LIGHTHOUSE CANTON

Not Just Generative - Agentic AI Is Coming, and It's Redefining Compute

Equity Insights
June 2025

Not Just Generative - Agentic AI Is Coming, and It's Redefining Compute

AI computing is becoming more specialized. Instead of relying on just one type of chip (like a GPU), tech giants are now assembling custom silicon stacks to handle the rising complexity and scale of AI. This is especially critical as we move from generative AI—where the focus is on producing content—to Agentic AI, where systems must reason, make decisions, and act independently across continuous workflows.

XPUs and custom ASICs are emerging as key enablers of this new phase, delivering the speed and efficiency needed for real-time, autonomous inference. They're not replacing GPUs—but augmenting them—powering a new generation of AI that doesn't just create, but gets things done.

From Generative to Agentic AI: A New Compute Frontier

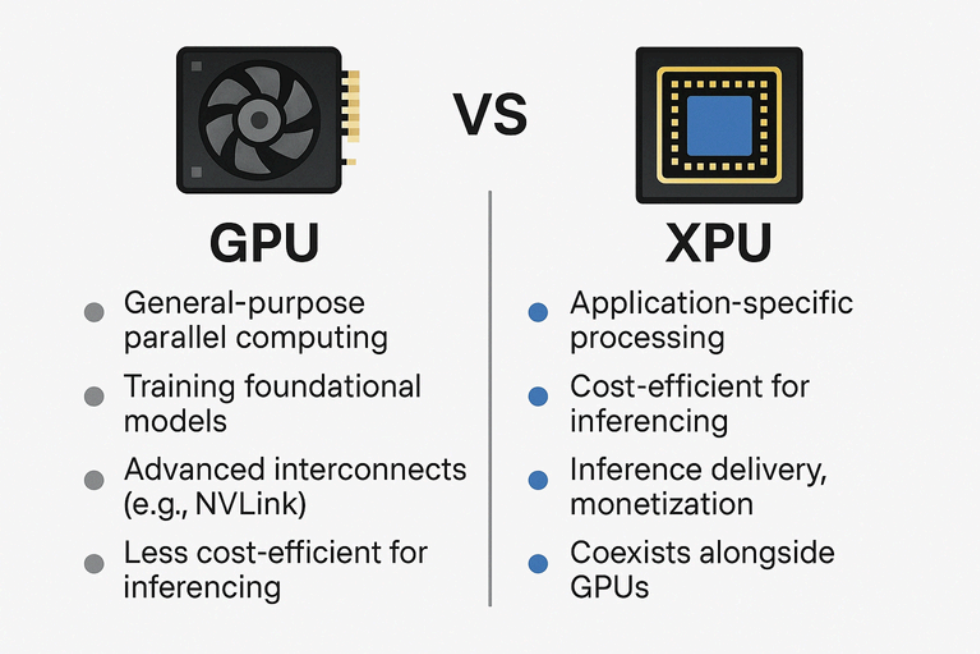
Just as the world was beginning to grasp the potential of generative AI — models that churn out text, images, and code — a more ambitious wave is quietly building: Agentic AI.

These aren't just systems that create; they think, plan, and act. An Agentic AI doesn't wait for your next prompt. It breaks down goals into subtasks, interacts with tools and APIs, monitors its own progress, and dynamically adjusts course — autonomously. It's the difference between a painter and a project manager.

And unlike static content generation, this requires relentless real-time inference, precise decision loops, and ultra-efficient hardware. The spotlight now shifts to application-specific silicon (ASICs) / XPUs — custom-built chips that can execute these tasks not just faster, but smarter. In this new AI paradigm, it's not just about intelligence — it's about execution at scale.

What Are GPUs and XPUs Anyway?

Think of GPUs as the all-rounders of computer chips. They're good at doing lots of things at once and are used heavily in building (or "training") AI models, like ChatGPT or image generators.XPUs, on the other hand, are more like specialists. They're designed to do just one thing very well — like answering AI questions, running smart assistants, or handling video recommendations. These chips go by different names — TPUs, DPUs, or just "custom AI chips" — but they all aim to do specific jobs fast and efficiently.



Who's Winning in This New AI Chip Race?

Cloud Provider	Custom Chip (XPU)	Partner Companies
AWS	Inferentia, Trainium	Marvell, Alchip (rumored)
Google	TPU, Axion CPU	Broadcom, TSMC
Microsoft	Athena, Maia	Marvell, possibly Broadcom
Meta	MTIA, custom DPUs	Marvell, Broadcom
ByteDance	AI inference chip	Broadcom
OpenAI	Unconfirmed development	Broadcom (rumored)

- Marvell: Co-developing chips with AWS, Microsoft, and Meta.
- Broadcom: Supplying AI chip tech to Google, Meta, and others.
- TSMC: Building nearly all these chips in its factories.
- Synopsys & Cadence: Providing tools to design these chips.

Why Are We Talking About XPUs Now?

Because the big tech companies (called hyperscalers) like Google, Amazon, and Microsoft are all making or using these XPUs to save money and improve performance. These chips are:

- Cheaper to run for specific tasks
- More energy-efficient
- Built for real-time AI uses, like chatbots or video recommendations

Training AI models with GPUs is expensive and slow. But once the AI is trained, XPUs take over to deliver fast, cost-effective results to millions of users.

Nvidia's Big Shift: Working With XPUs

Nvidia, the top GPU maker, recently introduced something called NVLink Fusion. It's a way for their chips to work better with XPUs from other companies. This is a big change: instead of doing everything alone, Nvidia is helping different types of chips work together.

This move shows that even Nvidia sees the future as a team sport — with GPUs handling the complex model training, and XPUs running the everyday, money-making AI services.

What's the Difference Between Training and Inference?

Training is teaching the AI how to think. It takes a long time and lots of expensive GPUs. Inference is when the AI uses what it's learned to give you answers or suggestions. It happens quickly, many times a second.

Inference is where the money is made — when you ask Siri a question, or YouTube suggests a video, or a chatbot answers you. This is where XPUs shine.

The Big Picture

AI computing is becoming more specialized. Instead of just using one type of chip (like a GPU), companies are building custom chip teams to handle specific jobs faster and cheaper.

XPUs aren't here to replace GPUs — but to work alongside them. They're making AI services more scalable, efficient, and profitable. As this shift happens, companies like Marvell, Broadcom, and TSMC are becoming critical players behind the scenes.

LIGHTHOUSE CANTON

Singapore

Lighthouse Canton Pte Ltd

16 Collyer Quay #11-02
Collyer Quay Centre
Singapore 049318

+65 67130570

Dubai

Lighthouse Canton Capital DIFC Pte Ltd

The Exchange Gate Village 11, Unit 802 Dubai International Financial Centre PO Box 507026 Dubai, UAE

+971 45 861500

India

Lighthouse Canton India Private Limited

Unit No 104A, First Floor, Worldmark 2,Asset Area No. 8
Hospitality District Aerocity, IGI Airport, New Delhi 110037
+91 9650473961

Unit 507 & 508, INS Tower,
Bandra Kurla Complex,
Bandra East,
Mumbai - 400051
+91 9650473961

1st Floor, WeWork37, Cunningham Cross Rd,
SRT Road, Vasant Nagar, Bengaluru-560001

+91 9900096873

info@lighthouse-canton.com service@lighthouse-canton.com Lighthouse Canton

DISCLAIMER

The contents of this document are confidential and are meant for the intended recipient only. If you are not the intended recipient, please delete all copies of this document and notify the sender immediately.

This document, provided as a general commentary, is for informational purposes only and is not to be construed as an offer to sell or solicit an offer to buy any financial instruments in any jurisdiction. This does not constitute any form of regulated financial advice, and your independent financial advisor should be consulted prior to taking any investment decision(s).

This document is based on information from Sources which are reliable but has not been independently verified by Lighthouse Canton Pte Ltd and its affiliate companies ("LC"). LC has taken the reasonable steps to verify the contents of this document and accept no liability for any loss arising from the use of any information contained herein. Please also note that past performances are not indicative of future performance.

Information contained herein are those of the author(s) and does not represent the views held by other parties. LC is also under no obligation to update you on any changes made to this document.

This document is prepared by Lighthouse Canton Pte Ltd and its affiliate company, Lighthouse Canton Capital (DIFC) Pte Ltd, which are regulated by Monetary Authority of Singapore (“MAS”) and Dubai Financial Services Authority (“DFSA”) respectively. MAS and DFSA has no responsibility for reviewing, verifying and approving the contents of this document and/or other associated documents. The contents of this document may not be reproduced or referenced, either in part or in full, without prior written permission from LC.

This document is confidential and is only intended for Accredited Investors and/or Professional Clients, as defined by MAS and DFSA.