

大模型计算的挑战和未来展望

朱晓明

2024年9月

3. 计算的发展，概念层出不穷

数值计算

- 使用数字计算机求数学问题的近似解
- 数值计算的结果是离散的，并一定存在误差

模拟计算

- 以连续变化的模拟量(电流、电压)作为运算对象
- 以并行计算为基础，计算速度快，程序简单，但灵活性较差

模糊计算

- 依据模糊规则，从控制变量的输入得到最终输出
- 包含模糊规则库、模糊化、推理方法和去模糊化

神经网络计算

- 通过构造人工神经网络，以模拟人类神经系统的结构和功能
- 并行分布处理、非线性映射

生物计算

- 生物学原理计算，含蛋白质计算、DNA/RNA计算等
- 存储容量大、计算并行度高、计算功耗低

类脑计算

- 借鉴人脑的各种机制
- 结构层次模仿脑、器件层次逼近脑、功能层次超越脑

群智计算

- 广泛分布性、移动性、连接性的大规模感知
- 对低质冗余、碎片化感知数据的优选和增强理解

人机共生计算

- 人机共生，实时交互
- 人机混合增强智能，突破单独人类或者机器智能局限

超级计算机

- 计算机中功能最强、运算速度最快、存储容量最大的一类计算机
- 国家科技发展水平和综合国力的重要标志

云计算

- 利用互联网实现随时随地、按需、便捷地使用共享计算设施、存储设备、应用程序等资源的计算模式
- 具有超大规模、高可靠、弹性可扩展、按需服务等特点

雾计算

- 由大量异构的分布式设备，通过有线/无线的方式协同完成任务的计算与存储
- 更接地气的云计算，数据延时低、地理分布广、移动性好

边缘计算

- 利用靠近数据源的边缘地带来完成数据的处理、分析和存储
- 可实现快捷且近乎实时的分析和响应

5. 大模型体现的AI能力，正在快速推动产业变革

用户层

应用层

能力层

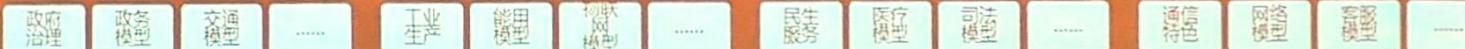
技术层

基础层

行业智能化应用

↑支撑

L1 行业大模型



↑衍生

L0 基础大模型

语言 / 大模型

视觉 / 大模型

语音 / 大模型

结构化数据 / 大模型

多模态 / 大模型

AI模型生产工具

深度学习框架 / 开源模型

预训练大模型

模型训练 / AI开发平台

AI算力基础

AI芯片

云计算与云服务

智能计算平台

智能服务器

数据资源

数据整合

第三方合规数据

外部合规数据

数据标注结构化

硬件设施

算力资源

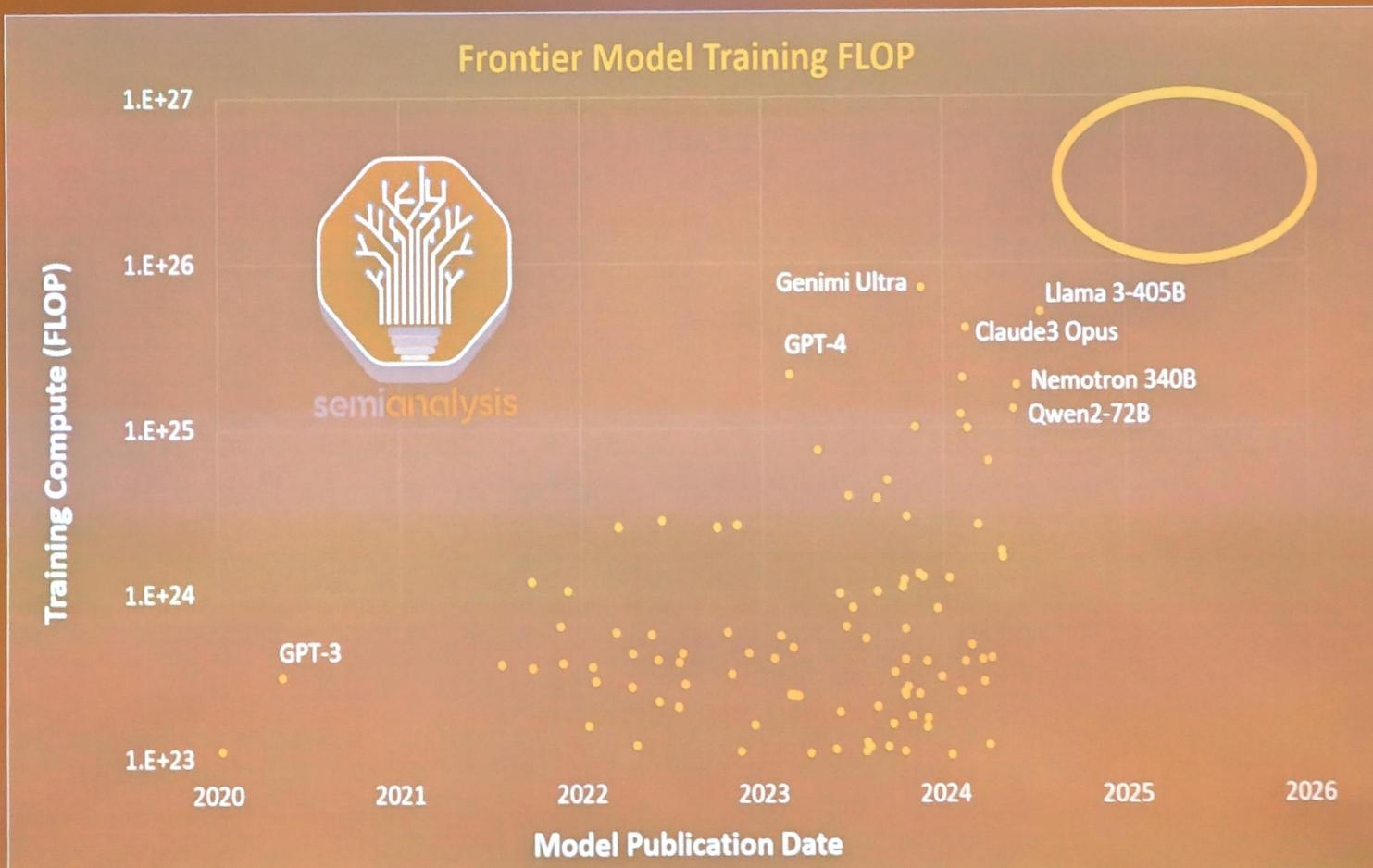
储存资源

网络资源

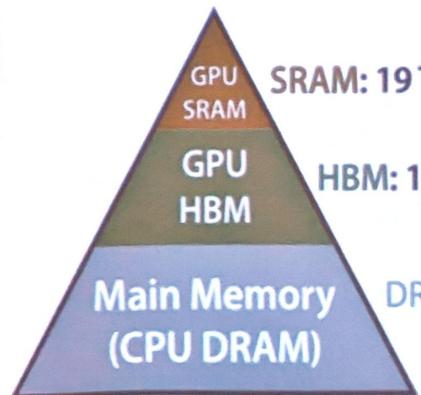
安全资源

直观展现了产业数字化价值，吹响了信息化到智能化的冲锋号

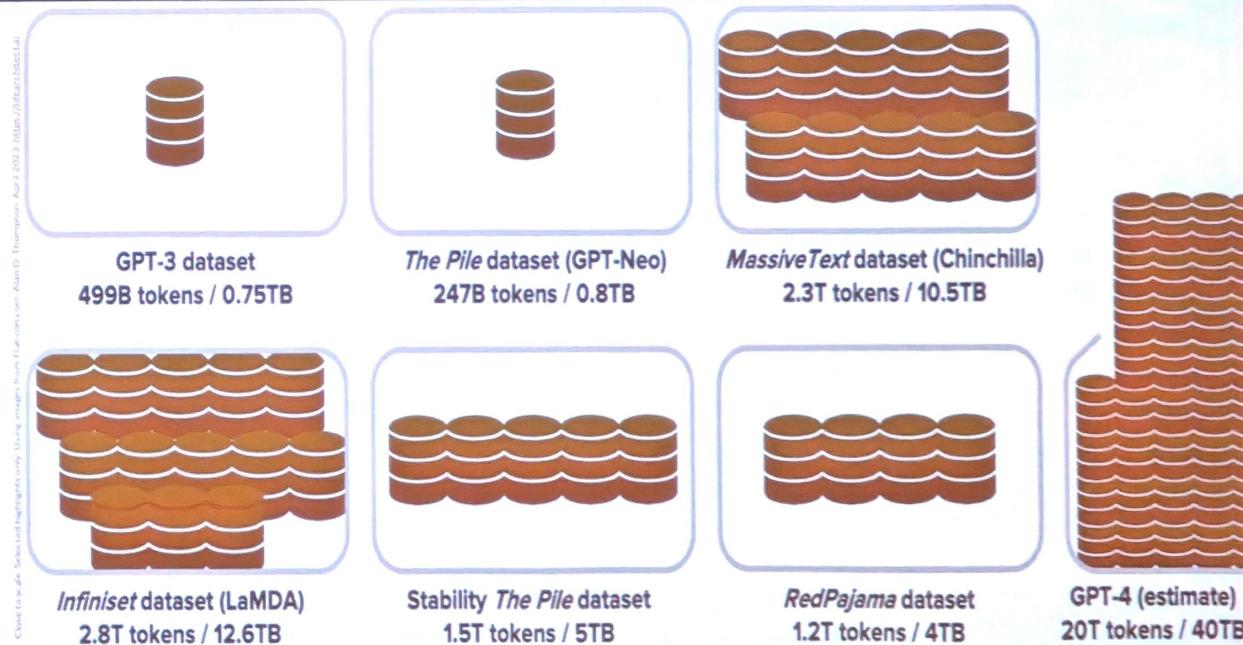
1. 大模型计算的算力挑战



1. 大模型计算的算力挑战



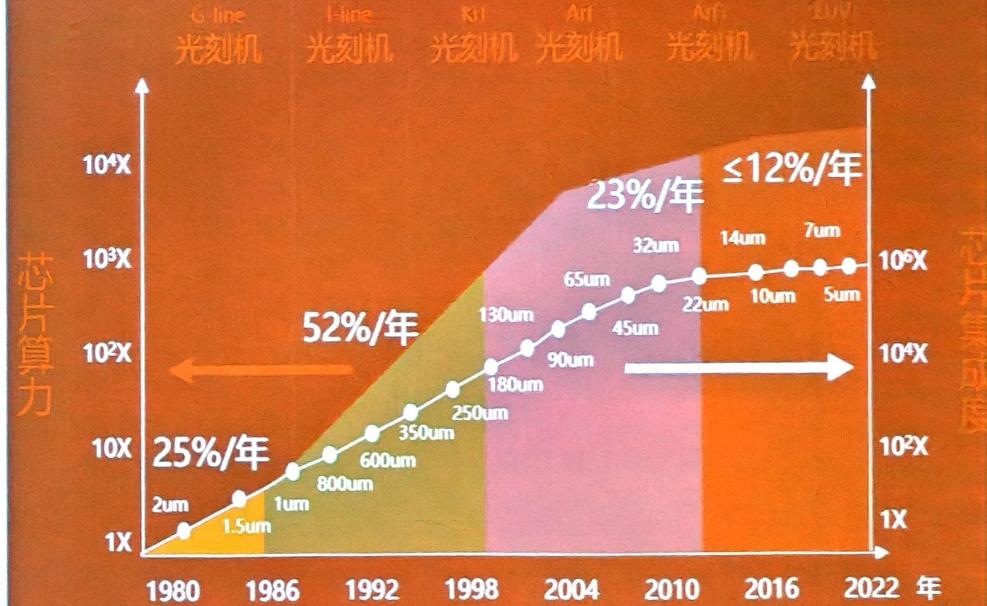
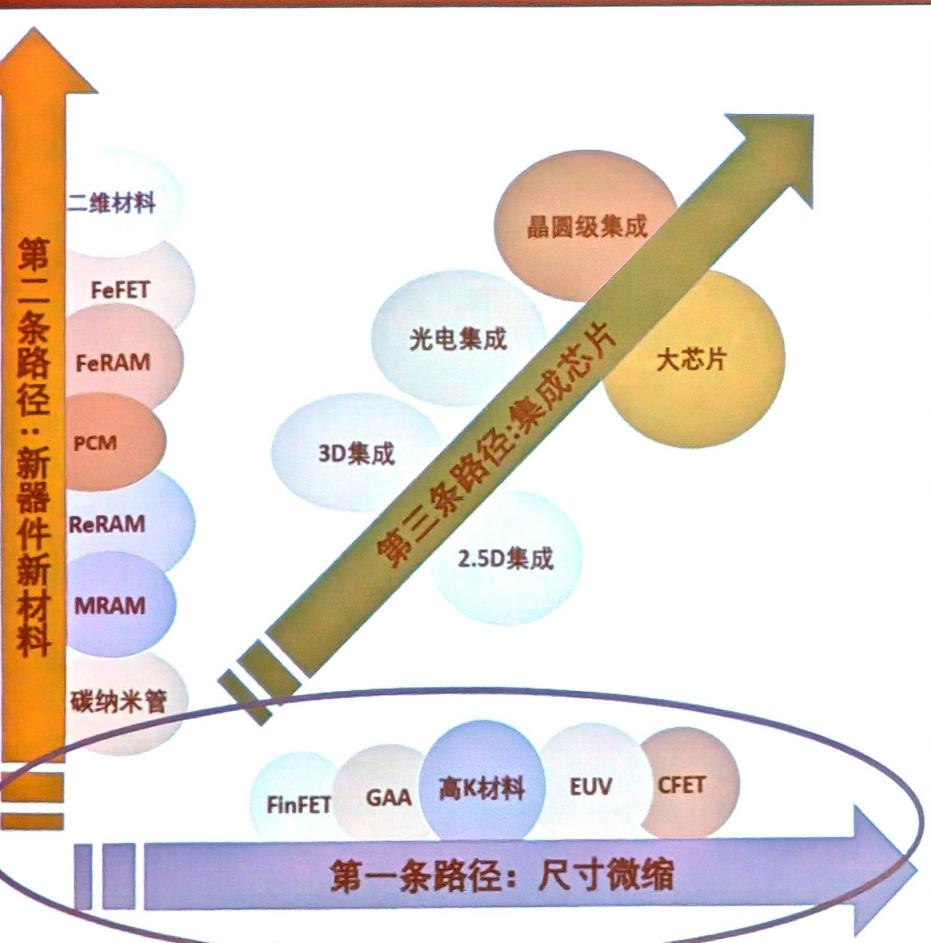
Memory Hierarchy with Bandwidth & Memory Size



训练大模型所需的内存...

大模型海量的训练数据...

1-1. 算力的核心是芯片技术



遵循摩尔定律提升算力的路径，随着尺寸微缩接近物理极限，未来将会失效

1-1. 奔跑的“中国芯”

数据中心



摩尔线程
MOORE THREADS



壁仞科技
BIREN TECHNOLOGY



天数智芯
Iluvatar CoreX

INNOGRIT

云豹智能
Micro

百度昆仑

METAX
沐曦集成电路

HYGON



赛昉科技
StarFive



MEMBLAZE



兆芯

BIWIN.

汽车



地平线
Horizon Robotics



JWJOULWATT

AXERA



禾赛科技



制造&封测



CXMT



CanSemi



Innoscence
英诺森科

材料&设备



FerroTec



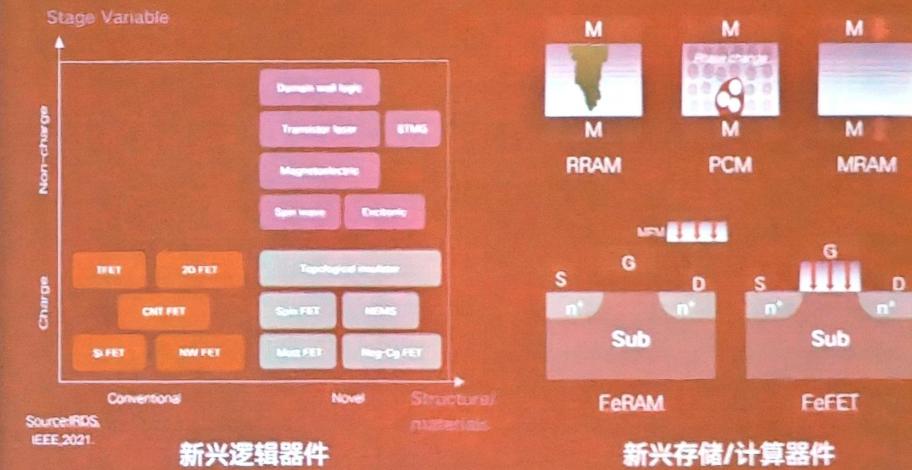
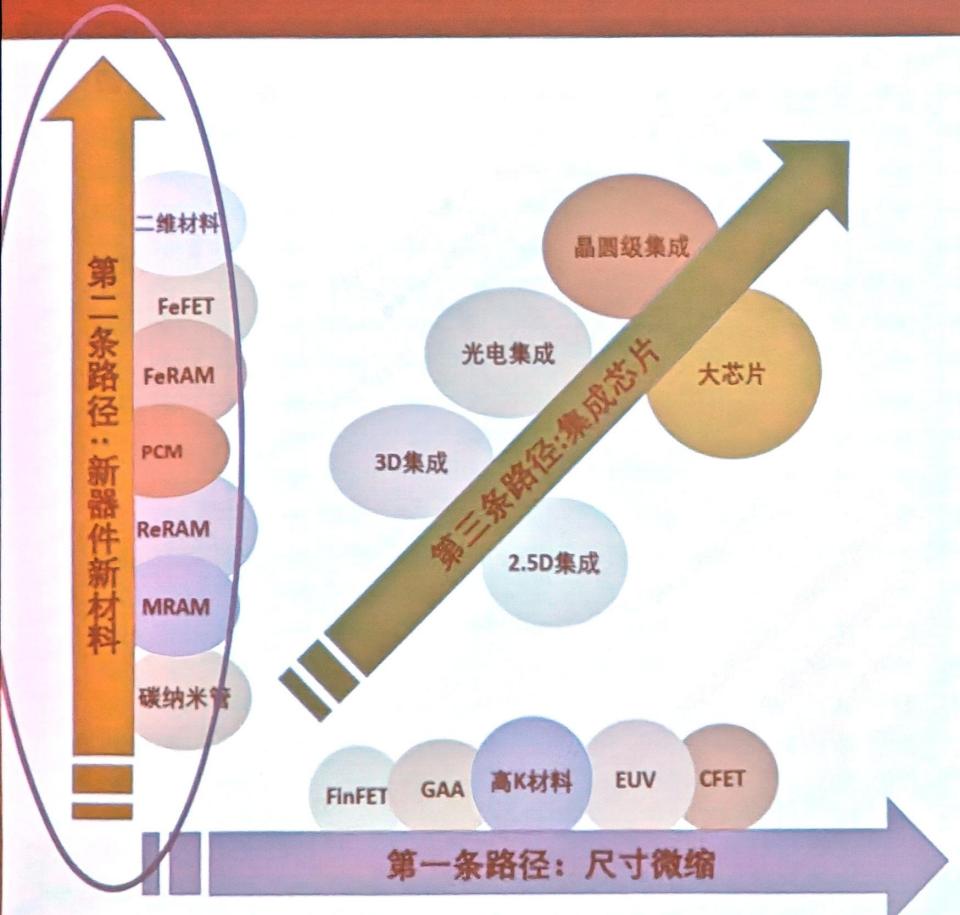
JAST

SICC.

EDA&IP

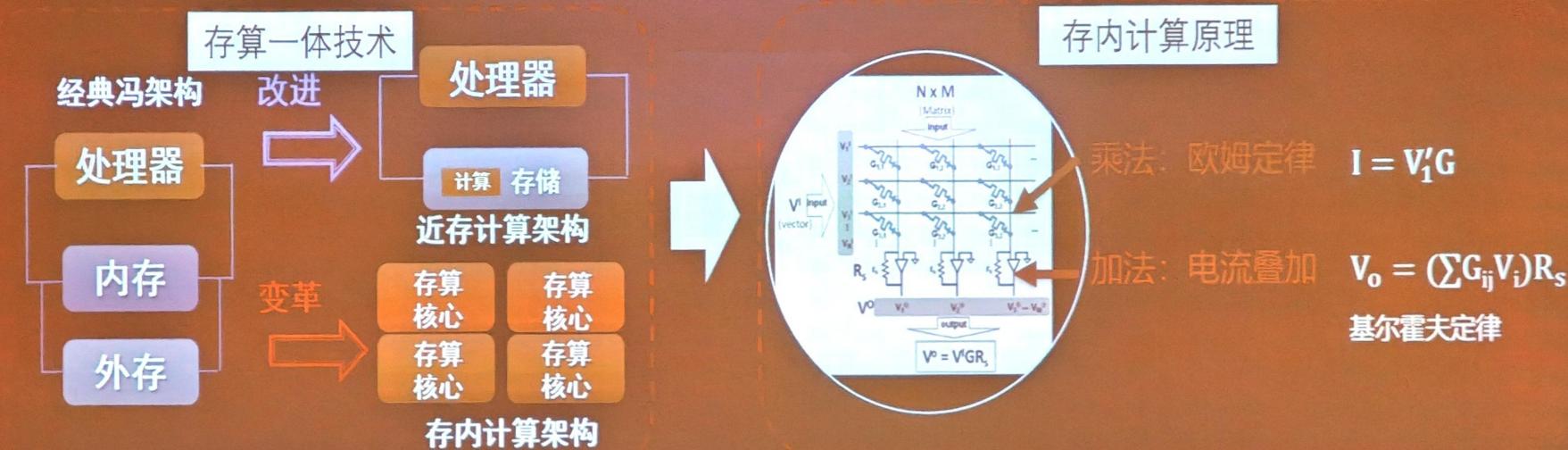


1-2. 新原理器件、突破冯诺依曼架构



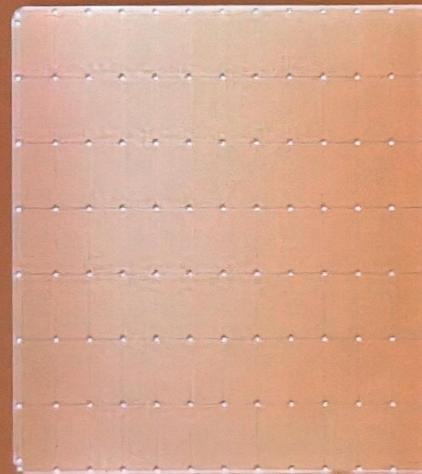
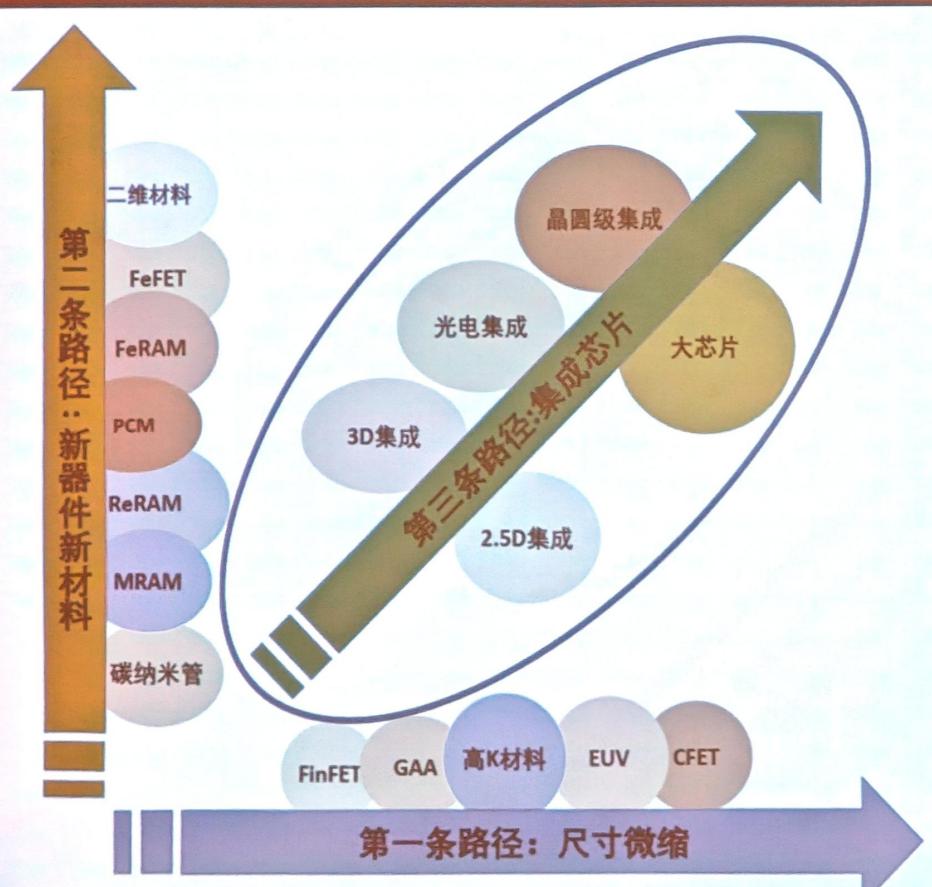
新原理器件，突破既有计算架构限制，
提升先进计算效率，大幅提高能效

1-2. 存算一体的非冯芯片架构，提升计算能效



| | Our work '2023 | Our work '2022 | Stanford '2022 | Michigan &Facebook '2020 | Tsinghua Tianjic '2019 | Alibaba NPU '2020 | MediaTek '2020 | IBM '2021 | Cambricon '2022 | Nvidia A100 '2020 | Nvidia H100 '2022 |
|------------|----------------|----------------|----------------|--------------------------|------------------------|-------------------|----------------|-----------|-----------------|-------------------|-------------------|
| 计算架构 | 存算一体（存内计算） | | | | 存算一体（近存计算） | | | | 传统 | | |
| 工艺 | 28 nm | 180 nm | 40 nm | 22 nm | 28 nm | 12 nm | 7 nm | 7 nm | 7 nm | 7 nm | 4 nm |
| 功耗(W) | 0.8 | 0.0328 | 0.135 | 0.128 | 0.95 | 280 | 0.174-1.053 | 3.88-11.5 | 75 | 400 | 350 |
| 能效(TOPS/W) | 7-10 | 1.56 | 2.2 | 0.96 | 1.28 | 2.95 | 3.42-13.32 | 8.9-16.5 | 2.56 | 1.56 | 8.65 |

1-3. 解决互连问题，通过集成芯片技术提升算力算能



Cerebras WSE-2

2.6 Trillion Transistors
46,225 mm² Silicon



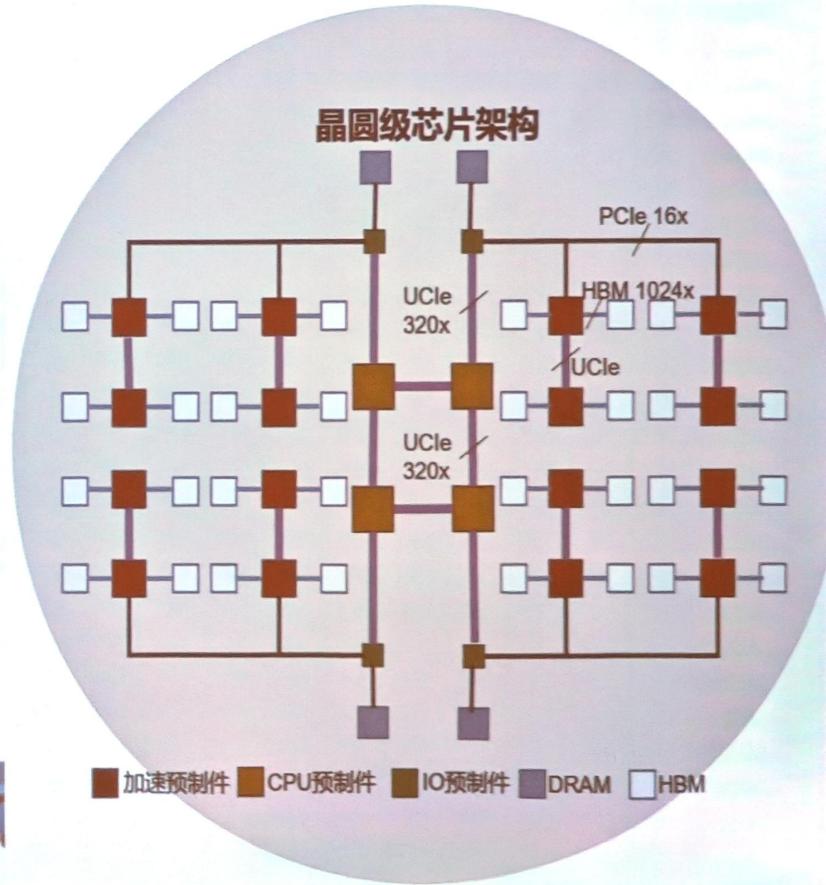
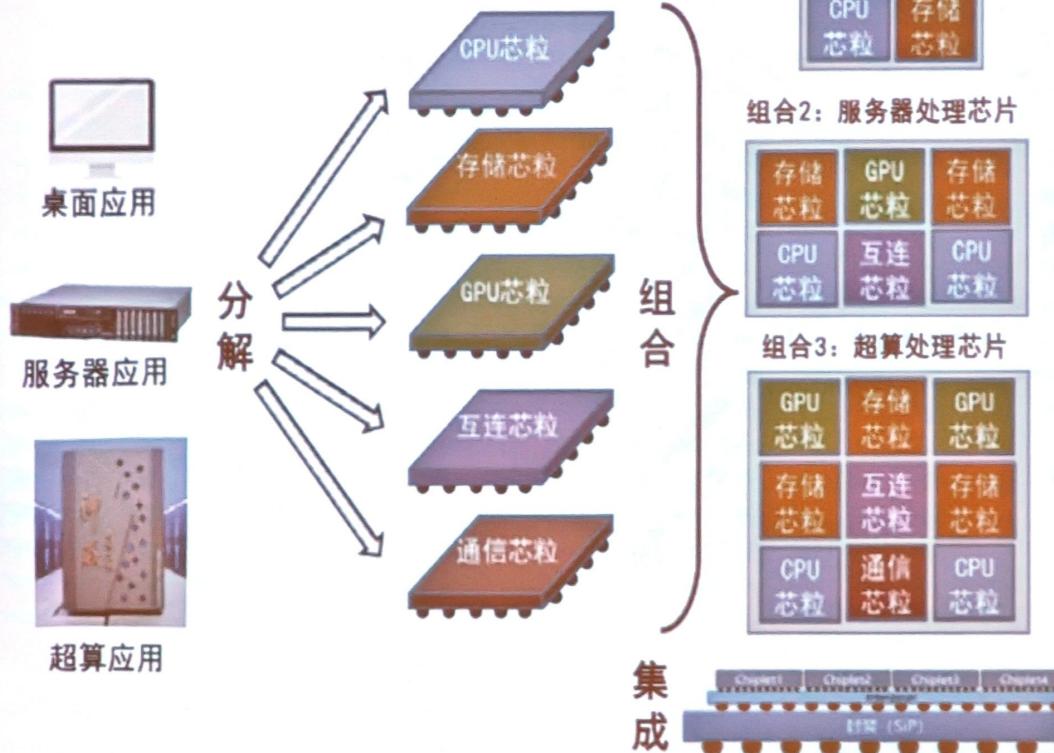
Largest GPU

54.2 Billion Transistors
826 mm² Silicon

通过集成芯片技术提升算力算能

1-3. 集成芯片技术的愿景

应用需求驱动



2. 大模型计算的数据挑战

高质量数据

DATA

Text



Images



Speech



Structured Data



3D Signal



Training

Foundation Model

Adaptation

TASKS



Question
Answering



Sentiment Analysis



Information
Extraction



Image Captioning



Object
Recognition

2-1. 行业的数据意识

基于深入的行业应用场景，首先构建自监督数据集以满足特定领域需求，随后通过大模型的微调和监督训练，实现对数据的高效利用和模型性能的优化。在工程部署和性能优化阶段，采用合适的部署策略和技术，确保模型在实际环境中的高效运行。同时，不断积累领域数据并进行模型迭代优化，以适应行业变化和不断提升模型的精度和适应性。



2-1. 行业合作案例：老板电器“食神”大模型

ROBAM 老板 爱是味道

提供基于更自然更具情感温度的烹饪体验

从 此 爱 上 烹 饪

交互更流畅简单
从指令到对话



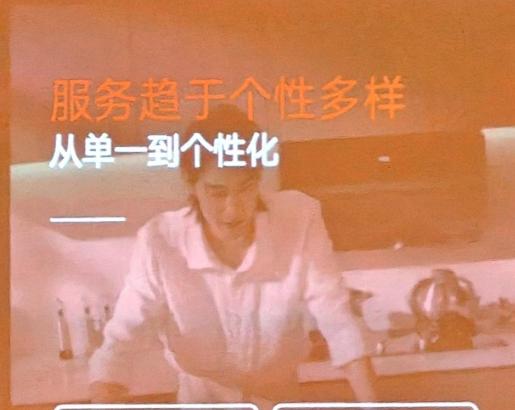
多模态交互

烹饪口语化交互

复杂意图理解

自我学习进化

服务趋于个性多样
从单一到个性化



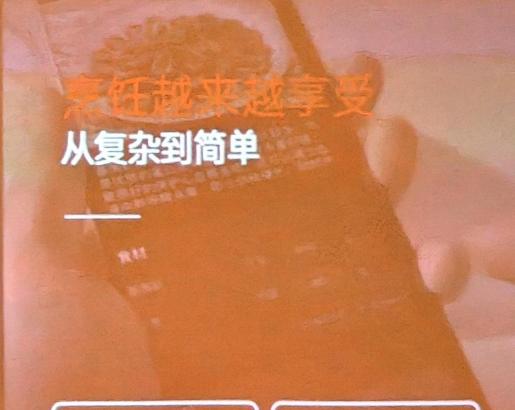
情绪共情

安全感知

主动记忆与服务

个性化服务和内容推荐

烹饪越来越享受
从复杂到简单



多餐协同

个性化菜谱

食材成熟度判断

智能实时烹饪指导

2-2. 构建行业数据的“闭环”

ROBAM老板 畅享创造

流畅的自然语言交互、最懂中式烹饪的助理、个性化主动服务

应用



数字厨电设备



ROKI 小程序



ROKI APP



ROKI IoT平台

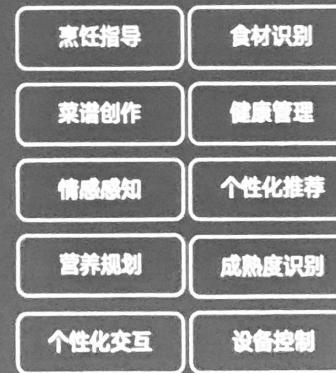
烹饪大模型

对话增强

多模检索增强

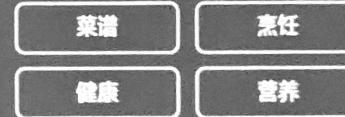
有监督微调

反馈学习



行业知识

知识图谱

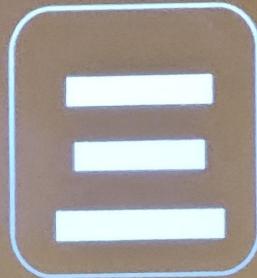


行业数据



多模态数据





未来展望

1. 计算的战略意义



万物数化的智慧时代

数据是统一的公共要素资源

计算技术支撑着数据的全生命周期，是数字经济时代核心生产力

获取

传输

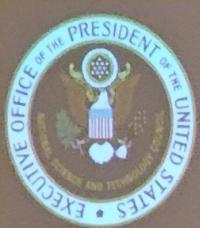
存储

处理

应用

计算技术

2. 先进计算已成为国际竞争焦点



NATIONAL STRATEGIC COMPUTING INITIATIVE UPDATE: PIONEERING THE FUTURE OF COMPUTING

A Report by the

FAST-TRACK ACTION COMMITTEE ON STRATEGIC COMPUTING
NETWORKING & INFORMATION TECHNOLOGY
RESEARCH & DEVELOPMENT SUBCOMMITTEE
COMMITTEE ON SCIENCE & TECHNOLOGY ENTERPRISE
of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

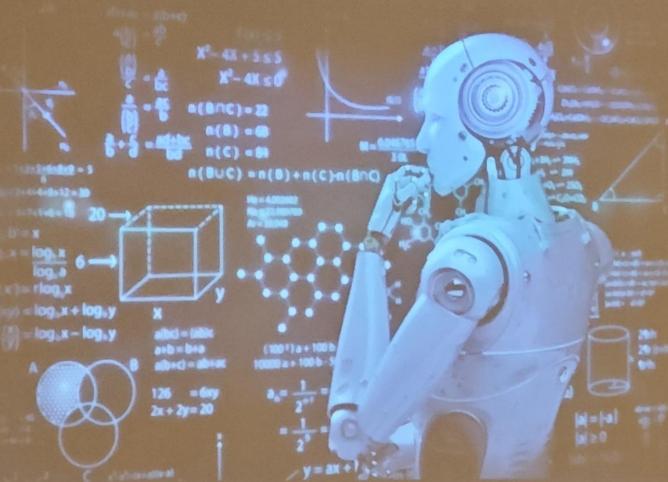
NOVEMBER 2019

美国国家战略计算计划：引领未来计算

- 2016、2019年两次发布
- 发展未来计算软硬件
- 打造计算战略基础设施
- 推动国际标准，实现数据发现、访问、兼容性和可重用性
- 强调网络安全对计算生态的重要性
- 建设学科、政策、产业生态系统；确保全球领先地位
- 奠定在科学及工程竞争力和国家安全等方面领跑的基础

4. 智能计算

智能计算是大数据、云计算和人工智能时代的计算变革，是以“计算密集（Computational Intensive）、数据驱动（Data Driven）、基于模型（Model Based）”为主要特征的计算形态



国家创新能力的核心基础

Computing by Intelligence

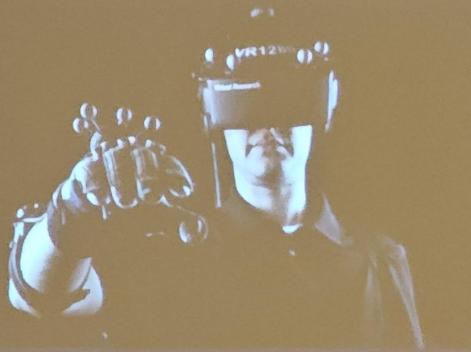


推动产业变革和开放创新的新动力

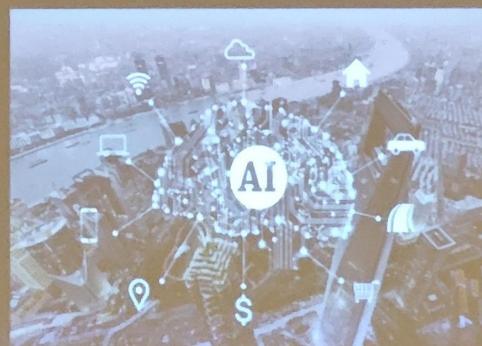
Computing for Intelligence

5. 智能计算的发展愿景

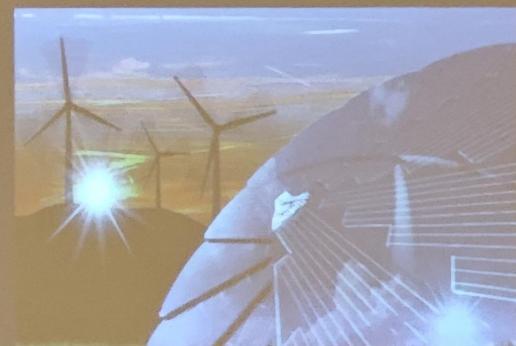
便捷化，泛在化，普惠化



自然人机交互，“傻瓜式”应用



面向应用适配，无处不在



优化资源消耗，经济适用

让计算像水和电一样广泛使用、便捷使用！