物理学打破摩尔定律？

刘子鸣[1]，段明阳[2]，刘玉鑫*

摘要

对于计算机性能的发展趋势这个问题，著名的摩尔定律似乎可以给出答案——芯片会每18至24个月小一倍。但是摩尔定律会一直保持正确吗？在本文中，我们结合了量子力学、热力学统计物理、信息论的知识，探索了计算机性能的物理极限，其中包括计算速度、存储空间和并行程度。我们将阐明，这些性能分别和计算机的能量、熵和几何尺寸有着密不可分的联系。另一个重要的问题是，什么样的物理体系便于实现这种“终极计算机”，文中探讨了几种有趣而重要的物理体系。

关键词

终极计算机　性能极限　物理学

[1]北京大学物理学院，*1600011313*　[2]北京大学物理学院，*1600011311*　***通讯作者：** yxliu@pku.edu.cn

1. 介绍

摩尔定律是由英特尔（Intel）创始人戈登·摩尔（Gordon Moore）提出来的。其具体内容为：当价格不变时，集成电路上可容纳的元器件的数目，约每隔18-24个月便会增加一倍，性能也将提升一倍。这一定律解释了信息技术进步的速度，体现了人类的智慧。

值得说明的是，摩尔定律虽然被称为定律，它却显然不能由更基本的原理推出。事实上，它仅仅是时代进步、人类智慧的产物。所以我们完全有理由相信，摩尔定律终将失效。为论证摩尔定律失效的必然性，可以简单地采取这样的论述：摩尔定律要求元器件越做越小，根据摩尔定律外推，当元器件小于人们认知的长度尺度之时，摩尔定律就会失效。这个论述还只是一个保守的论述。本文会说明，当元器件尺度小到量子效应不可忽略时，摩尔定律就已经受到了极大的挑战。人的智慧造就了摩尔定律，但人的智慧毕竟无法突破自然的极限。

为了讨论方便，我们介绍一位文中会多次出现的朋友，它叫做“终极计算机”，如图1所示。我们的这位朋友重1kg，体积为1L，它的所有性能都达到了物理上的极限。它仍然是数字式计算机（digital computer），用0和1来存储信息，并通过简单的逻辑操作来进行运算，比如NOT，AND，FANOUT。

我们在正文中会仔细讨论对这些极限的估计。正文分为5个部分，第2部分探讨“终极计算机”的能量和计算速度的关系；第3部分讨论“终极计算机”的熵和存储空间的关系；第4部分研究“终极计算机”的尺寸和并行程度的关系；第5部分我们提出了几种可能实现“终极计算机”的物理体系，并讨论了它们的可行性。

图 1. “终极计算机”朋友

2. 能量限制计算速度

经典计算机的数基于二进制表示，二进制中1和0分别对应逻辑中的true和false。几个典型的布尔运算有AND（和），NOT（非），FANOUT（扇出[1]）。

2.1 计算机的基本操作

Toffoli发现，AND，NOT，FANOUT这三种运算可以由一个具有一般性的结构实现[1]，如图2所示。它有X，Y，Z三个输入和X'，Y'，Z'三个输出。X'会直接复制X，Y'会直接复制Y，写成逻辑等式即$X' = X$，$Y' = Y$。Z'的表达式是$Z' =$

[1]它的功能是，将一个数复制为多个，并输出。因为输入只有一个，而输出有多个，结构像一把扇子，因此叫做扇出。

$XY\overline{Z}+\overline{XY}Z$。简单来说，$Z$只有在$X$和$Y$同时为1时才会翻转[2]。

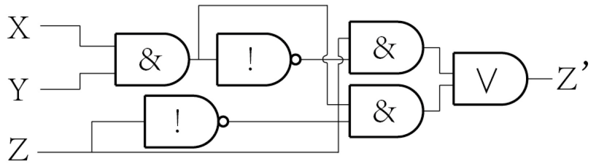

图 2. Toffoli装置

有趣的是，这个装置可以构造上面的三种运算。AND模式：设置$Z=0$，则有$Z=XY$；NOT模式：设置$X=Y=1$，则有$Z=\overline{Z}$；FANOUT模式：设置$Y=1$，$Z=0$，则有$Z'=X'$，且由于$X'=X$，相当于一个X复制了两份。

这三种运算的统一性，让我们有理由相信，计算机处理三种运算所需时间是差不多的(至少在同一量级上)，因此后面我们选取一个代表进行分析即可。下面我们分析，进行一次NOT运算所需的时间(下限)。

2.2 Not运算的时间下限[2]

从经典二进制来看，0和1是两个对立的态，因此NOT运算是良定义的。如果态指代的是量子力学中波函数所描述的状态，那么态处于无穷维的Hilbert空间中，这时就得重新思考NOT运算的意义了。我们还是尝试从经典情形获得灵感，经典0/1系统是一个双态系统，态矢量分别可以写作$(1,0)^{\mathrm{T}}$和$(0,1)^{\mathrm{T}}$，它们的一个显著特征是内积为0。由此，我们可以在量子世界中定义NOT运算：量子体系从初态演化到与初态正交的末态，称为一次NOT运算。根据不确定性原理，系统演化的速度受到能量的限制，因此NOT运算必然有一个时间下限。

一般的不确定性原理[3]告诉我们，对于两个力学量A和B，它们分别的标准差和对易子满足关系：$\Delta A\Delta B\geq\frac{1}{2}|\overline{[A,B]}|$。特别地，如果我们研究$A$的演化，取$B=H$，我们有$\Delta A\Delta E\geq\frac{1}{2}|\overline{[A,H]}|$。力学量平均$\overline{A}$随时间的变化关系是$\frac{\mathrm{d}}{\mathrm{d}t}\overline{A}=[A,H]/i\hbar$，带入前式并令特征时间$\tau_A=\Delta A/|\frac{\mathrm{d}}{\mathrm{d}t}\overline{A}|$，我们最终得到$\Delta E\cdot\tau_A\geq\hbar/2$。这个不等式告诉我们，$\Delta E$越小，$\tau_A$的值越大。

特别地，我们研究一个特殊情况，它可以非常简洁地说明NOT操作的意义。考虑一个二态系统，本征态和本征能量分别是$|\psi_1\rangle$，$|\psi_2\rangle$和E_1，E_2。假设在$t=0$时刻，系统处于$|\psi((0)\rangle=\frac{1}{\sqrt{2}}((\psi_1\rangle+|\psi_2\rangle)$的状态上。$t>0$时刻的态可以写成$|\psi(t)\rangle=\frac{1}{\sqrt{2}}(|\psi_1\rangle\exp(\mathrm{i}E_1t/\hbar)+|\psi_2\rangle\exp(\mathrm{i}E_2t/\hbar)$。何时$|\psi((t)\rangle$才会演化到$|\psi(0)\rangle$的正交态上呢？取两者的内积$|\langle\psi(0)|\psi(t)\rangle|=|\frac{1}{2}(\exp(-\mathrm{i}E_1t/\hbar)+\exp(-\mathrm{i}E_2t/\hbar))|=|\cos((E_1-E_2)t/2\hbar)|$。两态正交时内积为零，即$(E_1-E_2)t/\hbar=(2n+1)\pi(n\in\mathbb{Z})$。因此，一个状态演化到正交态，也就是翻转一比特所需的最短时间[3]为$t_{\min}=\pi\hbar/(E_1-E_2)=\pi\hbar/2\Delta E(\Delta E=(E_1-E_2)/2)$。

进一步地，我们只考虑正能量的体系，ΔE代表能量分布的标准差，而E代表能量分布的均值，则显然有$\Delta E<E$。将这个不等式带入上面$t_{\min}$的表达式，并且令频率$f=1/t_{\min}$，我们最终得到NOT操作频率为$f<2E/\pi\hbar$。为了获得直观的认识，我们借助“终极计算机”朋友来估算这个上界。取$m=1\mathrm{kg}$并有质能关系$E=mc^2$，计算出$f\lesssim 5.53\times10^{50}\mathrm{Hz}$，即1s内可以进行$\sim10^{50}$次NOT运算。

一个值得探讨的问题是：这个上限是否和计算机的工作模式有关？更确切地说，并行会不会比串行提高频率？令人失望的是，结论是否定的。因为我们给计算机分配多个任务，相当于对总能量进行划分，即$E=\sum_i E_i$，总频率$f=\sum_i f_i<\frac{2}{\pi\hbar}\sum_i E_i=\frac{2}{\pi\hbar}E$，得到了一样的上界。但不可否认的是，并行仍然是有意义的，不同的并行度会产生不同的能量的分散程度ΔE，这为将计算机视为一个正则系综提供了理论基础[4]。

2.3 和当今计算机的对比

当今的计算机计算速度差不多是$f\sim10^{10}\mathrm{Hz}$，远远小于$10^{50}\mathrm{Hz}$。原因是多方面的：第一，我们估算能量是用的质能关系，也就是认为锁在原子核内部的能量是可以释放出来的，但实际上我们很难利用这些内部能量；第二，在可用能量的部分里，我们也没有充分利用每一个自由度。举个例子，电路中的0/1都是依靠高低电位来实现的，而电位差是大量的电子堆积形成特定分布产生的后果，是电子的集体行为。

3. 熵限制了存储空间

[2]所谓翻转，是指把1变成0，把0变成1。也就是作NOT运算。

[3]另一个情况是，系统具有连续能谱，且能量均匀分布在E_1和E_2之间，计算表明最短的翻转时间是上面二态系统的两倍，即$t_{\min}=\frac{\pi\hbar}{\Delta E}$。

3.1 熵和信息量的关系

回顾一下熵的意义：熵描述了一个物理系统的混乱程度。所谓混乱程度，就是这个系统微观配型的数目。精确的数学描述来自于玻尔兹曼熵[5]，它定义熵S为微观配型数取对数，并乘以玻尔兹曼常数k_B：

$$S = k_B \ln W$$

对于经典的数字式计算机，假设它有m个位，则它一共能提供2^m种表示，即$W = 2^m$。它的玻尔兹曼熵为$S = k_B \ln(2^m) = mk_B \ln 2$。因此，一个熵为$S$的系统，它存储的信息量是$\frac{S}{k_B \ln 2}$比特。

3.2 单个比特的计算速度

上一部分我们计算得，对于“终极计算机”，总运算频率$f = \frac{2E}{\pi\hbar}$。我们不禁要问，单个比特的运算速度f_0是多少？将f除以比特数即得到f_0：

$$f_0 = \frac{2\ln 2 k_B E}{\pi\hbar S}$$

考虑到很多热力学系统满足如下关系[4]：

$$\frac{S}{E} \sim (\frac{\partial S}{\partial E})_V = T$$

所以我们得到$f_0 \sim \frac{k_B T}{\hbar}$。

3.3 Bekenstein关于最大存储空间的计算

限于篇幅的原因，这里略去Bekenstein原作[6][7][8]的推导过程，仅仅列出最重要的一些假设和结果：

- 考虑一个黑体辐射体系，体积为V，内部有不同种类的粒子；在温度T时，熵主要由质量小于$k_B T/2c^2$的粒子贡献；

- 第l种粒子贡献的总能量为

  $$E = r_l \pi^2 V (k_B T)^4 / 30\hbar^3 c^3$$

  其中r_l是和粒子有关的因子[5]；

- 第l种粒子贡献的熵为

  $$S = 2r_l k_B \pi^2 V (k_B T)^3 / 45\hbar^3 c^3 = 4E/3T$$

- 在以热波长λ_T为边长的正方体内，每种粒子贡献$(2\pi)^5 r_l / 90\ln 2 \approx 10^2$比特，其中热波长$\lambda_T = 2\pi\hbar c/k_B T$；

- 根据熵和信息量的关系，我们据此估算出信息量：

  $$S = (4/3)k_B(\pi^2 rV/30\hbar^3 c^3)^{1/4} E^{3/4} = k_B \ln 2 \cdot I$$

为了能够估计出数量级，我们考虑黑体辐射是由光子主导的($r = 2$)。为了让能量为$E = mc^2$，“终极计算机”需要达到$T = 5.87 \times 10^8$K的高温，同时熵为$S = 2.04 \times 10^8$J/K，比特数为$I = S/k_B \ln 2 = 2.13 \times 10^{31} \sim 10^{31}$。上一节我们估计出运算频率是$\sim 10^{50}$次，因此一个比特一秒内能进行$\sim 10^{19}$次翻转。

3.4 怎么维持“终极计算机”的运行？

一个热力学系统为了保持热平衡，外界输入的能量会以热的形式释放出来。所以关键的问题是，进行一次运算，外界需要多大的能量？Laudauer[9]曾对这个问题进行过细致的研究，他发现对于可逆逻辑操作(一对一或一对多，比如NOT，FANOUT)，外界不需要提供任何能量；但是对于不可逆操作(多对一，比如AND，ERASURE)，系统会有热量耗散。仍然可以通过微观状态数来理解这个现象：进行不可逆操作后，系统变得无序，微观状态数变多，熵增大，需要通过散热来维持热平衡。以ERASURE为例，擦除一个比特使得系统的熵增大$k_B T \ln 2$，从而外界作功和散热为$k_B T \ln 2$。

如果一台计算机只进行可逆逻辑运算，那么理论上它不会有任何能量耗散。但即使是经典的计算机，它也会具有自动纠错机制。一旦发现错误进行改正，就会进行一次ERASURE操作，从而释放热量。对于“终极计算机”，它运算速度那么快，如果要进行纠错，会不会释放过多的热量呢？

斯特番-玻尔兹曼定律(Stefan-Boltzmann law)告诉我们，黑体单位时间单位面积发送给环境的比特数为：

$$B = \pi^2 k_B^3 T^3 / 60\ln 2 \hbar^3 c^3 = 7.2 \times 10^{42}$$

“终极计算机”体积为1L，表面积为$\sim 10^{-2}\mathrm{m}^2$，因此它1s内可以向外界输送$\sim 10^{40}$比特的信息量。同时，它每秒进行10^{50}次运算，那么“终极计算机”可以容忍的错误率上界为$\sim 10^{40}/10^{50} = 10^{-10}$。也就是说，如果错误率高于这个极限，并不是所有错误都能得到改正，因此我们的“终极计算机”变得不可靠。

最严重的问题还是输入自由能的问题。辐射的功率为4.04×10^{26}W，因此输入的自由能功率也是

[4] 事实上，很普遍地有$T = CE/S$。黑体辐射$C = 4/3$，理想气体$C = 3/2$，黑洞$C = 1/2$

[5] r_l是粒子/反粒子数(光子1，电子/正电子2)乘上极化方向(光子2，电子/正电子2)乘上统计因子(玻色子1，费米子7/8)

等量的4.04×10^{26}W。然而，1kg物质所具有的能量仅为$mc^2=10^{17}$J！换句话说，为了维持1kg“终极计算机”的正常运转，每秒需要消耗10^9kg物质对应的自由能！这在现在看来是不现实的。

3.5 和当今计算机的比较

“终极计算机”的存储量是10^{31}比特，比较而言，当今计算机存储量10^{10}比特就远远没达到这个极限。造“终极计算机”的困难还是多重的：第一，前面讨论能量提到过的，经典计算机采用高低点位进行表示逻辑0/1，这种表示有冗余自由度；第二，“终极计算机”所需要的技术非常之高，比如超高温等离子体的产生，对系统稳定性的控制，以及维持正常运转所需要输入的能量；第三，即使我们现在真造出了“终极计算机”，它也只是一个存储海量数据的“哑巴”；我们现在的技术允许我们一秒读写10^{12}比特的存储，读出“终极计算机”内部的10^{31}比特需要的10^{12}年！

4. 尺寸限制并行度

4.1 CPU和GPU的比较

说到并行的硬件，很多人的第一反应会是GPU(Graphics Processing Unit)，它的内部结构和CPU的比较可以从图3中看出来。CPU主要是一种串行架构，内部核心运算单元ALU功能强大，但是数目不多。相比之下，GPU主要是并行架构，有很多小而高效的ALU，它们可以同时处理不同的工作。可以想象，GPU内部的ALU越做越小，并行程度会越来越高；但如果小到量子效应不可忽略，即相邻的ALU之间会相互影响，并行程度就会饱和。为了考虑这个效应，我们引入了交流时间$t_{\rm com}$和翻转时间$t_{\rm flip}$以及用它们的比值来定义并行度P。

4.2 计算并行度的极限

对于尺度为R的计算机，从一端通过信号传递到另一端，至少需要时间$t_{\rm com}=\frac{2R}{c}$(这个下限对应信号以光速传播)。而翻转1比特的时间在前面已经计算出，$t_{\rm flip}=\frac{\pi\hbar S}{2\ln2k_BE}\sim\frac{kT}{\hbar}$，因此并行度$P$：

$$P=t_{\rm com}/t_{\rm flip}\sim k_BRT/\hbar c=2\pi R/\lambda_T$$

其中$\lambda_T=\frac{hc}{k_BT}$。一个简单的理解方式是，在$\lambda_T$附近内部的比特，由于量子效应，它们相互不可分辨，所以不能并行工作。在长度R中有R/λ_T个热波长，每个热波长内部的比特串行工作，不同热波长内的

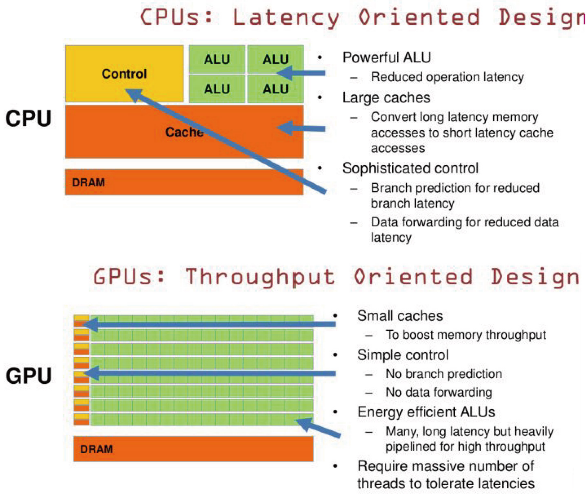

图 3. CPU和GPU的比较

比特相互不影响地并行工作，因此并行度可以理解为热波长的个数。

4.3 物质密度对并行度的影响

对于我们的“终极计算机”，利用$2R=10^{-1}$m，$2E/\pi\hbar\approx10^{51}$，$S/k_B\ln2\approx10^{31}$，计算出它的并行度为$\sim10^{10}$，即它是高度并行的。

如果保持质量不变，但压缩它的体积，直至把它压缩到史瓦西半径R_s，并行度会发生什么变化？它的史瓦西半径为$R_s=2Gm/c^2=1.5\times10^{-27}$m。黑洞的熵正比于其表面积$S=k_BA/4$，相应地信息量为$I=S/k_B\ln2=4\times10^{16}$比特。一比特翻转的时间和交流的时间[6]分别为$t_{\rm flip}=\pi\hbar I/2E=\pi^2R_s/c\ln2$和$t_{\rm com}=\pi R_s/c$，并行度$P=t_{\rm com}/t_{\rm flip}=\ln2/\pi<1$，因此史瓦西黑洞是高度串行的。

同等质量下，体积越小的计算机允许的并行度上限越低，这里还是量子效应在捣乱。

5. 搭建“终极计算机”

5.1 核磁共振(NMR)

在医学上，核磁共振已经具有了广泛的应用，但它的原理同样也适用于制造计算机。假设一个核子具有自旋μ，它有自旋向上和自旋向下两种本征态。如果把它放入磁场B中，两个本征态对应的本征能量分别是μB和$-\mu B$。一个核子的自旋方向记录了一个比特的信息，且它的翻转时间$t_{\rm flip}=$

[6]这里交流的时间取得是直径上的两端点，它们通过圆周运动的光线相互交流，而不是直接通过直径交流，所以这个因子是π而不是2。

$\pi\hbar/2\mu B$。可以发现，为了让计算机的计算速度变大，可以通过增加磁感应强度B。

5.2 重离子碰撞

考虑一个典型的碰撞（之所以称为典型，是因为CERN在大型强子对撞机中进行实验的参数和这里的例子接近），100个核子和100个核子对心碰撞，能量在200GeV，碰撞产生$\sim 10^4$个介子。一些重要的参数是：

- 翻转时间$t_{\text{flip}} \approx \pi\hbar/2E \approx 10^{-29}s$
- 熵$S \approx 4k_B \times 10^4$(每个介子携带有$4k_B$的熵)
- 信息量$I = S/k_B ln2 \approx 10^4 \sim 10^5$比特
- 100个核子的直径$D = 12 - 13\text{m}$
- 洛伦兹收缩因子$\gamma = 100$
- 碰撞时间$D/\gamma c \approx 10^{-25}\text{s}$
- 每秒进行的操作数量$f = 2E/\pi\hbar \approx 10^4\text{Hz}$

总结来说，对单次碰撞而言，1s内10^4个比特进行10^4次运算，看起来效率不高，而且对心碰撞比较难把控。但如果碰撞频率很高，性能也可能很可观。

5.3 库仑相互作用

考虑两个电子构成的计算机，它们相距的距离为r，则库伦相互作用势能为e^2/r，从而翻转需要的时间为$t_{\text{flip}} = \pi\hbar r/2e^2$。而交流的时间是$t_{\text{com}} = r/c$。令人有点意外的是，并行度是一个不随$r$变化的值，即$P = t_{\text{com}}/t_{\text{flip}} = 2\alpha/\pi$，其中$\alpha = e^2/\hbar c = 1/137$为精细结构常数。

两个电子利用库仑相互作用构建的计算机，是一个高度串行的架构，看起来似乎不是那么强大。

6. 结论

对于摩尔定律能否能一直继续这个问题，我们详细地讨论，发现物理学对摩尔定律构成极大的挑战——能量限制了计算机的运算速度，熵限制了计算机的存储空间，几何尺寸限制了并行度。重1kg，体积为1L的由光子气体组成的“终极计算机”，1秒内可以进行10^{50}次运算，并且有10^{31}比特的存储空间。虽然它的性能非常诱人，创造它和维持它运行是非常困难的事。文中提到的容易操纵的物理系统，性能上又远不如“终极计算机”。虽然如此，我们始终相信人类的智慧——一方面，在传统计算机架构下，技术会发展，计算机性能会逐渐逼近“终极计算机”；另一方面，我们可以不断探索新的理论、新的物理系统，甚至创造新的计算机架构，在全新的设置下探讨计算机性能的问题。

致谢

感谢刘玉鑫院长给我们提供的参考文献，以及量子讨论班的各位同学们的积极讨论。

参考文献

[1] Tommaso Toffoli Edward Fredkin. Conservative logic. *International Journal of Theoretical Physics*, 21, (1982).

[2] Y.Bohm Aharonov. D. time in the quantum theory and the uncertainty relation for the time and energy domain. *Phys.Rev*, pages 1649–1658.

[3] 曾谨言. 量子力学（卷一）. (2007).

[4] Seth Lloyd. Ultimate physical limits to computation. *Nature*, 406:1047–1054, (2000).

[5] E.T. Jaynes. Gibbs vs boltzmann entropies. *American Jornal of Physics*, 33:391–398, (1965).

[6] H.J. Bermermann. Minimum energy requirements to information transfer and computing. *Int.J.Theor.Phys.*, 21:203–217, (1982).

[7] J.D. Bekenstein. Universal upper bound on the entropy-to-energy ration for bounded systems. *Phys.Rev.D*, 23:287–298, (1981).

[8] J.D. Bekenstein. Energy cost of information transfer. *Phys.Rev.Lett.*, 46:623–626, (1981).

[9] R. Laudauer. Irreversibility and heat generation in the computing process. *IBM J.Res.Dev*, 5:183–191, (1961).