智源学者成果展示——智能体系架构与芯片

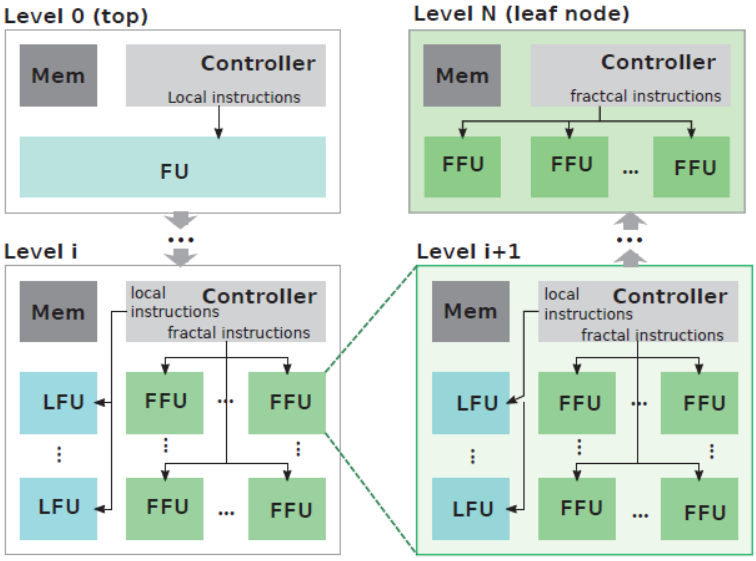
作者　陈云霁（中国科学院计算技术研究所研究员、智源首席科学家）
杜子东（中国科学院计算技术研究所副研究员、智源青年科学家）
蔡一茂（北京大学教授、智源研究员）
杨玉超（北京大学研究员、智源青年科学家）
刘　琦（中国科学院微电子研究所研究员、智源研究员）
陈文光（清华大学教授、智源研究员）
翟季冬（清华大学副教授、智源青年科学家）
张悠慧（清华大学研究员、智源研究员）
李国齐（清华大学副教授、智源青年科学家）
罗国杰（北京大学副教授、智源研究员）
尹首一（清华大学教授、智源研究员）
孙广宇（北京大学副教授、智源青年科学家）

2020 年 6 月

智能芯片：分形机器学习芯片

中国科学院计算技术研究所研究员、智源首席科学家陈云霁和**中国科学院计算技术研究所副研究员、智源青年科学家杜子东等**提出了一系列具有相同 ISA 的同类、序列、多层且相似的机器学习处理器架构 Cambricon-F，极大提高了编程效率，让不同设备具有同构的、统一的软硬件，可以共享同一套开发工具、同一套软件栈，运行同一套程序，从而极大地提高了机器学习计算机的处理效率。

Zhao, Yongwei & **Du, Zidong** & Guo, Qi & Shaoli, Liu & Xu, Zhiwei & Chen, Tianshi & **Chen, Yunji**. (2019). *Cambricon-F: machine learning computers with fractal von neumann architecture*. Proceedings of the 46th International Symposium on Computer Architecture.



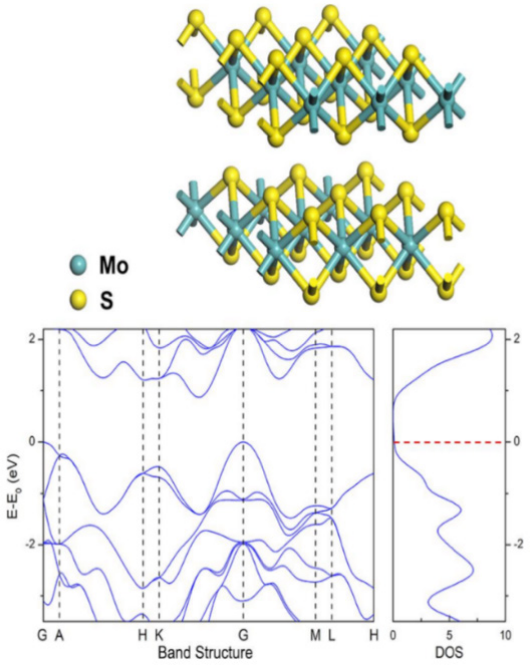
A typical fractal von Neumann architecture: level 0 (top node) ... level i node and its son node in level i+1 ... level N (leaf node).

图片来源：作者论文

智能芯片：类脑智能芯片

北京大学教授、智源研究员蔡一茂与**北京大学研究员、智源青年科学家杨玉超等**基于 MoS2 器件首次同时模拟实现了神经元和突触功能，使用双栅极晶体管结构的 MoS2 中性电阻，通过使用不同的驱动信号，MoS2 中性光片可以编程为神经元、突触或 n 型 MOSFET，可视为神经形态电路设计中的基本组件，为未来可重构神经形态系统提供了可行的解决方案。

Bao, Lin & Zhu, Jiadi & Yu, Zhizhen & Jia, Rundong & Cai, Qifeng & Wang, Zongwei & Xu, Liying & Wu, Yanqing & **Yang, Yuchao & Cai, Yimao** & Huang, Ru. (2019). Dual-Gated MoS2 Neuristor for Neuromorphic Computing. ACS Applied Materials & Interfaces. 2019.

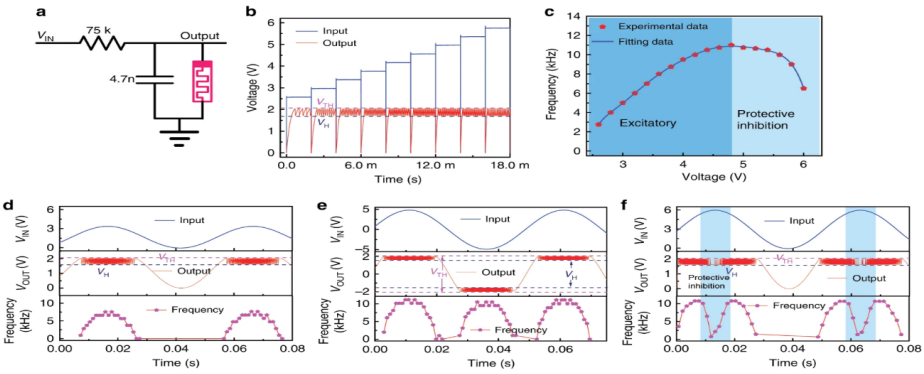


The band structure and density of states (DOS) of intrinsic MoS2
图片来源：作者论文

智能芯片：类脑智能芯片

中国科学院微电子研究所研究员、智源研究员刘琦等首次提出利用忆阻器构建人工脉冲传入神经元，实现了模拟信号到脉冲信号的转换，并搭建了零静态功耗的机械脉冲感受系统，可用于实现脉冲神经形态机器人与外界环境的实时交互。

Xumeng Zhang, Ye Zhuo, Qing Luo, Zuheng Wu, Rivu Midya, Zhongrui Wang, Wenhao Song, Rui Wang, Navnidhi K. Upadhyay, Yilin Fang, Fatemeh Kiani, Mingyi Rao, Yang Yang, Qingfei Xia, **Qi Liu***, Ming Liu*, J. Joshua Yang*. *An Artificial Spiking Afferent Nerve Based on Mott Memristors for Neurorobotics.* Nature Communications 11, 51(2020).



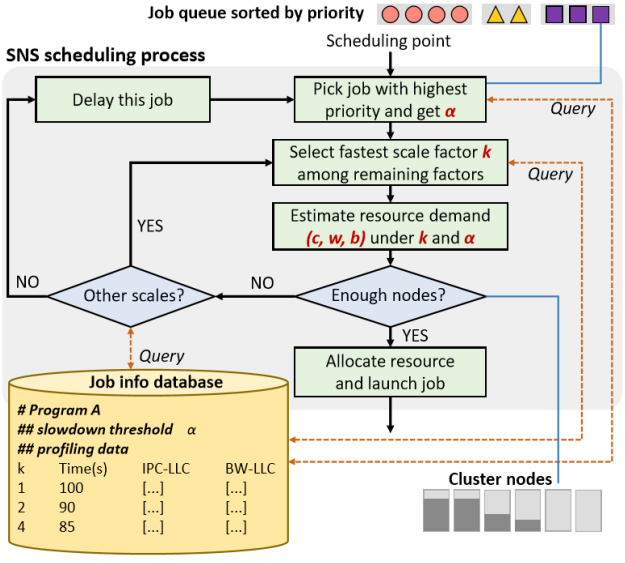
SNS scheduling process

（右）图片来源：作者论文

智能整机：智能超算任务调度

清华大学教授、智源研究员陈文光和**清华大学副教授、智源青年科学家翟季冬等**首次创新性地提出了一种能自动地将资源受限的应用程序扩展到更多的节点并以资源兼容的方式共同定位作业的批处理调度策略 Spread-n-Share（SNS），该新策略会自动地将资源受限的应用程序分散到更多结点上以缓解程序的性能瓶颈，并将资源互补的作业放置于共同的结点上，实现了验证 SNS 的原型调度器 Uberun，提高了 19.8% 的总体系统吞吐量，同时实现了平均 1.8% 的单任务加速。

Xiongchao Tang, Haojie Wang, Xiaosong Ma, Nosayba El-Sayed, **Jidong Zhai**, **Wenguang Chen**, Ashraf Aboulnaga. *Spread-n-share: improving application performance and cluster throughput with resource-aware job placement.* In Proceedings of International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2019.

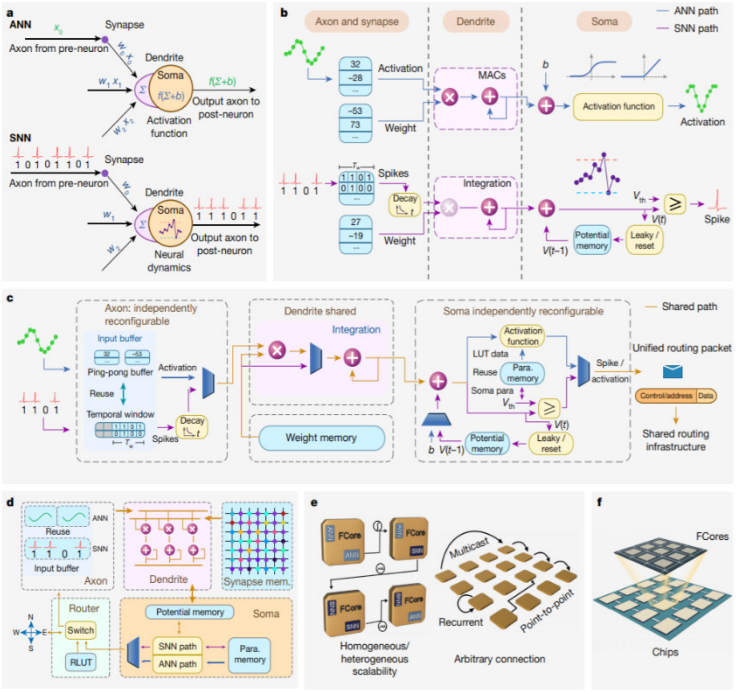


SNS scheduling process
图片来源：作者论文

智能芯片编程编译：异构融合架构、系统软件及应用

清华大学研究员、智源研究员张悠慧和**清华大学副教授、智源青年科学家李国齐等**首次提出了面向人工通用智能的 ANN 与 SNN 异构融合的发展路线，并在智能运行平台上完成认知 / 感知等任务，实现中国在芯片和人工智能两大领域《自然》论文零的突破。

Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, **Youhui Zhang（共同一作）**, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, Feng Chen, Ning Deng, Si Wu, Yu Wang, Yujie Wu, Zheyu Yang, Cheng Ma, Guoqi Li, Wentao Han, Huanglong Li, Huaqiang Wu, rong Zhao, Yuan Xie, Luping Shi. (2019) *Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature*.

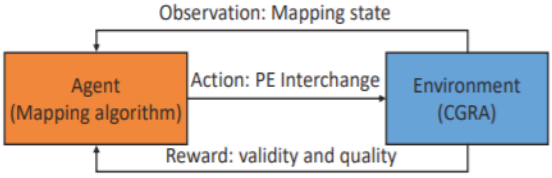


Design of the Tianjic chip. 图片来源：学者论文

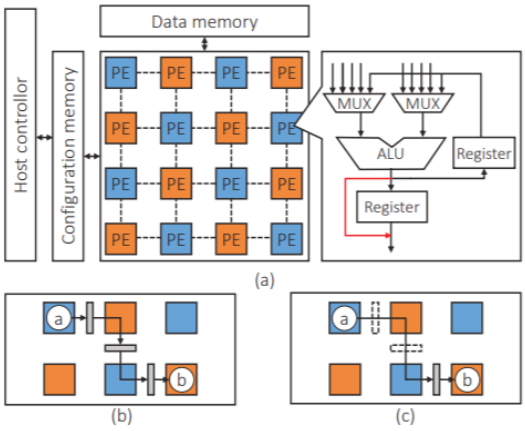
智能芯片设计方法：CGRA 数据流图映射方法

北京大学副教授、智源研究员罗国杰和**清华大学教授、智源研究员尹首一等**首次提出基于深度强化学习的 CGRA 数据流图映射优化技术，将 CGRA 上的 DFG 映射形式化为强化学习中的代理，该解决方案通过代理的交换行为来统一布局，路由和处理元素插入。实验结果表明，该方法的映射质量与最新的启发式算法相当，可以适应不同的体系结构，且能快速收敛。

Dajiang Liu, **Shouyi Yin, Guojie Luo**, Jiaxing Shang, Leibo Liu, Shaojun Wei, Yong Feng, and Shangbo Zhou. *Data-Flow Graph Mapping Optimization for CGRA with Deep Reinforcement Learning*. IEEE Transactions on Comput-er-Aided Design of Integrated Circuits and Systems (IEEE TCAD), 2019.



An overview of RL based DFG mapping on CGRA



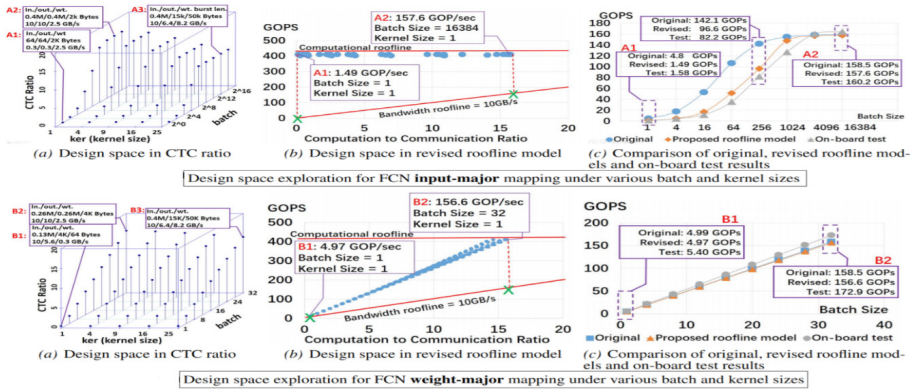
(a) A general CGRA architecture (b) A registered routing (c) Anunregistered routing

图片来源：学者论文

智能芯片设计方法：神经网络 FPGA 自动映射

北京大学副教授、智源青年科学家孙广宇等首次提出将 Caffe 框架输出的神经网络模型自动化映射到 FPGA 上，实现了和手工设计类似甚至更好的性能，达到了国际领先水平，并获得 TCAD 2019 最佳论文奖。

Chen Zhang, **Guangyu Sun**, Zhenman Fang, Peipei Zhou, Peichen Pan, Jason Cong. *Caffeine: Towards Uniformed Representation and Acceleration for Deep Convolutional Neural Networks*. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems , Vol. 38, No. 4（IEEE TCAD）, 2019.



图片来源：学者论文

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院