorherrschaft wird über Marketing-Nebelkerzen geführt, weil die eigentlichen Leistungsunterschiede kleiner werden.

Das ist der historische Reflex: Wenn eine Ebene der Kette Preisdruck bekommt, verlagert sich die Marge — meist nach unten, in Richtung des Nadelöhrs mit dem stabilsten physischen Monopol.

4 • Was jetzt zählt

Für die kommenden Quartale lohnt sich der Blick auf drei Bruchstellen:

- **Effizienz-Chips** wie OpenAIs „Jalapeño“ könnten die Rechenkosten pro Token senken — gut für Anwender, potenziell schlecht für die Margen der Chip-Designer, die bislang von hohem Rechenhunger profitieren.
- **Open-Source-Modelle aus China** drücken den Preis auf der Modell-Ebene. Das hilft Hyperscalern durch billigere Inputs, stellt aber Nvidias Preissetzungsmacht indirekt unter Beobachtung, sobald weniger High-End-Rechenleistung pro Antwort nötig ist.
- **ASML bleibt der stabilste Punkt** in der Kette — solange es keinen zweiten Anbieter für modernste EUV-Technik gibt, ist diese Ebene gegen Preisdruck von unten weitgehend immun.

Am Rand spielt auch die politische Großwetterlage mit hinein: Wenn Bundesstaaten in den USA und Wahlbehörden wie in Brasilien strengere Regeln für synthetische Inhalte durchsetzen, steigt der regulatorische Druck auf die gesamte KI-Wertschöpfungskette. Das betrifft zwar primär Anwendungen, nicht Chips, ist aber ein Hinweis darauf, dass der KI-Boom nicht nur an technischen, sondern auch an politischen Grenzen gemessen wird.

5 • Die Investoren-Brille

Handwerk statt Hype heißt hier: Nicht fragen „Ist Nvidia teuer?“, sondern „**Welche Ebene der Kette halte ich eigentlich — und wie stabil ist ihr Nadelöhr?**“

Drei Fragen für den Depot-Check:

- Wie viel Prozent meines KI-Exposures hängt allein an Nvidia als Designer — und wie viel an den weniger sichtbaren Ebenen wie ASML (Werkzeug) oder TSMC (Fertigung)?
- Was passiert mit meiner Position, wenn die Kosten pro Token durch Effizienz-Chips oder Open-Source-Modelle spürbar fallen — wird meine Beteiligung eher zum Gewinner durch mehr Nutzung oder zum

Verlierer durch weniger Marge pro Einheit?

- Habe ich nur die „Luxusmarge“ im Depot oder auch die Ebene mit dem stabilsten physischen Monopol?

Der kopierbare Prompt für deine eigene Recherche:

Analysiere anhand der mir vorliegenden, öffentlich bekannten Zahlen [Zahlen hier einfügen], welche Ebene im KI-Stack – Werkzeughersteller, Foundry, Chip-Designer oder Cloud-Anbieter – aktuell die stabilste Preissetzungsmacht hat. Erkläre für jede Ebene ein konkretes Risiko, falls günstigere Modelle oder Effizienz-Chips den Preis pro Rechenschritt drücken. Gib keine Kauf- oder Verkaufsempfehlung, sondern eine nüchterne Pro-Contra-Liste je Ebene.

Checkliste vor der nächsten Entscheidung:

- Kenne ich die Rolle jeder Position in meinem Depot innerhalb der Kette — Werkzeug, Fertigung, Design oder Cloud?
- Habe ich mir angesehen, wie viele Anbieter es auf dieser Ebene realistisch gibt: einen, zwei oder einen offenen Markt?
- Verstehe ich, wodurch die Marge auf „meiner“ Ebene gefährdet werden könnte — Wettbewerb, Regulierung oder technischer Fortschritt bei der Effizienz?

Kein Vergleich, kein Chart und kein Quartalsbericht ersetzt diese einfache Übung: die Kette entlanggehen und fragen, wer wirklich das Nadelöhr hält. Genau das ist der Unterschied zwischen Hype folgen und Handwerk betreiben.

Quellen

Nvidia-Earnings: Wenn Rechenzeit zur Lizenz zum Geldddrucken wird · ASML: Ohne EUV keine KI — warum der Engpass strategischer wird · Zhipu GLM-5.3-Flash und Qwen3.8-Flash: Chinas Open-Source-Konter auf Claude & Co. · Qwen3.8-Flash: Alibabas effizientes Multimodal-Workhorse für Unternehmen · Brasilien definiert Deepfake-Härtetest — US und EU unter Zugzwang · US-Bundesstaaten im KI-Wahlkampf: Deepfake-Gesetze zwischen Schutz und Zensurgefahr · Ox Alpha entpuppt sich als GLM-5.3-Flash — was das über Benchmark-Wettläufe verrät

Transparenzhinweis: Dieses Dossier wurde redaktionell erstellt und verantwortet; KI-Tools wurden zur Recherche und Strukturierung eingesetzt. Es dient ausschließlich Bildungs- und Informationszwecken und stellt keine Anlageberatung, Kauf- oder Verkaufsempfehlung dar (WpHG). Kurs- und Größenangaben nach bestem Wissen zum angegebenen Stand; keine Gewähr.

EGEKID — „Es gibt eine KI dafür“ · Daily AI & Finance News, werktags 6:00 Uhr · kostenlos abonnieren: rockstrategie.com/egekid/

Transparenzhinweis (EU AI Act, Art. 50): Recherche & Text mit KI-Unterstützung — geprüft und verantwortet von Udo A. Schacht.

Keine Anlageberatung (WpHG): Dieses Dossier dient ausschließlich Bildungs- und Informationszwecken und stellt keine Kauf- oder Verkaufsempfehlung dar. Angaben nach bestem Wissen zum angegebenen Stand; keine Gewähr.

Udo A. Schacht · Rockstrategie (Einzelunternehmer) · Gartenstrasse 10a · 31675 Bückeburg · rockstrategie.com · Impressum: <https://ki-trading.ai/impressum/>