

NEXTSILICON



# ***BREAKING THE CPU AND GPU BOUNDARIES***

In Simulation & Modelling



ModSim 2026

## PRESENTER

# HI! I'M ILAN

VP architecture  
& Co-Founder

## A bit about me

I co-founded **NextSilicon** in **2018** and am its **VP Architecture**.

My 30 years' experience include senior staff engineer (software) at **Mellanox Technologies**, distinguished architect at **Integrity Project**, and senior engineer at Nocturnus Systems and Cogniview Systems.

I have authored many patents in computer architecture, and hold a B.Sc. degree in **physics and computer science** from the Hebrew University of Jerusalem.



The background of the image is a solid blue color. Scattered around the central text are several Next Silicon components. At the top, there are three square modules in white frames, each with a black square in the center. One of these frames has the text 'ARCHITECTURE BY NEXTSILICON' and a part number 'NXT232R96GB'. To the right, there is a long, thin module with a gold connector. On the right side, there is a module with a black square and the text 'MAVERICK2' and 'NXT20400'. At the bottom, there is a green circuit board with various components and a black square, labeled 'MAVERICK2'. To the right of it is another square module in a white frame, labeled 'MAVERICK2' and 'NXT20300KV232R96GB'.

NEXTSILICON

# ***THE NEXTGEN COMPUTE PLATFORM***

**MAVERICK·2**

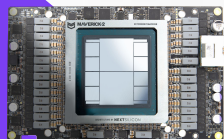
**MAVERICK·2**

# **DATAFLOW ACCELERATOR**



# ACCELERATING VANGUARD

## MAVERICK-2



**128x Maverick-2** installed on  
Sandia National Labs'  
Vanguard system





**EVERYONE  
MAKES ONE**

Usual Suspects

**nVIDIA®** **intel®**

**AMD** **Qualcomm**



# **EVERYONE MAKES ONE**

## Usual Suspects

**nVIDIA** **intel**

**AMD** **Qualcomm**

## Hyperscalers

**G** **amazon**

**∞** **Meta** **Microsoft**



# EVERYONE MAKES ONE

## Usual Suspects

**nVIDIA** **intel**

**AMD** **Qualcomm**

## Hyperscalers

**G** **amazon**

**∞** **Meta** **Microsoft**

## AI labs

 **OpenAI**

**ANTHROPIC**



# WHY?

## Usual Suspects

**nVIDIA** **intel**

**AMD** **Qualcomm**

## Hyperscalers

**G** **amazon**

**∞** **Meta** **Microsoft**

## AI labs

 **OpenAI**

**ANTHROPIC**

## Global Tech Giants



 **Alibaba**



**SQUEEZE  
A LITTLE  
MORE**



**OPTIMIZE**



**SPECIALIZE**

### Usual Suspects

**nVIDIA** **intel**

**AMD** **Qualcomm**

### Hyperscalers

**G** **amazon**

**∞ Meta** **Microsoft**

### AI labs

 **OpenAI**

**ANTHROPIC**

### Global Tech Giants



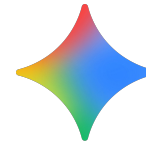
 **Alibaba**



# ***SPECIALIZE***

EXTREME EXAMPLES

## **Google Frozen V2**



**Gemini-specific ASIC**  
bakes the model into the silicon

## **Taalas** (Acq by AMD)



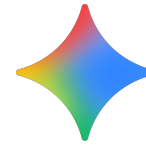
**ROM weights**  
on last mask



## ***SPECIALIZE***

EXTREME EXAMPLES

### **Google Frozen V2**



**Gemini-specific ASIC**  
bakes the model into the silicon

### **Taalas** (Acq by AMD)



**ROM weights**  
on last mask

---

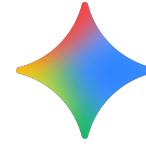
Market and industry, not research and innovation



## **SPECIALIZE**

EXTREME EXAMPLES

### **Google Frozen V2**



**Gemini-specific ASIC**  
bakes the model into the silicon

### **Taalas** (Acq by AMD)



**ROM weights**  
on last mask

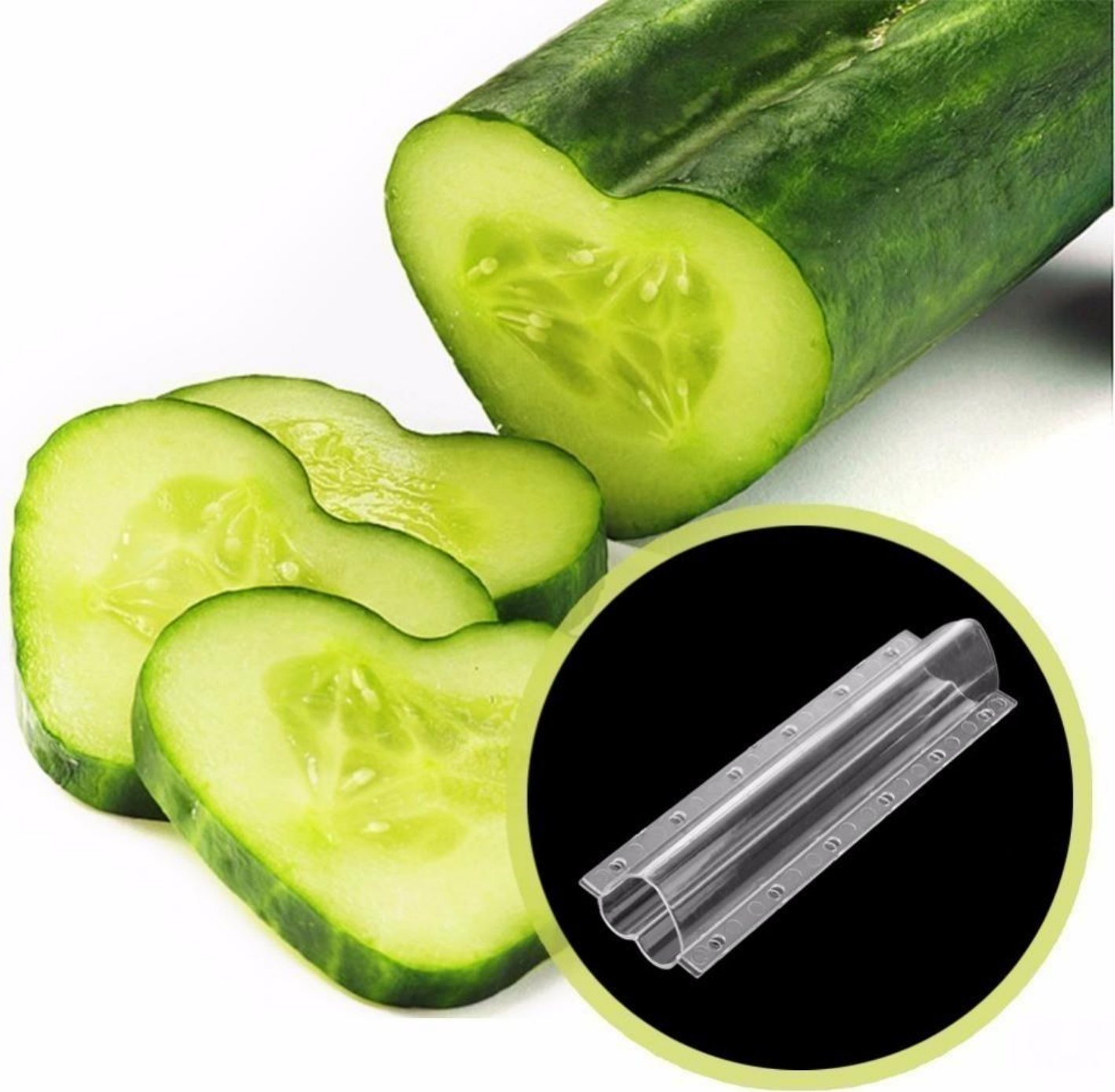
---

Market and industry, not research and innovation

**Leads to stagnation, not to progress**



**BREAK THE  
MOLD**





# NOVEL MODELS

The **most common** use cases:

- Casual Generative AI
- Chatbots
- Agentic uses

Depend on **fast-changing novel models**

**Public ones**, tip of the iceberg:

- **DeepSeek V4** (April 2026)
- **Kimi K3** (July 2026)
- **Minimax H3** (July 2026)
- **Qwen 3.8** (August 2026) - just last week!



# NOVEL MODELS

The **most common** use cases:

- Casual Generative AI
- Chatbots
- Agentic uses

Depend on **fast-changing novel models**

**Public ones**, tip of the iceberg:

- **DeepSeek V4** (April 2026)
- **Kimi K3** (July 2026)
- **Minimax H3** (July 2026)
- **Qwen 3.8** (August 2026) - just last week!

**\$100M ASIC Makes sense financially?**

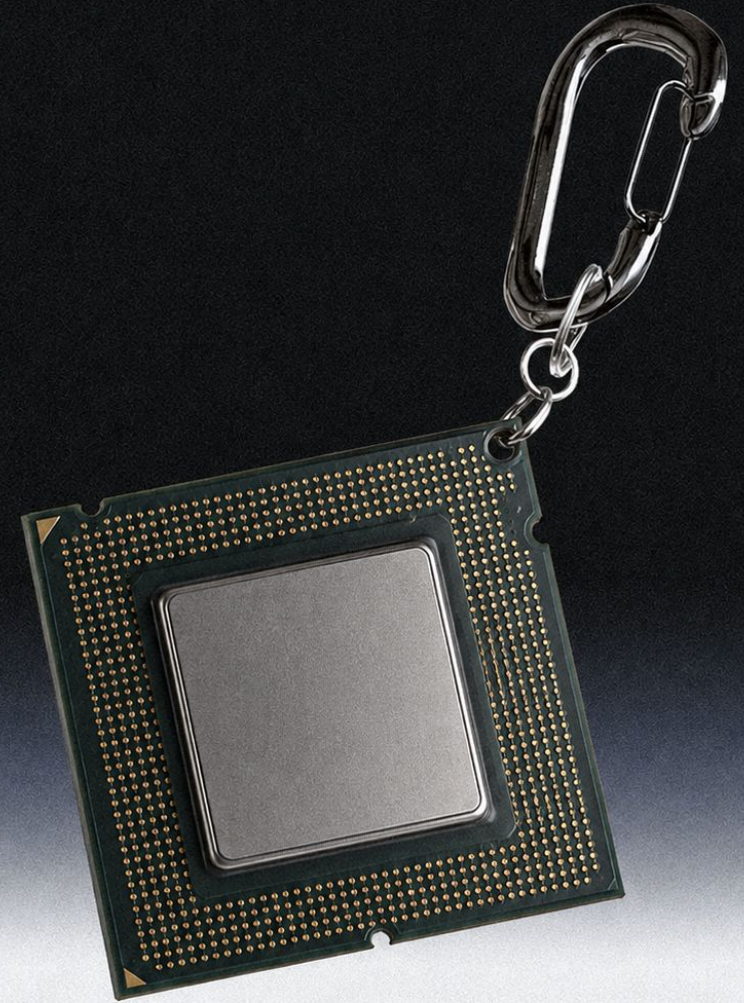


NOVEL MODELS

# MAKE SENSE FINANCIALLY?

Today's algorithm is  
tomorrow's keychain

Re-do **every time** the algorithm  
changes





# GENERAL PURPOSE HARDWARE

## Investors circa 2018:

> *Why is HPC and general-purposeness interesting at all? Everything is just matrix multiplications*





# GENERAL PURPOSE HARDWARE

## Investors circa 2018:

> *Why is HPC and general-purposeness interesting at all? Everything is just matrix multiplications*

Recent AI advancements are **far from it**

Activations, Sparse MoE, Hybrid attention, Gated DeltaNet, 'Adaptive Thinking', 'Dreaming V3'

The trend: Future of AI depends on **algorithm variety**





# AGENTIC MODSIM

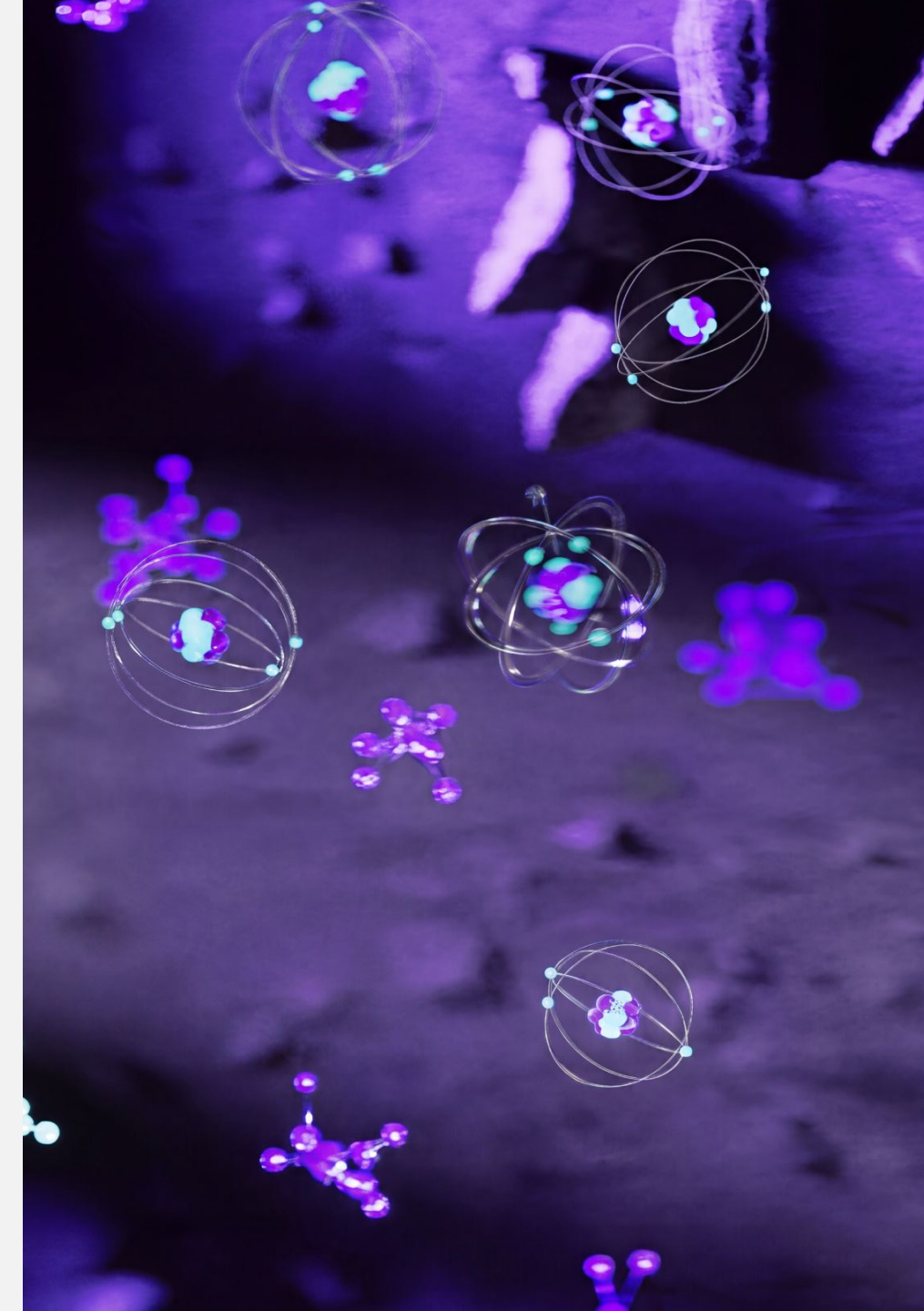
## Agentic AI in ModSim

Chemistry question -> script to run GROMACS

Accelerators don't run GROMACS... or even the script.

Leading to complex systems

A system that can do both!





# **FUTURE:** **AGI**

AGI Holy Grail

Changing rapidly.

**What would it look like 2 years from now?**

The future-proof approach allows you to be better prepared for what's next in AGI.



## TAKEAWAY

- > **Mixed HPC and AI workloads are here:** surrogate models, physics-informed ML, Agentic AI using HPC modelling
- > Systems of today have much **to improve**
- > **Big opportunities** in general purpose hardware for AI
- > Very few contestants - it's **super hard.**



NEXTSILICON

# MAVERICK·2

INTELLIGENT COMPUTE ARCHITECTURE®





# NEXTSILICON

© 2026 Next Silicon Ltd., all rights reserved. Next Silicon retains the copyright in all of the material in this document as a collective work under copyright laws. You may not copy, republish, redistribute, or exploit in any manner any material from these pages without the express written consent of Next Silicon. All trademarks, trade names, or logos included in this document are owned by Next Silicon or its business affiliates or third parties, as the case may be. All rights in such names, marks, or logos are reserved by Next Silicon and/or respective holders or Next Silicon's business affiliates or third parties, as the case may be.