

# 开源大模型创新背后的 RISC-V 算力架构革新

演讲人：苏中

知合计算 / 首席AI科学家

**AiCon**

全球人工智能开发与应用大会

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询

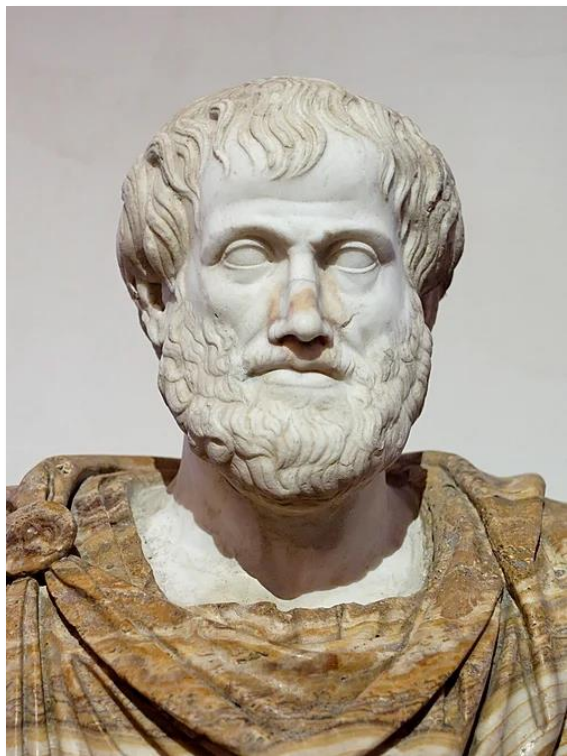


查看会议



# 人工智能发展：从符号主义到链接主义

符号主义的鼻祖：亚里士多德与他的三段论

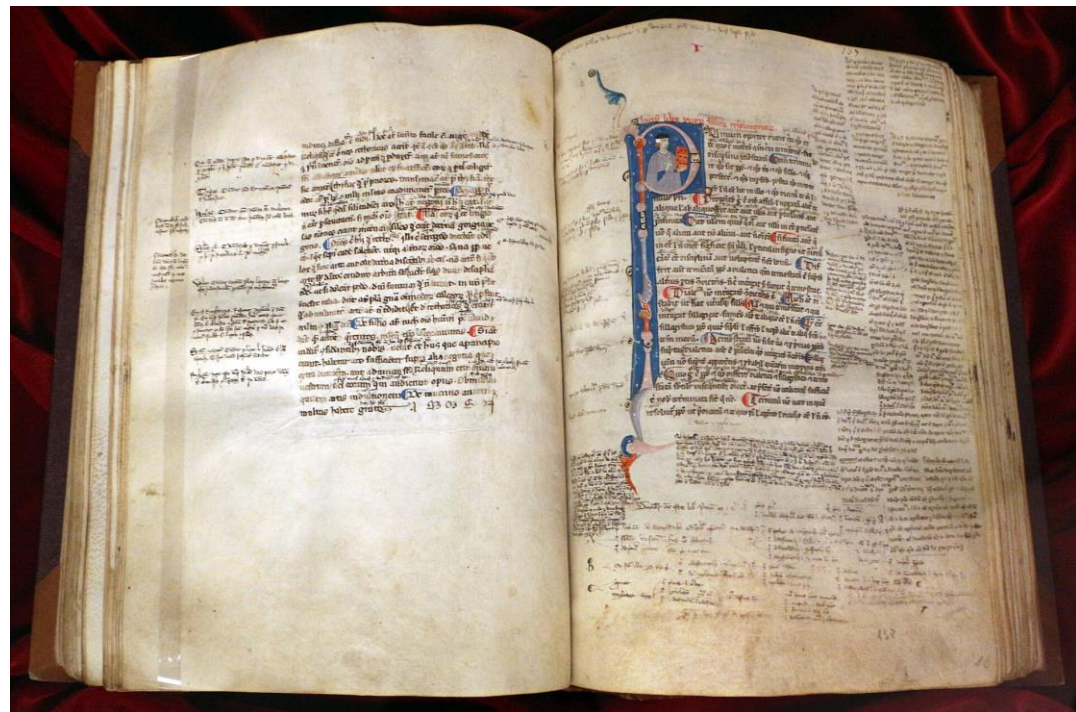


All men are mortal.  
Socrates is a man.  
Therefore, Socrates is  
mortal

所有人都会死。  
苏格拉底是人。  
因此，苏格拉底会死。

亚里士多德 Ἀριστοτέλης Aristotél  
(古希腊 公元前384–公元前322年)

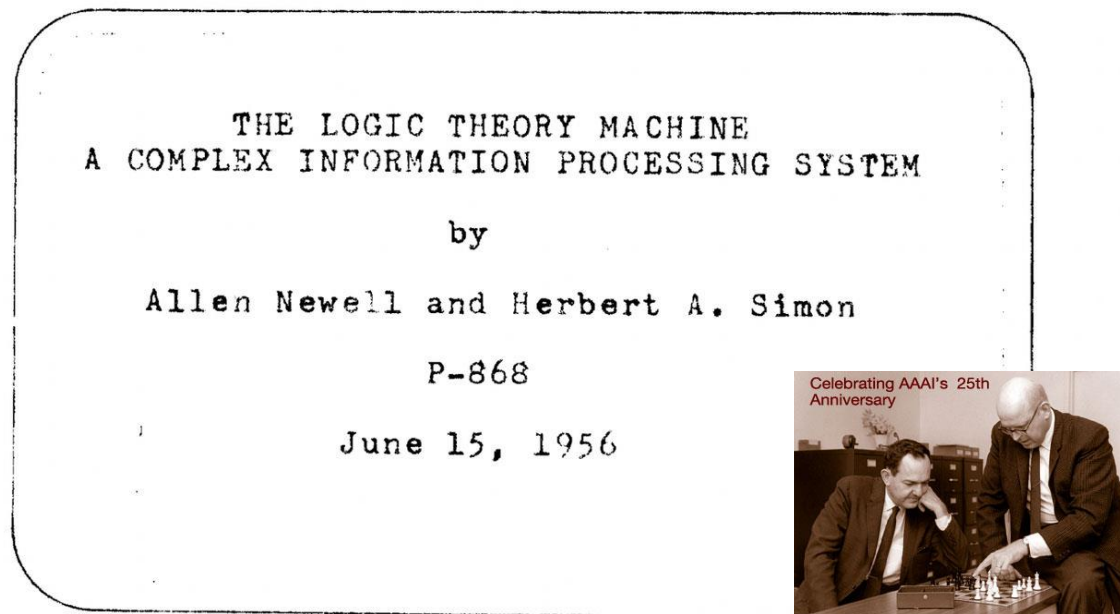
<https://en.wikipedia.org/wiki/Aristotle>



亚里士多德《Prior Analytics》拉丁文版  
约1290年，佛罗伦萨劳伦齐亚纳医学院图书馆

# 符号主义的发展

符号主义的发展：从第一个AI程序Logic Theorist，专家系统的成功到日本第五代计算机计划

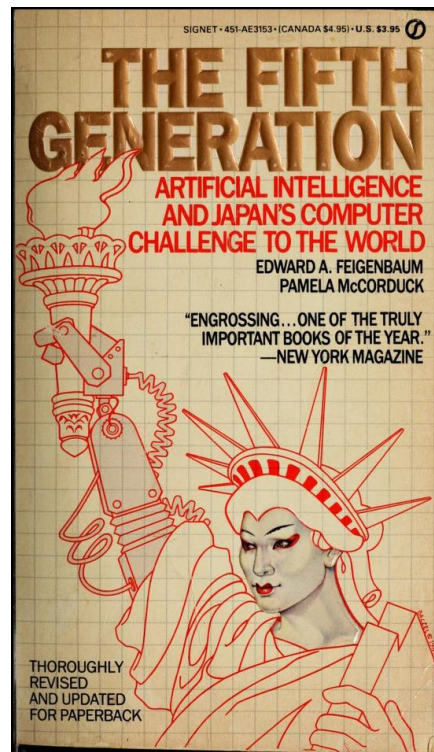


**Logic Theorist:** 第一个人工智能程序(第一个特别设计用于模仿人类解决问题能力的程序)由赫伯特·西蒙(Herbert Simon),艾伦·纽厄尔(Allen Newell)和约翰·肖(John Shaw)于1955年至1956年创建,在数学定理证明方面取得突破。

[https://en.wikipedia.org/wiki/History\\_of\\_artificial\\_intelligence](https://en.wikipedia.org/wiki/History_of_artificial_intelligence)

<https://www.aikatana.com/p/legacy-japans-fifth-generation-computer-systems-fgcs-project-ai>

<https://archive.org/details/fifthgeneration00edwa/mode/2up?view=theater>

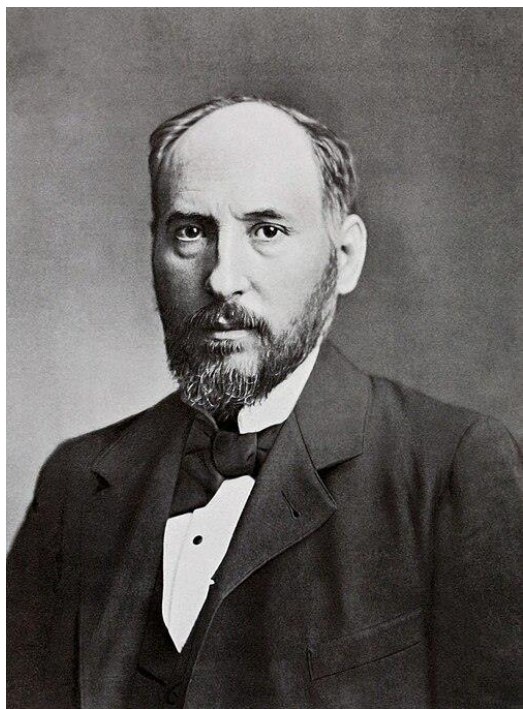


第五代电脑是日本通商产业省于1982年的一个大型研发计划,其目的为开发一部划时代的电脑,利用大量平行计算,使它拥有超级电脑的运算效能和可用的人工智能能力。

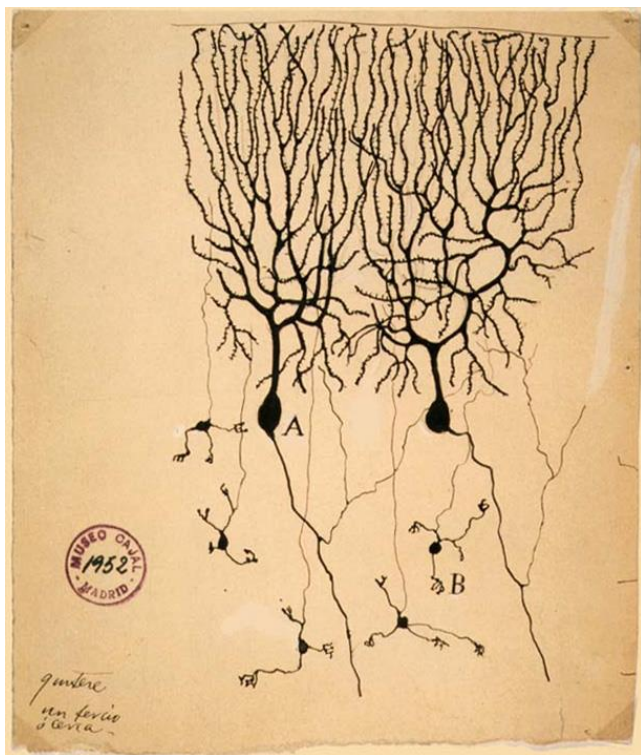


# ■ 链接主义的起步，来自于人类对于脑的认识

智能的来源：结构与规模

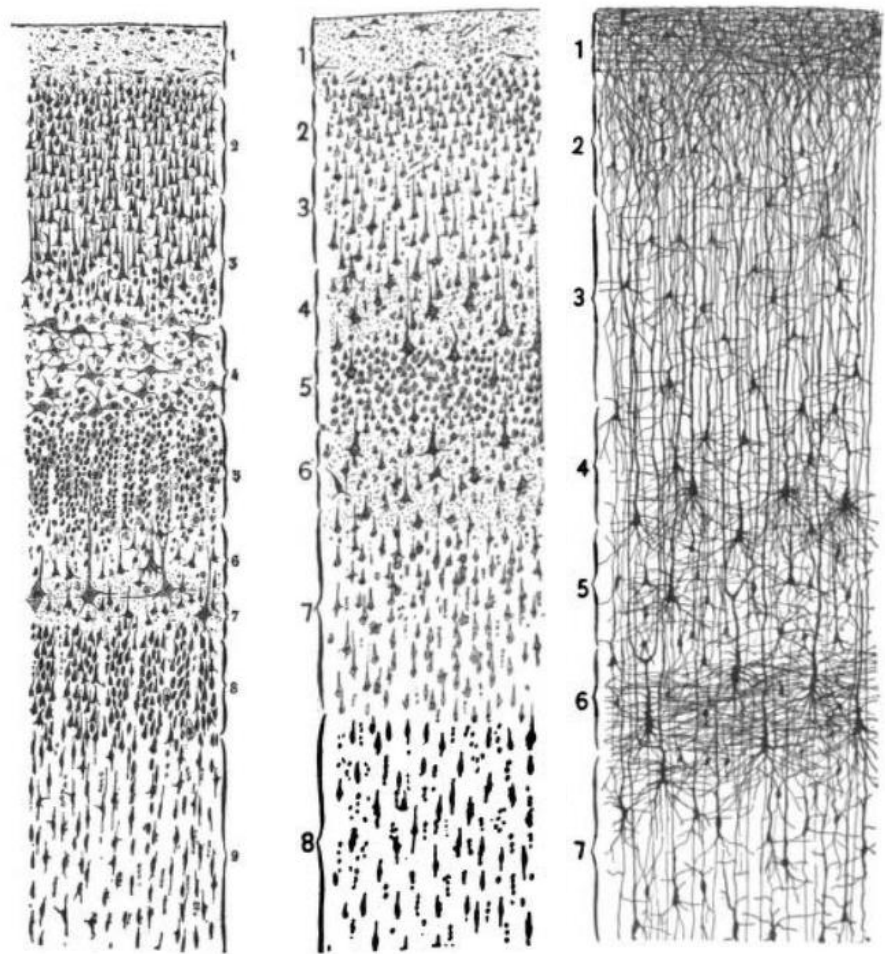


桑地亚哥·拉蒙·伊·卡哈尔  
(1852-1934、西班牙)  
Ramón y Cajal in 1899  
1906年诺贝尔生理学/医学奖



1899年绘制的鸽子小脑普金耶细胞 (A) 和颗粒细胞 (B) 图

[https://en.wikipedia.org/wiki/Santiago\\_Ram%C3%B3n\\_y\\_Cajal](https://en.wikipedia.org/wiki/Santiago_Ram%C3%B3n_y_Cajal)



《comparative study of the sensory areas of the human cortex》, 1900



# 人工智能的评测标准：图灵实验



艾伦 图灵 Alan Mathison Turing  
(英国 1912–1954)

VOL. LIX. No. 236.] [October, 1950

## MIND A QUARTERLY REVIEW OF PSYCHOLOGY AND PHILOSOPHY

### I.—COMPUTING MACHINERY AND INTELLIGENCE

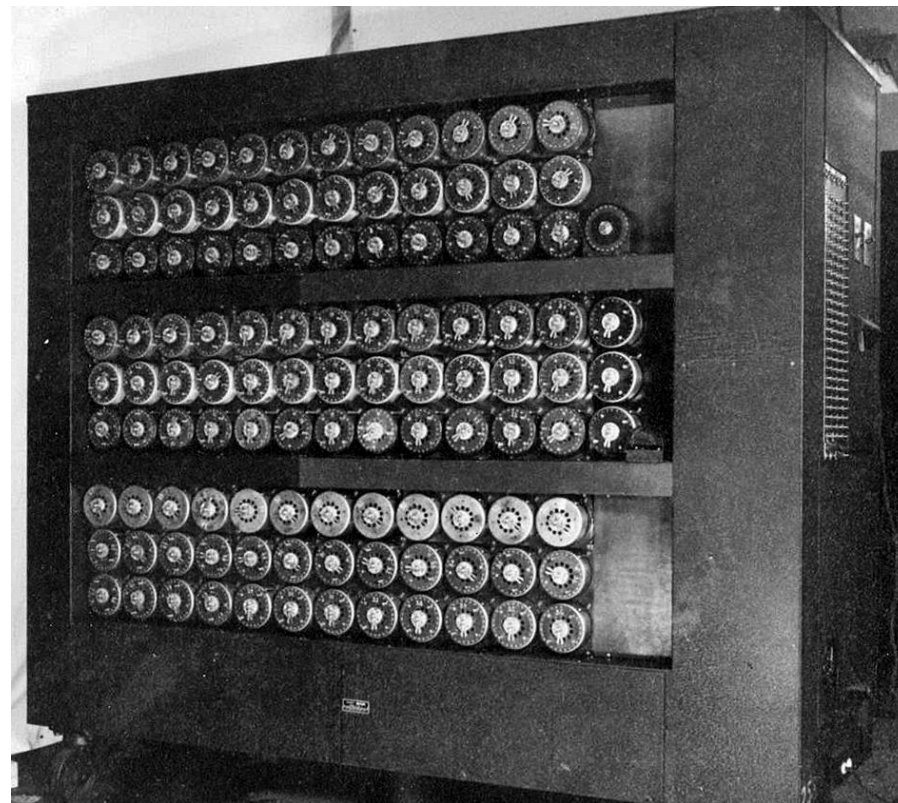
By A. M. TURING

#### 1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?  
Now suppose X is actually A, then A must answer. It is A's  
28 433



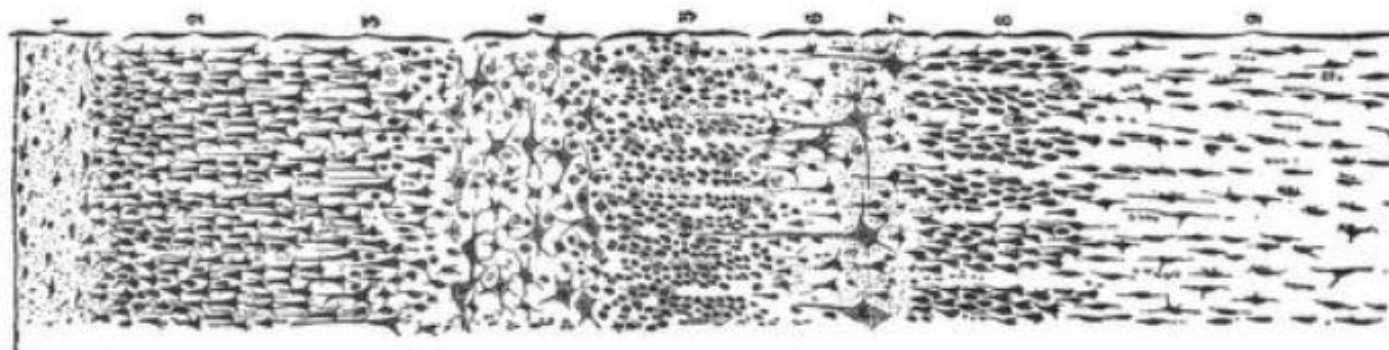
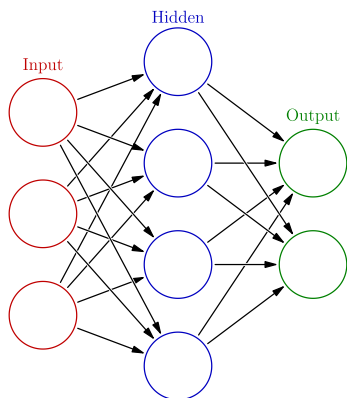
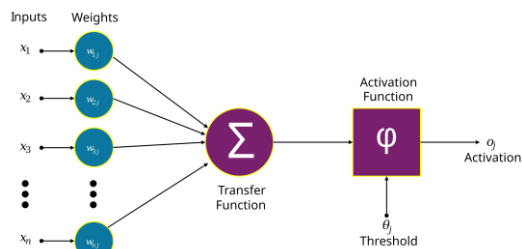
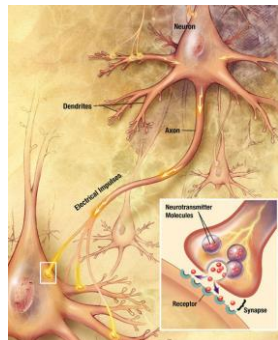
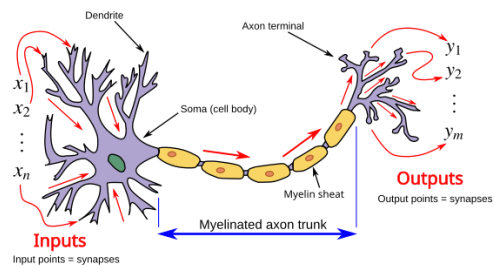
图灵发明的“炸弹”，它是一种机电计算装置，二战期间成功破译了德国恩尼格玛机加密的信息

[https://en.wikipedia.org/wiki/Alan\\_Turing](https://en.wikipedia.org/wiki/Alan_Turing)

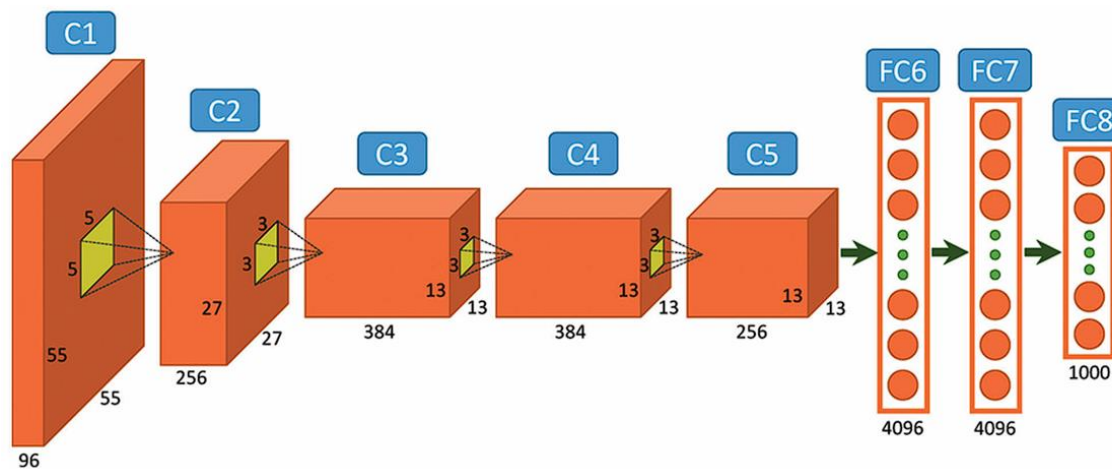
[https://en.wikipedia.org/wiki/Bombe#/media/File:Wartime\\_picture\\_of\\_a\\_Bletchley\\_Park\\_Bombe.jpg](https://en.wikipedia.org/wiki/Bombe#/media/File:Wartime_picture_of_a_Bletchley_Park_Bombe.jpg)

# 人工神经网络：从感知机到深度学习

人工智能的发展：规模与结构



人类视觉皮层的纵向切面（横放）桑地亚哥·拉蒙·伊·卡哈尔，1900



深度学习技术的引爆点：图像识别的AlexNet 架构。  
它由 8 层组成：5 个卷积层和 3 个全连接层



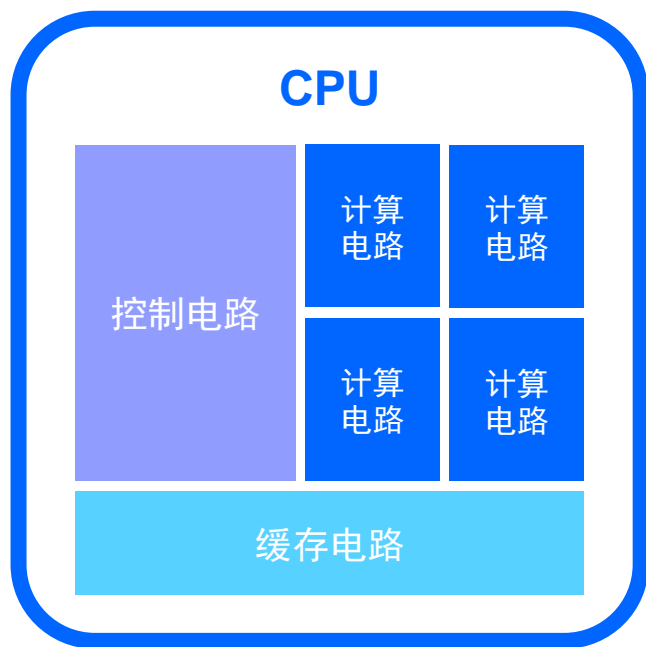
Geoffrey Hinton



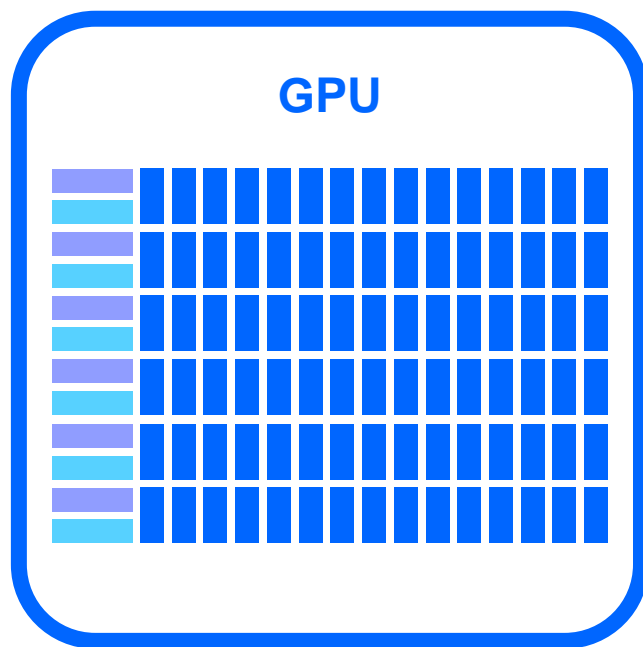
Yann LeCun

<https://viso.ai/deep-learning/alexnet/>  
<https://awards.acm.org/about/2018-turing>

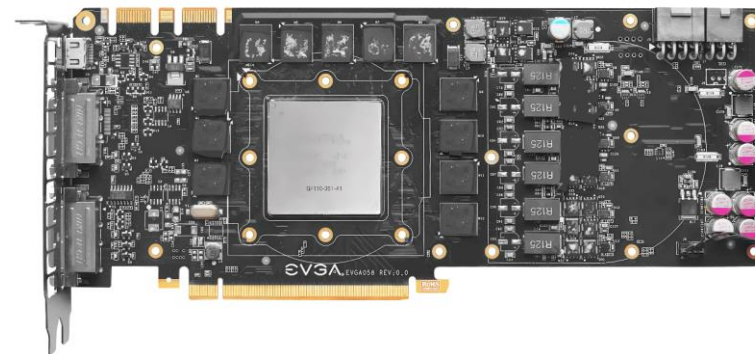
# AlexNet: AI算力架构从CPU转向GPU的起点



- 计算密度相对低
- 复杂的控制电路，应对复杂的运算
- 庞大的缓存电路，缓存数据



- 计算密度高
- 能做的运算的计算复杂度低
- 内存访问的带宽高



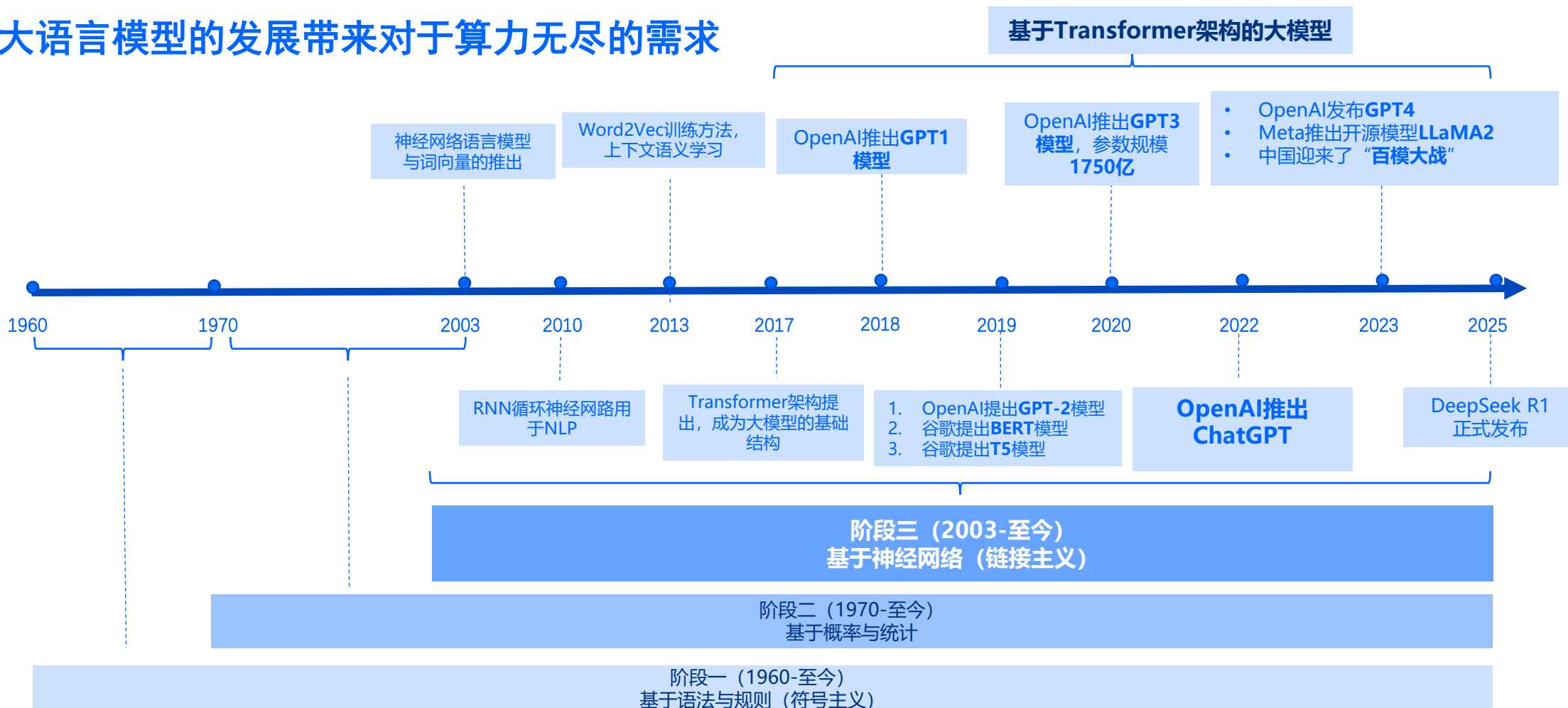
AlexNet=  
6000万个参数  
+65万个神经元  
+卷积神经网络的高性能C++/CUDA实现  
+2\*GTX580 GPU

该模型由Alex Krizhevsky与多伦多大学的Ilya Sutskever和博士顾问Geoffrey Hinton于2012年合作开发@University of Toronto



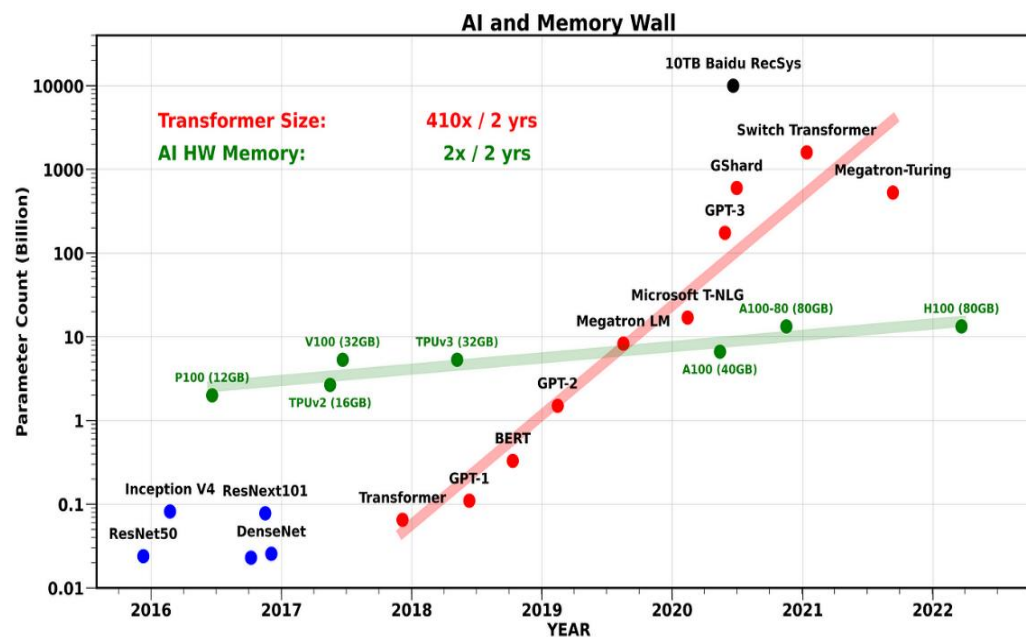
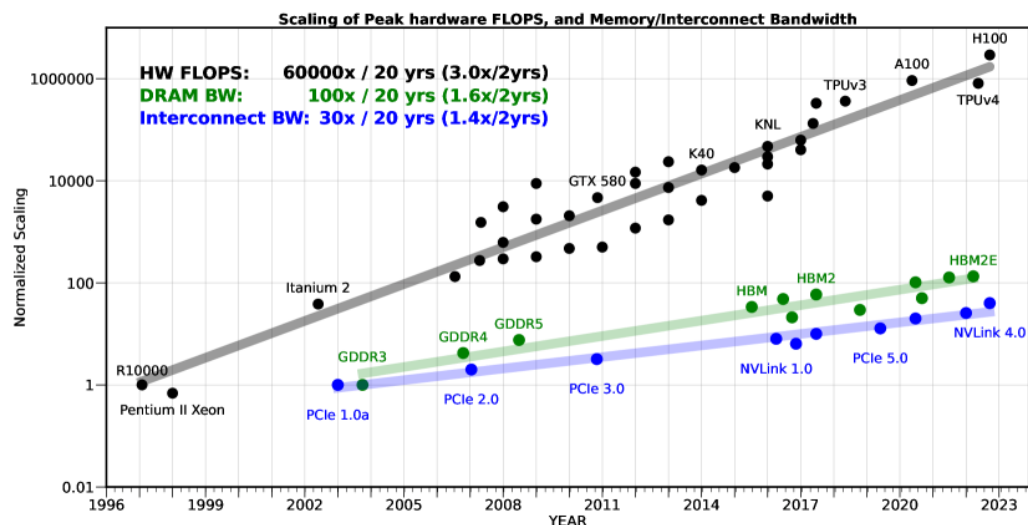
# 人工智能的发展：从感知到认知

大语言模型的发展带来对于算力无尽的需求



# ■ 当前计算架构无法满足未来的需求

从大语言模型进一步发展到多模态，AI的“读万卷书”和“行万里路”

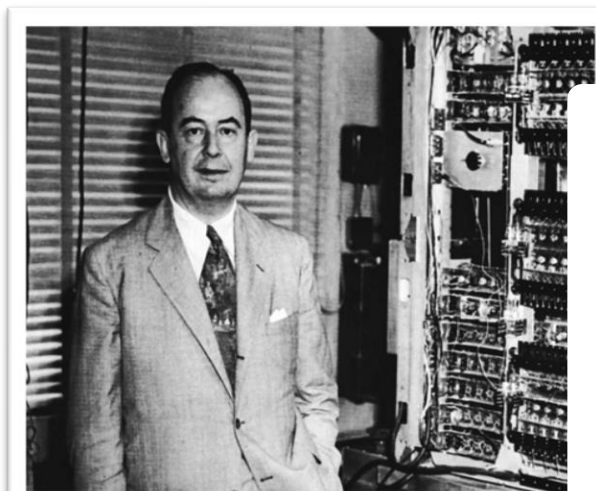


- 人工智能的发展，完全颠覆了算力实际需求增长速度，计算架构重构成为唯一出路

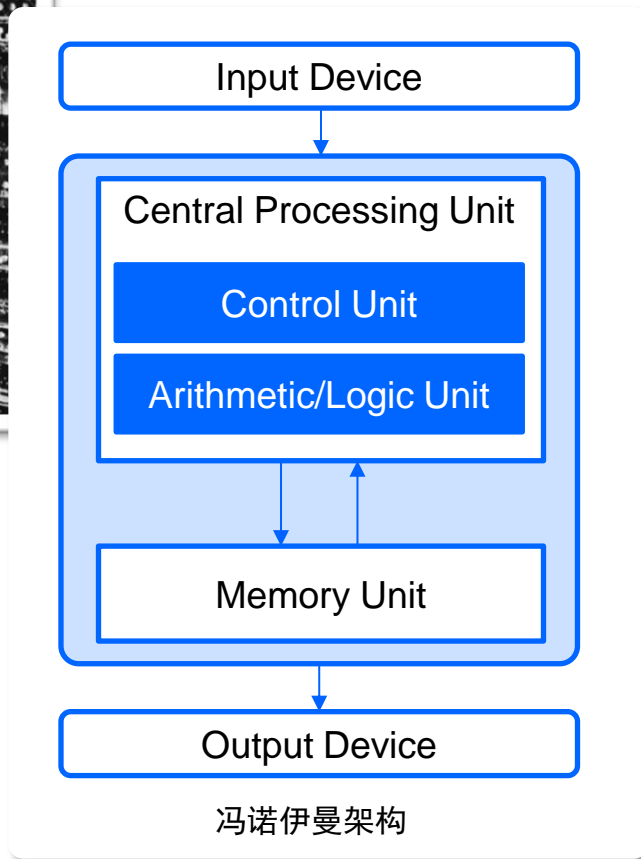


# ■ 未来十年是计算架构创新的黄金10年

算力瓶颈、存力瓶颈、互联瓶颈亟待突破



现代计算机之父 约翰·冯·诺伊曼



技术趋势：面向领域的架构（DSA）将支撑计算机体系结构黄金十年

- ❑ RISC架构先驱、图灵奖获得者John Hennessy和David Patterson教授预测，随着摩尔定律的放缓，计算机体系结构方面形成突破变得尤其重要，未来十年是DSA架构的黄金十年；
- ❑ 新突破方向有：面向领域软硬件协同设计，增强的安全技术，**开源的指令架构**和敏捷开发
- ❑ **正式提出了RISC-V是新架构创新的底座**

# RISC-V：最适合AI时代高性能CPU打造的架构

开放技术底座加上可扩展的能力是实现AI时代最佳算力

PC时代  
(1980年代 – 2000年代)

- 通用计算能力要强
- 以X86架构成为主流
- X86架构完全闭源，架构不可定制

移动时代  
(2000年代 – 2020年代)

- 功耗要低
- ARM架构成为主流
- ARM架构采用授权模式，但由单独公司掌握，定制化程度低

- 经过多年发展，X86架构与ARM架构在PC、移动场景各自拥有丰富的软件生态
- 闭源模式的中心化驱动，对AI的支持需要代际间的大版本更新
- CPU IP根技术依存于少数企业，缺乏自主权
- 传统架构对于异构可扩展性的支持仍然有待加强

AI时代  
(2020年代-)

- 新架构的演进能力、定制化能力要强
- 以RISC-V为代表的开源精简指令集成为CPU架构创新主要力量
- 开源、开放的特性，为基于RISC-V架构的产品带来以下优势：

## 硬件架构

- ❑ 可扩展的模块化指令集体系
- ❑ 针对AI不断完善的RISC-V Vector / Matrix / Tensor扩展
- ❑ 开源、自主可控的指令集

## 软件生态

- ❑ 开源软件体系
- ❑ 日趋成熟的基础生态
- ❑ 快速覆盖的应用生态

## 应用落地

- ❑ 灵活度更高的商业模式
- ❑ 易于根据实际需求平衡计算、内存和互联能力
- ❑ Scale Up & Scale Out潜力巨大



# 全球都在推动RISC-V架构在AI方面的演进

打开开源硬件之门，以无限潜力获得更多成功可能

## RISC-V更适配AI高性能计算场景



成本更低

无需购买CPU IP授权  
每代新产品无需重新购买



历史包袱小

X86架构：3600+条指令  
ARM架构：1000+条指令  
RISC-V架构：47条基础指令



功耗小

硬件逻辑设计相对简洁  
能效比高且功耗小




灵活性好

支持可拓展指令  
支持模块化指令子集

## RISC-V架构创新尝试

 **tenstorrent**  
AI+CPU

 **esperanto.ai**  
AI

 **VENTANA**  
CPU

 **RAIN**  
AI (PIM)

 **Rivos**  
CPU+AI

 **Meta**

 **SiFive**  
CPU IP

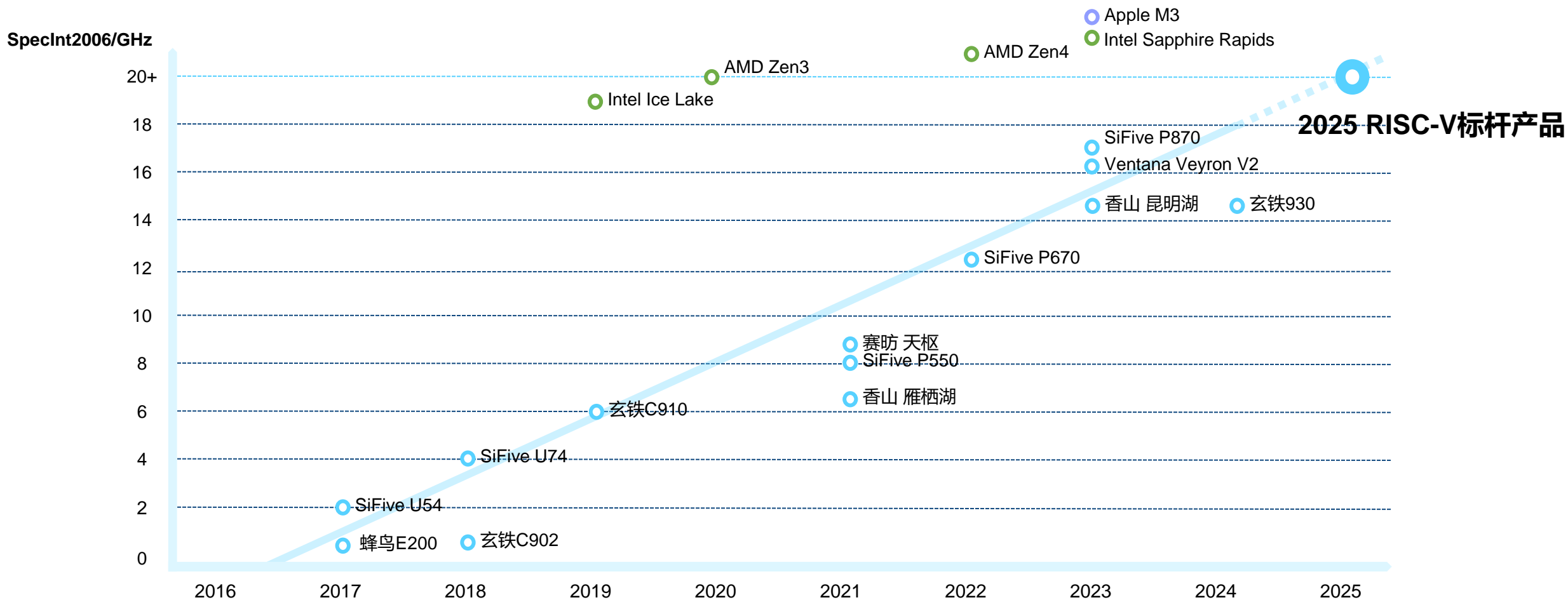
AI

## RISC-V创造了开源硬件的无限潜力

- 从通用计算到AI计算、图形计算、隐私计算、科学计算等各种计算场景，RISC-V以其开源特性带来的创新颗粒度，拥有更多的成功可能

# RISC-V处理器性能不断提升

SPECint2006评分每两年提升2/GHz



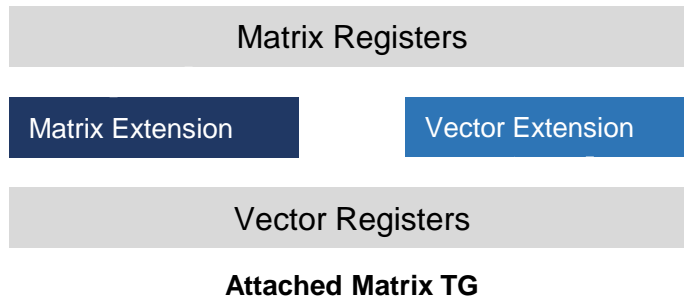


# ■ 立足中国走向世界：AI国际标准建设竞争激烈

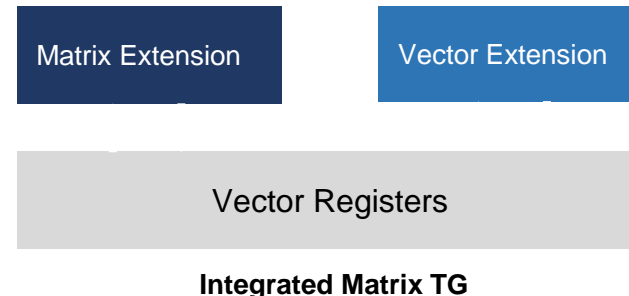
Matrix的本质优势

Scalar: 1 element --- 1 operation  
Vector: N elements --- N operation  
Matrix:  $N^2$  elements ---  $N^3$  operation

架构1：与vector寄存器独立  
( Intel、Apple、DAMO Academy、  
Streamcomputing )



架构2：复用vector寄存器资源  
( SiFive )

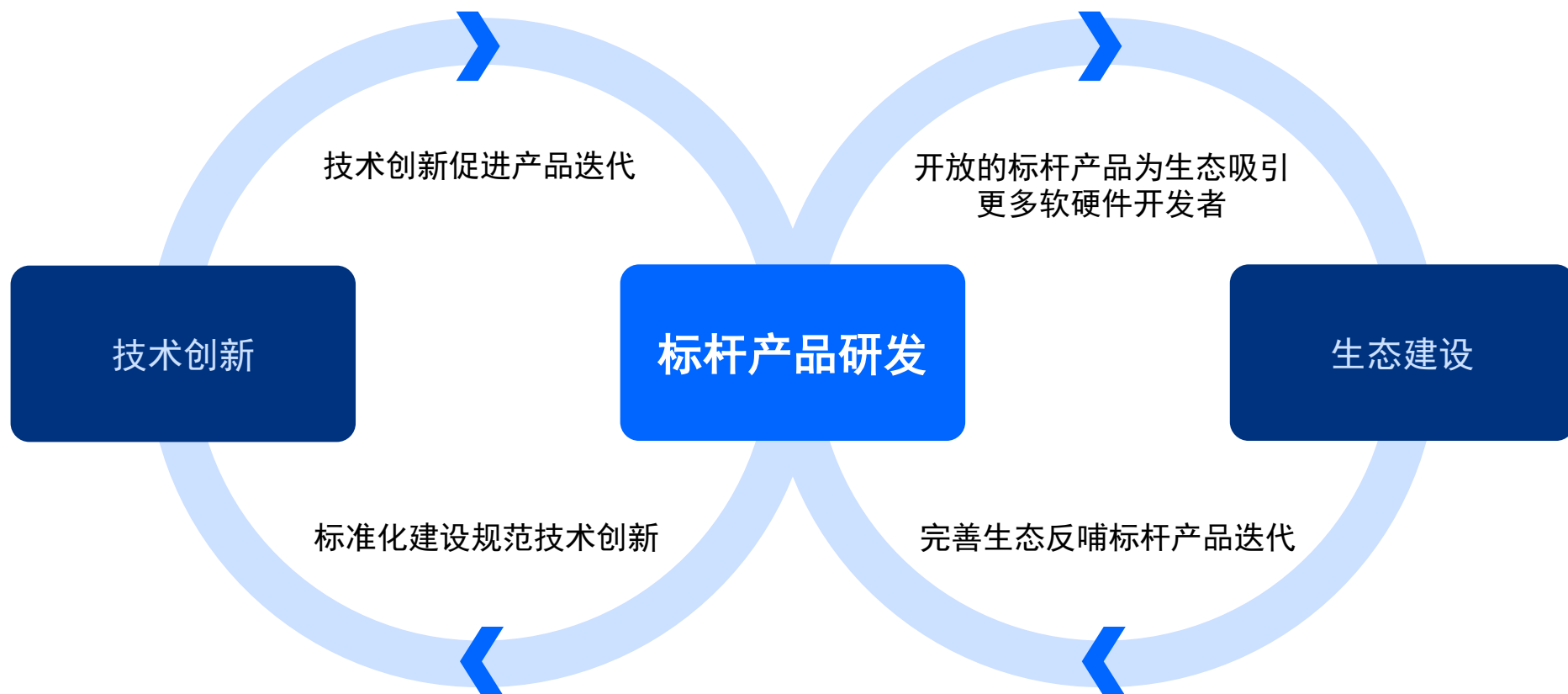


两种Matrix架构特点对比

	Attached Matrix	Integrated Matrix
算力	与Vector解耦，Matrix算力可灵活配置，可方便实现大算力	受限于Vector算力，Matrix/Vector算力配比相对固定
拓扑	支持灵活拓扑结构，包括多核共享Matrix算力	仅支持单核独享
硬件实现	与Vector松耦合，频率支持异步，时序功耗更加友好	与vector紧耦合，与CPU同频，大算力下频率功耗物理实现不易收敛
前端带宽占用	指令力度大，对前端取指压力小	指令力度较小，前端取指带宽占用大
资源	需要额外Matrix寄存器资源，执行单元可以设计更加高效	无需专用Matrix寄存器资源，同算力下执行单元较attached架构更大
软件编程	需增加额外Matrix相关context维护	复用vector资源

# RISC-V从技术走向产品再到生态

全产业链协作，聚焦标杆产品研发，  
并形成“技术创新-产品研发”与“产品研发-生态建设”的正面双循环





📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

# THANKS

∏ 知合计算

让计算更高效

MAKE COMPUTING MORE EFFICIENT

AiCon

全球人工智能开发与应用大会