北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

并行编程原理与实践

1. 并行编程概述

王一拙、计卫星

北京理工大学计算机学院

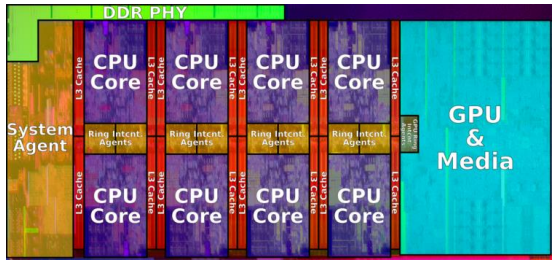
目录 | CONTENTS

1 为什么要进行并行编程

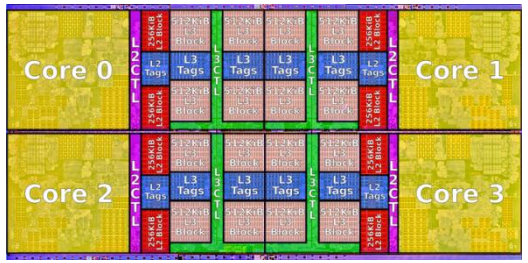
计算机系统结构的发展

并行是计算机系统结构发展的趋势

如何充分发挥现代计算机系统的性能



Intel Core i9

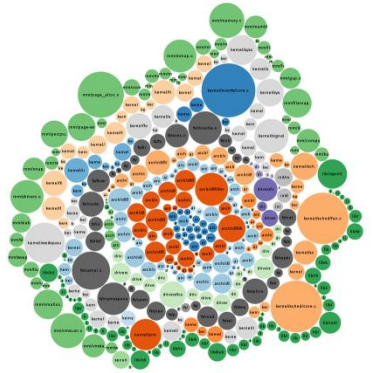


AMD Zen 7

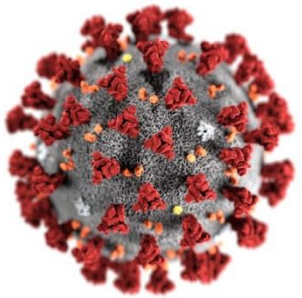
软件应用的发展

应用程序的规模不断扩大、复杂性不断提高

大数据、机器学习、生物信息处理等领域的发展



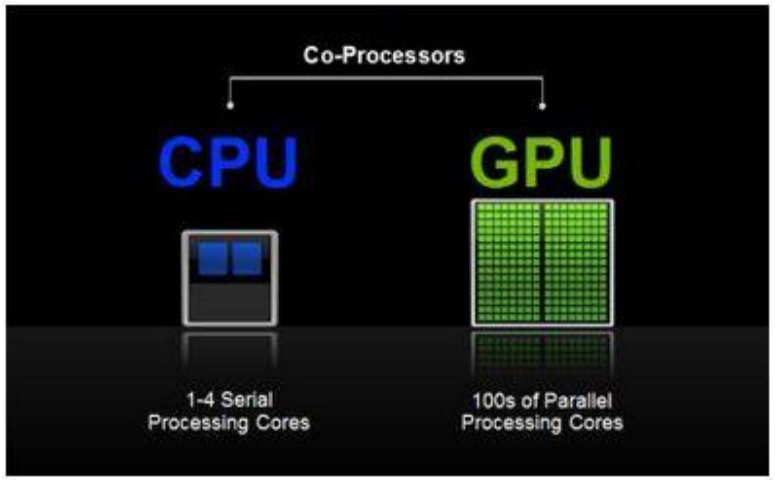
Linux kernel

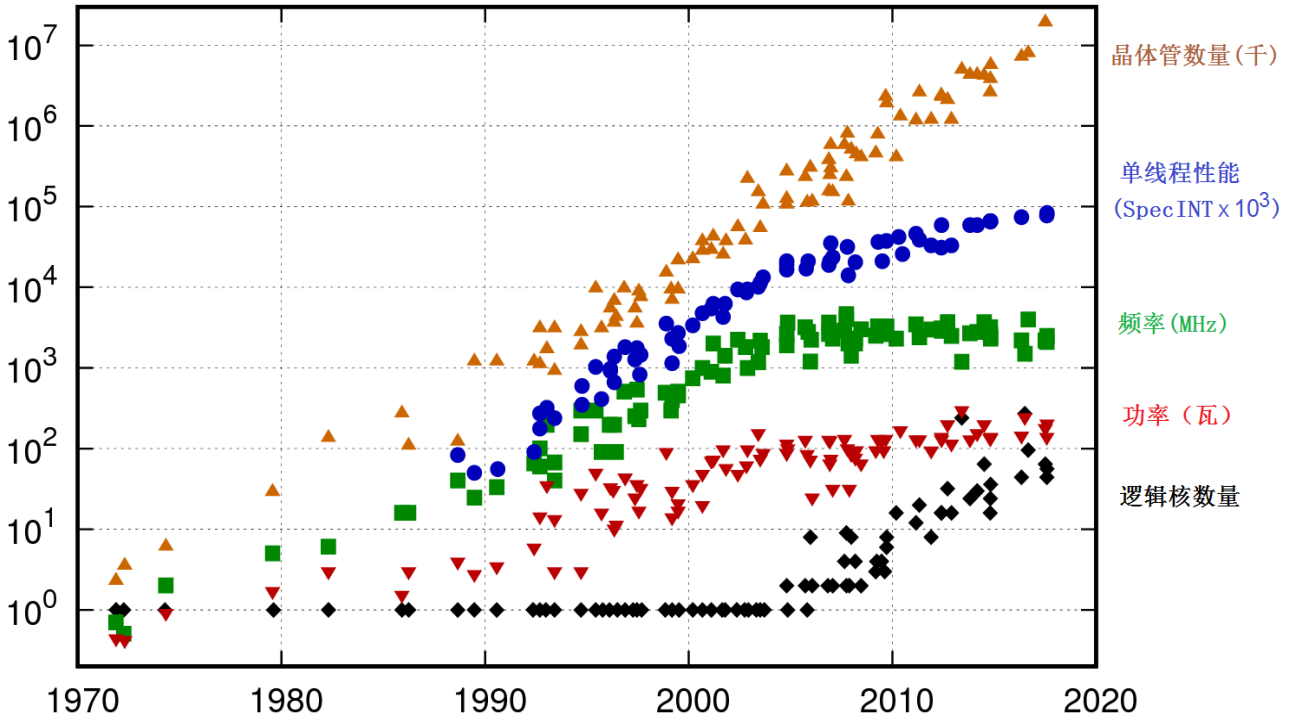


COVID-19

计算机系统结构的发展趋势

- 从单核到多核/众核
- 从同构到异构





Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

为什么要进行并行编程

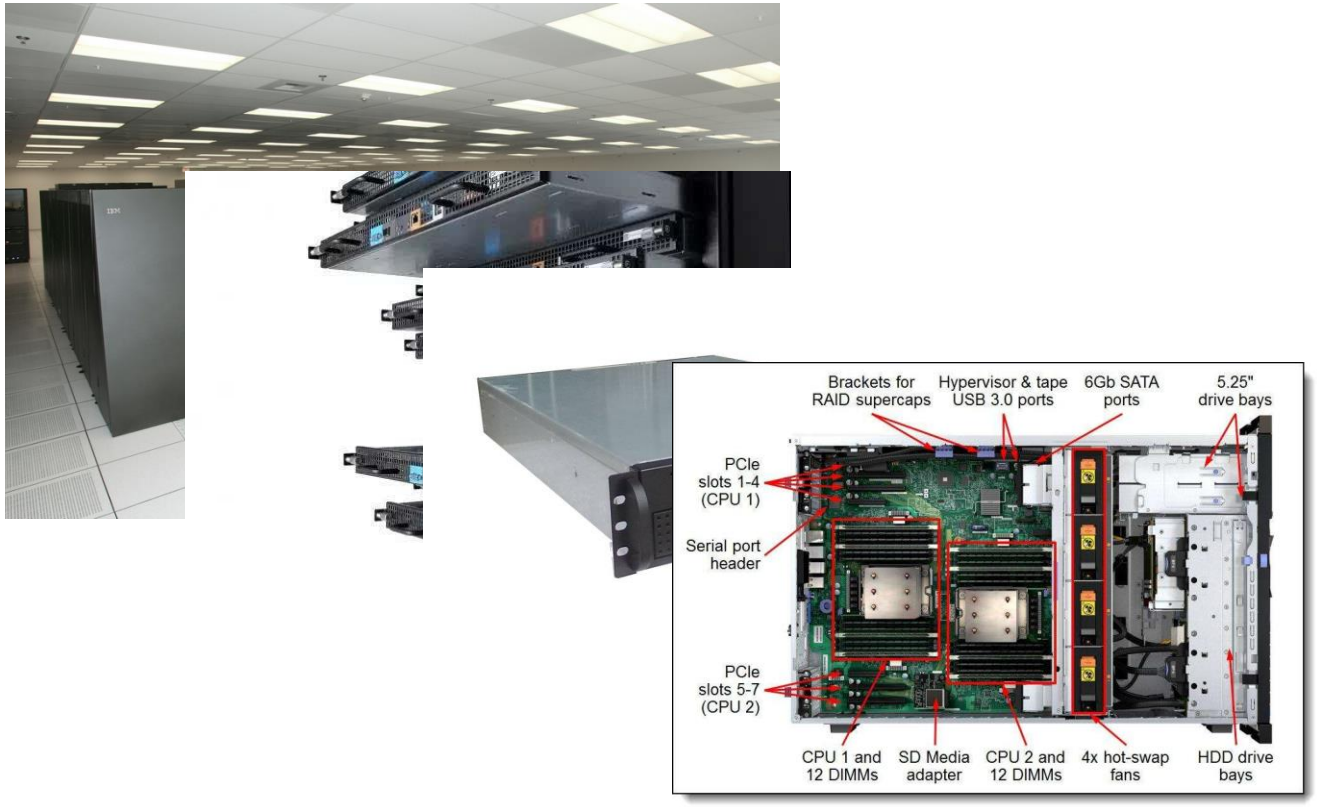
无所不在的并行计算机系统

无所不在的并行计算机系统

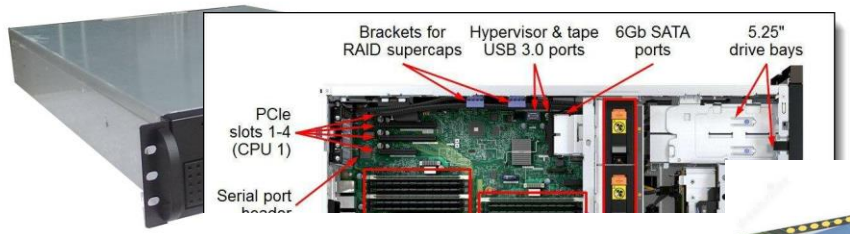
为什么要进行并行编程

无所不在的并行计算机系统

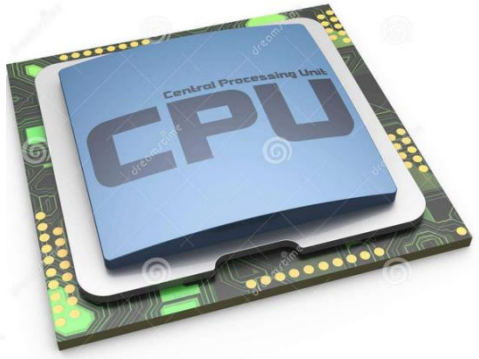
无所不在的并行计算机系统



无所不在的并行计算机系统

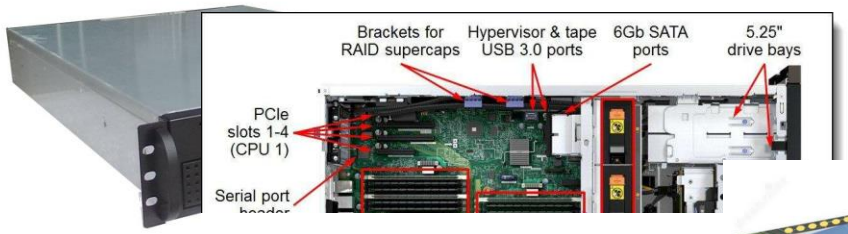


NVIDIA GeForce Titan Black
发布时间：2014年
首发价格：7999元

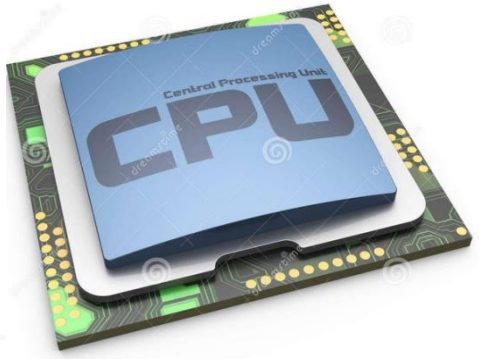


为什么要进行并行编程

无所不在的并行计算机系统

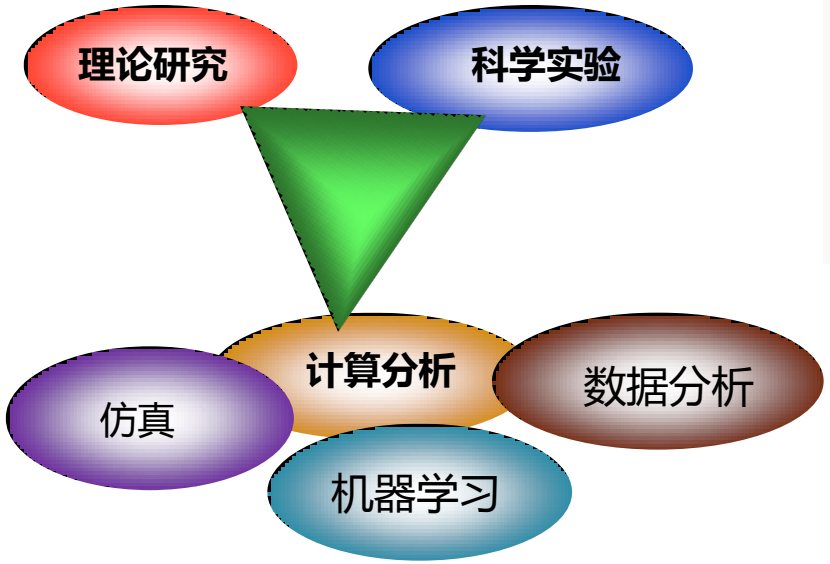


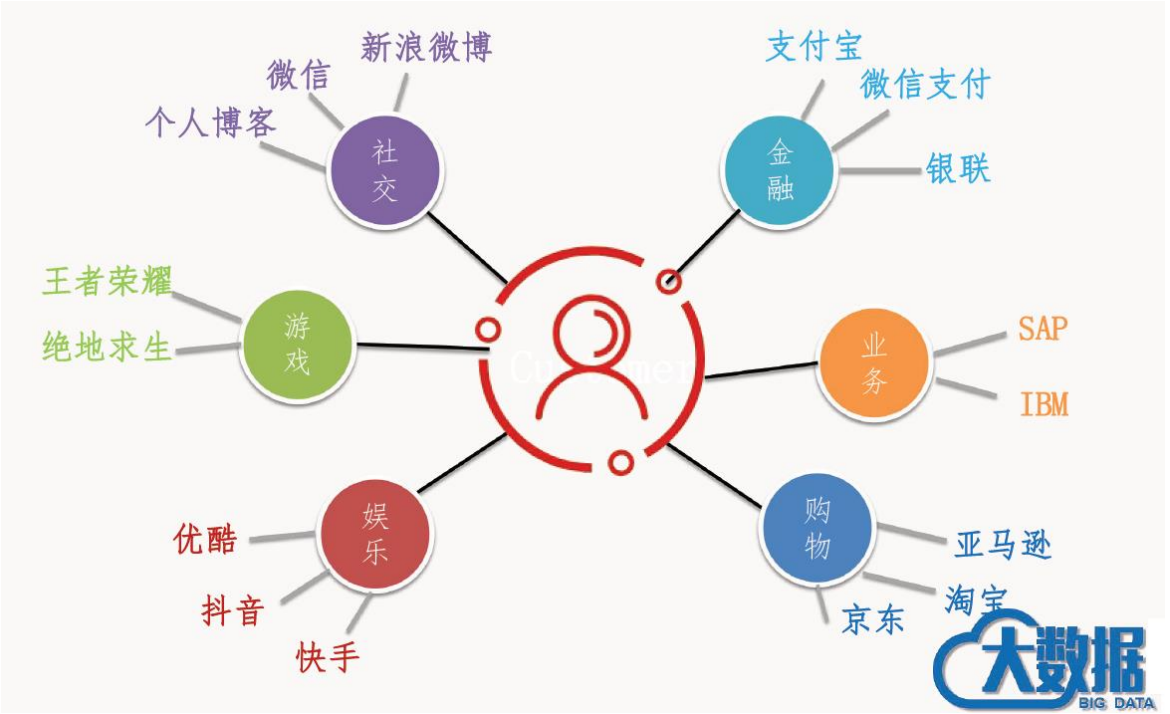
NVIDIA GeForce Titan Black
发布时间：2014年
首发价格：7999元



软件应用的发展趋势

- 数据规模越来越大
- 复杂程度越来越高

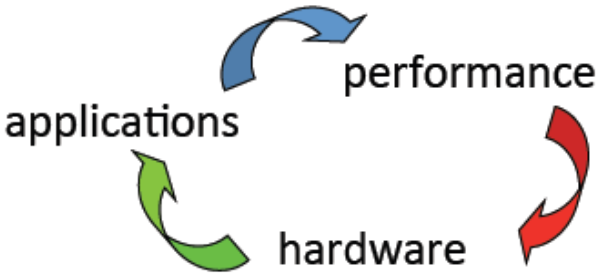




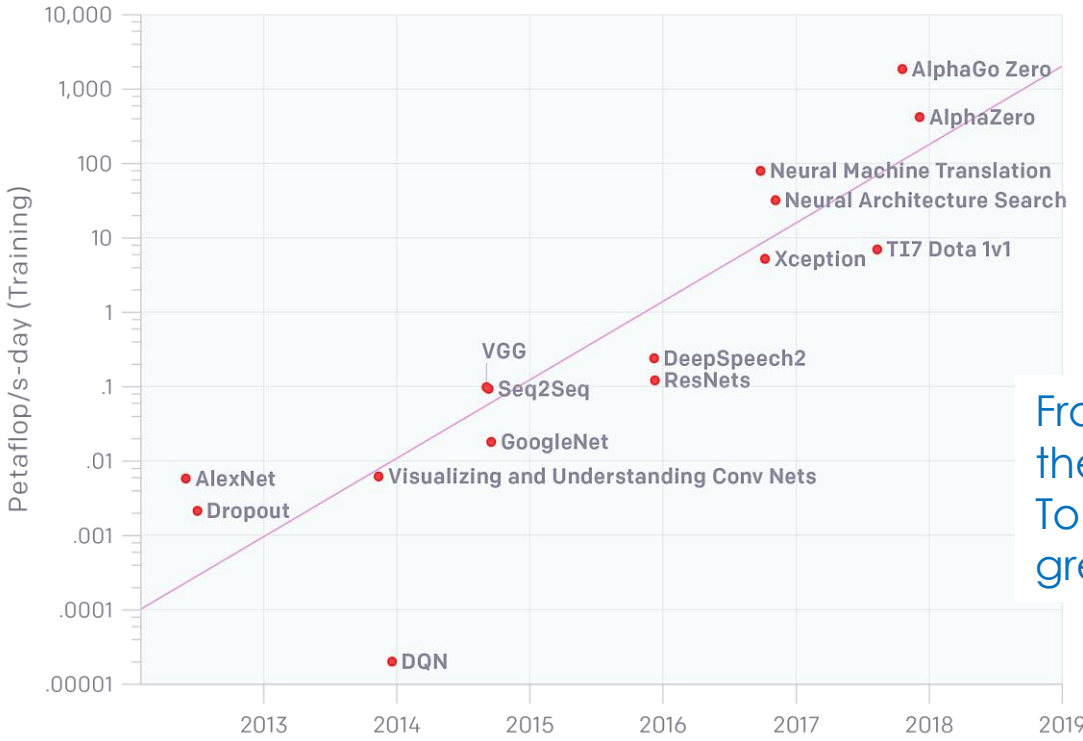
为什么要进行并行编程

软件应用的发展趋势

- 数据规模越来越大
- 复杂程度越来越高



300,000x increase from 2011 (AlexNet) to 2018 (AlphaGoZero)



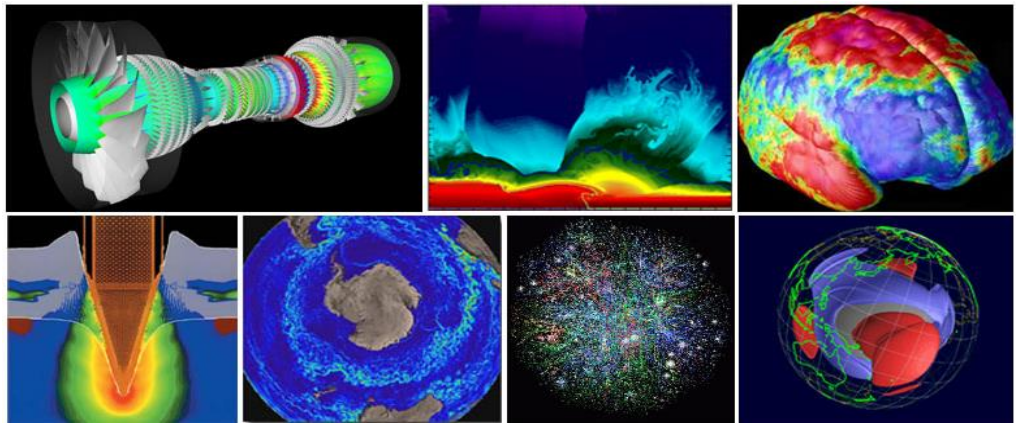
From 2011-2017 the fastest Top500 machine grew < 10x

https://blog.openai.com/ai-and-compute/

并行计算的应用领域

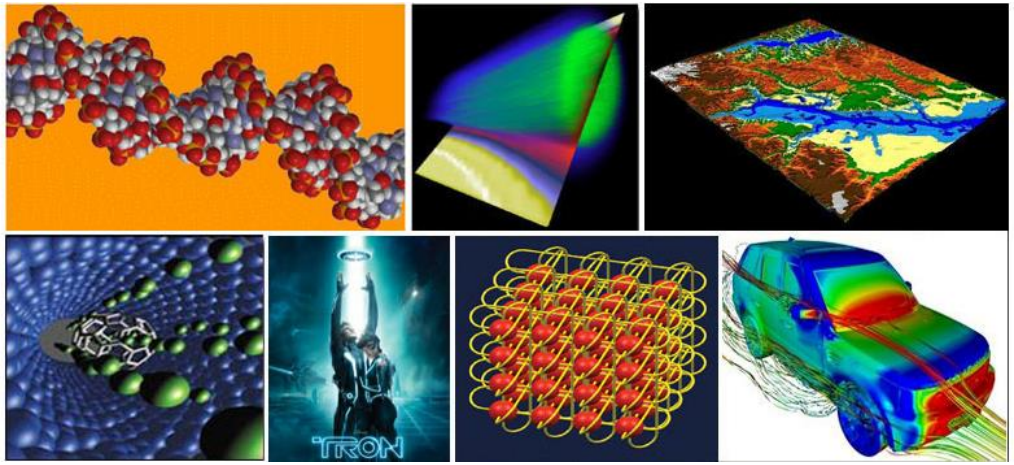
科学和工程计算

- 大气、地球、环境
- 应用物理、核能、粒子、高压、聚变、光学
- 生命科学、生物技术、遗传学
- 化学、分子科学
- 地质、地震



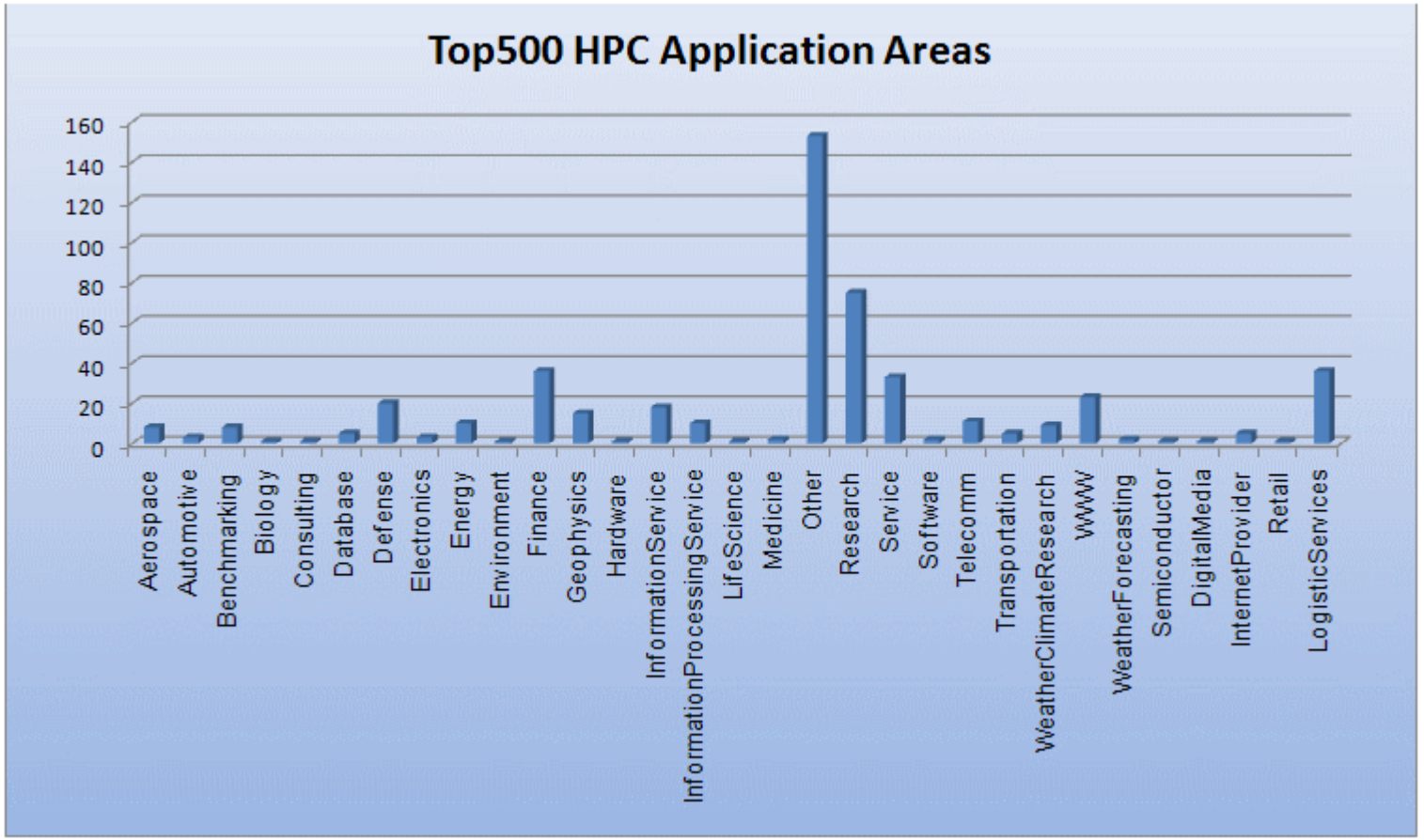
工业和商业应用

- 大数据分析、数据库、数据挖掘
- 石油勘探
- 网络搜索引擎、基于网络的商业服务
- 医学影像与诊断、药物设计
- 财务和经济建模、跨国公司管理
- 先进的图形和虚拟现实，尤其是在娱乐行业
- 网络视频和多媒体技术



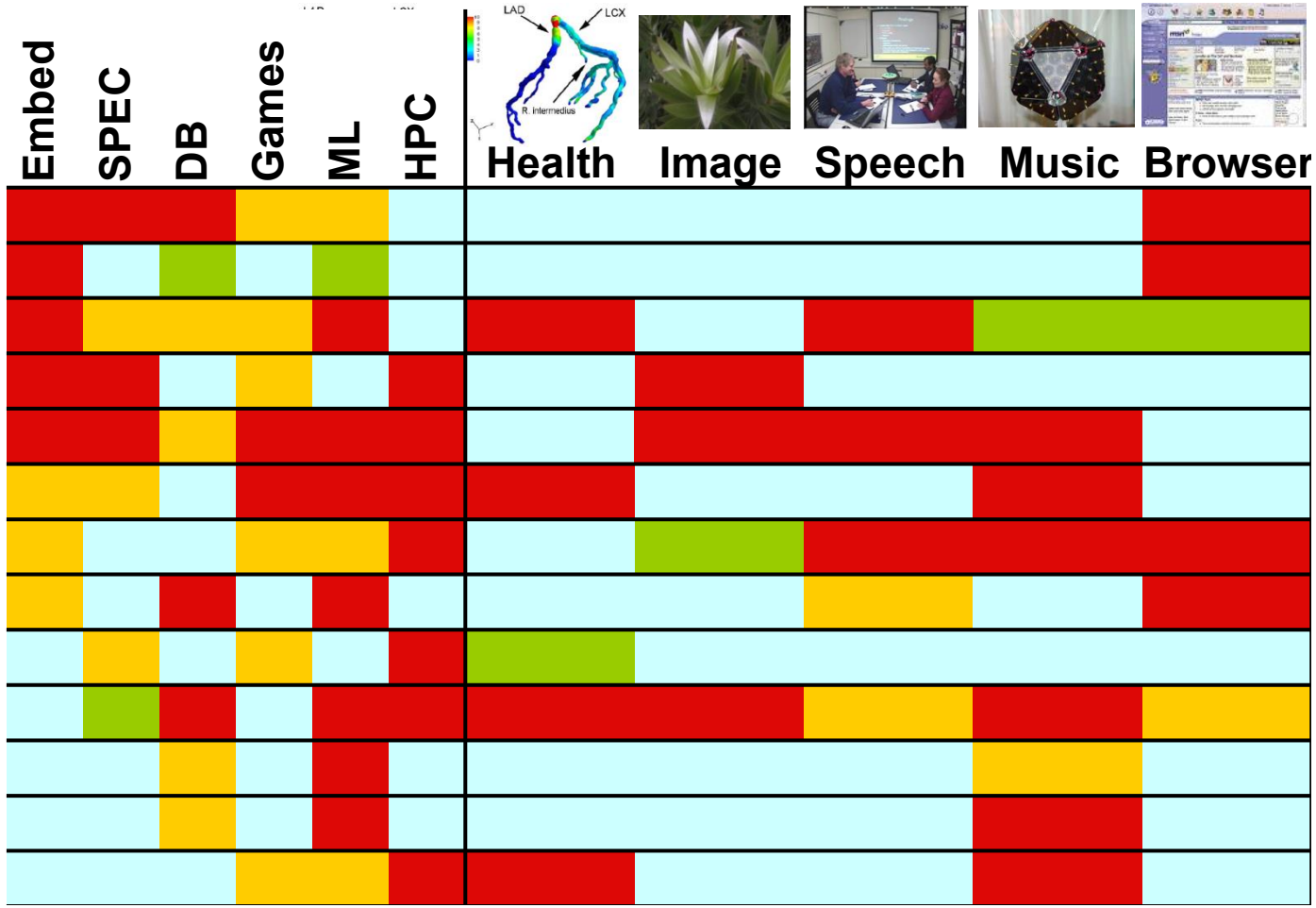
并行计算的应用领域

- 航空航天
- 机械与自动化
- 生物、制药
- 国防、电力、金融
- 互联网、信息服务
- 天气预报
- 。。。



为什么要进行并行编程

- 应用中的核心算法



并行计算带来的挑战

- 新的算法设计、算法分析技术
- 新的编程语言、模型、开发和调试工具
- 操作系统、底层服务和I/O接口对并行的支持
- 新的硬件结构
- 程序员需要了解存储系统和互连网络的结构等

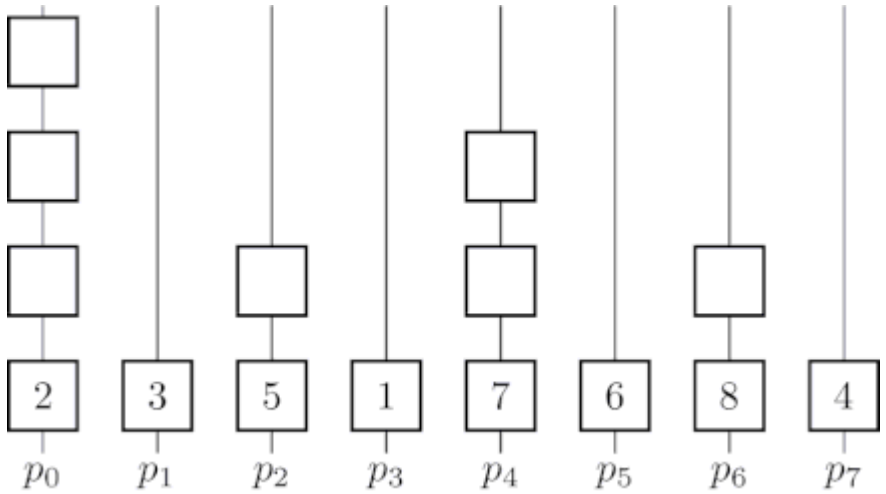
2 如何实现并行程序

示例1：序列求和

- 对一维数组V[1...n]，求 $s = \sum_{i=1}^{n} V_i$
- 串行算法：O(n)

```
S = 0
do i = 1, n
   S = S + V(i)
end do
```

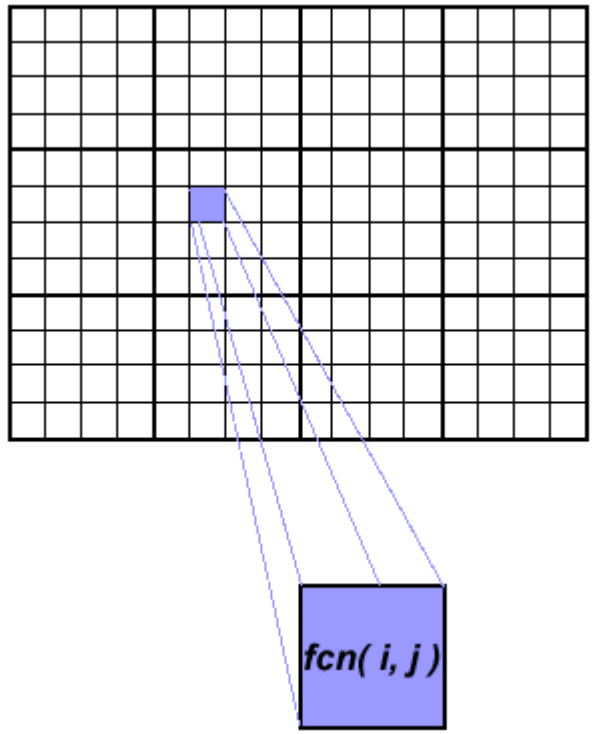
- 并行算法：O(log n)



示例2：数组处理

- 对二维数组的每个元素施加一个处理函数
- 串行算法

```
do j = 1,n
  do i = 1,n
    a(i,j) = fcn(i,j)
  end do
end do
```

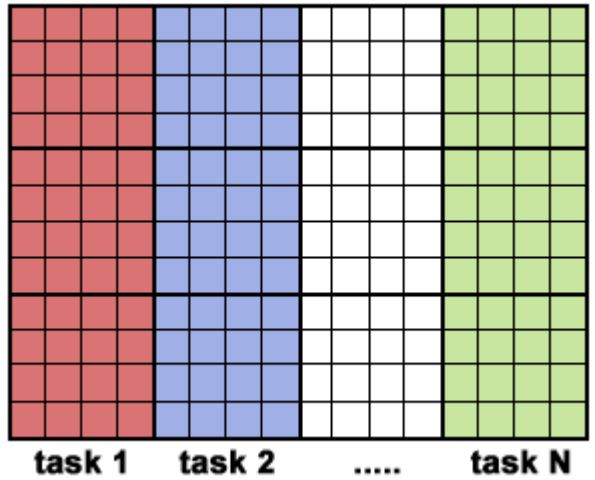


示例2：数组处理

- 并行算法：每个处理单元（PE, Processing Element）计算数值中一部分
  - 数组元素平均分配给多个PE ---- 负载均衡
  - 数组元素之间没有数据依赖 ---- 不需要同步
  - 在共享存储系统上实现 ---- 不需要通信（数据传输）

```
do j = mystart, myend
  do i = 1, n
    a(i,j) = fcn(i,j)
  end do
end do
```

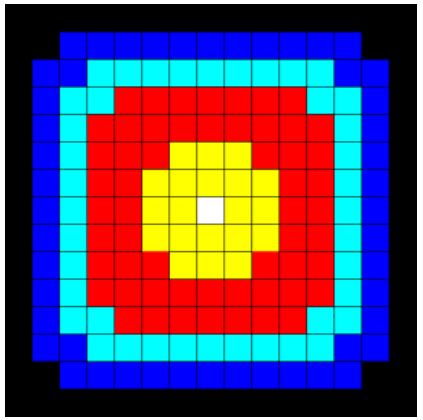
```
#pragma omp parallel for
for (i = 0; i < n; i++){
  for (j = 0; j < n; j++){
    a[i][j] = fcn(i,j);
  }
}
```



示例3：热传导方程计算

- 热传导方程描述了给定初始温度分布和边界条件下的温度随时间变化

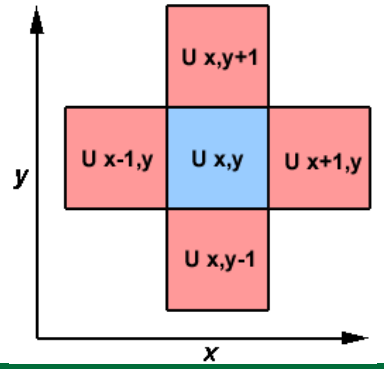
$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x_1}(k_1 \frac{\partial u}{\partial x_1}) + \cdots + \frac{\partial}{\partial x_n}(k_n \frac{\partial u}{\partial x_n}) + F(x,t)$$



- 用有限差分方法求解2-D热方程

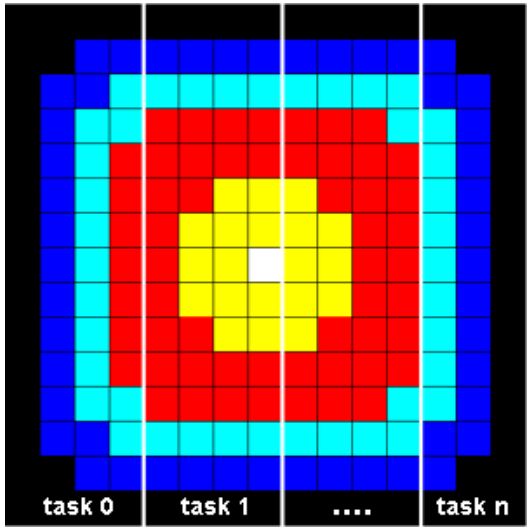
$$U_{x,y} = U_{x,y} + C_x * (U_{x+1,y} + U_{x-1,y} - 2 * U_{xy}) + C_y * (U_{x,y+1} + U_{x,y-1} - 2 * U_{x,y})$$

```
do iy = 2, ny - 1
  do ix = 2, nx - 1
    u2(ix, iy) =  u1(ix, iy)  +
         cx * (u1(ix+1,iy) + u1(ix-1,iy) - 2.*u1(ix,iy)) +
         cy * (u1(ix,iy+1) + u1(ix,iy-1) - 2.*u1(ix,iy))
  end do
end do
```



示例3：热传导方程计算

- 并行算法：每个处理单元计算一部分区域
  - 区域平均分配给多个PE ---- 负载均衡
  - 区域边界点的计算需要旁边区域的数据 ---- 需要通信（数据传输）
  - 下个时间步的计算依赖于上一步的计算结果 ---- 需要同步



示例3：热传导方程计算

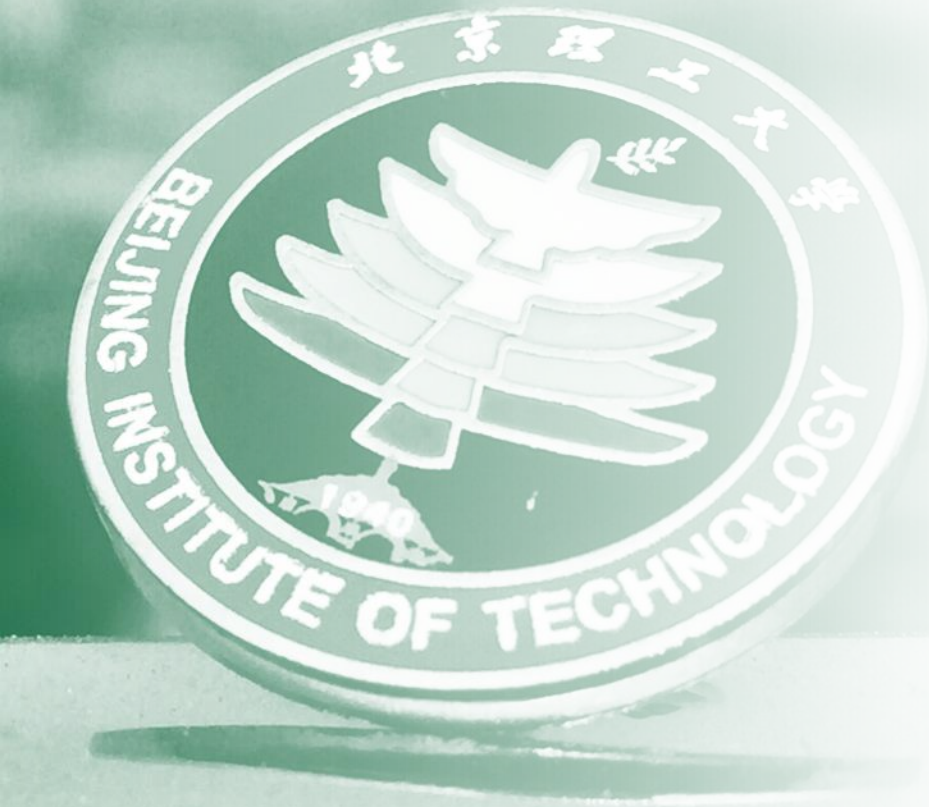
- SPMD方式并行实现
  - 主进程（MASTER）和工作进程（WORKER）都承担部分计算
  - 每个时间步的计算要同步（starting info）
  - 进程之间需要传输数据

```
find out if I am MASTER or WORKER

if I am MASTER
  initialize array
  send each WORKER starting info and subarray
  receive results from each WORKER

else if I am WORKER
  receive from MASTER starting info and subarray

  # Perform time steps
  do t = 1, nsteps
    update time
    send neighbors my border info
    receive from neighbors their border info
    update my portion of solution array

  end do

  send MASTER results
endif
```

- **实现并行程序要考虑：**
  - 问题能否（适不适合）被并行化？
  - 问题如何进行划分？ 任务划分 数据划分
  - 并行处理单元之间数据如何通信？
  - 数据之间的依赖如何？
  - 任务之间需不需要同步？
  - 负载是否均衡？
  - 。。。

如何实现并行程序

- **发掘硬件上各个层次的并行性**
  - 指令级并行 ---- 单个处理核内部（指令流水、VLIW、超标量、SIMD）
  - 线程级并行 ---- 多个处理核之间
  - 进程级并行 ---- 多个处理器之间
  - 循环并行
  - 任务并行
  - 访存并行
  - 。。。

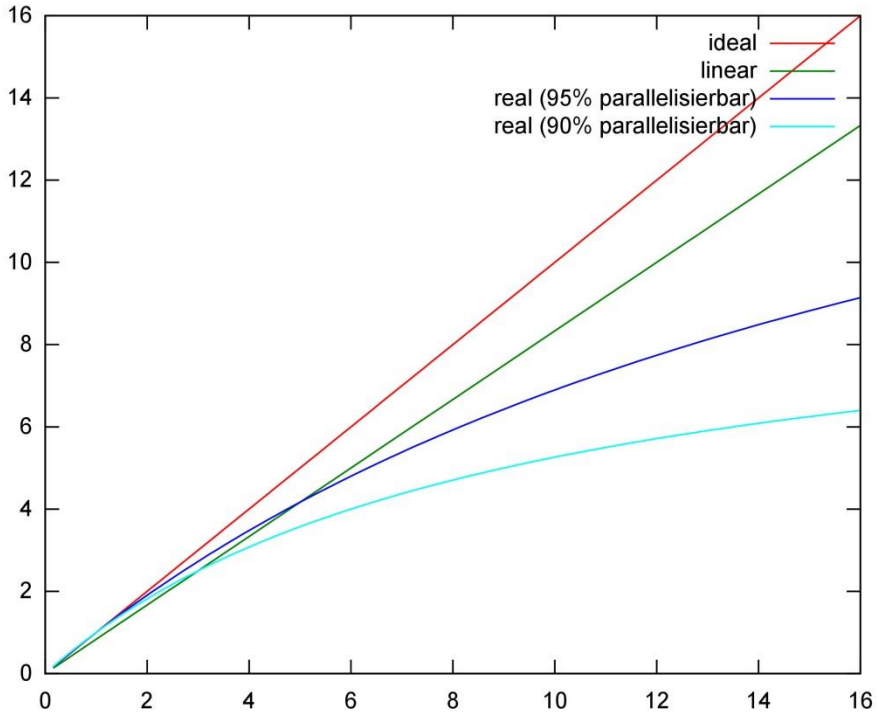


3 并行程序的性能评估

加速比（Speedup）

$$S_p = \frac{T_1}{T_p}$$

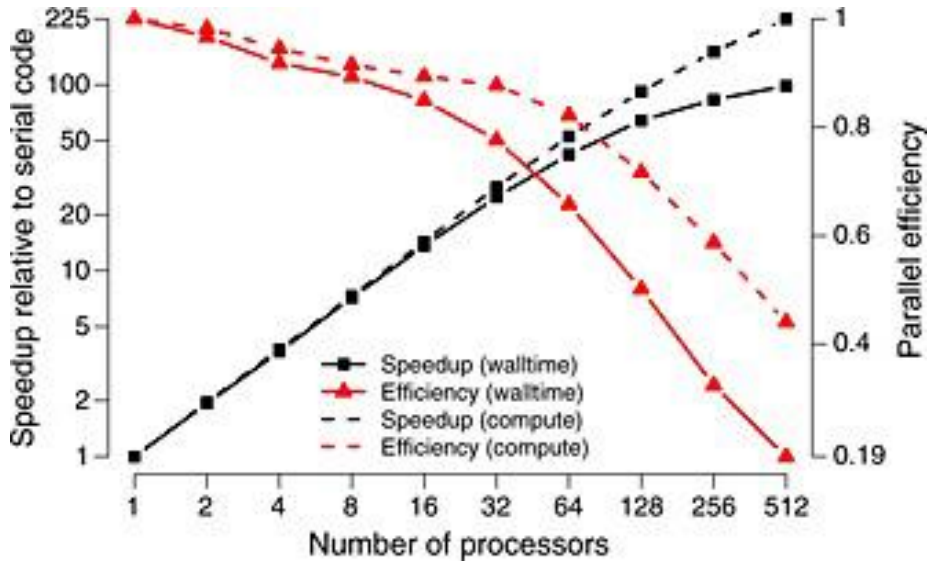
- p指CPU数量
- T1指顺序执行算法的执行时间
- Tp指当有p个处理器时，并行算法的执行时间



并行效率（Parallel Efficiency）

$$E_p = \frac{S_p}{p} = \frac{T_1}{pT_p}.$$

- Ep的值一般介于0～1之间，用于表示在解决问题时，相较于在通信与同步上的开销，参与计算的处理器得到了多大程度的充分利用。
- 由定义易见，拥有理想加速比的算法并行效率为1。



可扩展性（Scalability）

- 如果一个技术（算法/程序）可以处理规模不断增长的问题，那么它就是可扩展的。
- 对并行程序，如果输入规模增大，同时增加进程/线程个数（PE个数），能够使得并行效率保持不变，那么该程序就有很好的可扩展性。
- 强可扩展：问题规模固定，增加进程/线程个数时，效率不变。
- 弱可扩展：增加进程/线程个数时，只有以相同倍率增大问题规模才能使效率值不变。

Amdahl定律

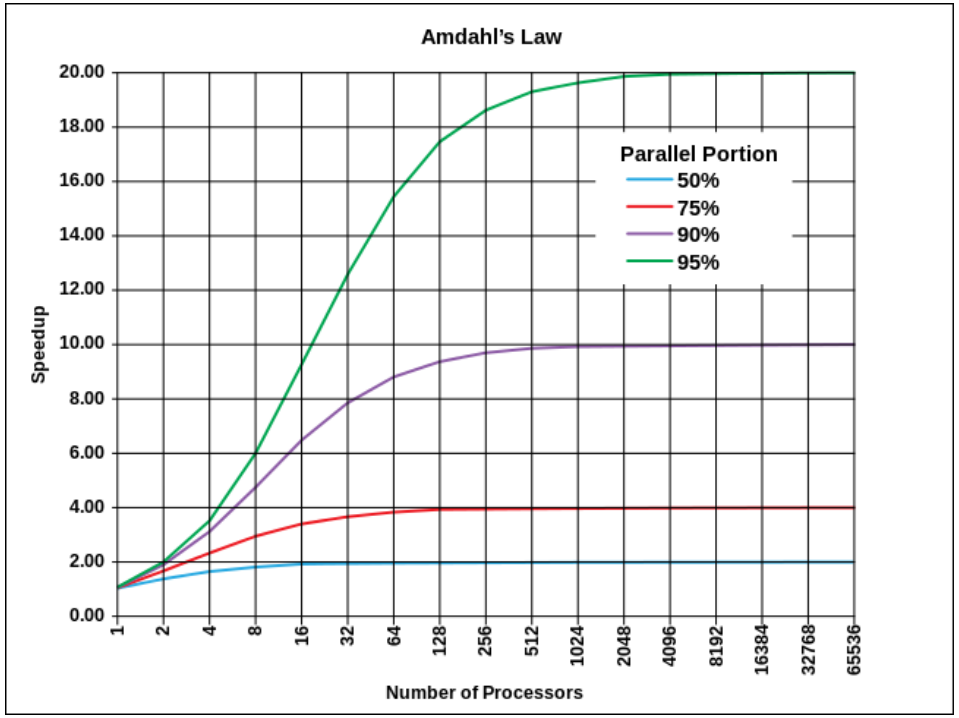
- 系统执行某应用程序需要的时间为Told，该应用程序某部分执行时间与总时间比例为α，若将该部分性能提升k倍，总的执行时间为：

Tnew = (αTold / k) + (Told - αTold) = Told [(1 - α) + α/k]

则加速比S = Told/Tnew为：

$$S = \frac{1}{(1-\alpha) + \alpha/k}$$

- 对系统某部分加速时，其对系统整体影响取决于该部分重要性和加速程度。
- 要想显著加速整个系统，必须提升全系统中相当大的部分的速度。

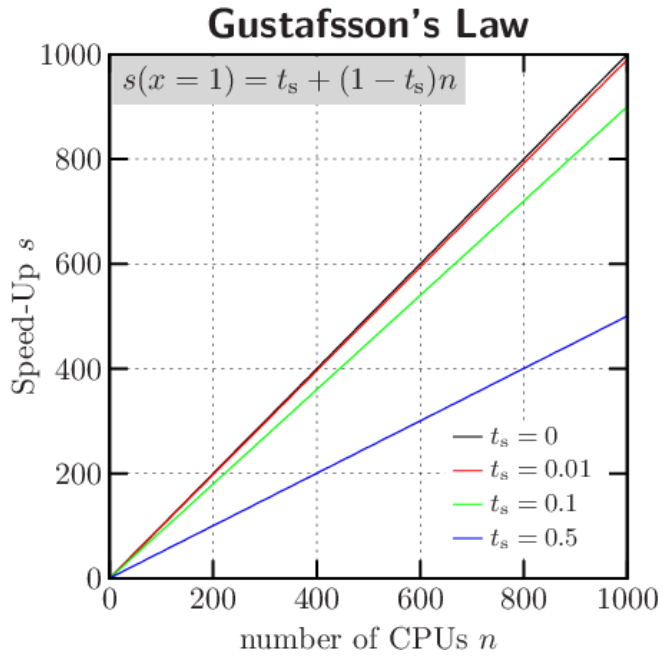


Gustafson定律

- 定义a为系统串行执行时间，b为系统并行执行时间，n为处理器个数，那么系统执行时间（串行时间+并行时间）可以表示为a+b，系统总执行时间（串行时间）可以表示为a+nb， 串行比例F=a / (a + b).

  则加速比为：

$$Sn = \frac{a+nb}{a+b} = \frac{a}{a+b} + \frac{nb}{a+b} = F + n(\frac{a+b-a}{a+b}) = F + n(1-F) = n - F(n-1)$$



- 说明处理器数量，串行比例和加速比之间的关系。
- 只要有足够的并行化任务，那么加速比和处理器个数成正比。

程序计时

- 进程的三种状态为阻塞、就绪、运行。
- 墙上时钟时间 = 阻塞时间 + 就绪时间 + 运行时间
- 用户CPU时间 = 运行状态下用户空间的时间
- 系统CPU时间=运行状态下系统空间的时间
- 用户CPU时间+系统CPU时间=运行时间
- 通常使用墙上时钟时间衡量性能
- 使用API提供的计时接口
- 注意时间的单位

```
clock_t begin = clock();
    /*do your time-consuming job*/
clock_t end = clock();
double time_spent = (double)(end -
       begin)/ CLOCKS_PER_SEC;
```

```
#define MILLION 1000000
struct timespec tpstart;
struct timespec tpend;
clock_gettime(CLOCK_MONOTONIC,&tpstart);
      /*do your time-consuming job*/
clock_gettime(CLOCK_MONOTONIC,&tpend);
long timedif = MILLION*(tpend.tv_sec-
     tpstart.tv_sec)+(tpend.tv_nsec
     -tpstart.tv_nsec)/1000;
```

4 超算系统动态和排名

什么是超级计算机

- 超级计算机是指由数千甚至更多处理器组成、能计算普通计算机和服务器不能完成的大型复杂课题的计算机，被誉为“计算机中的珠穆朗玛峰”。
- 多数形态都是计算集群，主要特点包含两个方面：极大的数据存储容量和极快速的数据处理速度。

超级计算机的价值和研究意义

- 超算是解决国家经济建设、社会发展、国防建设等领域重大挑战性问题的重要手段，已成为世界各国争夺的一个战略制高点。在这方面的研制水平标志着一个国家的科学技术和工业发展的程度，体现着国家经济发展的实力。
- 在交通工具制造、气候问题、生物信息、地震监测、地球科学、天体物理、公共健康、材料科学、人类/组织系统研究等领域，超级计算机都发挥着无可替代的作用。
- 例如：2018年，中国气象局安装了国产“派-曙光”超级计算机，峰值计算性能8Pflops，计算能力跃居气象领域世界第3位。

超算系统国际排名

https://www.top500.org

- TOP500

  超级计算机性能排行榜。1993年，由美国与德国超级计算机专家联合编制，以LinPack BenchMark性能测试值为序进行排名，确定全球超级计算机500强。

- Green500

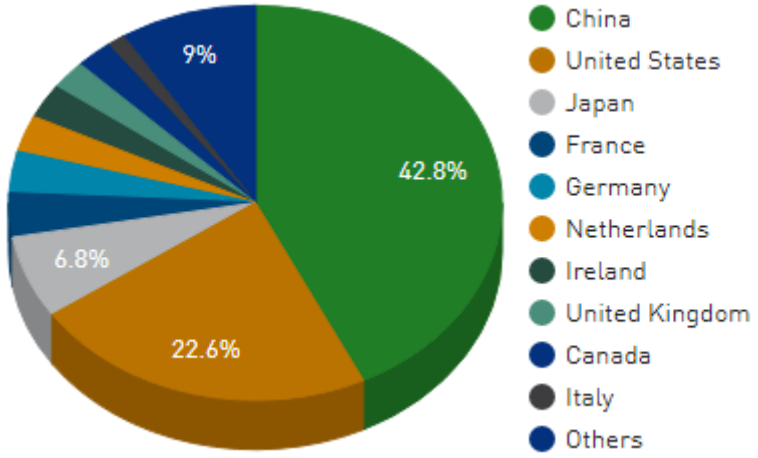
  超级计算机能效排行榜。基于 LinPack BenchMark，衡量计算机每瓦功耗可以提供的计算速率—每瓦性能（Gflops/watts）。

- “戈登·贝尔”奖

  超级计算应用领域的诺贝尔奖。设立于1987年，由美国计算机协会（ACM）于每年全球超级计算大会（SC）上颁发，由当年TOP500排行名列前茅的计算机系统上的应用获得。

超算系统动态和排名

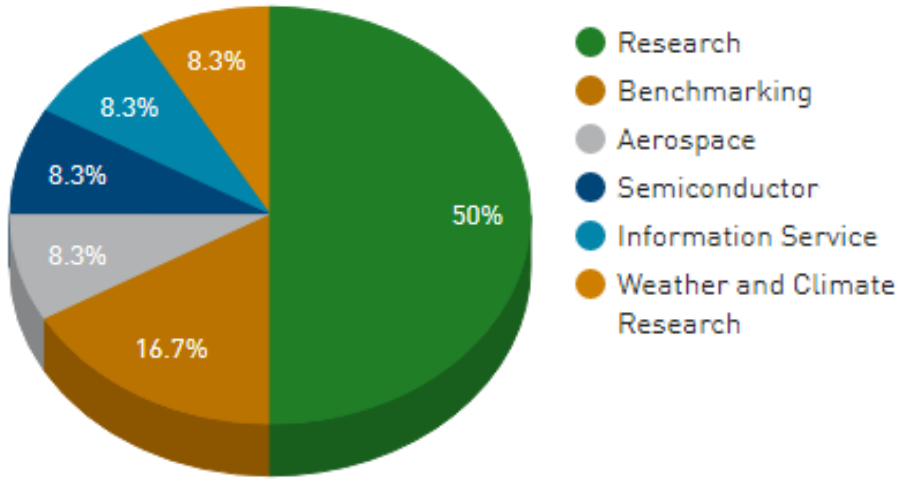
TOP500　2020年11月



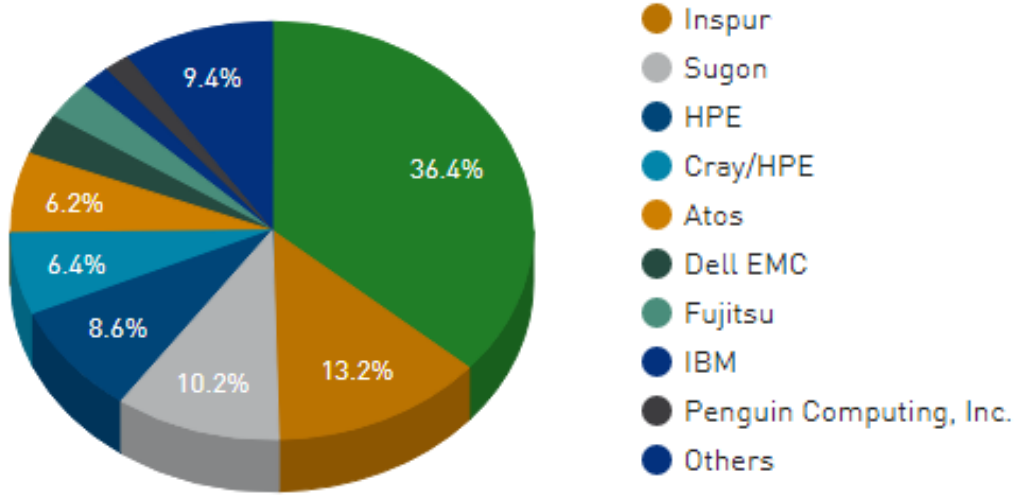
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646
6	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482

TOP500 2020年11月

Application Area System Share



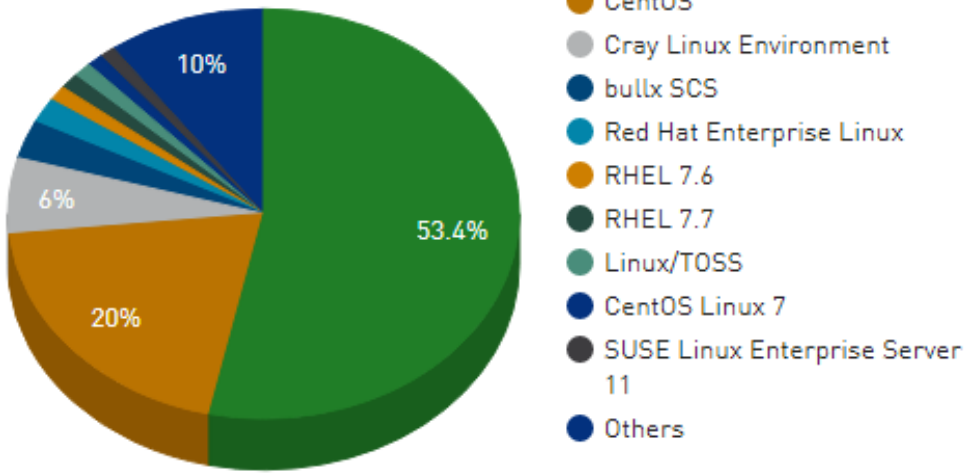
Vendors System Share



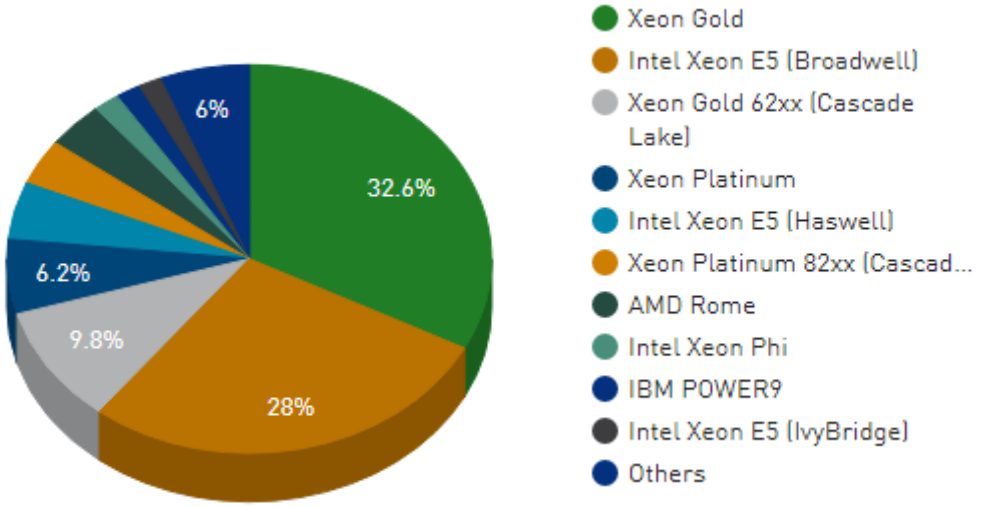
超算系统动态和排名

TOP500　2020年11月

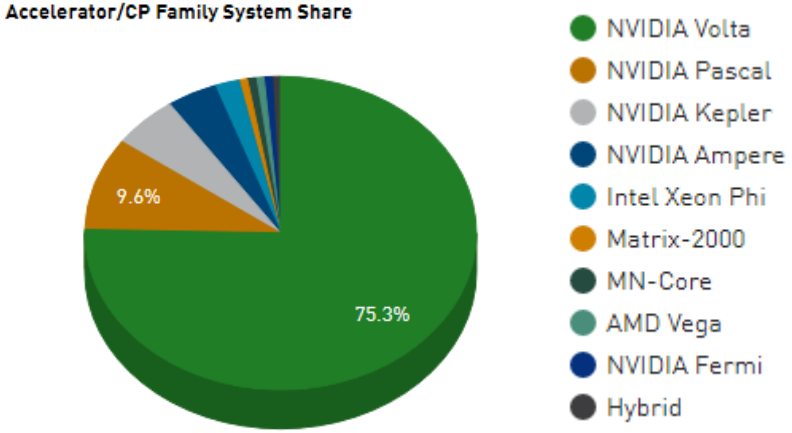
Operating System System Share



Processor Generation System Share



Accelerator/CP Family System Share



我国超算系统的发展情况

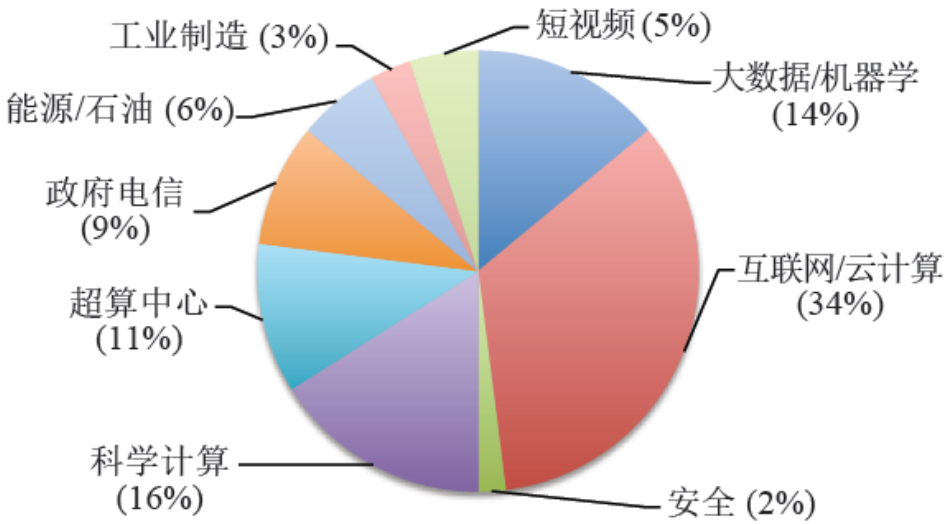
- 联想深腾1800万亿次机群在2002年世界TOP500中排第43名
- 联想深腾6800 万亿次机群名列2003年世界TOP500 第14 名
- 曙光4000A 10万亿次机群2004年世界TOP500中排第10名
- 国防科大研制的天河一号A千万亿次超级计算机2010 年11 月TOP500 排第1名
- 国防科大研制的天河二号2013 年6 月开始6连冠
- 国家并行计算机工程技术研究中心研制、安装在国家超级计算无锡中心的神威·太湖之光超级计算机（Sunway TaihuLight）2016年紧接着开始4连冠
- 2016 年，6月份中国大陆超算系统数量在历史上首次超过美国，排名世界第一

- 2016年11月中国科学院软件研究所杨超等基于神威·太湖之光完成的“千万核可扩展全球大气动力学全隐式模拟”获戈登贝尔奖
- 2017年11月清华大学地球系统科学系付昊桓等基于神威·太湖之光完成的“非线性大地震模拟”获戈登贝尔奖

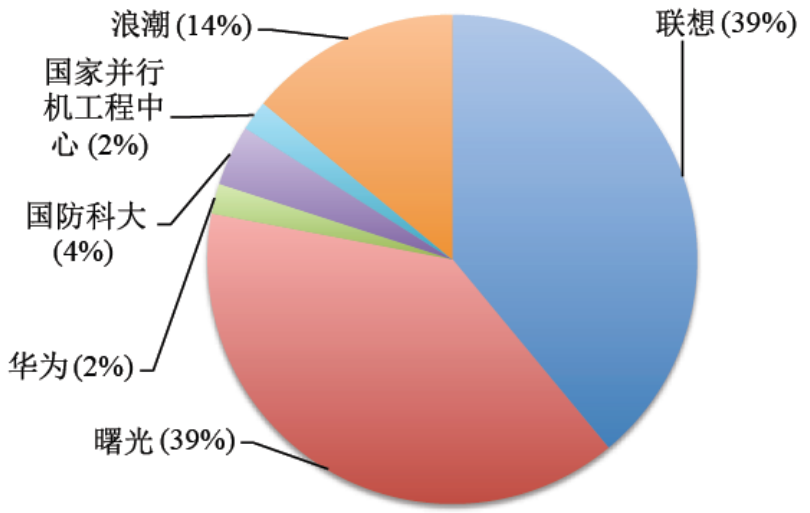
我国超算系统的发展情况

- 就在我国超算奔跑在快速路上时，遇到了新问题 — 美国芯片禁运。
  - 2015年2月，美国商务部将我国国家超级计算长沙中心、国家超级计算广州中心、国家超级计算天津中心和国防科技大学四家机构列入芯片限售之列。
- 国家“十三五”高性能计算国家重点研发专项课题中，中科曙光（曙光E级原型机）、国防科技大学（天河三号E级原型机）和江南计算技术研究所（神威E级原型机）同时获批进行E级超算的原型机系统研制项目，形成了中国E级超算“三头并进”的局面。
- 神威·太湖之光全面采用国产众核芯片“申威”和“神威睿思” 操作系统。
- 天河三号采用国产“飞腾”处理器、“迈创”众核处理器（Matrix-2000+）和“麒麟” 操作系统

我国超算系统的发展情况



中国TOP100行业应用领域机器系统份额图（2019.11）



中国TOP100 制造商机器数量份额图（2019.11）

袁国兴, 张云泉, 袁良. 2019年中国高性能计算机发展现状分析[J]. 计算机工程与科学, 2019(12).


北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY
谢谢!
德以明理 学以精工