冯·诺伊曼《计算机与人脑》要点归纳及启发

刘宇航
中国科学院计算技术研究所

关键词：计算机　智能

引言

冯·诺伊曼在20世纪50年代中期著有《计算机与人脑》(*The Computer and the Brain*)[1]一书。这是他为原定1956年春天在耶鲁大学举办讲座准备的讲稿，但由于在1955年10月被查出患有癌症[1]，未能如期赴会演讲，讲稿也没有写完。全文虽然篇幅不长，只有83页，但蕴含着一些对今天仍然有启发意义的重要思想。用美国电气与电子工程师协会计算机学会(IEEE-CS)前任主席戴维·阿兰·格里尔(David Alan Grier)的话说，“从书中的字里行间，可以感受到冯·诺伊曼在争分夺秒地整理自己的想法”[2]。

我们仔细阅读了冯·诺伊曼的这部英文原著，并提炼了十个要点，或许能为今天的计算机基础理论、芯片设计、系统开发提供一些启发性的指导。这些思想、方法、技术，直至今天不但没有过时，而且变得越来越重要。它们都是在计算机诞生不足10年的时候由一个人在一本书中，而且是在重病中撰写完成的，是非常难能可贵的。我们对提炼的每一个要点结合新时代特征进行了分析，并指出了参考意义。

通过这些要点，我们对计算机科学的重要基石有了一些新的认知，有些之前我们认为可能比较新颖的东西（比如层次化存储），实际上在计算机诞生初期就被提出甚至进行了量化分析，每年顶级会议上出现的一些新成果都是这些思想的实现；有些之前我们认为可能比较陈旧的东西（比如虚拟化），实际上换一个角度可能是一种新的研究思路。真正具有本质的重要性的东西，无所谓“新”与“旧”，应该在历史发展中传承和保持下来。

要点与分析

1. 提出应对经验进行形式化

在原著引言部分的第2页，冯·诺伊曼提到理想的状态不是停留在经验(body of experience)层次上，而是要对经验进行形式化(formalized)分析。

在今天新时代条件下，人工智能、大数据、云计算、边缘计算等领域正在蓬勃发展，在实践中累积了越来越多的经验，但是理论基础都相对薄弱，都需要构建各自领域中有较强针对性的基础理论。以人工智能为例，数学家丘成桐教授2017年在中国计算机大会上指出[3]，人工智能需要一个坚实的理论基础，否则它的发展会有很大困难。人工智能

[1] 1957年，冯·诺伊曼因患癌症去世，年仅53岁。他曾以观察员身份参加了比基尼岛环礁核试验，很多人认为他得癌症与这次试验有关。——编者注

在工程上取得了很大发展，但理论基础仍非常薄弱。现代以神经网络为代表的统计方法及机器学习是一个黑箱算法，可解释性不足，需要一个可被证明的理论作为基础。计算机科学家、图灵奖得主姚期智教授在2017年表示，中国想在2030年实现人工智能世界创新中心的战略目标，首先要解决人工智能发展的理论问题[4]。2018年3月，美国国家科学基金会(NSF)投入了1000万美元支持加州大学伯克利分校研发实时、智能、安全、可解释(RISE)的人工智能系统[5]，可见世界对这一领域的“争夺”。

将更多的以形式化为特征的数学理论应用到计算机科学中，不仅能有效地提出各种计算机算法，而且能给出新兴领域的理论基础，即构建新时代计算机科学。例如，我们已开展针对云计算领域的可计算理论的研究，以图灵、邱奇、波斯特等创立的传统可计算理论为基础，考虑云计算领域的系统特点（比如有限资源共享）和应用特点（比如服务质量要求有差异），对实现“可计算”的资源要求和时间要求做了限定。这也是一种对新兴的云计算领域更加实用、更精准有力的理论，被称为“实用可计算理论”[6]。

2. 提出由一组“基本操作”通过组合和反馈实现“复杂操作”

冯·诺伊曼在第一部分第一章第5页表达了这一思想。这一思想可以分解为三层含义：(1)将操作区分为“基本操作”和“复杂操作”两类，基本操作只有少量的、固定的几种，复杂操作则可以有大量的、灵活的多种；(2)由“基本操作”构成“复杂操作”的方式只有两种，就是组合和反馈，这对应着组合逻辑电路和时序逻辑电路（这是电路的两大基本类型）；(3)实现给定的“复杂操作”的“基本操作”组合可能不唯一，从功能上看，可能有多组“基本操作”都可以实现“复杂操作”，但效率可能有差别。

这一思想的三层含义，对于今天的计算机，无论是科学基础理论，还是处理器芯片和计算系统设计，都是非常重要的。

对科学基础理论来说，可计算性理论是计算机科学最核心的基础理论，如果没有可计算性理论，计算机将难以称为计算机科学。递归函数是可计算性理论的核心概念，因为图灵可计算函数类就是递归函数类，两者完全等价。递归函数的定义，是从少量的、简单的基本函数开始的，这些基本函数有三个，一个是后继函数 $s(x)=x+1$，一个是零函数 $o(x)=0$，一个是射影函数 $U(x_1, x_2, \cdots, x_n)=x_j$。由这些基本函数通过代入、递归构成复杂函数。这里值得注意的是，代入、递归与组合、反馈是对应的。冯·诺伊曼的这一思想与可计算性理论有非常一致的对应，或许表面上看起来很简单，但它的实质内容处于计算机科学理论的核心位置。

对处理器芯片设计来说，将指令集和程序区分开来，可以以不变的少量指令构成万变的应用程序。指令集（比如x86, MIPS, RISC-V）中的不同类型的指令都是有限的，但可以编写的不同程序的数量极其庞大，这样硬件上的固定性与软件上的任意性的矛盾就得到了解决。同样一个功能的程序，用不同的指令实现，效率会有比较大的差异，所以指令集的设计非常重要，比如寒武纪芯片对机器学习领域的应用程序设计了专门的指令集[7]，以获得比传统通用指令集更大的效率提升。

对计算系统设计来说，需要将处理器核、高速缓存、内存、网络等硬件的特征与已有的和即将开发的应用程序的特征区分开来。计算系统包含处理器芯片，所以计算系统比处理器芯片高一个层次，设计层次相应地从微体系结构上升为体系结构。硬件的特征决定了它满足应用程序需求的能力。当前设计少量几种加速器配合通用处理器的异构计算成为一种流行模式。加速器比指令更宏观，但本质是一样的，都是完成复杂操作的基本操作。现在需要研究的问题是，应该设计哪些加速器（对应着定义基本操作），如何协同这些加速器和通用处理器（对应代入、递归，或者说组合、反馈）。

3. 提出人造计算机和人脑在本质上都是自动机

冯·诺伊曼在第一部分第二章第7页表达了这一思想。自动机是计算机和人脑的基类。他将计算机的存储器称为机器的记忆器官 (memory organs)，将运算器称为机器的活跃器官 (active organs)。“器官”一词的使用，反映了冯·诺伊曼对计算机的认识，他采用一种类比的方法来认识计算机。如果人脑和计算机之间在客观上存在本质上的联系，而这种本质联系，是设计出具有智能的计算机所必需的。那么这种类比就不是可有可无的，而是必须的，否则就无法认识这种联系。

我们通过表1对比了人类大脑与处理器芯片的一些指标，其中处理器芯片的数据参考了龙芯3B 1500 [8]。整体上看，人脑比计算机内部单元多，能效高，内部互连密集，便于并行处理，这些优点都值得计算机借鉴。

4. 提出存储程序的思想

冯·诺伊曼在第一部分第三章第17页表达了这一思想。在实现存储程序 (memory-stored control) 之前，以插入方式使用控制序列点描述要计算的问题，一个求解问题对应一种插入方式，求解问题改变之后，插入方式也要改变。具体分析来说，存储程序的思想可以分解为三层含义：(1) 指令在逻辑意义上描述一种操作，也仅仅是在逻辑意义上指令与数据才有区别；(2) 指令在物理意义上与数据一样，都用数字表示，都存放在存储器中，在相同的计算机硬件上，只须改变软件来描述不同的应用问题；(3) 指令要有确定的后继者，这样一个指令序列可以被连贯地、自动地执行，作为一个整体完成一个功能。实现存储程序的思想，是现代计算机的标志。

5. 提出摩尔定律的雏形

冯·诺伊曼在第一部分第七章第30页预测，器件速度每10年可能提高两个数量级。这个预测的表述，虽然不像1965年摩尔用两页纸表述的那样正式，但是与实际是相符的。如果每两年翻一番，10年将有32倍的速度提升 ($2^{10/2}$=32)；如果每18个月翻一番的话，10年将有约100倍的速度提升 ($2^{10/1.5}$=101)。无论是两年还是18个月翻一番，32和101都是两个数量级。翻一番的周期受各种因素影响具有波动性，摩尔原来预测是18个月，后来修改为24个月，有延长的趋势 [9]。冯·诺伊曼的表述有两大优势：(1) 提出的时间比摩尔早约10年；(2) 以10年的尺度考察速度提升，结果是个数量级，避免了直接提出具有波动性的具体的翻一番的周期长度。

器件工艺创新和体系结构创新，是计算机系统性能提升的两大动力。其中，器件工艺创新服从摩尔定律。超级计算机性能的发展遵循千倍定律，即每隔10年超级计算机的性能就会提高三个数量级，根据上面的分析，这里很快就有一个推论，这三个数量级中有两个数量级是由于器件工艺创新的贡献，有一个数量级是由于体系结构创新的贡献。以数量级为单位，两者的贡献比为2 : 1。随着摩尔定律的逐渐失效，这个比例会逐渐变为1 : 1，最后趋近于0 : 1。

6. 提出在记忆器官附近要有“活跃”器官提供服务和管理

冯·诺伊曼在第一部分第七章第31页提出这一思想，大意是：“记忆”集合需要有辅助的子组件，它由“活跃”器官组成，用以服务和管理记忆集合。这一思想可以解读为：在存储部件附近可以而且应该配有运算部件。

在过去60年中，人类逐渐地做到了这一点，有两点体现：(1) 存储体附近配有相应的控制器，在控制器中记录元数据 (metadata)，并做服务和管理，对系统内部操作进行管控，使系统从完全无序状态

表1 人类大脑与处理器芯片的比较

	器件数量	功耗	频率	通信	体积	能力
人类大脑	10^{10}~10^{11}个神经元	20W	100Hz	10^{14}个突触	1.4×10^{3} cm^3	可处理数学上无法严格定义的问题
处理器芯片	10^{10}个晶体管	40W	10^{9}Hz	稀疏互连网络	3.2 cm^3	只能处理数学上严格定义的问题

转变为有序状态（即低熵状态[6]），实现诸如高速缓存替换策略、预取策略、调度策略等，每年ISCA、HPCA、Micro、ASPLOS等计算机体系结构或系统领域的顶级会议上都会有这样一些论文，其中标签化体系结构[10]是一个典型代表。(2) 在过去几十年，逐渐出现了PIM(Processor In Memory)形式的附带有计算功能的存储器，这可以有效地应对大数据的容量大、价值密度低的挑战。长期以来，一种广泛传播的说法是：冯·诺伊曼结构是一种计算与存储分离的结构，由此导致了数据供应能力成为瓶颈（即存储墙问题）。这种说法不仅仅是对冯·诺伊曼本人原意的误解，也是对冯·诺伊曼结构的误解，也因此影响了我们设计最优的冯·诺伊曼结构。

结合下一节我们将介绍的存储系统层次化原理，我们发现，冯·诺伊曼始终对计算（活跃器官）和存储（记忆器官）采取一种中立的不偏不倚的视角，没有以其中一个为中心。现在越来越强调以存储为中心，其实是在抵消早些年以计算为中心，这是时间上的对称。计算与存储是对称的，比如，有存储层次结构 (memory hierarchy) 就有计算层次结构 (computing hierarchy)；有PIM就有MIP(Memory In Processor)，MIP就是存储层次化结构。通过PIM或MIP，实现了“存算一体化”，存算没有分离。总之，对称与分离是两回事。

7. 提出了存储系统（记忆器官）的层次化原理

冯·诺伊曼在第一部分第七章第32~36页论述了这一原理。此原理可以分解为三层含义：(1) 将存储层次划分为多级；(2) 随着存储级序数的增加，存储容量增大，存取时间增加，单位容量的经济成本会随之下降；(3) 即使在同一存储层次上，存取时间也有变化性，其平均值受到待解算问题（统计性质）的影响。

作为这个学科最重要、最核心的内容之一，这一原理在帕特森 (Patterson) 所著的计算机体系结构教科书的各个版本中都保留了下来。近年来各种新兴的存储器件（如NVM）让我们感觉到存储层次化可能是比较新的思想，但实际上在60年前就被冯·诺伊曼正式地提出了。据我们目前所知，《计算机与人脑》是被广泛使用的平均存储访问时间模型 (Average Memory Access Time, AMAT) 的最早出处。

对应上述含义的第3点，冯·诺伊曼用专门的一节讨论了存取时间的复杂性，这对今天的研究尤为重要。现在做体系结构研究，模拟器是进行性能评估的一个重要工具，但有一个突出的缺点就是速度慢，其中一个重要原因是对存储系统的精细模拟引发的时间开销。比如在常用的GEM5模拟器中，存储系统可以采用不同精度的模型，最简单的一种是假设同一层次的所有存储访问的存取时间都为常数，这时模拟速度比较快，但是与实际存储系统的偏离很大，模拟结果的误差也很大。总之，这一原理对于今天来说越来越重要，冯·诺伊曼在计算机诞生初期能够一般化地提出这一原理，其思考力和原创精神，非常值得我们学习。

8. 提出Amdahl定律的雏形

冯·诺伊曼在第二部分第九章第51页指出，“不是任何串行运算都是能够直接变为并行的，因为有些运算只能在另一些其他运算完成之后才能进行，而不能同时进行（即它们必须运用其他运算的结果）”。这个思想略微量化一下，就是Amdahl定律了。在Amdahl定律基础上，Gustafson定律和Sun-Ni定律后来分别于1988年和1990年建立。这三个定律是超级计算的三大基本定律。当我们每半年看到超算Top500排行榜时，应该想到那些速度不断倍增的超级计算机背后的规律，冯·诺伊曼60年前就已经大体认识到了。

9. 提到图灵的工作作为虚拟化的雏形

冯·诺伊曼在第二部分第十三章第72页提到图灵在1937年证明的“有可能发展一种代码指令系统，使得一台计算机像另一台计算机那样操作”。长期以来，波佩克 (Popek) 1974年的论文[11]被很多人认为是虚拟化技术的基础理论的源头，实际上这个源头应该至少向前推37年。

这里虚拟化的含义有两层：一是像现代工业界所做的那样，例如，在使用MIPS指令集的龙芯处理器上运行x86指令构成的程序，就要经过一个以指令翻译为实质内容的虚拟化过程，这个基于龙芯处理器的计算机就像一个基于Intel处理器的计算机那样工作；二是计算机和人脑，能否相互像对方那样工作？

以虚拟化的观点去理解计算机和人脑，是一个非常独到新颖的角度。相比第一层含义，第二层含义的意义更大。现在学术界对于弱人工智能和强人工智能划分的界限，对强人工智能能否实现、如何实现、是否应该实现、是否应该研究，都还在讨论之中。一个确定的事实是，冯·诺伊曼在60年前就已经研究了“计算机能否思考”这个问题，他的思路和观点在今天仍值得审视和参考。

在集合论里，若集合A包含集合B中的每个元素，并且集合B包含集合A中的每个元素，那么这两个集合就是等价的。1974年的虚拟化理论论文的核心理论基础就是这一原理[12]。冯·诺伊曼需要证明人脑可以做计算机能做的所有事情（仅考虑功能，不考虑效率），还需要证明计算机可以做人脑能做的所有事情[2]。用虚拟化的语言来说，人脑的指令集是X，计算机的指令集是Y，当计算机的任意一条指令在人脑上执行时，能否用X中的一组指令模拟这条指令的功能？当人脑的任意一条指令在计算机上执行时，能否用Y中的一组指令模拟这条指令的功能？

冯·诺伊曼在文献[12]中使用生物学的例子，作为描述计算机的基础，说明计算机所有的部件在本质功能上都能在人脑中找到对应物，这样就证明了人脑可以做计算机能做的每件事情，也就是人脑可以虚拟化计算机。但是关于计算机是否可以虚拟化人脑，目前还没有得到证明，原因就是人脑不是人设计的，人目前还不知道人脑的指令集。

Amdahl定律、Gustafson定律、Sun-Ni定律

1960年代人们开始用互连网络将多个处理器连接起来以实现更快的计算机，但对于这种多处理器系统能达到的加速效果却无法判断。1967年Gene Amdahl在美国信息处理学会(AFIPS)的年度学术会议上提出了一个并行计算加速模型，解决了困扰人们的问题。Amdahl巧妙地把程序分为只能串行执行部分s与可被并行加速部分p($s+p=1$)，于是在N个处理器的并行计算机上就可以得到如下加速比：

$$Speedup(N) = \frac{1}{s+\frac{p}{N}}$$

这就是著名的Amdahl定律。该定律描述了一个非常简洁优美的并行程序加速模型，但却推导出一个悲观的结论——不管N有多大，任何程序的加速比都存在上限，也就是$1/s$。这导致在很长一段时间内人们都认为没有必要研制大规模并行计算机。1988年，John L. Gustafson在*Communications of the ACM*上发表论文，提出了Gustafson定律。该定律的核心在于打破了Amdahl定律中任务量固定的假设（即$s+p=1$），提出计算固定时间完成的任务量($s+p*N$)来得到可扩展的并行加速比：

$$Speedup(N) = \frac{s+p*N}{s+p} = N + (1-N)*s$$

根据Gustafson定律，由于不限制任务量，所以可以通过扩大问题规模，增加$p*N$工作量来提高并行计算加速比，可以说是解放了人的思想。1988年以后，并行计算机的规模越来越大，解决问题的规模也越来越大。例如美国能源部“百亿亿次级计算计划”(ECP)指出，从P级计算到E级计算，晶体管性能只能提高50%，而并行度要提高670倍，也正是Gustafson定律的体现。

1990年，密西根州立大学的博士生孙贤和（现为伊利诺理工大学教授）与倪明选（现为澳门大学教授）在Supercomputing会议上提出了一种Memory-Bounded并行加速模型，成功地统一并扩展了Amdahl定律与Gustafson定律。该模型后被称为“Sun-Ni's Law”，写入并行计算教科书。

10. 提出以“语言”的角度理解人脑，人脑的语言不是数学的语言

冯·诺伊曼在第二部分第十四章第 81~83 页提出这一思想。他指出，语言的出现是历史的偶然。人类有很多很多语言，这种多样性说明任何一种语言中都没有绝对的和必要的东西。比如希腊语或梵语只是历史的事实，而不是绝对的逻辑上的必要。人类的逻辑和数学在表达形式上都是历史的偶然，它们也可以存在于现有形式之外的其他形式之中。

人类进行逻辑推理或算术演算时，可以延续很多步数（这里步数称为逻辑深度或算术深度），因此人类的数学表现为很深的逻辑深度或算术深度。但是神经系统采用的是很小的逻辑深度或算术深度，比如人类的视网膜对于眼睛所感受到的图像进行重新组织，是由三个顺序相连的突触实现的，即只有三个连续的逻辑步骤，逻辑深度为 3，可以限制误差的累积和传播。因此，神经系统中的逻辑结构与人类的逻辑和数学中的逻辑结构是不同的。

把神经系统所使用的语言称为第一语言，人类在讨论数学时，是在讨论建立在第一语言之上的第二语言。神经系统中的数学和逻辑，当被视为语言时，与人类通常所说的语言有本质上的不同，两者要有一个翻译的过程，也就是虚拟化的过程。

相关工作

2017 年 3 月，戴维·阿兰·格里尔在《中国计算机学会通讯》上就冯·诺伊曼的《计算机与人脑》撰文，指出“计算机科学家们常常只喜欢读近期发表的文献。IEEE 和 ACM 数字图书馆的统计表明，我们很少阅读两三年前发表的文献。然而，这样的行为局限了我们的视野，使我们忽视了本领域积淀的财富。我们只注意到我们现在关心的那些问题，反而遗漏了**影响计算机领域研究 60 年的那些重要思想**。”

戴维强调了阅读经典文献的重要性，介绍了冯·诺伊曼著作撰写的历史背景，提出不应该遗漏“影响计算机领域研究 60 年的那些重要思想”，但是没有具体指出这些思想。本文可以与戴维的短文相互补充。

结束语

本文按照原著表述顺序总结了冯·诺伊曼著作的 10 个要点，这些对今天计算机的理论创新、工艺创新、结构创新都是非常值得参考的重要思路。 ■

刘宇航

CCF 专业会员。中国科学院计算技术研究所副研究员。主要研究方向为计算机体系结构、高性能计算、大数据、智能并发系统。

liuyuhang@ict.ac.cn

参考文献

[1] Neumann J V. *The Computer and the Brain*[M]. Yale University Press, 1958.

[2] 戴维·阿兰·格里尔 . 老与新: 计算机与人脑（吴茜媛、李姝洁，译）[J]. 中国计算机学会通讯 , 2017, 13(3): 67-68.

[3] 丘成桐 . 现代几何学与计算机科学 [J]. 中国计算机学会通讯 , 2017, 13(12):8-13.

[4] 新华社 . 姚期智：人工智能当前缺少理论　中国有望实现突破 [OL].(2017-08-24).http://www.xinhuanet.com/2017-08/24/c_1121538019.htm.

[5] NSF invests $30 million to pursue transformative advances at frontiers of computing and information science[OL].(2018-02-27).https://www.nsf.gov/news/news_summ.jsp?cntn_id=244648&org=OAC&from=news.

[6] 徐志伟 , 李春典 . 低熵云计算系统 [J]. 中国科学 - 信息科学 , 2017,47(9):1149-1163.

[7] Shao Li, Du Z, Tao J, et al. DianNaoYu: An Instruction Set Architecture for Neural Networks[C]// *Proceedings of the 43rd ACM/IEEE International Symposium on Computer Architecture (ISCA'16)*, 2016.

[8] Hu W, Zhang Y, Yang L, et al. Godson-3B1500: A 32nm 1.35GHz 40W 172.8GFLOPS 8-core processor[C]// *IEEE International Solid-State Circuits Conference Digest of Technical Papers*. IEEE, 2013:54-55.

[9] Waldrop M. More than Moore[J]. Nature, 2016(530):145-147.

[10]Ma J, Sui X, Sun N, et al. Supporting Differentiated Services in Computers via Programmable Architecture for Resourcing-on-Demand (PARD)[J]. Acm Sigarch Computer Architecture News, 2015, 43(1):131-143.

[11]Popek G J, Goldberg R P. Formal requirements for virtualizable third generation architectures[J]. Acm Sigops Operating Systems Review, 1974, 17(4):412-421.

[12]Neumann J V. First draft of a report on the EDVAC[J]. IEEE Annals of the History of Computing, 2002, 15(4):27-75.