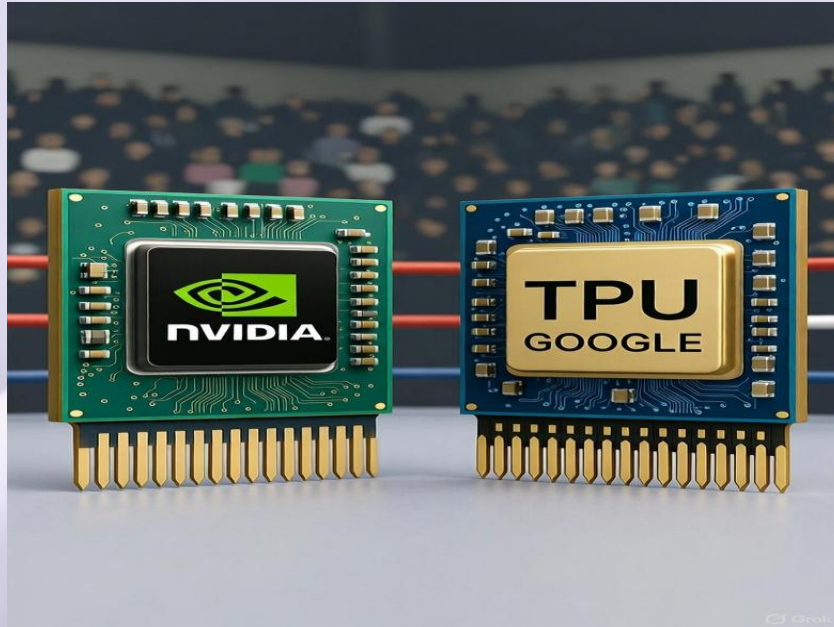




הקרב על מאיצי ה-AI: מדוע ה-TPU של Google אינו מהווה איום ממשי על שליטת NVIDIA



שוק ה-AI בצמיחה מסחררת, אך הקרב על החומרה אינו סימטרי

- שוק החומרה לעולם הבינה המלאכותית גדל בקצב שמעט חסר תקדים, ולצד הביקוש העצום למאיצי NVIDIA, קמה שאלה חדשה: האם שבבי ה-TPU של Google יכולים להפוך למתחרה אמיתית, ואולי אף לשבש את שליטת ה-GPU של NVIDIA?
- לכאורה מדובר בתחרות בין שני סוגי שבבים, אך בפועל זהו עימות בין שתי פילוסופיות עסקיות שונות לחלוטין, והשפעתו על שוק שבבי ה-AI אינה פשוטה להבנה.
- הוויכוח סביב יכולתה של Google להתחרות ב-NVIDIA באמצעות שבבי ה-TPU שלה מקבל משנה תוקף ככל שהביקוש העולמי למחשוב AI עולה. אך מאחורי הכותרות מסתתר פער מהותי בין שני סוגי השבבים, ומה שמשיקיעים רבים מפספסים הוא שה-TPU נולד כמוצר פנימי של Google, בעוד ה-GPU הוא מוצר גלובלי שנועד לשוק הרחב. Google אמנם משתפת פעולה עם Broadcom בפיתוח השבב, אך שומרת לעצמה את כל התכנון הארכיטקטוני הרגיש. Broadcom נחשפת רק לשלב הסופי של המימוש, ללא הידע התכנוני שמאפשר שכפול או מסחור חיצוני. בכך נוצר מבנה פעולה שבו ל-TPU אין נתיב טבעי להפוך למוצר תעשייתי פתוח.
- במקביל, הצמיחה ב-TPU מונעת כמעט לחלוטין מצרכים פנימיים של Google: שילוב AI במנוע החיפוש, בפרסום, ב-Youtube, ב-Gemini ובשירותי הענן. ה-TPU מספק לגוגל שליטה מלאה על תשתית המחשוב שלה ומפחית תלות ב-NVIDIA אך אינו מהווה חלופה מעשית לשוק כולו. מנגד, הביקוש ל-GPU נובע מהתעצמות עולמית בלמידה עמוקה: משלב האימון המתמשך של מודלים מתקדמים, דרך עומסי חישוב מורכבים של מודלים "מחשבים", ועד התפשטות ה-AI לעולמות רובוטיקה, רכב אוטונומי ותעשיות חכמות. אלו תחומים שבהם נדרש ציוד אחיד, גמיש וסטנדרטי – תכונות שבשלב זה רק GPU מסוגל להציע בקנה מידה עולמי.
- ל-TPU עצמו קיימים חסמי צמיחה משמעותיים: הוא דורש תשתיות חשמל וקירור ייעודיות, ארון שרתים ייחודי ורשת פנימית סגורה שאינה קיימת מחוץ לדאטה-סנטרים של Google. בניגוד ל-GPU, למשתמשים אין אפשרות לרכוש ולהתקין TPU באופן עצמאי, והגישה היחידה אליו היא דרך Google Cloud. גם ברמת התוכנה, רוב עולם ה-AI בנוי כיום על PyTorch ו-CUDA, בעוד TPU מבוסס על-XLA, מה שהופך את המעבר לטכנולוגיה של Google לתהליך יקר, מסוכן וקשה ליישום בארגונים. לכך מתווסף מתח עסקי בתוך Google עצמה: צוותי הענן רוצים להשכיר TPU ללקוחות, בעוד צוותי Gemini מעוניינים לשמור את עוצמת המחשוב לעצמם, מחשש לסייע למתחרות שיכולות לפגוע בעסקי הפרסום של החברה.
- התוצאה למשיקיעים ברורה: למרות עלייה חדה בהיקפי השימוש של Google ב-TPU, אין מדובר במתחרה שמסוגל לערער את מעמדה של NVIDIA. ה-TPU ימשיך לשמש כעמוד תווך פנימי וככלי ענן אסטרטגי המיועד למספר מצומצם של לקוחות גדולים, אך שוק ה-AI הרחב יישאר תלוי ב-GPU כפתרון הגמיש, הפתוח והסטנדרטי היחיד. בכך NVIDIA שומרת על יתרון מבני עמוק, בעוד Google מחזקת את יכולותיה, אך כאמור, לא מייצרת איום אמיתי על מאיצי ה-AI הגלובליים של החברה בעשור הקרוב – במיוחד ככל שה-AI ממשיך להגדיל את הביקוש לאנרגיה בעולם.