摩尔定律及计算前沿

周尚博

2024,12,25

Outline

- 摩尔定律及其影响
- 高性能计算
- 量子计算
- 碳纳米管计算机

摩尔定律

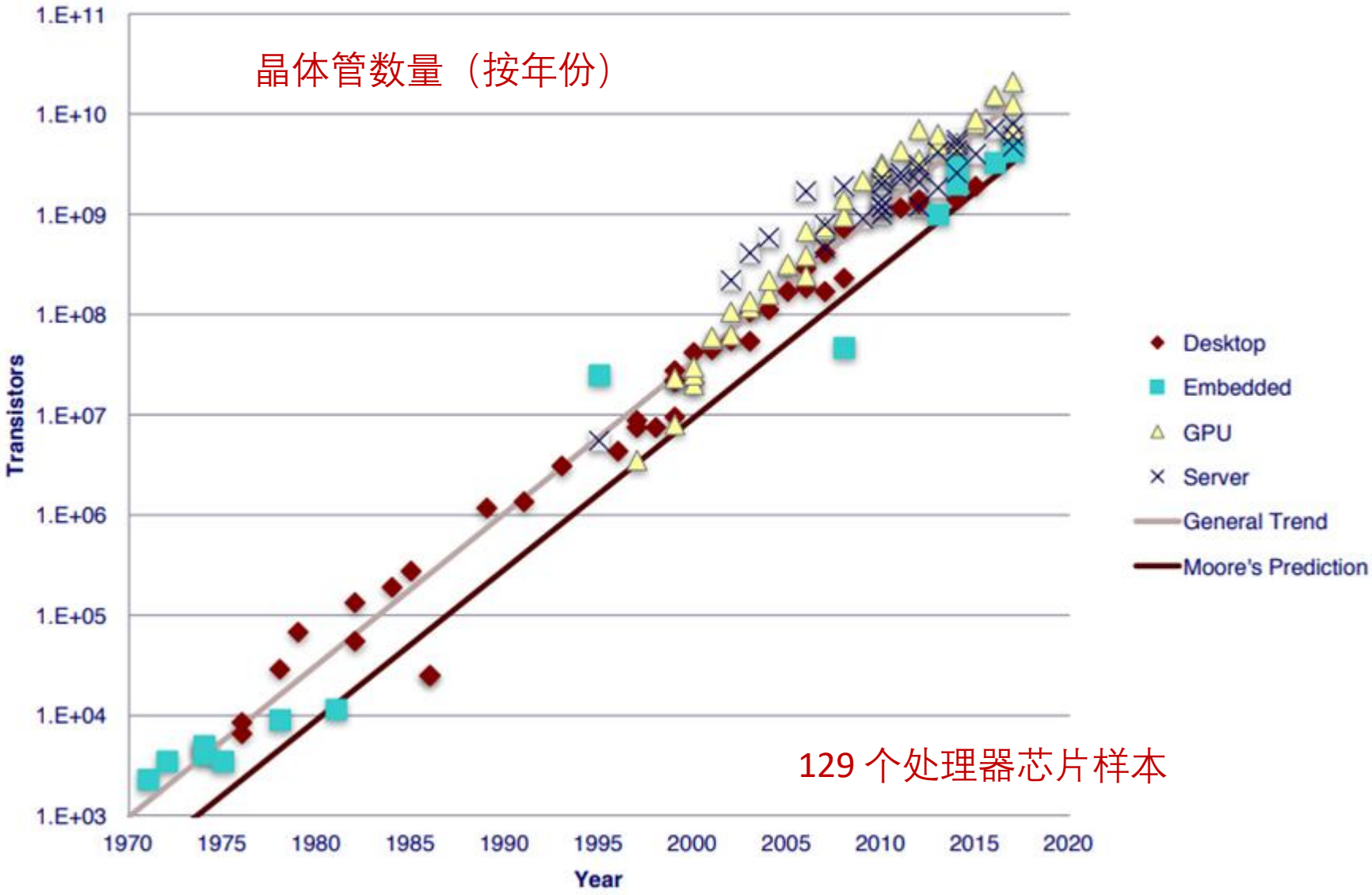
摩尔定律

- **晶体管数量翻倍**：摩尔定律主要描述的是集成电路中晶体管数量的增长。每两年，集成电路上晶体管的数量大约会翻一番。

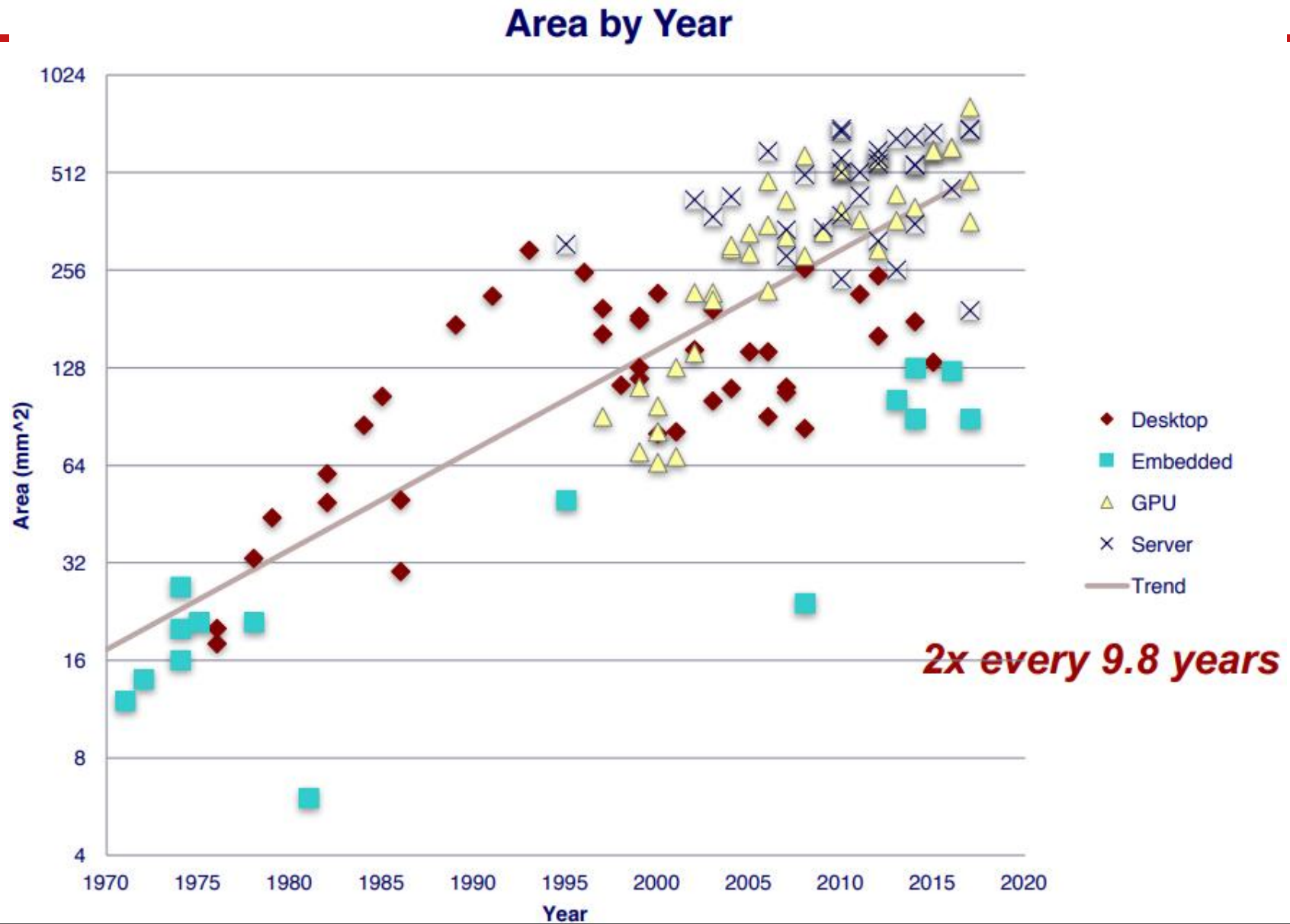
- **性能提升**：晶体管数量的增加意味着计算机的计算能力和处理速度也会随之提升，因为更多的晶体管可以并行处理更多的信息。

- **成本下降**：随着技术进步，制造成本逐渐降低。每单位晶体管的成本变得更低，这使得计算机设备的价格得以降低，计算机普及率不断提高。

摩尔定律


晶体管数量（按年份）
1.E+11
1.E+10
1.E+09
1.E+08
1.E+07
1.E+06
1.E+05
1.E+04
1.E+03
Transistors
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year
Desktop
Embedded
GPU
Server
General Trend
Moore's Prediction
129 个处理器芯片样本

摩尔定律


Area by Year
1024
512
256
128
64
32
16
8
4
Area (mm^2)
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year
Desktop
Embedded
GPU
Server
Trend
2x every 9.8 years

摩尔定律的影响

- **计算机性能的飞速提升**

处理能力：随着晶体管数量的增加，计算机的计算能力显著提高。多核处理器和并行计算的引入使得计算任务可以在多个核心之间分配，进一步提升了处理速度。

小型化与便捷性：晶体管的微型化使得计算机、智能手机、笔记本电脑等设备越来越小巧，甚至嵌入到日常生活中的各类电子产品（如家电、汽车、穿戴设备等）中。

- **成本降低与普及**

更低的成本：由于晶体管密度不断增加，每个芯片上可以集成更多的功能，这使得单个芯片的制造成本显著降低。

消费电子普及：降低的成本使得计算机和信息技术设备逐渐普及，促进了互联网、云计算、智能手机等技术的发展，改变了人们的生活和工作方式。

- **技术创新推动**

新技术的出现：摩尔定律的推动力之一是不断创新的半导体制造技术，如更小的光刻技术、3D晶体管和纳米技术等。每次技术的突破都为集成电路的微型化和性能提升提供了新的可能性。

芯片设计的进步：随着晶体管数量的增加，芯片设计变得更加复杂。为了充分利用晶体管数量的提升，设计者不断创新，推动了芯片架构的优化，提升了计算效率。

- **经济和社会影响**

科技产业发展：摩尔定律推动了信息技术产业的快速发展，特别是在计算机硬件、移动通信、网络和云计算等领域。

全球化和互联网化：随着技术成本的降低，计算机和网络技术成为各行各业的重要工具，推动了全球化进程和互联网经济的崛起。

摩尔定律的局限

- **物理极限：**随着晶体管尺寸不断缩小，逐渐接近原子级别，摩尔定律面临着物理上的极限

量子效应：在纳米尺度下，量子效应（如隧穿效应）开始显著影响电子流动。传统的硅晶体管依赖于电子的经典行为，当晶体管的尺寸变得越来越小，电子穿越极薄的隔离层变得更加困难，导致晶体管不再完全“开”和“关”，造成漏电流，影响芯片的性能。

热效应：当晶体管尺寸变小，单个晶体管的功率密度增加，导致热量产生。小尺寸下的晶体管更容易产生过热问题，传统的冷却方法也变得不再有效，这对芯片的工作稳定性和性能是一个严重挑战。

- **制造成本的上升**

制造技术的复杂性：随着晶体管尺寸不断减小，制造过程需要更精细的技术。例如，光刻技术必须能够精确地将光线聚焦到极小的尺寸。这对制造设备提出了更高的要求，导致生产成本大幅增加。

摩尔定律的“翻倍”速度放缓：过去，半导体制造厂商能够以较低成本每两年生产更多晶体管。然而，随着工艺逐渐接近物理极限，继续大规模增加晶体管数量的成本显著上升，且进展逐渐放缓，很多工厂需要投入数十亿美元进行设备升级。

- **能效和功耗**

功耗问题：晶体管数量的增加，意味着计算机的电路变得更加复杂，功耗和散热问题日益突出。尽管可以通过更小的晶体管来提高密度，但功耗和散热问题并未得到解决，尤其是在高性能计算和数据中心环境中，散热成为一个严重问题。

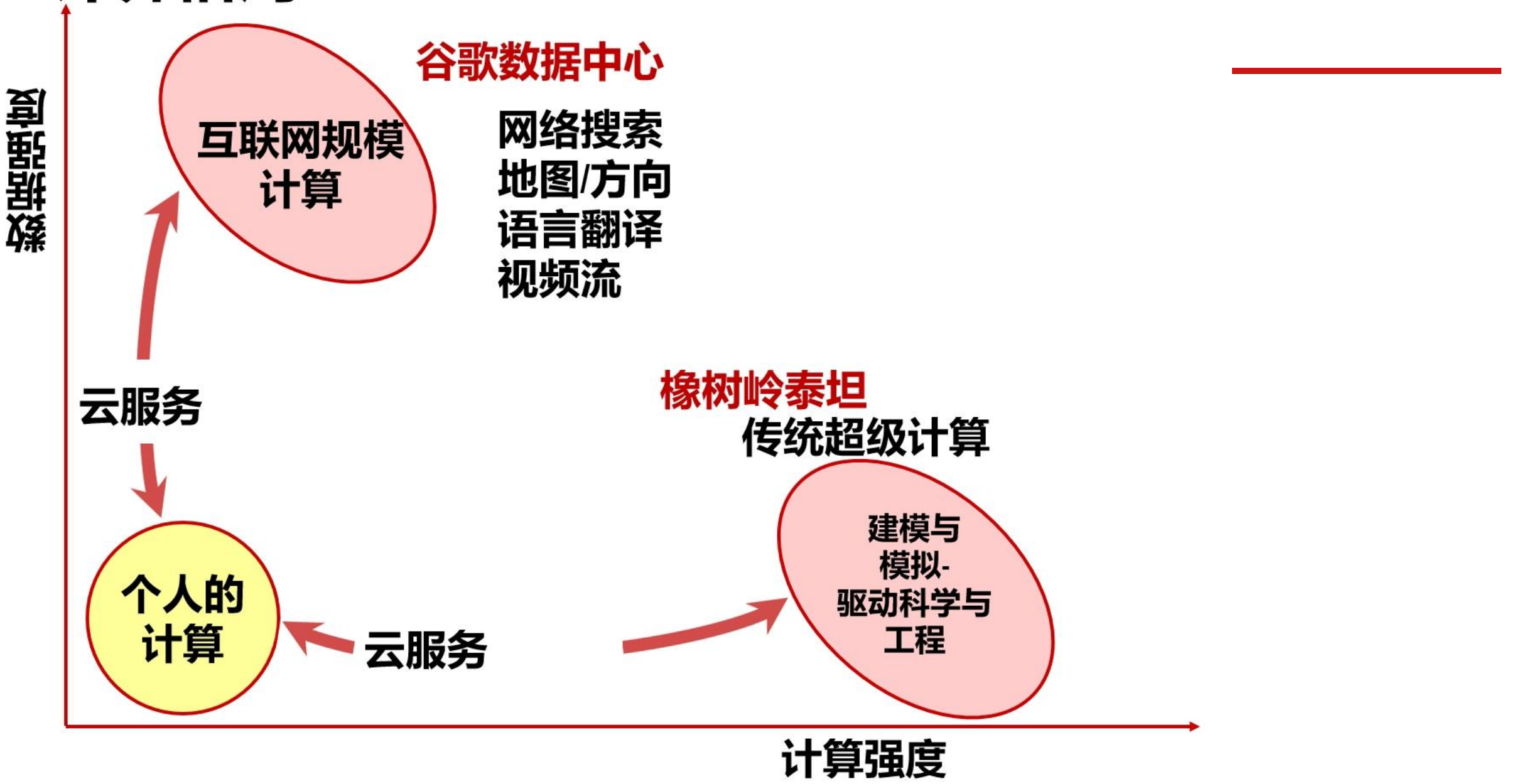
功率密度：随着晶体管尺寸的减小，单个芯片上的功率密度不断增加，传统的散热技术难以应对这些挑战。这导致芯片在高负载下容易过热，从而影响稳定性和性能。

- **技术创新的瓶颈**

摩尔定律假设了通过不断的小型化和缩小尺寸来实现性能的提升，但当晶体管变得越来越小，新的创新空间变得越来越窄。尽管有新的制造技术和材料不断出现，但突破原有的半导体技术已变得更加困难。例如，使用三维晶体管和鳍式场效应晶体管等技术虽然提升了晶体管的性能，但面临的技术挑战仍然很多，且进展不如以往那样迅速。

高性能计算

计算格局


数据强度
互联网规模
计算
谷歌数据中心
网络搜索
地图/方向
语言翻译
视频流
云服务
橡树岭泰坦
传统超级计算
建模与
模拟-
驱动科学与
工程
个人的
计算
云服务
计算强度

Titan超级计算机

- **Titan的计算架构**

Titan超级计算机的主要计算单元由两种不同类型的处理器组成：

CPU（中央处理单元）：Titan使用了基于AMD的“Opteron 6274”处理器。每个处理器有16个核心，并且每个核心支持超线程技术，使得每个CPU能够同时处理多个线程。

GPU（图形处理单元）：Titan采用了NVIDIA的Tesla K20x GPU，这些GPU使用了NVIDIA的Kepler架构。每个GPU包含许多核心（上千个小核心），可以并行处理大量数据，特别适合进行浮点运算和大规模并行计算。

这两种处理器的组合形成了一个**异构计算系统**，即同时使用CPU和GPU进行计算，以充分利用GPU在并行计算中的优势，同时利用CPU在串行计算和复杂控制逻辑中的优势。

- **并行计算原理**

任务级并行：将大任务分解为多个较小的任务，分配给不同的计算单元（CPU或GPU）进行处理。这种方法能够显著提高计算效率，尤其适用于大规模模拟和复杂计算问题。

数据级并行：每个计算单元处理数据的不同部分。GPU尤其擅长处理大量数据的并行计算，例如矩阵运算、数据密集型计算等。

- **Titan计算机的性能**

Titan的计算能力非常强大。其配置为：

计算节点：Titan的计算节点是由16核的AMD Opteron CPU和两个NVIDIA Tesla K20x GPU组成的。每个节点能够提供极高的并行处理能力。

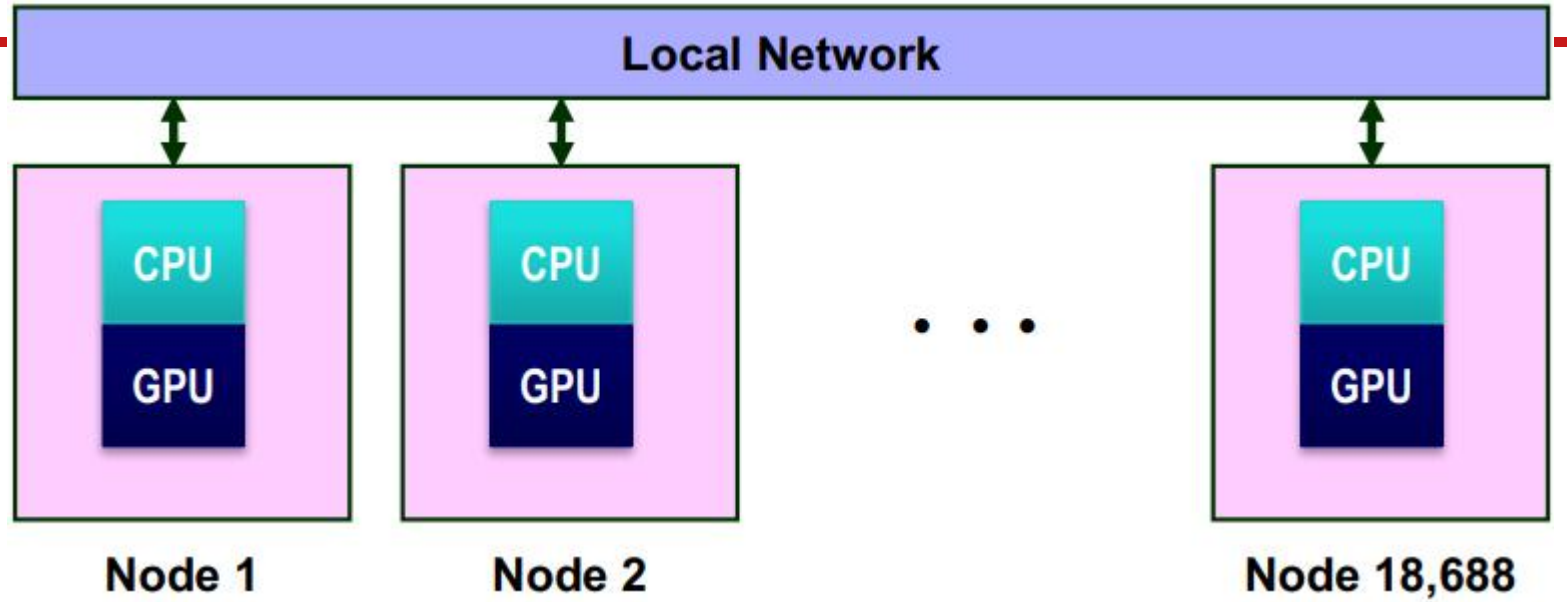
峰值性能：Titan的理论峰值性能为 27 petaflops（每秒27千万亿次浮点运算），使其在当时成为世界上最快的超级计算机之一。

Titan的性能不仅得益于多核CPU和大量GPU的结合，还得益于其高效的网络架构。Titan使用了**Cray XE6系统和Cray Gemini interconnects**，可以快速地将大量节点连接起来，确保在进行大规模计算时不会受到通信延迟的限制。

- **内存系统**

Titan配备了一个高效的内存系统，以确保各个计算节点之间的数据传输速度不会成为瓶颈。每个计算节点有两级内存：CPU内存：每个CPU节点有64GB的内存。GPU内存：每个Tesla K20x GPU有6GB的内存，专门用来存储需要进行GPU计算的数据。这些内存结合高速互联网络（如Cray Gemini）使得数据可以在不同节点之间高效流动，从而保持整个系统的高吞吐能力。

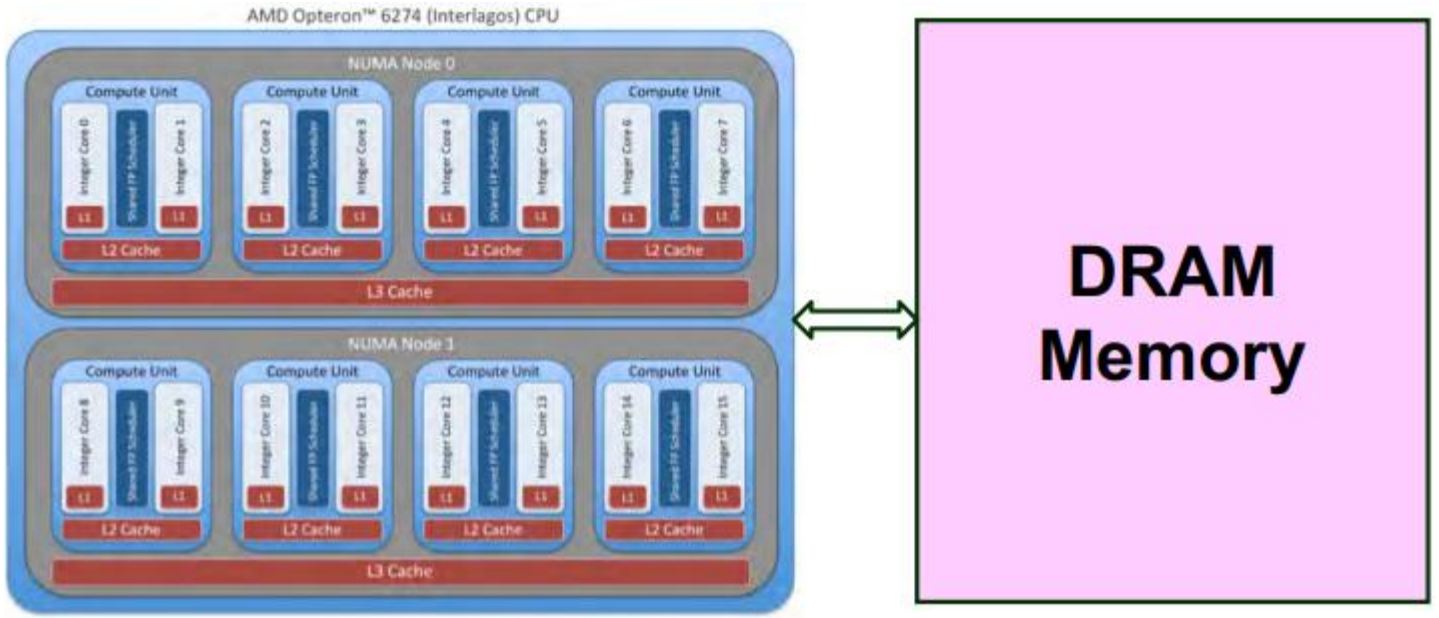
泰坦硬件



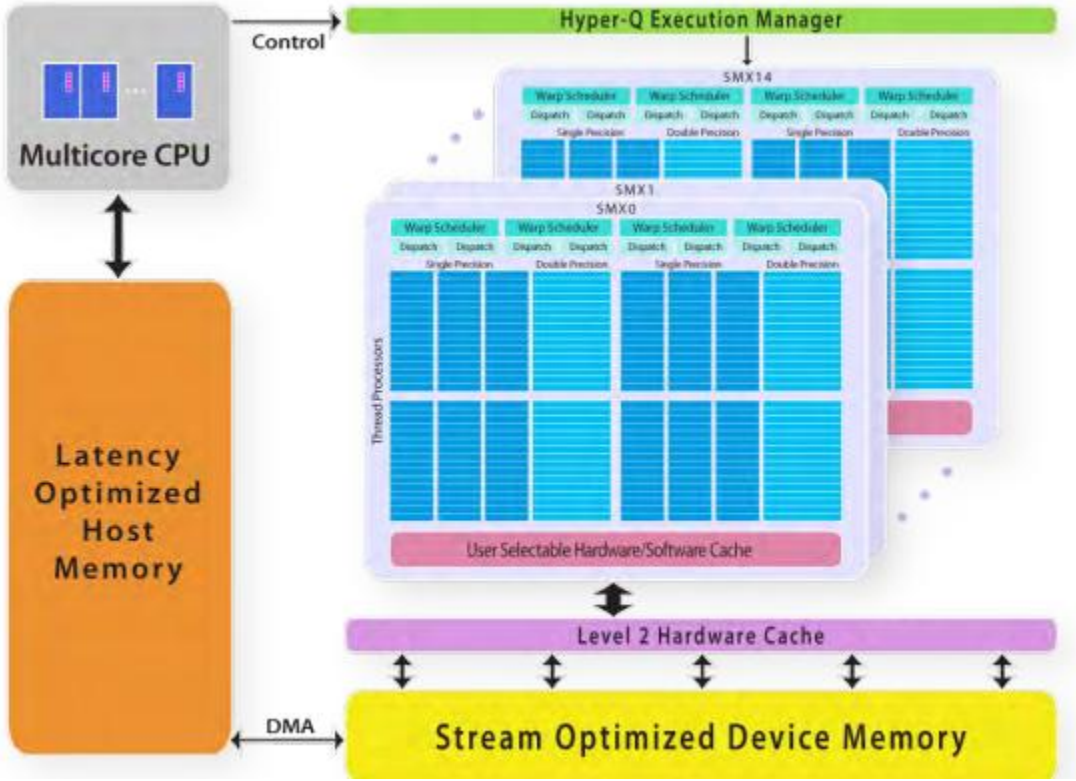
- **每个节点**
  - AMD 16 核处理器
  - nVidia图形处理单元
  - 38 GB DRAM
  - 没有磁盘驱动器
- **全面的**
  - 7兆瓦，2亿美元

Titan 节点结构：CPU

- **中央处理器**
  - 16 个核心共享内存
  - 支持多线程编程
  - 约为每秒 0.16 x 10^{12} 次浮点运算 (FLOPS)



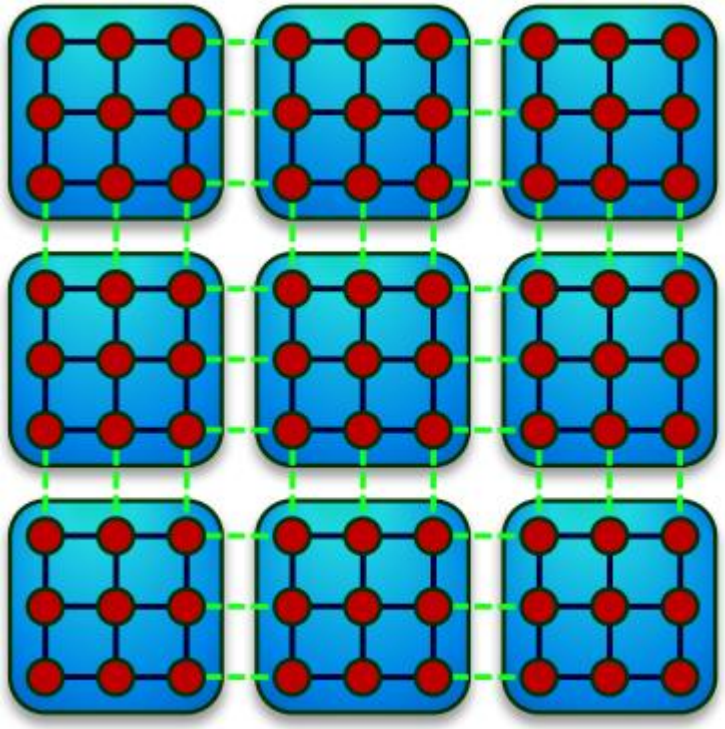
Titan 节点结构：GPU



- **开普勒 GPU**
  - 14 个多处理器
  - 每组12个16个流处理器
    - 14×12×16=2688
  - 单指令、多数据并行
    - 单条指令控制组内所有处理器
  - 4.0 x 1012 FLOPS 峰值性能

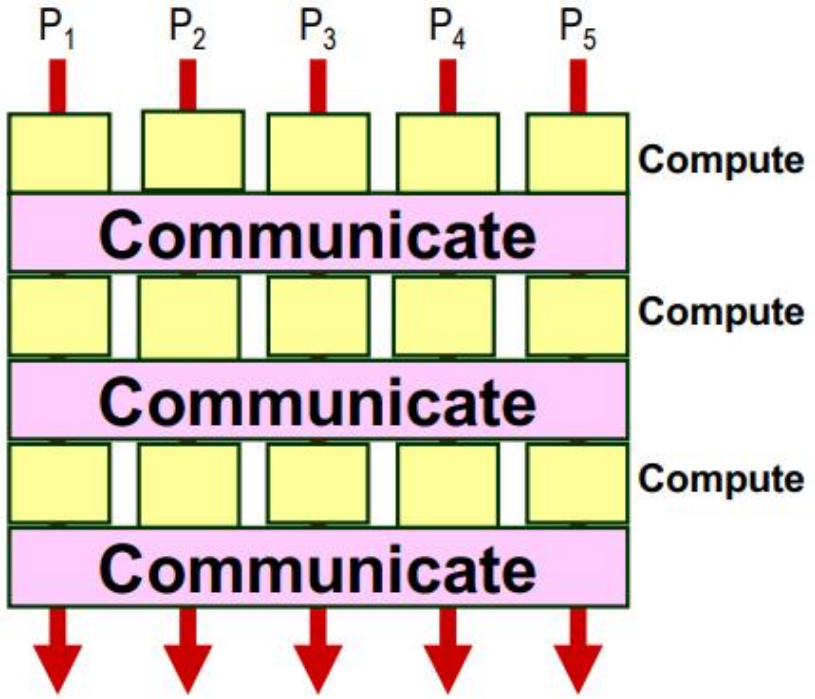
泰坦编程：原理

- **通过网格解决问题**
  - 例如有限元系统
  - 模拟随时间推移的运行
- **批量同步模型**
  - 分区
    - p 节点机器的 p 区域
  - 每个处理器的映射区域



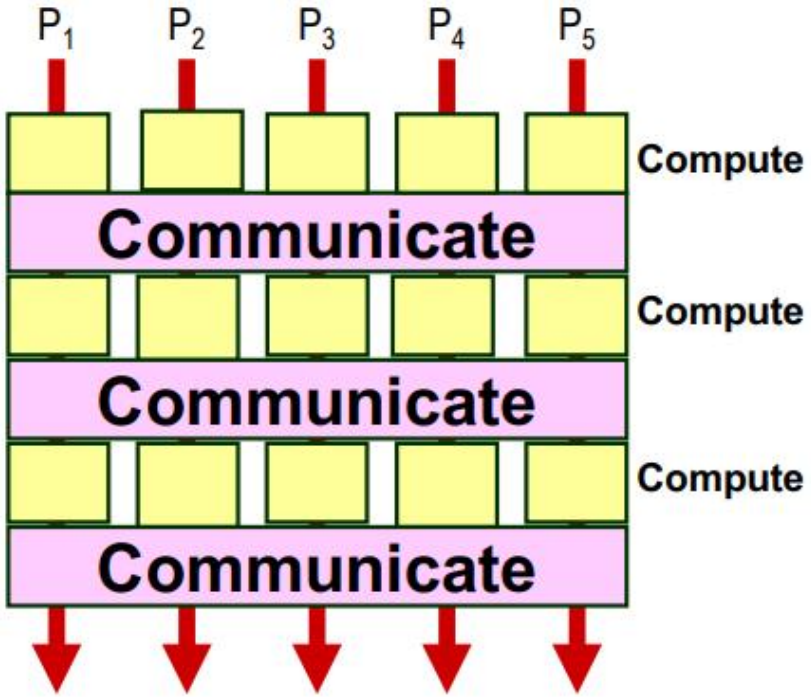
Titan 编程：原理（续）

- **批量同步模型**
  - 每个处理器的映射区域
  - 备用
    - 所有节点计算区域行为
      - 在 GPU 上执行
    - 所有节点在边界处传达价值

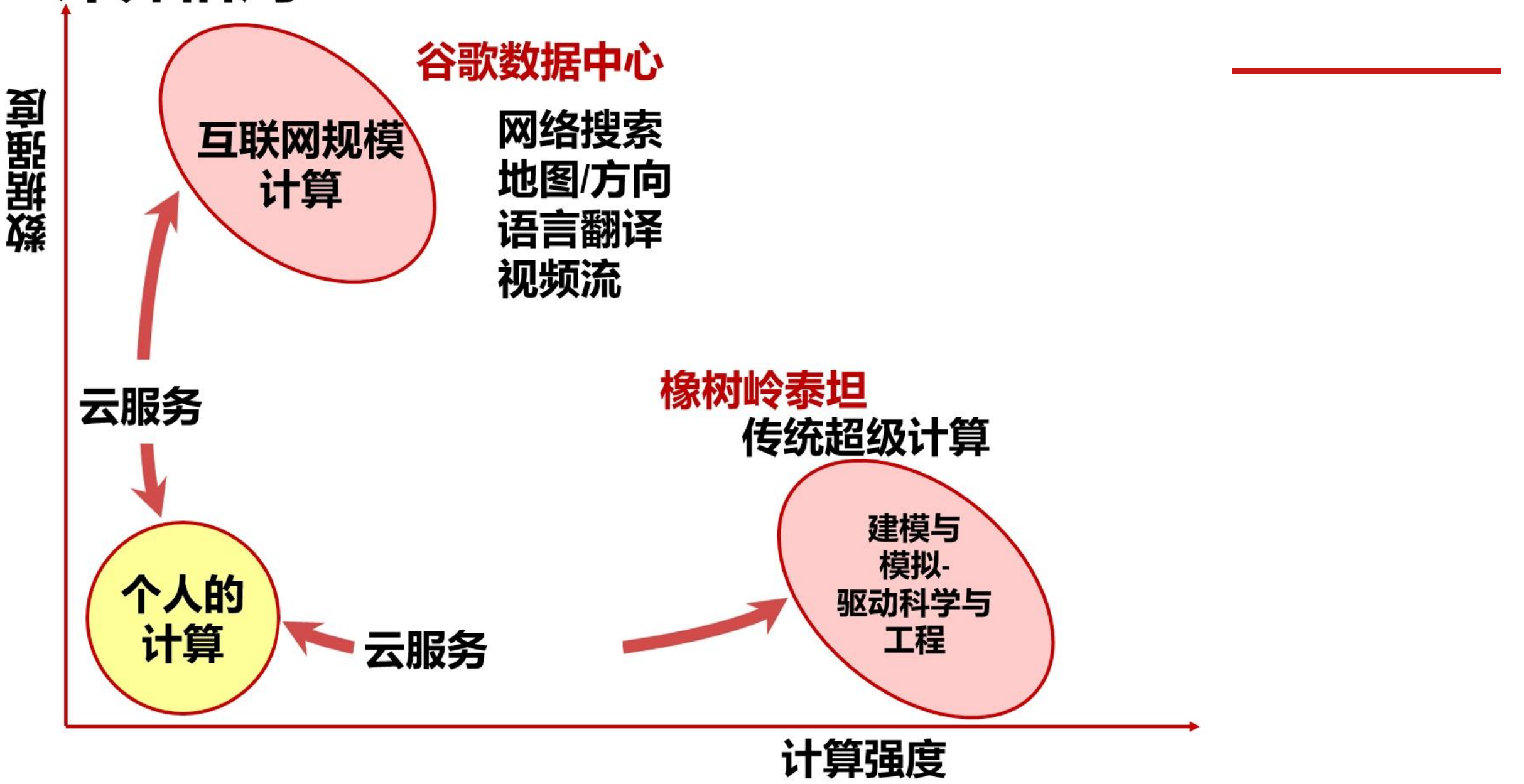


批量同步性能

- 受最慢处理器性能限制
- **努力保持完美平衡**
  - 设计高度可靠的硬件
  - 调整软件以使其尽可能规律
  - 消除“噪音”
    - 操作系统事件
    - 无关的网络活动



计算格局


数据强度
互联网规模
计算
谷歌数据中心
网络搜索
地图/方向
语言翻译
视频流
云服务
橡树岭泰坦
传统超级计算
建模与
模拟-
驱动科学与
工程
个人的
计算
云服务
计算强度

谷歌数据中心

- **谷歌数据中心**是由谷歌公司管理和运营的大型服务器场，提供**云计算资源、数据存储以及全球互联网服务支持。**它们是Google Cloud Platform（GCP）、Google Search等产品的物理基础设施。这些数据中心分布在全球不同的地点，提供高效、可靠、可扩展的数据存储和计算资源。

- **数据中心的架构**

服务器硬件：谷歌自定义设计服务器，能够最大化利用硬件资源，降低成本，提升效率。这些服务器通常包括自有设计的硬盘和处理器，以优化数据存储和处理能力。

电力供应：为确保数据中心的持续运行，谷歌数据中心配备了多重电源冗余，包括本地电网和备用发电机。谷歌还利用可再生能源（如风能、太阳能）来运行部分数据中心，推动绿色数据中心的发展。

冷却系统：数据中心的冷却设计非常重要，谷歌在冷却系统上投资了大量创新技术。通过使用外部空气（风冷）或液冷技术，谷歌最大化能源效率，同时降低运营成本。

高可用性：为了保证服务的持续稳定，谷歌的数据中心具有冗余的硬件和网络设施。所有重要设备都具有备份设施，一旦发生故障，流量会自动转移到其他节点。

- **全球分布**

谷歌在全球多个地区都建立了数据中心，以确保其服务的全球可用性和低延迟。谷歌的数据中心位于美国、欧洲、亚洲等地，每个数据中心都能够独立运行，支持全球范围的流量负载。例如，美国：谷歌在美国有多个数据中心，主要位于乔治亚州、南卡罗来纳州、爱荷华州等地。欧洲：在比利时、芬兰和荷兰等地建立了数据中心。亚洲：谷歌也在新加坡、台湾等地设有数据中心，以支持亚太地区的需求。

- **安全性与隐私**

谷歌的数据中心采取了极为严格的安全措施，包括：物理安全：访问控制严格，只有授权人员可以进入。数据中心的设施包括24小时监控、障碍物、警报系统等。网络安全：谷歌使用先进的加密技术，保护数据在传输过程中的安全。同时，谷歌对外部访问进行防火墙隔离，并持续进行漏洞检测。数据隐私保护：谷歌遵守全球范围内的数据隐私法规，确保用户数据的安全性和合规性。

谷歌数据集群

- **谷歌数据集群**是由多个服务器组成的计算资源池，它们协同工作以完成各种计算任务。谷歌的数据集群构成了谷歌云计算平台的核心基础设施，支持谷歌的各种服务，包括搜索引擎、YouTube、Google Cloud和其他在线服务。
- 数据集群的架构谷歌的数据集群是一种分布式计算架构，具有以下特点：

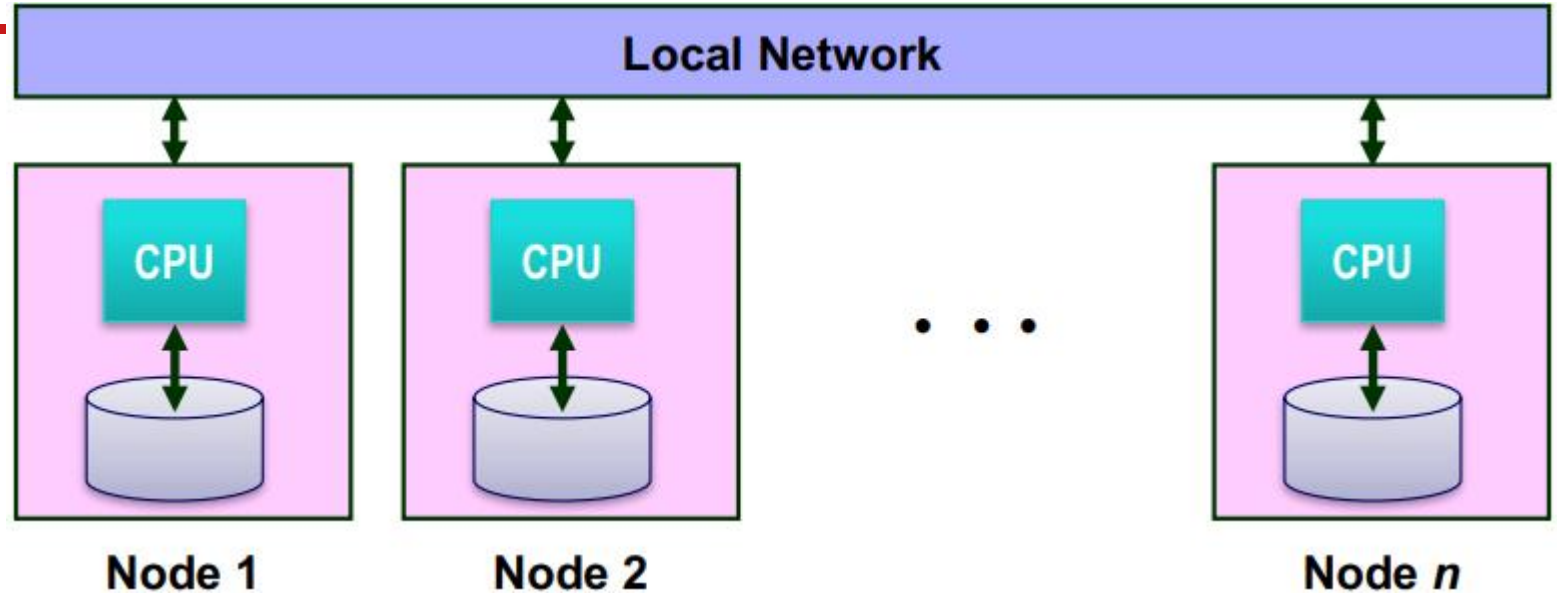
分布式计算：每个数据集群由多个计算节点（服务器）组成，节点之间通过高速网络连接，协同处理数据请求。分布式架构使得集群能够提供弹性扩展、负载均衡和高可用性。

大规模并行处理：每个数据集群都具备强大的并行计算能力，适合处理大规模数据分析、机器学习、人工智能等计算密集型任务。谷歌通过将任务分解成多个子任务，并在不同的计算节点上并行执行，提高计算效率。

存储和计算结合：数据集群内包含强大的存储资源，支持高速数据读取和写入。谷歌使用自己的分布式存储系统（如Google File System（GFS）和Colossus）来管理数据存储和检索。

- **数据集群与云计算谷歌的数据集群与其云计算服务紧密结合。**谷歌云计算平台（Google Cloud Platform）利用这些数据集群为全球的企业和开发者提供计算、存储、数据库等多种服务。
- 分布式计算框架谷歌的云计算平台和数据集群运行着多个**分布式计算框架**，旨在提高计算和数据处理效率，如MapReduce、Google File System (GFS)、Spanner。
- **负载均衡与自动扩展谷歌数据集群采用了负载均衡技术，能够根据实时流量和计算负载，动态分配计算资源。**
- **容错性与高可用性谷歌的数据集群具备容错能力。**如果某个计算节点出现故障，系统能够自动将任务迁移到其他节点，确保服务不中断。数据在多个节点和存储设备上进行冗余备份，以确保数据的持久性和安全性。

Google 集群

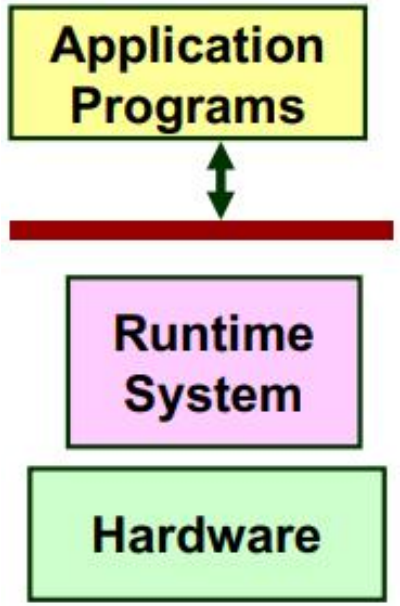


- **通常为 1,000–2,000 个节**
- **节点包含**
  - 2 个多核 CPU
  - 2 个磁盘驱动器
  - 动态随机存取记忆体

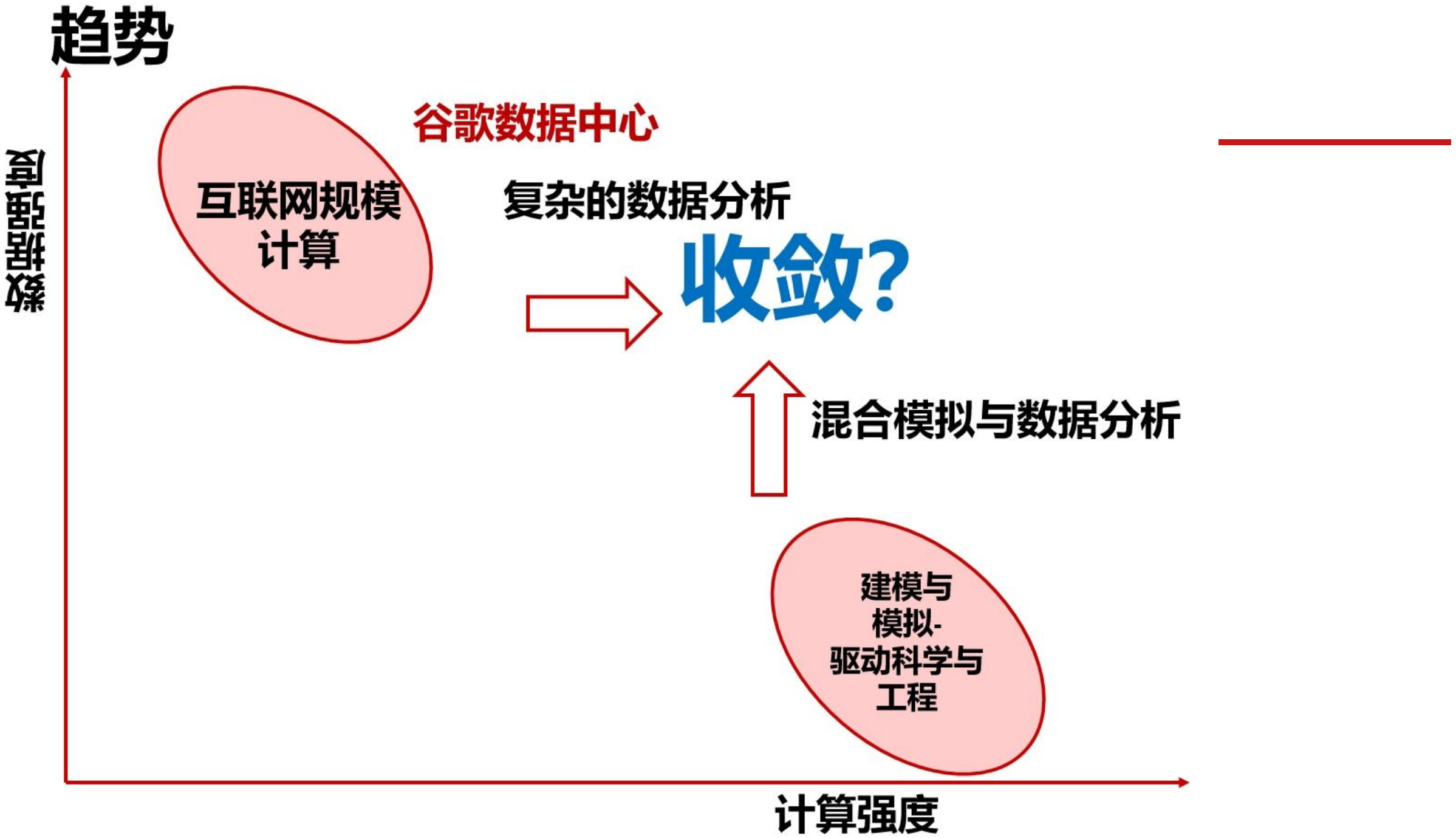
集群编程模型

- 以高级数据操作编写的应用程序
- 运行时系统控制调度、负载平衡……

独立于机器的编程模型



- **扩展挑战**
  - 集中式调度器形成瓶颈
  - 复制到磁盘/从磁盘复制成本非常高
  - 难以限制数据移动
    - 重要的性能因素


趋势
数据强度
谷歌数据中心
互联网规模
计算
复杂的数据分析
收敛?
混合模拟与数据分析
建模与
模拟-
驱动科学与
工程
计算强度

计算/数据融合

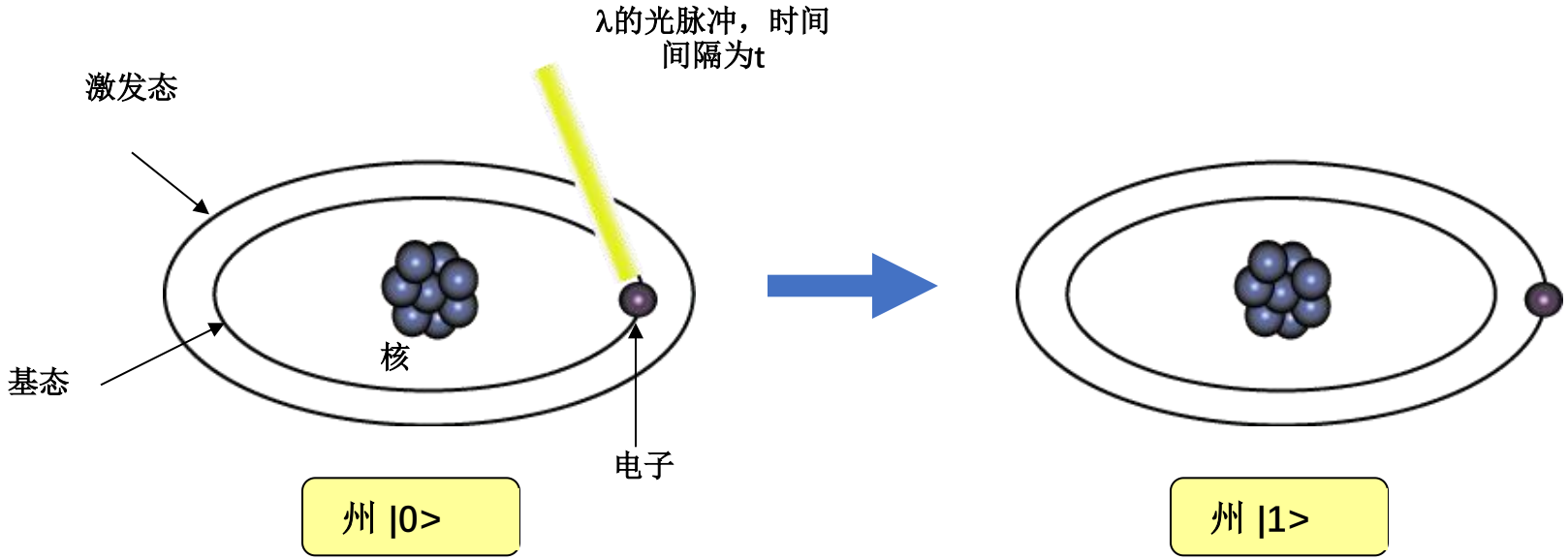
- **两类重要的大规模计算**
  - 计算密集型超级计算
  - 数据密集型处理
  - 互联网公司+众多其他应用
- **遵循不同的进化路径**
  - 超级计算机：充分利用现有硬件的性能
  - 数据中心集群：最大化各种数据中心任务的成本/性能
  - 产生了不同的硬件、运行时系统和应用程序编程方法
- **融合将带来重要益处**
  - 计算和数据密集型应用
  - 但不清楚该怎么做

量子计算

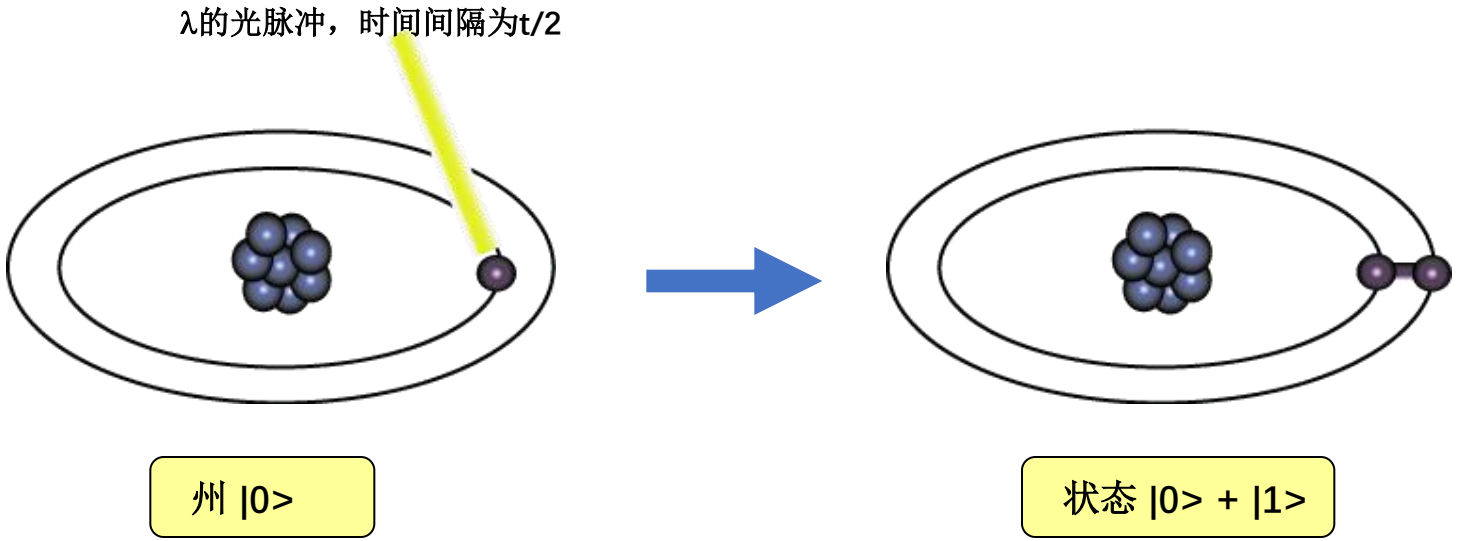
数据表示——量子比特

一个数据位由一个原子表示，该原子处于|0>和|1>表示的两个状态之一。这种形式的单个比特称为*量子比特*

量子比特的物理实现可以使用原子的两个能级。激发态代表 |1>，基态代表 |0>。



数据表示 - 叠加



考虑一个 3 比特量子位寄存器。所有可能状态的等权重叠加将表示为：

$$|\psi\rangle = |000\rangle + |001\rangle + \ldots + |111\rangle$$

$$\frac{1}{\sqrt{8}} \qquad \frac{1}{\sqrt{8}} \qquad \frac{1}{\sqrt{8}}$$

量子计算原理

- **量子叠加（Superposition）**量子叠加是量子计算的核心之一，它指的是量子比特（qubit）可以同时处于多个状态，而不仅仅是传统计算机中的“0”或“1”两种状态。通过叠加，一个量子比特可以同时代表“0”和“1”，这使得量子计算机在并行计算方面具有巨大的潜力。

  经典比特：传统计算机中的信息是以“0”和“1”表示的。

  量子比特：量子比特（qubit）可以处于“0”和“1”的叠加态，即它可以同时表示“0”和“1”，直到测量时才会坍缩为一个确定的状态。

  例如，一个量子比特可以表示：$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ 其中，α α 和 β β 是复数概率振幅，满足 $|\alpha|^2+|\beta|^2=1$ $|\alpha|^2 + |\beta|^2 = 1$。

- **量子纠缠（Entanglement）**量子纠缠是一种量子力学现象，当两个或更多的量子比特相互纠缠时，它们的状态变得不可分离，哪怕它们相隔很远。改变其中一个量子比特的状态，另一个量子比特的状态也会立即发生变化，这种现象超越了经典物理的直觉。量子纠缠为量子计算提供了强大的并行处理能力。在某些计算中，纠缠态可以实现比经典计算更快的计算速度。
- **量子干涉（Interference）**量子干涉是指量子系统中的波函数在某些路径上相互增强或相互抵消，从而影响计算结果。在量子算法中，干涉用来增强正确答案的概率并减少错误答案的概率。通过精确控制量子叠加态的干涉，量子计算机能够快速筛选出正确的解。
- **量子测量（Measurement）**量子测量是量子计算的一个重要步骤。当我们对一个量子比特进行测量时，它的状态会“坍缩”到某一个确定的值（“0”或“1”）。这种测量具有概率性，即每次测量的结果是随机的，但通过多次测量，可以获得概率分布。

量子计算技术

量子比特的实现方式

- 量子比特可以通过多种物理系统来实现，常见的实现方式包括：
- 超导量子比特：使用超导电路在极低温下生成量子比特，是目前最为主流的量子比特实现方式之一（如IBM、Google的量子计算机）。
- 离子阱量子比特：通过使用激光控制带电粒子（如离子）来创建量子比特，这种方法具有较高的精度和稳定性（如量子计算公司IonQ）。
- 光量子比特：利用光子的量子态（如偏振态、相位态）来表示量子比特，具有传输距离远等优势。
- 拓扑量子比特：一种理论上可以具有更高错误容忍度的量子比特，正在探索中。

量子门和量子电路

- 量子计算机通过量子门（Quantum Gates）来操作量子比特，进行逻辑运算。常见的量子门包括：
- Hadamard门（H门）：实现量子比特从基态到叠加态的转换。
- Pauli-X、Pauli-Y、Pauli-Z门：分别对应经典的NOT、Y和Z运算。
- CNOT门：一种重要的量子控制门，通常用于生成量子纠缠。
- Toffoli门：三比特量子门，是量子计算中实现某些计算任务的基础。

量子计算的应用

- **优化问题**：量子计算能够帮助解决一些传统计算无法高效解决的优化问题。比如，旅行商问题、最短路径问题、资源调度等。
- **密码学**：量子计算对于传统加密算法（如RSA）构成威胁。量子计算机能够通过量子算法（如Shor算法）快速破解大数分解，从而打破当前基于因式分解的加密系统的安全性。为此，研究人员正在开发量子安全加密算法（量子加密和量子密钥分发等）。
- **模拟物理和化学系统**：量子计算非常适合模拟量子力学系统，因为量子计算机本身基于量子力学原理。比如，量子计算可以模拟分子和材料的行为，为药物设计、新材料研发等提供重要支持。
- **机器学习与人工智能**：量子计算可以加速机器学习算法，尤其是大数据集的训练和处理过程。量子机器学习（Quantum Machine Learning）正在成为量子计算的重要研究方向。
- **量子化学和药物发现**：量子计算能够模拟分子和化学反应过程，这对新药物的发现和复杂分子合成有着巨大的潜力。通过精确的分子模拟，量子计算能够揭示传统计算机无法触及的分子结构和反应路径。

量子计算的优势与挑战

优势

- **并行性**：经典计算机是线性执行任务，而量子计算机通过量子叠加能够在同一时刻并行处理大量数据。这使得量子计算机在解决某些问题时，比经典计算机快得多。
- **信息存储**：经典计算机使用比特存储信息，而量子计算机使用量子比特存储信息，量子比特可以同时存储“0”和“1”的叠加状态。
- **计算速度**：对于某些问题，量子计算机能够显著提高计算速度。例如，量子计算在某些优化、因式分解和模拟等任务上，可以比经典计算机更高效。

挑战

- **量子退相干**：量子比特非常容易受到环境干扰，导致量子态“坍缩”丧失信息，这对量子计算的稳定性和可靠性构成挑战。
- **量子错误纠正**：量子计算需要有效的错误纠正机制，来应对量子比特容易受到噪声的影响。
- **量子硬件的可扩展性**：目前的量子计算机的规模相对较小，如何扩展到上千、上万量子比特并保持稳定性是一个重要的技术难题。

碳纳米管计算机

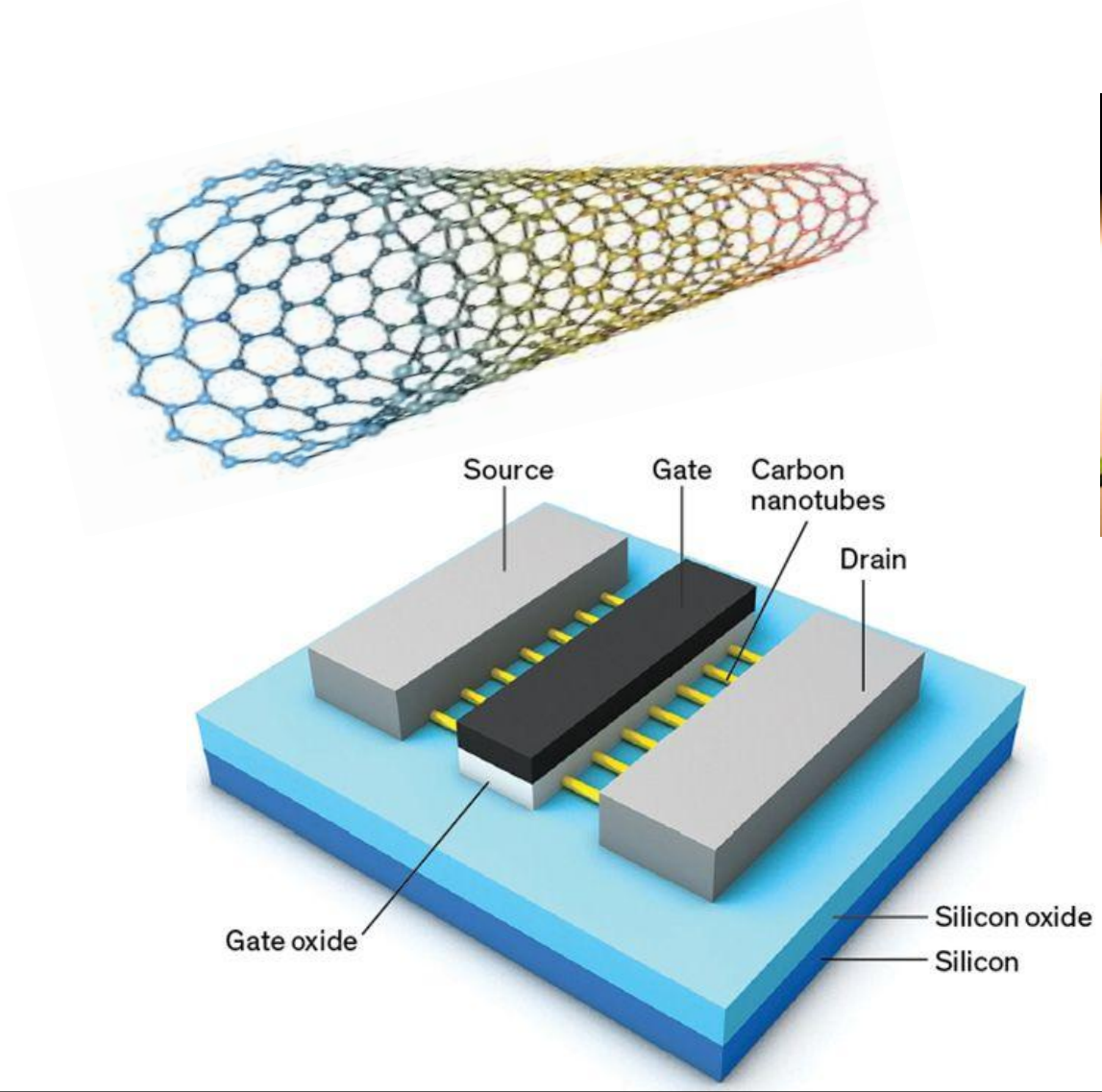
碳纳米管（CNT）介绍

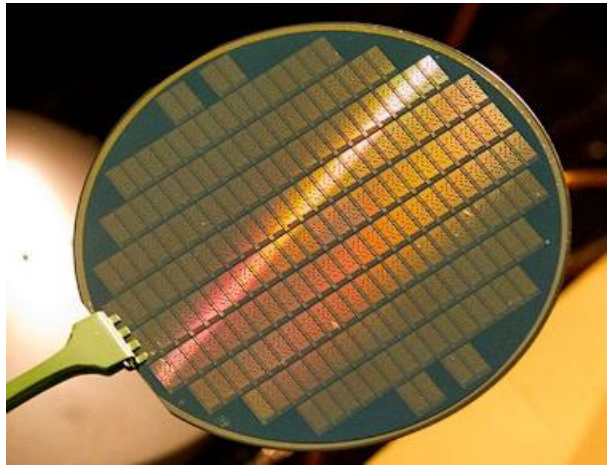
- **碳纳米管（Carbon Nanotube, CNT）计算机**是利用碳纳米管作为核心材料来构建计算机硬件的一种新型计算机架构。碳纳米管是一种由单层或多层碳原子构成的管状分子，具有优异的电学、机械和热学性质，这使得它们成为下一代计算机芯片和电子器件的理想候选材料。
- 碳纳米管是一种由单层石墨烯卷曲成管状的结构，具有**极高的强度和优异的导电性**，接近铜的导电性，远超硅和其他传统材料。碳纳米管能够承载非常高的电流密度，这使得它们在微型化的电子器件中具有巨大的应用潜力。
- 此外，碳纳米管的**热导性**非常强，是已知材料中热导率最高的之一。这意味着碳纳米管能够高效地散热，避免电子器件在高功率运行时过热，解决了硅基半导体在高频率运行下的散热问题。
- 碳纳米管可以根据其卷曲方式（单壁碳纳米管、双壁碳纳米管、多壁碳纳米管等）和直径的不同，呈现不同的带隙特性。某些碳纳米管是金属性的（无带隙），而某些是半导体性的（具有带隙）。这种可调的带隙特性使得碳纳米管在**制造高效开关元件**方面具有潜力。

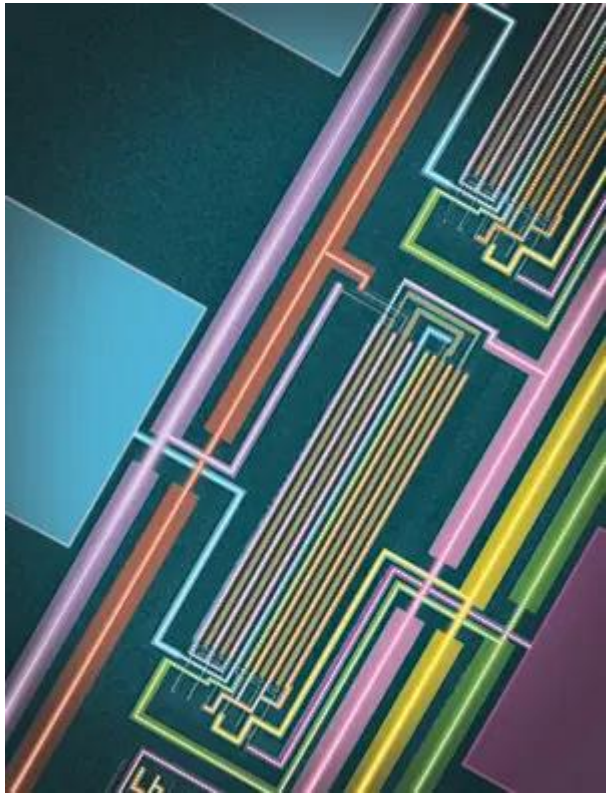
碳纳米管计算机架构

- **碳纳米管作为晶体管：**碳纳米管可以作为基本的开关元件，替代传统的硅晶体管。碳纳米管晶体管（CNTFET，Carbon Nanotube Field Effect Transistor）已经被提出并在实验中验证。与传统的MOSFET（金属氧化物半导体场效应晶体管）相比，碳纳米管晶体管具有更高的电子迁移率、更小的尺寸和更低的功耗。电子迁移率：碳纳米管的电子迁移率高得多，这意味着它们能够更快地传输信息，提高计算速度。低功耗：由于碳纳米管具有较低的开关电流和较少的功耗，它们有助于减少功耗问题，尤其是在高性能计算时。
- **碳纳米管交叉点（Interconnects）：**传统计算机芯片中，金属电路互联是限制芯片性能的瓶颈之一。碳纳米管可以作为芯片之间的高速互联组件。碳纳米管能够提供更低的电阻和更高的带宽，从而有效提高芯片间的通信速度。
- **碳纳米管作为存储介质：**除了作为开关元件外，碳纳米管还可以用于开发新型存储器件。研究者已经提出了基于碳纳米管的存储器件，如碳纳米管量子点存储器（Quantum Dot Memory），它利用量子力学效应存储信息，比传统存储器更高效。
- **碳纳米管集成电路（IC）:**碳纳米管集成电路（CNTIC）是将大量碳纳米管排列、连接在一起，构建复杂的电路系统。由于碳纳米管的优越性质，碳纳米管集成电路有望大幅提高集成度、降低功耗和增加计算速度。

CNT 晶体管及电路

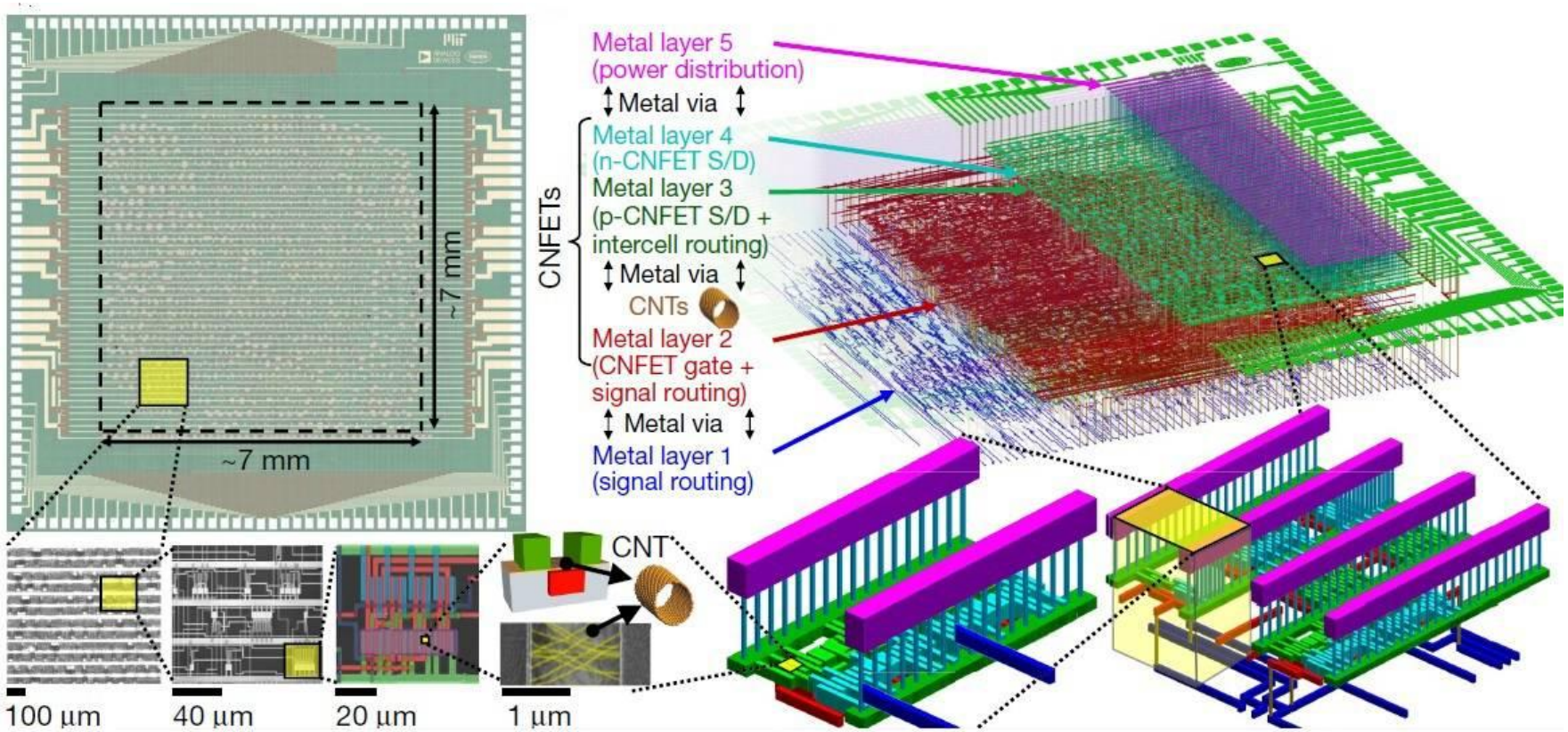

Source
Gate
Carbon nanotubes
Drain
Gate oxide
Silicon oxide
Silicon





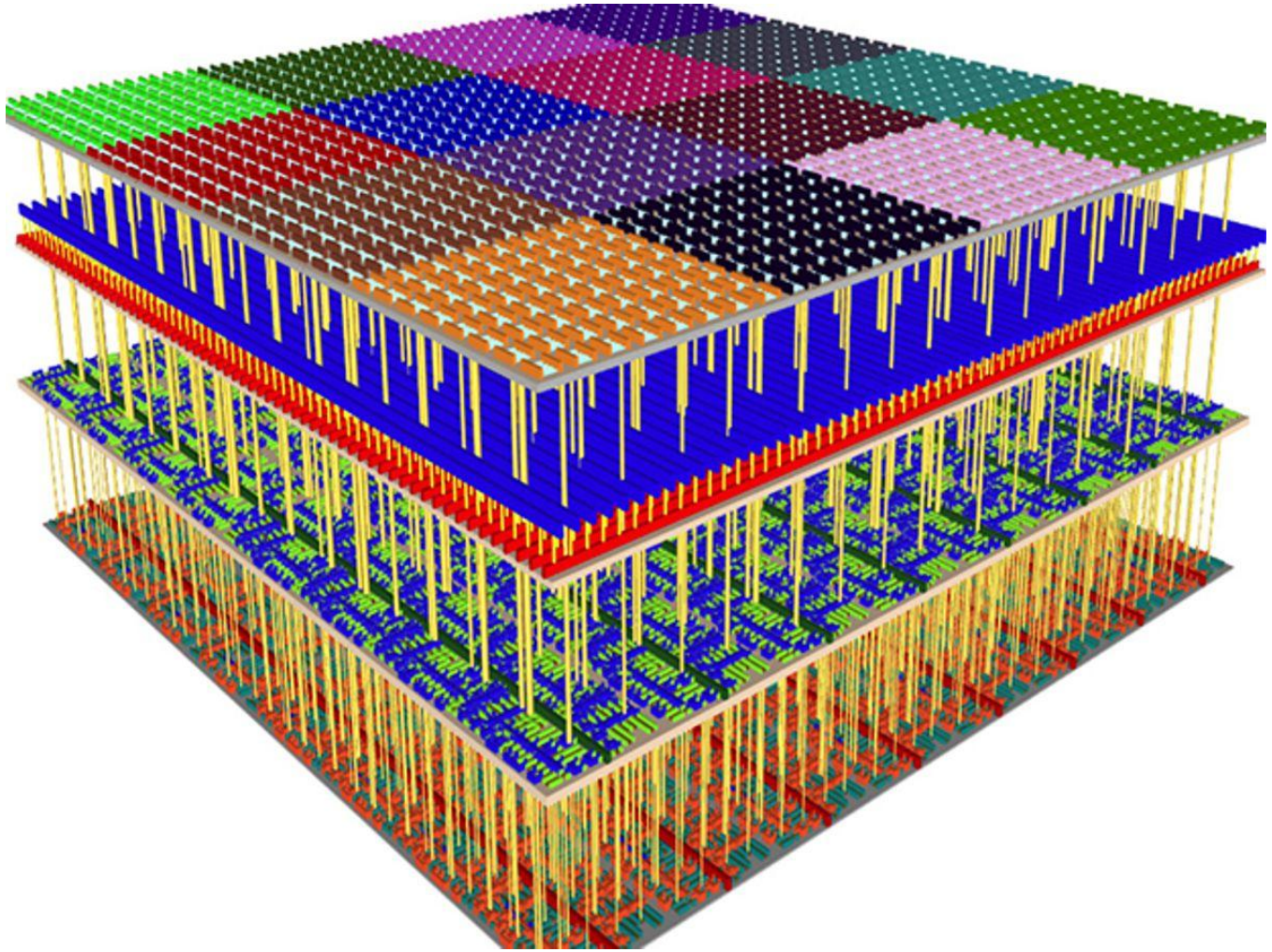
CNT微处理器

- **CNT** 的直径只有约 **1-2nm**
- 优异的电子性能



RV16X-NANO。来源：Nature 572,595-602 (2019)

3D CNT 计算机



纳米技术的三维集成，用于在单个芯片上进行计算和数据存储

碳纳米管计算机的优势与挑战

优势

- 高速和高效能
- 低功耗
- 微型化
- 散热能力强

挑战

- 生产规模化问题
- 电路集成和兼容
- 量产时的缺陷
- 使用寿命

Thank you!