中国科学院大学
夏季强化课程
2022

# Fast Solvers for Large Algebraic Systems

## Lecture 1. Large-scale numerical simulation

Chensong Zhang, AMSS

http://lsec.cc.ac.cn/~zhangcs

1 大规模数值模拟

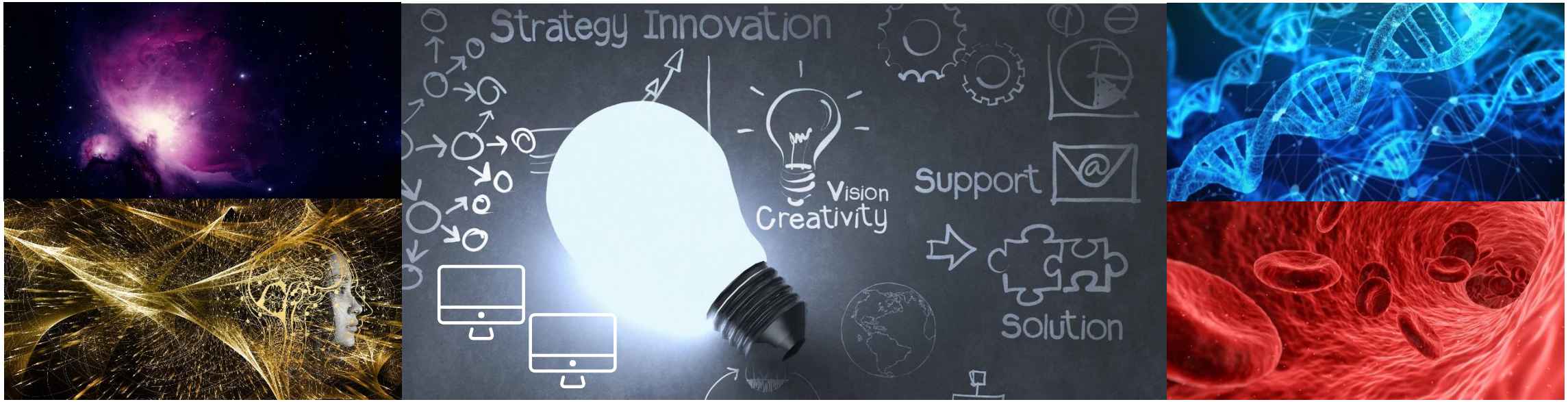
# Table of Contents

# Introduction

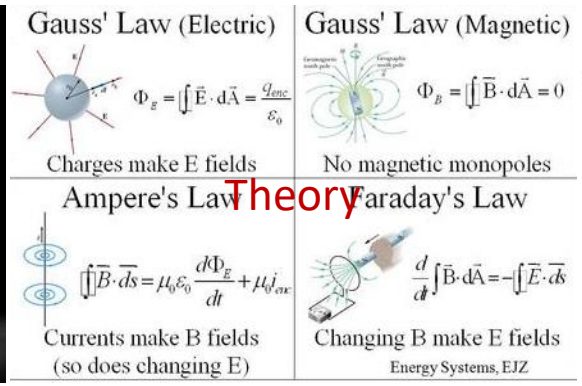
Some applications of computational mathematics

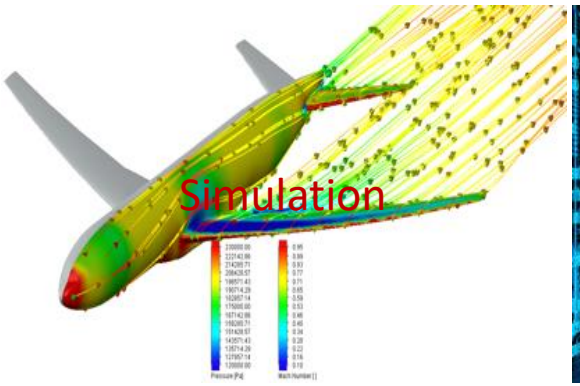
/01

# Four Paradigms of Scientific Discovery

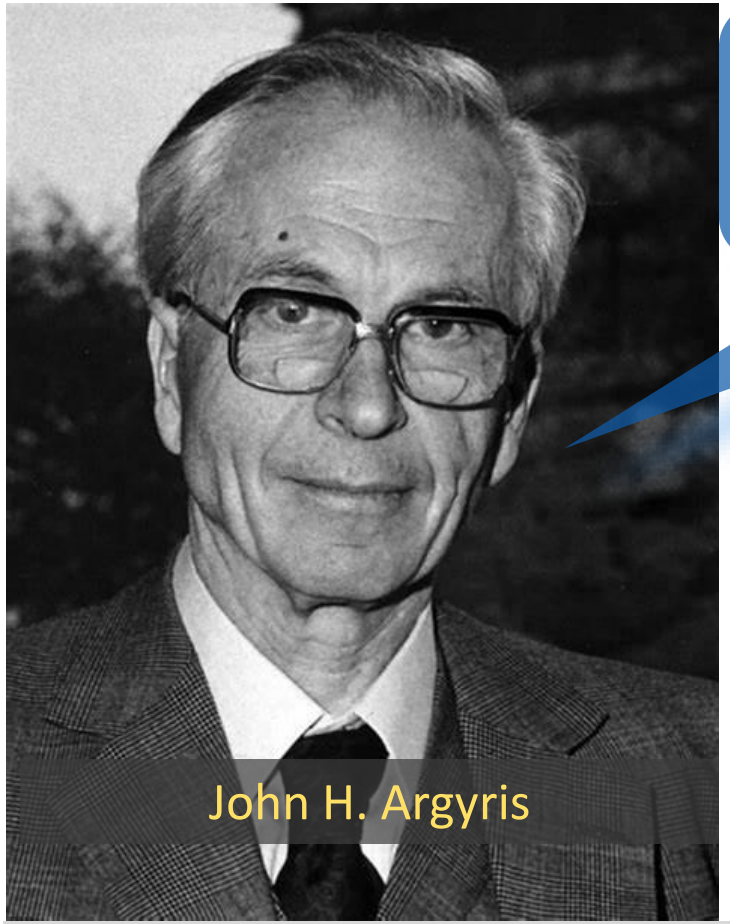



**The Fourth Paradigm: Data-Intensive Scientific Discovery, Hey, Tansley, Tolle, Microsoft, 2009**

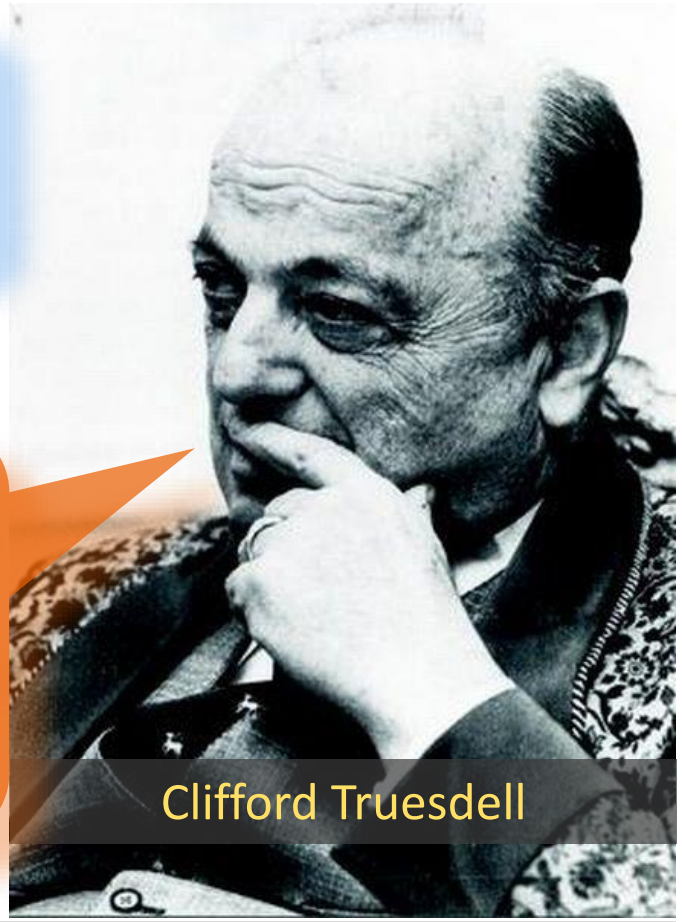




# Computer and Science


John H. Argyris

"The computer shapes the theory", Paper read to the Royal Aeronautical Society on May 18, 1965

"The computer: ruin of science and threat to mankind", in: An Idiot's Fugitive Essays on Science: Methods, Criticism, Training, Circumstances. Springer-Verlag, New York, 1984


Clifford Truesdell

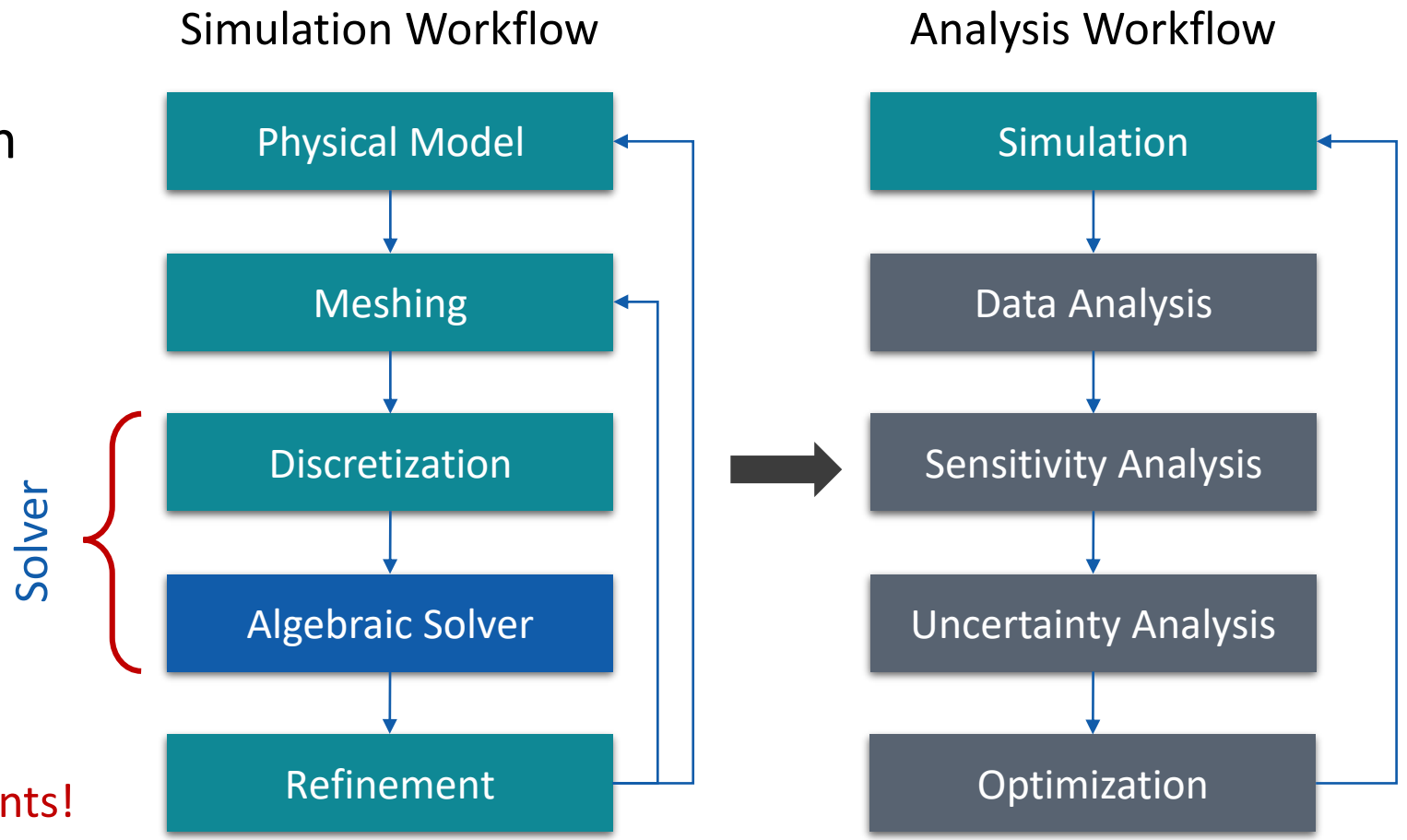
Ref: J.A. Cottrell, A. Reali, Y. Bazilevs, T.J.R. Hughes, "Isogeometric analysis of structural vibrations", Computer Methods in Applied Mechanics and Engineering, 195, Issues 41–43, 2006,

# Why Simulation So Important

In many situations, we have very limited theories and can not do experiments:

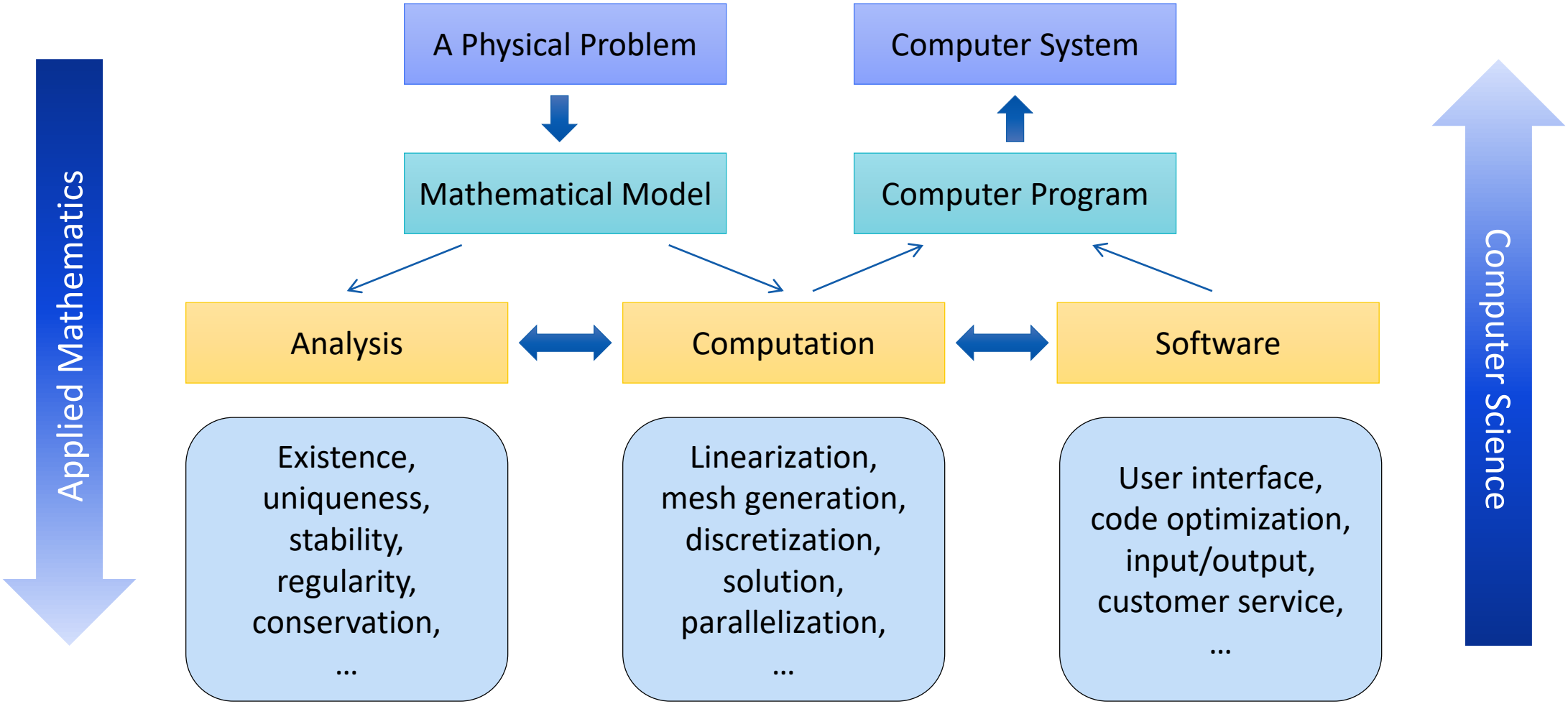
- Too slow
- Too difficult
- Too expensive
- Too dangerous

Simulation ≠ Numerical experiments!



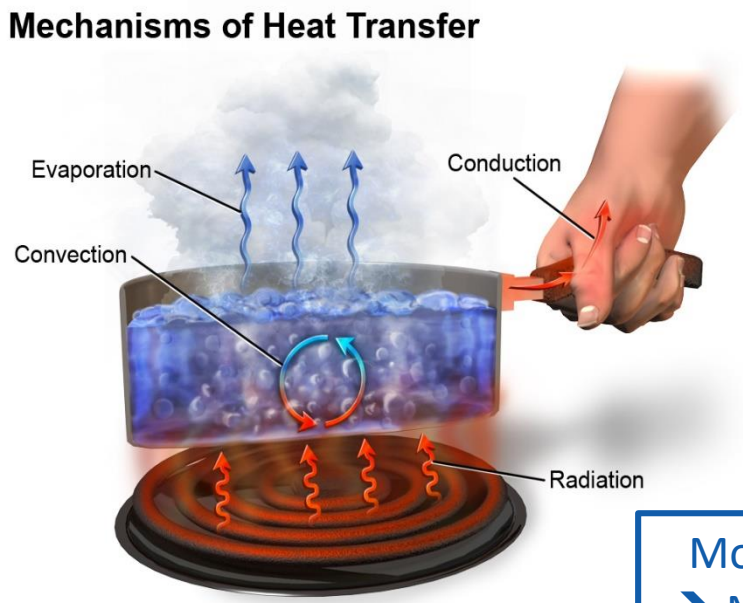

In this lecture, Solver := Algebraic Solver (Solution Method / Algorithm / Software)

# The Third Paradigm of Scientific Discovery

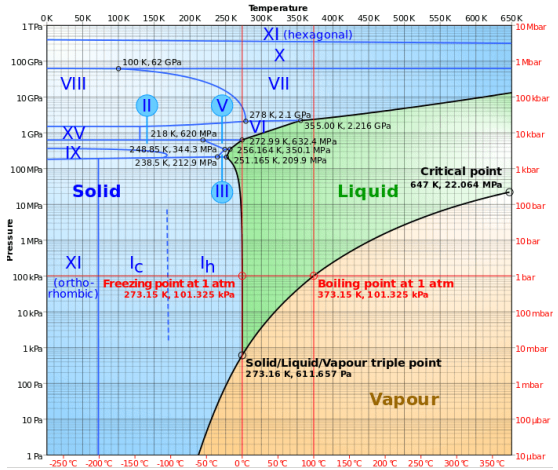



Applied Mathematics & Scientific Computing

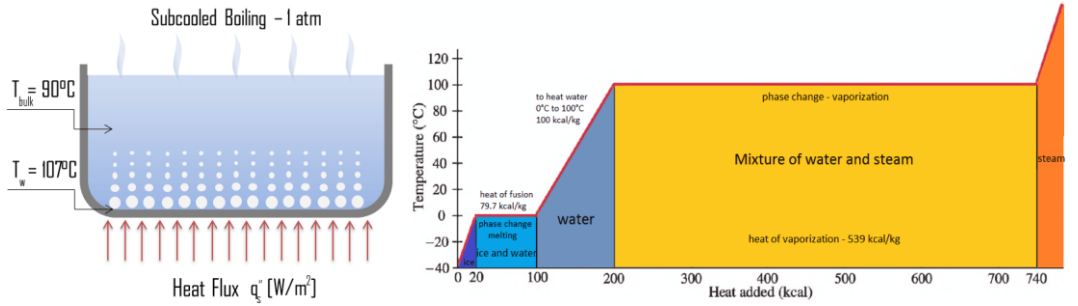
# Boiling Water

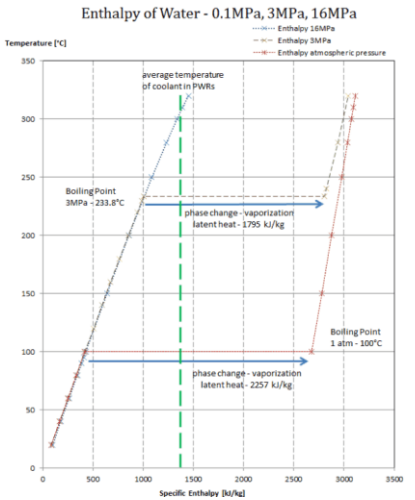



Source: https://thermtest.com/




Monophase
➔ Multiphase

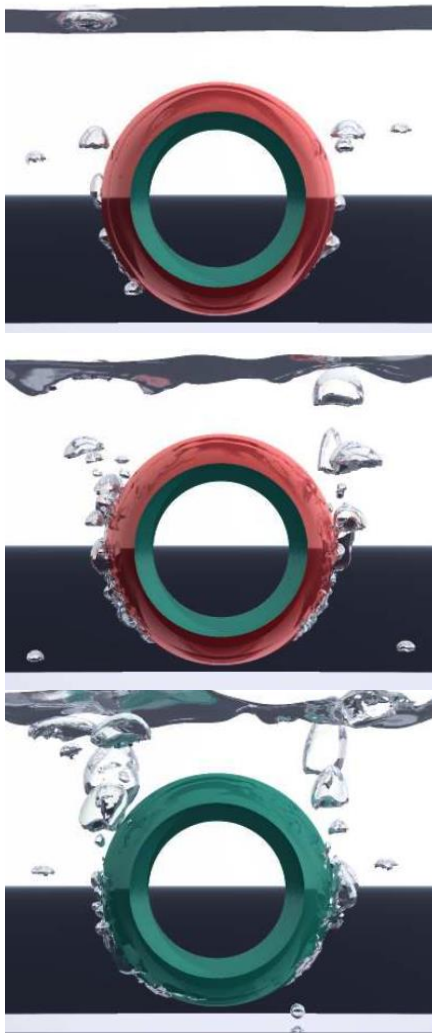


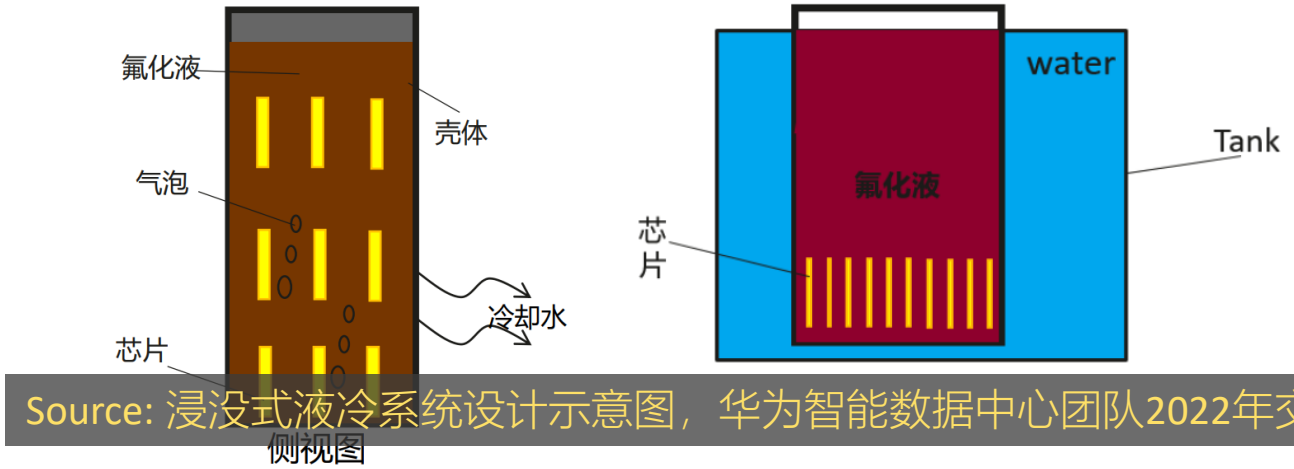




Source: https://www.thermal-engineering.org/

# Boiling Some More

- 传热
- 传质
- 相变
- 多相流
- 边界层
- 自由界面
- 气泡成核
- 气泡融合
- 气泡破灭



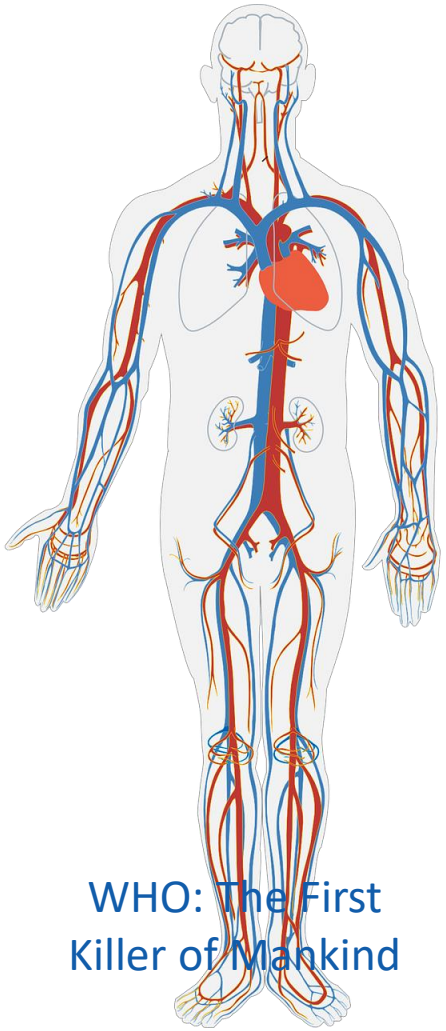



Source: 浸没式液冷系统设计示意图，华为智能数据中心团队2022年交流

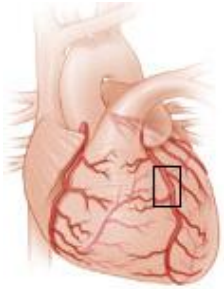
- 我国数据中心用电量以每年超10%的速度递增，2020年耗电量破2000亿千瓦时，占全国总电量的2.71%
- 冷却系统是数据中心提高能源效率的重点环节，它所产生的功耗约占数据中心总功耗的40%，有较大的优化空间

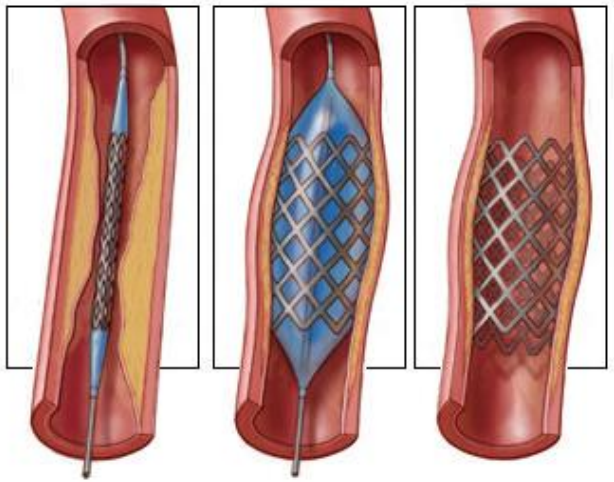
Source: Viorel, Unlusu, Metaxas, Sussman, Hussaini. “Physics based boiling simulation”, Eurographics/ACM SIGGRAPH Symposium on Computer Animation (2006)

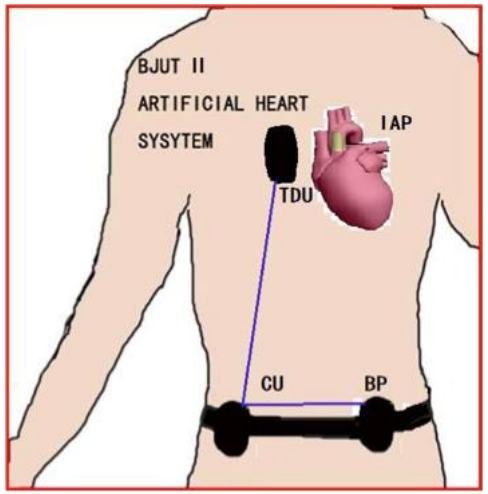
# Cardiovascular Diseases (CVDs)

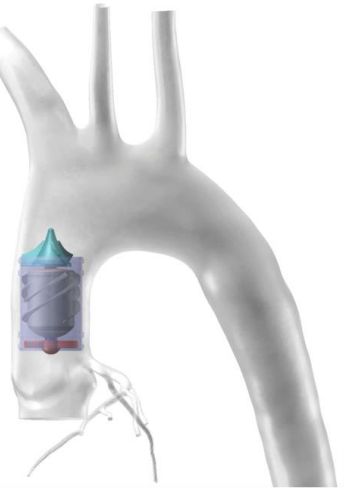


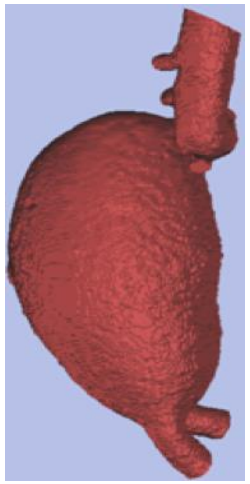
WHO: The First Killer of Mankind

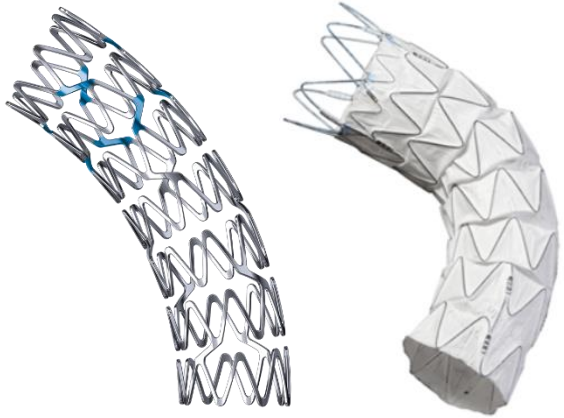


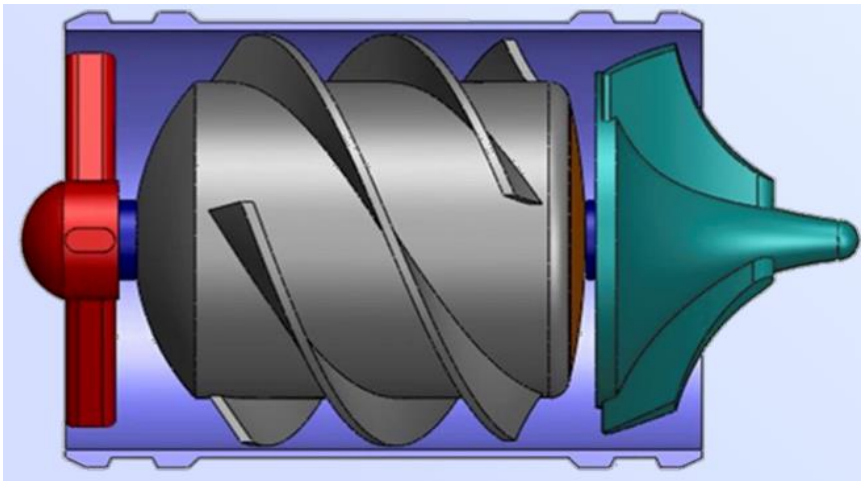




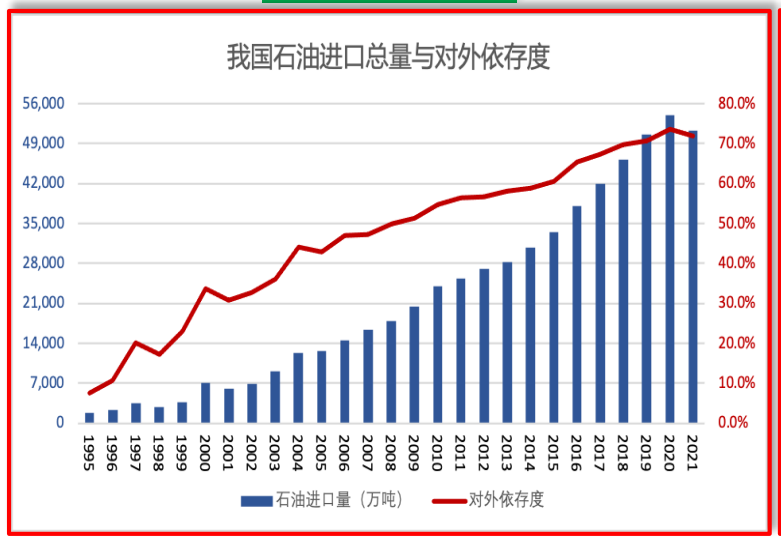






# Compositional Flows in Porous Media

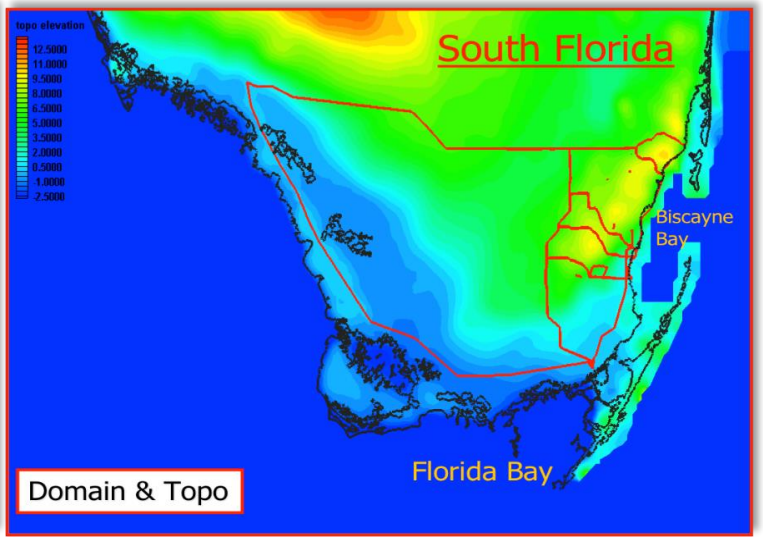
能源困局

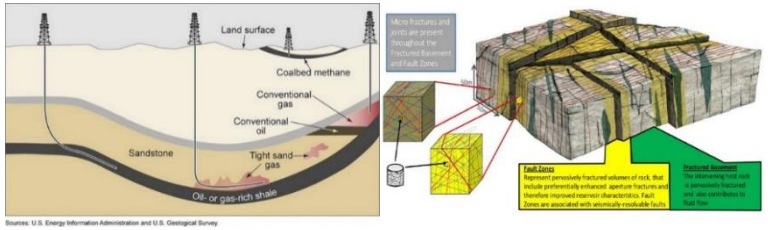



双碳目标

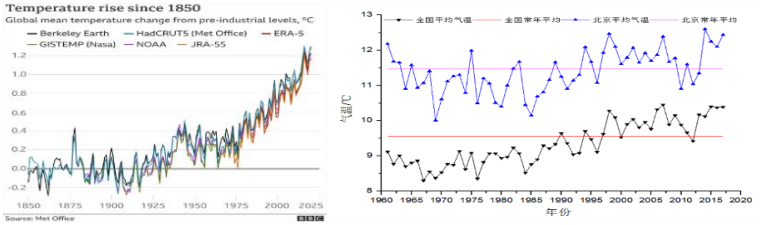

BPEnergy，中国生态环境部，中金公司研究部

污染治理



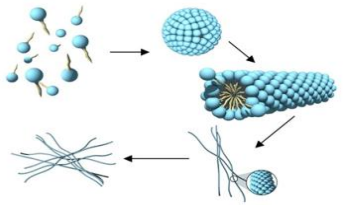

传统油藏、页岩油气、凝析气藏

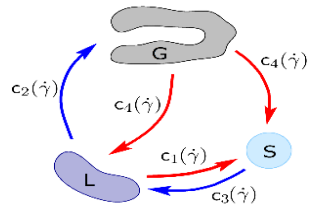






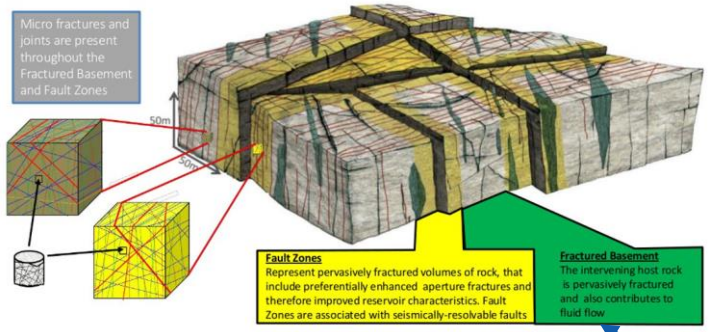
地热开发、二氧化碳埋存

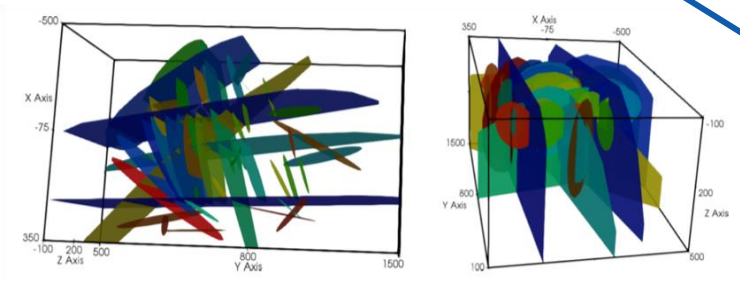


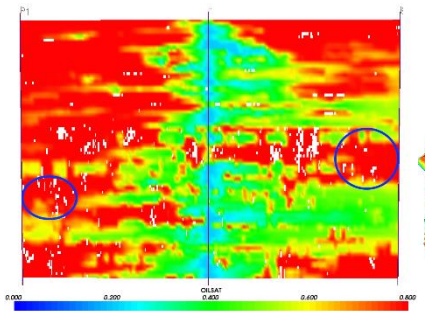


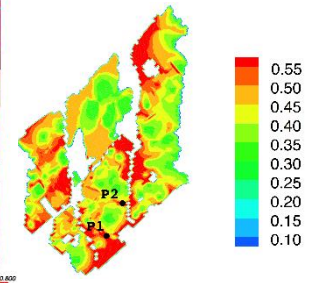

地下水污染处理、核废料埋存

# Petroleum Reservoir Development



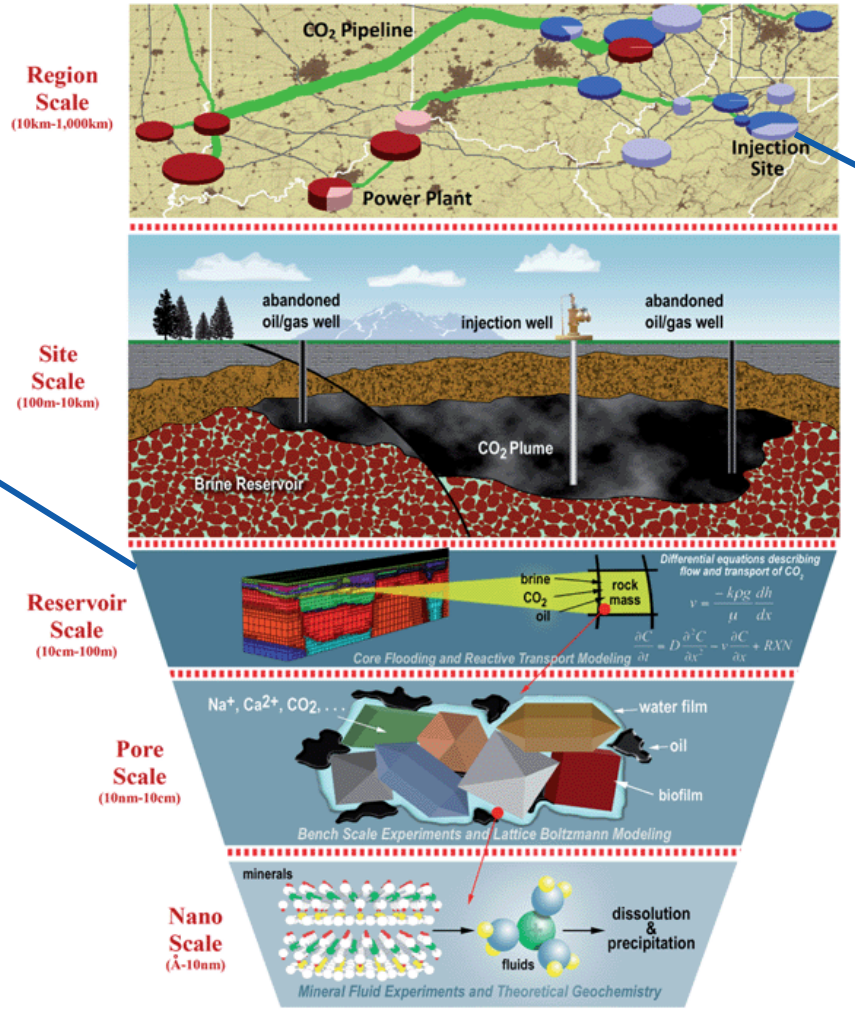



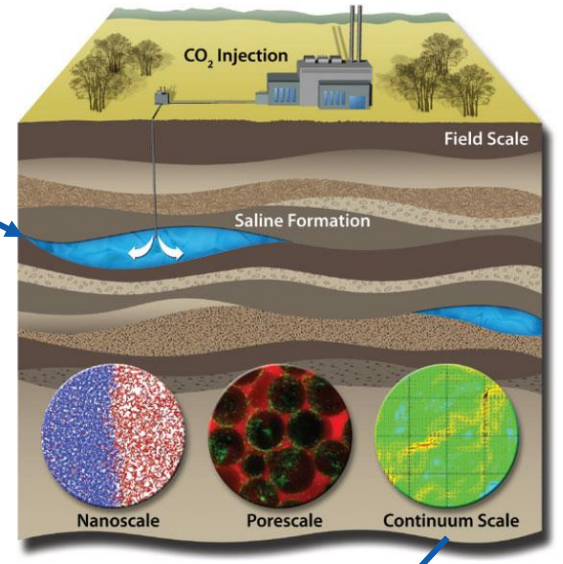



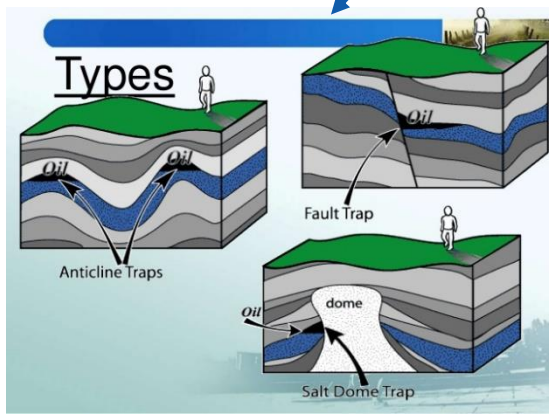









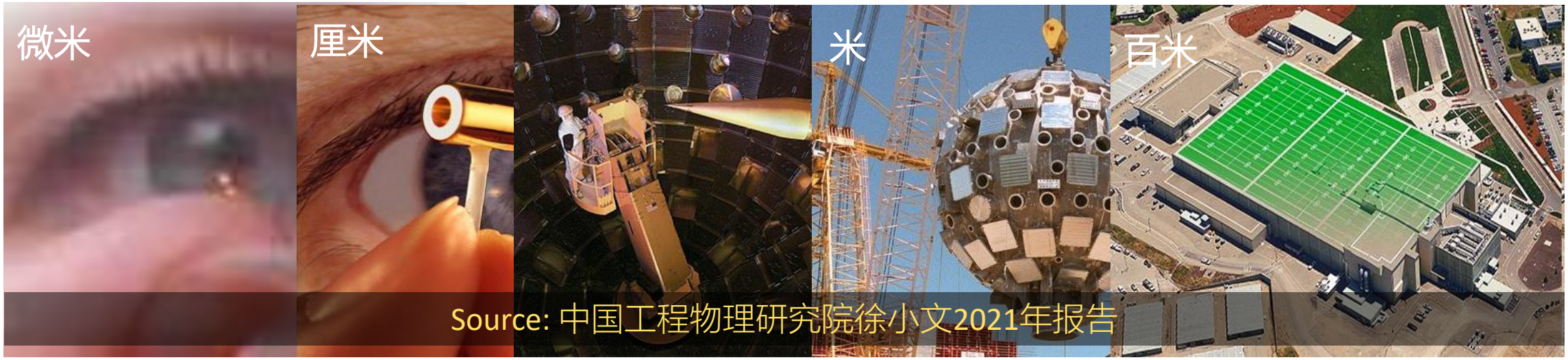




Source: Center for Petroleum & Geosystems Eng, UT Austin

# Fusion Energy

美国国家点火装置NIF(National Ignition Facility)于2021年8月8日突破记录，实现了1.35兆焦以上的能量产率，达到了触发它的激光脉冲的能量的70%，成为了最接近“点火”的实验。当核聚变反应产生的能量大于其消耗的能量时，就会发生“点火”现象，燃料可以继续自行“燃烧”，产生的能量超过引发初始反应所需的能量。

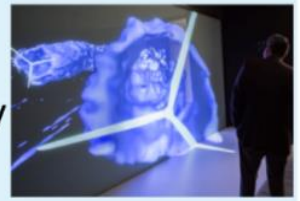
- 1994年提出
- 1997年动工
- 2008年建成于LLNL
- 2010年开始实验
- 设计总能量：1.9MJ
- 总投入已超：$35亿



# Some DoE Applications

### ExaWind: Turbine Wind Plant Efficiency

Harden wind plant design and layout against energy loss susceptibility; higher penetration of wind energy



Lead: NREL
DOE EERE

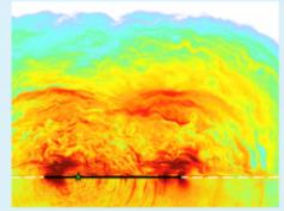
### ExaAM: Additive Manufacturing of Qualifiable Metal Parts

Accelerate the widespread adoption of AM by enabling routine fabrication of qualifiable metal parts

Lead: ORNL
DOE NNSA / EERE

### EQSIM: Earthquake Hazard Risk Assessment

Replace conservative and costly earthquake retrofits with safe purpose-fit retrofits and designs

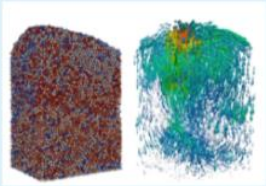


Lead: LBNL
DOE NNSA / NE, EERE

Source: DoE report, https://www.exascaleproject.org/

### MFIX-Exa: Scale-up of Clean Fossil Fuel Combustion

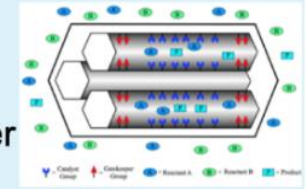
Commercial-scale demo of transformational energy technologies - curbing $CO_2$ emissions at fossil fuel power plants by 2030



Lead: NETL
DOE EERE

### GAMESS: Biofuel Catalyst Design

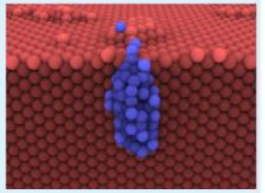
Design more robust and selective catalysts orders of magnitude more efficient at temperatures hundreds of degrees lower



Lead: Ames
DOE BES

### EXAALT: Materials for Extreme Environments

Simultaneously address time, length, and accuracy requirements for predictive microstructural evolution of materials



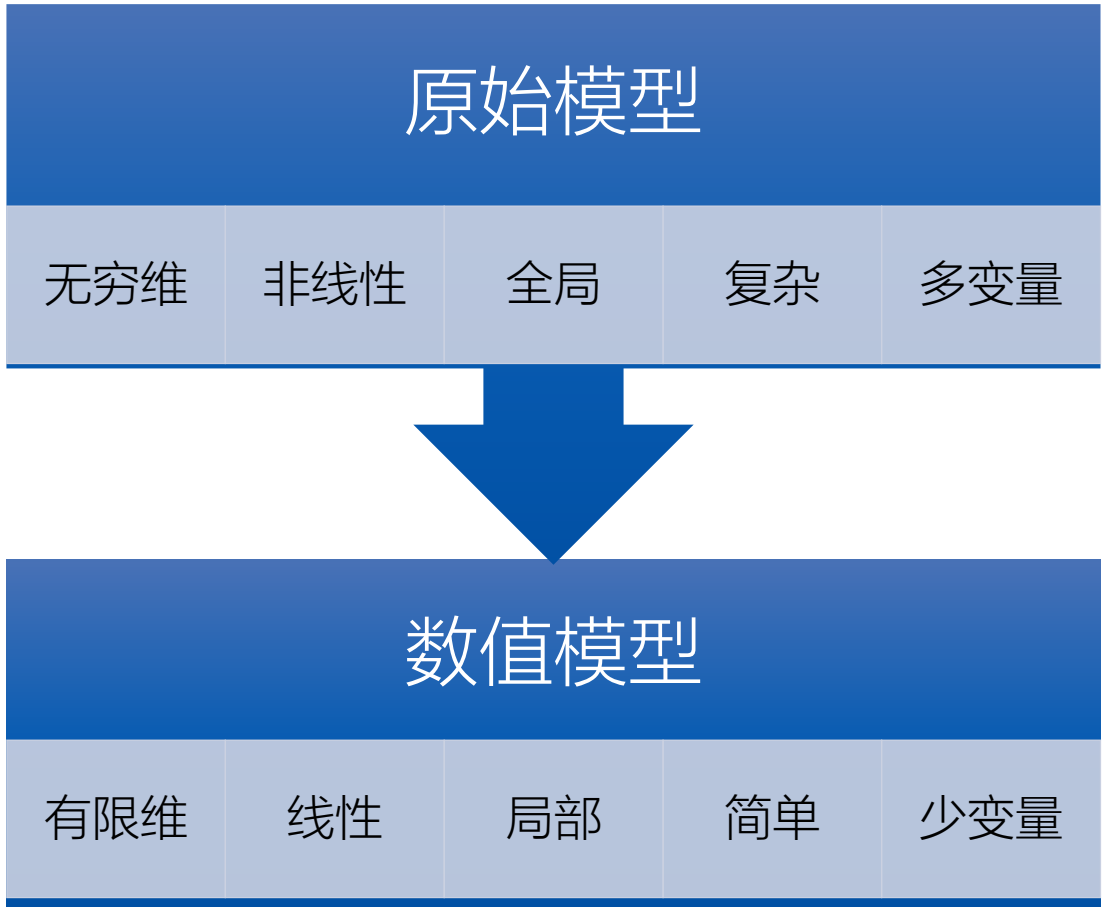
Lead: LANL
DOE BES, FES, NE

# Efficient Numerical Simulation

| | | | |
|---|---|---|---|
| 建筑设计 | 结构力学 | 复杂区域 | 可计算建模 |
| 聚变裂变 | 流体力学 | 复杂方程 | 网格生成 |
| 汽车设计 | 渗流力学 | 多物理场 | 自适应方法 |
| 航空航天 | 热力学 | 强非线性 | 高阶方法 |
| 集成电路 | 电动力学 | 高维问题 | 快速解法器 |
| 油气开发 | 天体力学 | 超大规模 | 并行计算 |
| 生物制药 | 声学 | 不确定性 | 不确定性量化 |
| 环境污染 | 化学反应 | 反复计算 | 代理模型 |
| ⋮ | ⋮ | ⋮ | ⋮ |

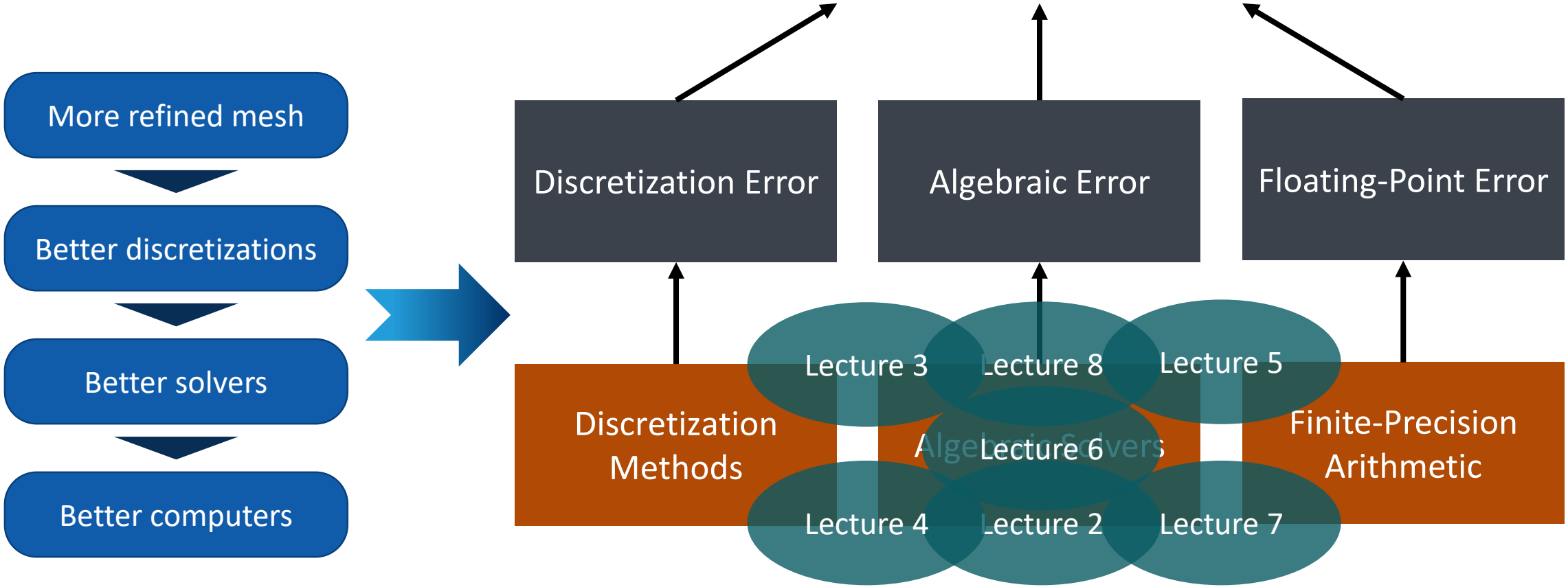
# Wisdoms in Numerical Simulation

## 数值方法设计思路




# Sources of Error in Simulation

Approximation: $u(x) = U_h(x) + \mathcal{E}_{\text{dis}} + \mathcal{E}_{\text{alg}} + \mathcal{E}_{\text{fp}}$

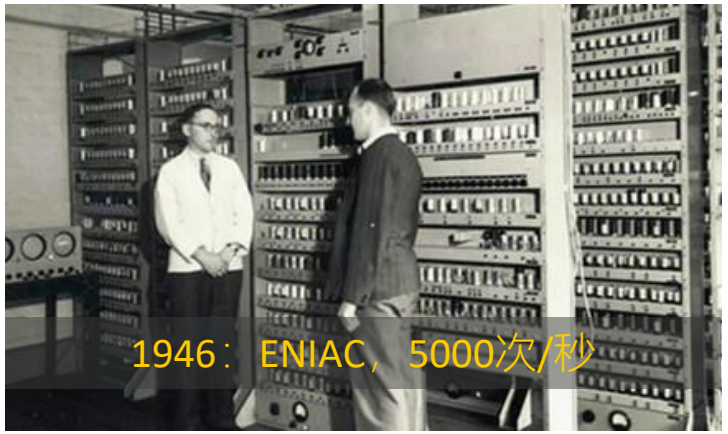


# Challenges in Large Simulation

Main difficulties and concerns in large-scale simulation

/02

# 75 Year, 14 Magnitudes


1946：ENIAC，5000次/秒


1976：Cray-I，亿次/秒


2020：富岳，超50亿亿次/秒


1959：104机，1万次/秒


1983：银河-I，亿次/秒

百亿亿次（1018 次/秒）花落谁家？

Source: 中国工程物理研究院徐小文2021年报告

# Performance “Walls”


Road To Exascale

- Hardware
- Software
- Applications

| | Today's Systems | Predicted Exascale Systems* | Factor Improvement |
|---|---|---|---|
| System Peak | $10^{16}$ flops/s | $10^{18}$ flops/s | 100 |
| Node Memory Bandwidth | $10^{2}$ GB/s | $10^{3}$ GB/s | 10 |
| Interconnect Bandwidth | $10^{1}$ GB/s | $10^{2}$ GB/s | 10 |
| Memory Latency | $10^{-7}$ s | $5 \cdot 10^{-8}$ s | 2 |
| Interconnect Latency | $10^{-6}$ s | $5 \cdot 10^{-7}$ s | 2 |

*Sources: from P. Beckman (ANL), J. Shalf (LBL), and D. Unat (LBL)

- Memory wall: bandwidth/latency of the channel between CPU and RAM
- Power / energy wall: the chip’s overall temperature and power density is high

$$\text{Dynamic Power} = (\text{Activity Factor}) \cdot (\text{Capacitance}) \cdot (\text{Voltage})^2 \cdot (\text{Frequency})$$ **+ Power Leakage**

- Instruction-level parallelism (ILP) wall: availability of parallel instructions for a processor

# Dramatic Architecture Changes for Exascale


Source: 中国工程物理研究院徐小文2019年报告

- 2009-2014：围绕E级计算，DOE组织数十次战略研讨，影响了后续100P/1000P系统的布局以及相关研究计划的部署
- 2015：奥巴马签署国家战略计算规划 (NSCI) 总统令，要求保持美国在超级计算领域的核心竞争力，开始对中国的超算中心实施禁运
- 2016：DOE和NNSA启动百亿亿次计算攻关计划（ECP），全面推进百亿亿次"应用-软件-硬件"协同研发，确保2023年左右实现国家战略安全领域的百亿亿次计算能力
- 2019：DOE签订了三台E级机的采购合同（$18亿硬件研发费用 + $18亿软件开发费用）

K: infrastructure, culture, portable, reusable, composable, interoperable, reusable

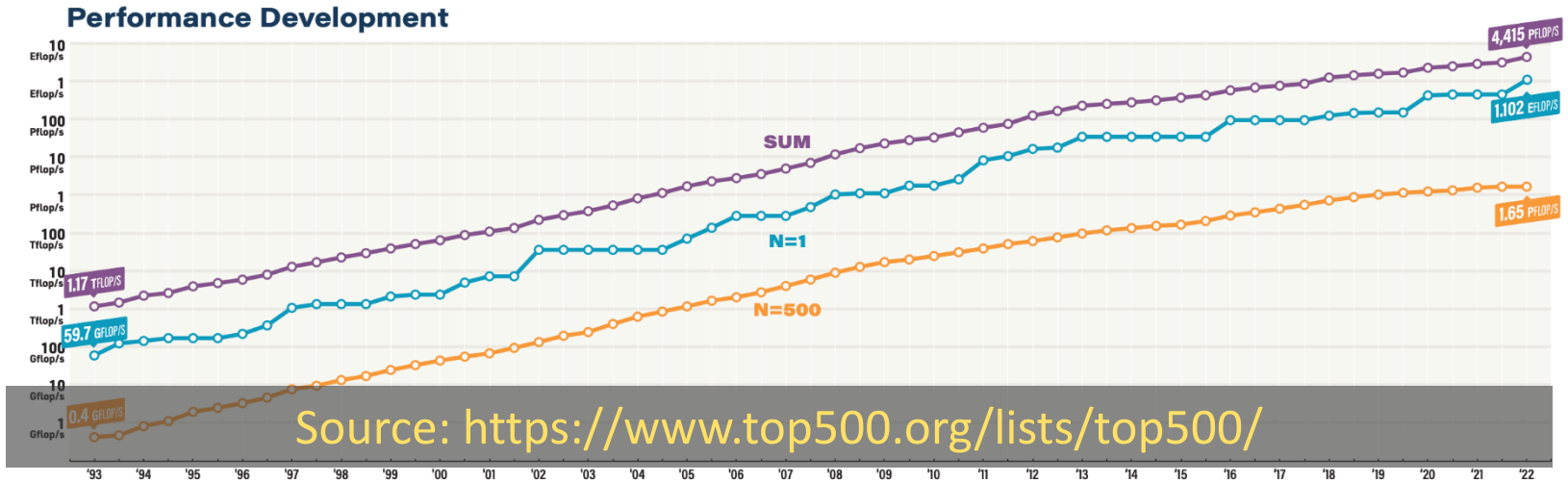
# HPC Top500 List 2022.06

TOP 500 The List.

PRESENTED BY ICL INNOVATIVE COMPUTING LABORATORY · BERKELEY LAB Lawrence Berkeley National Laboratory · ISC GROUP Moving Forward.

FIND OUT MORE AT top500.org

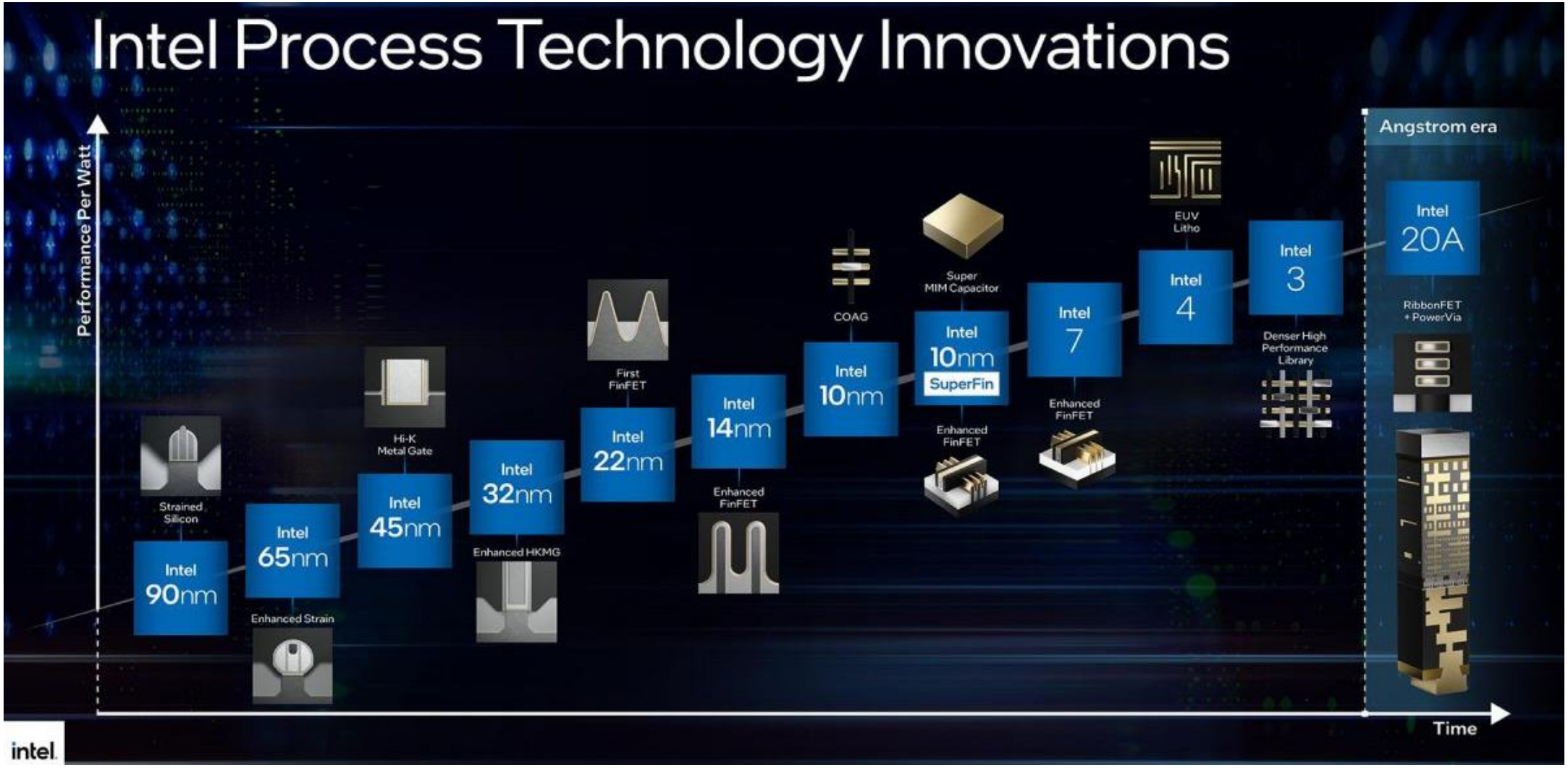
| JUNE 2022 | SYSTEM | SPECS | SITE | COUNTRY | CORES | RMAX PFLOP/S | POWER MW |
|---|---|---|---|---|---|---|---|
| 1 | Frontier | HPE Cray EX235a, AMD Opt 3rd Gen EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-10 | DOE/SC/ORNL | USA | 8,730,112 | 1,102.0 | 21.3 |
| 2 | Fugaku | Fujitsu A64FX (48C, 2.2GHz), Tofu Interconnect D | RIKEN R-CCS | Japan | 7,630,848 | 442.0 | 29.9 |
| 3 | LUMI | HPE Cray EX235a, AMD Opt 3rd Gen EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-10 | EuroHPC/CSC | Finland | 1,268,736 | 151.9 | 2.94 |
| 4 | Summit | IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband | DOE/SC/ORNL | USA | 2,414,592 | 148.6 | 10.1 |
| 5 | Sierra | IBM POWER9 (22C, 3.1GHz), NVIDIA Tesla V100 (80C), Dual-Rail Mellanox EDR Infiniband | DOE/NNSA/LLNL | USA | 1,572,480 | 94.6 | 7.44 |



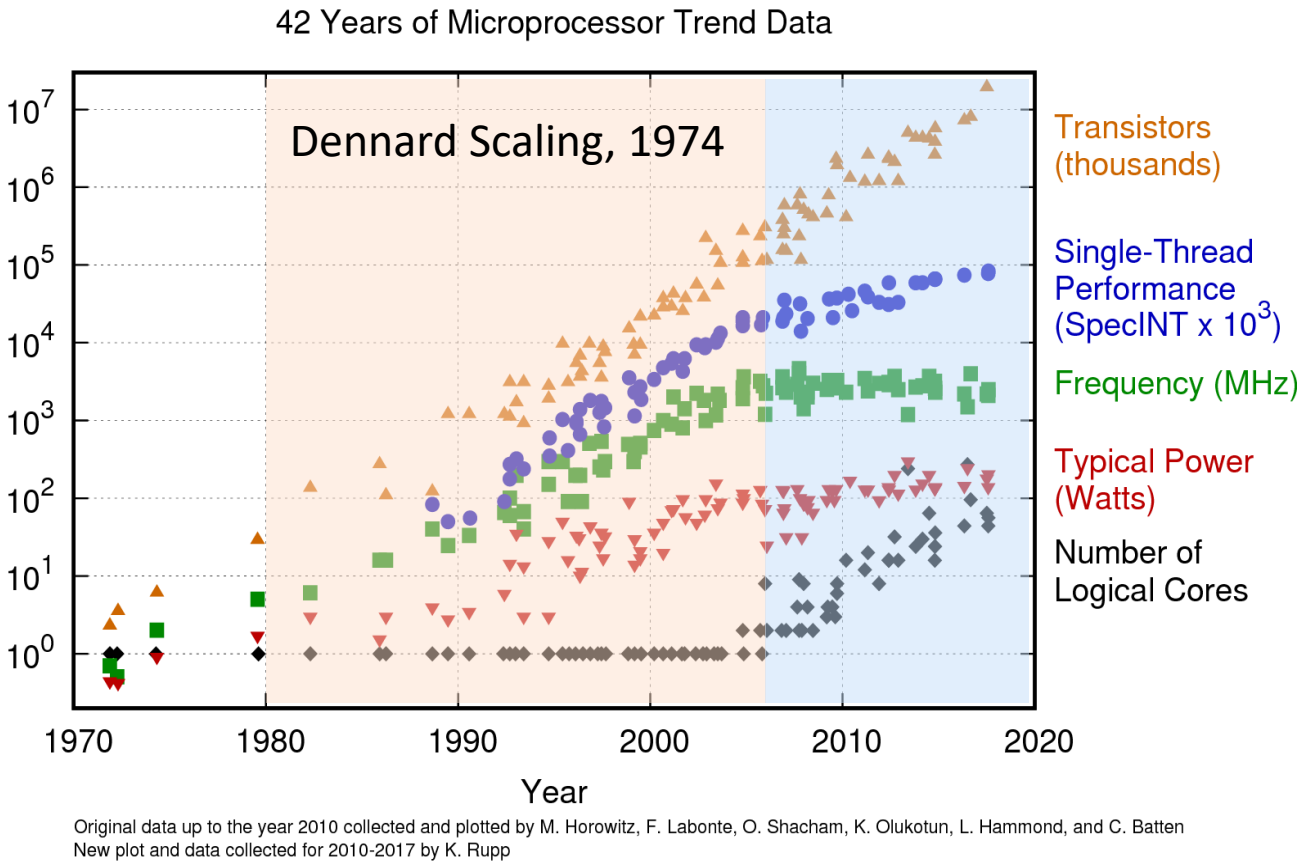

Source: https://www.top500.org/lists/top500/

- OK. Fancy!
- A lot of cores!!
- Cost a lot of money!!!
- Can this trend continue?
- Do we need HPC?
- Can we use HPC well?
- How to use HPC well?
- HPL (dense/direct)
- HPCG (sparse/iterative)
- HPL-AI (low precision)
- Green 500 (energy)

# Intel® Roadmap

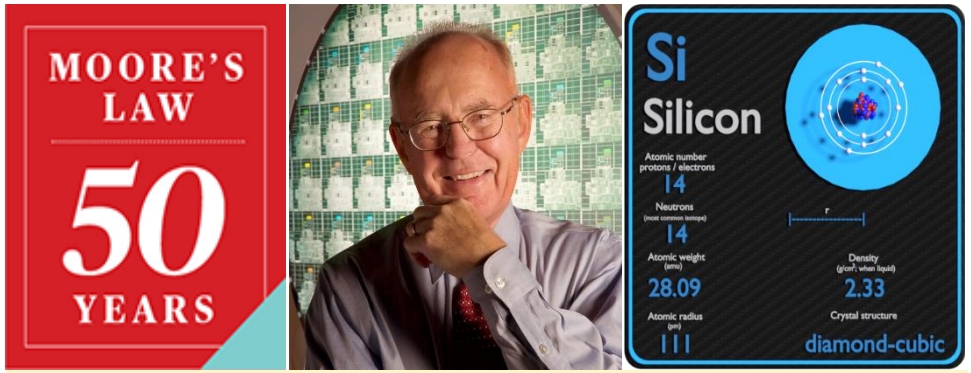


# Moore's Law




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

- "There's plenty of room at the bottom", R. Feynmann, 1959
- Moore's law, G. Moore, Intel® Co-founder, 1975

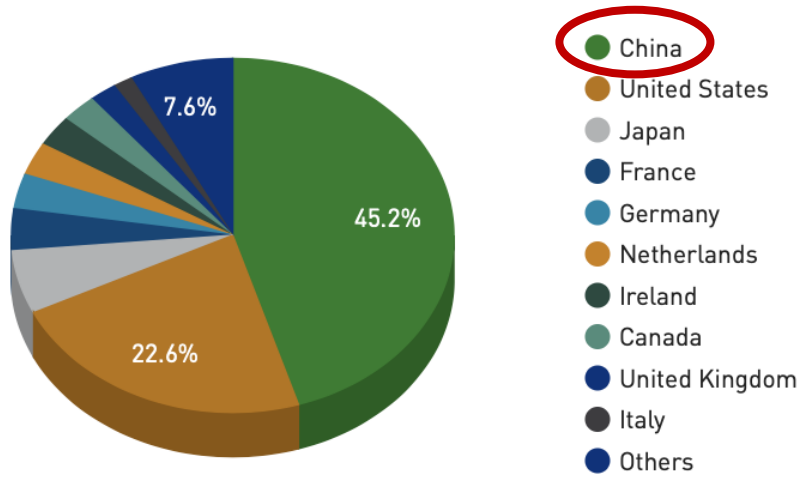



- 由于半导体技术的限制，传统CPU技术路线不能满足需要
- 瓶颈：Power、ILP、latency …
- 异构多核被广泛采用，芯片开发成本大幅提升
- sub-nano时代？硬件厂商路线图：2024年20Å，2040年代**2Å**

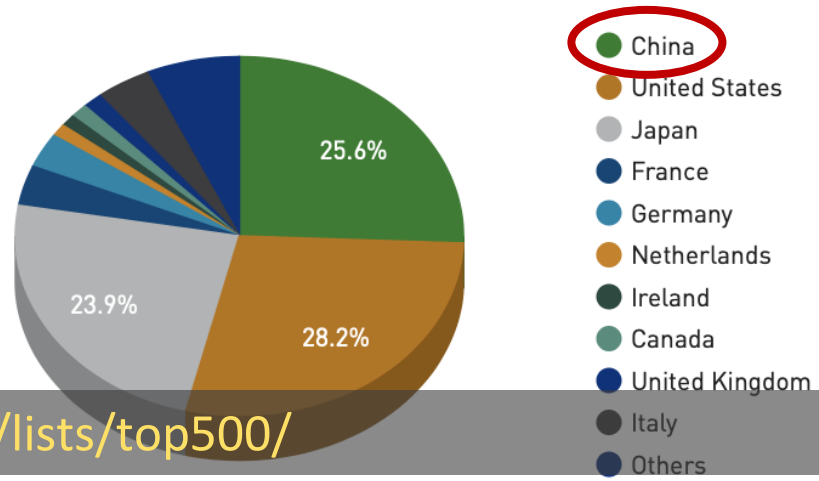
# HPC Top500 List 2021.11

| Rank | System | Cores | Rmax (TFlop/s) | Rpeak (TFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>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<br>DOE/SC/Oak Ridge National Laboratory<br>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<br>DOE/NNSA/LLNL<br>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<br>National Supercomputing Center in Wuxi<br>China | 10,649,600 | 93,014.6 | 125,435.9 | 15,371 |
| 5 | **Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE<br>DOE/SC/LBNL/NERSC<br>United States | 761,856 | 70,870.0 | 93,750.0 | 2,589 |

**Countries System Share**

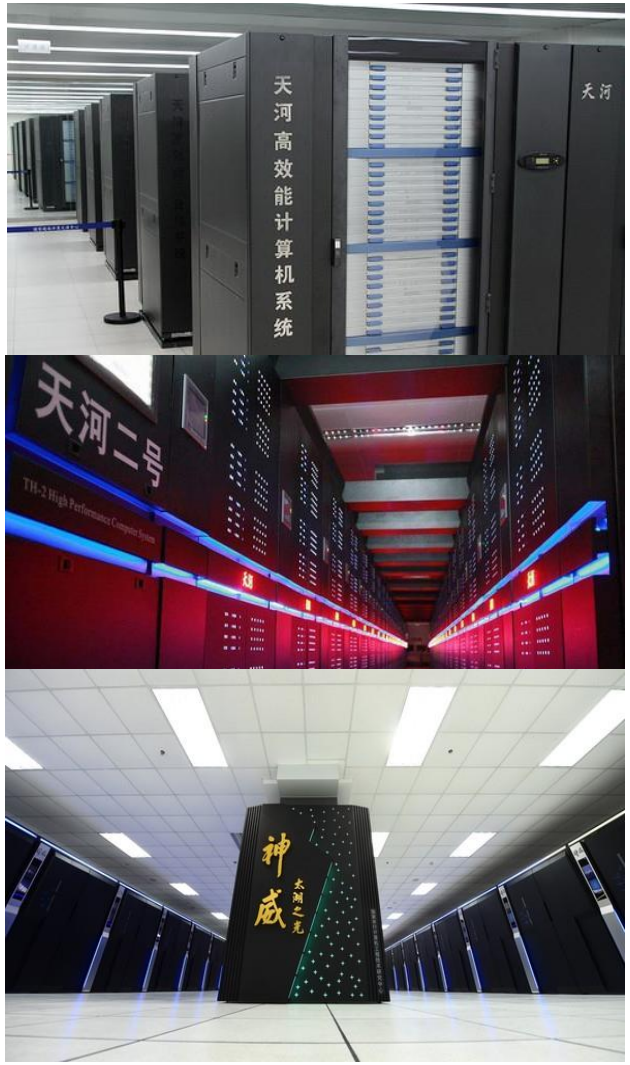


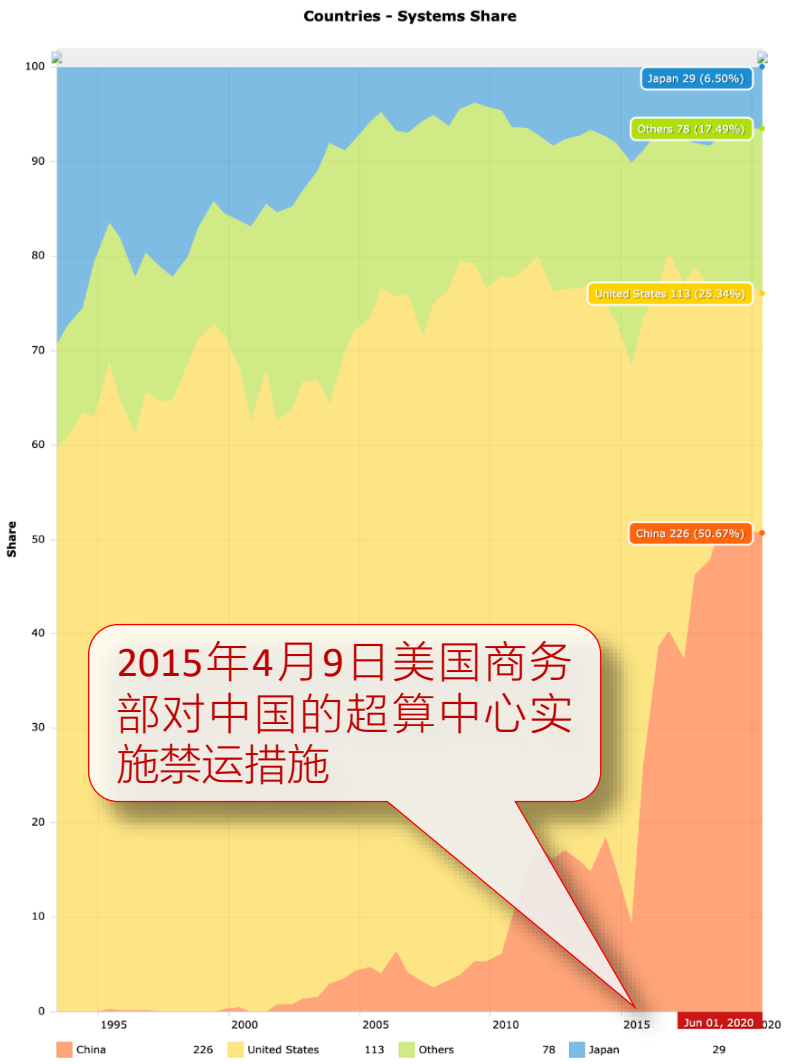

**Countries Performance Share**




Source: https://www.top500.org/lists/top500/

# Why Worried



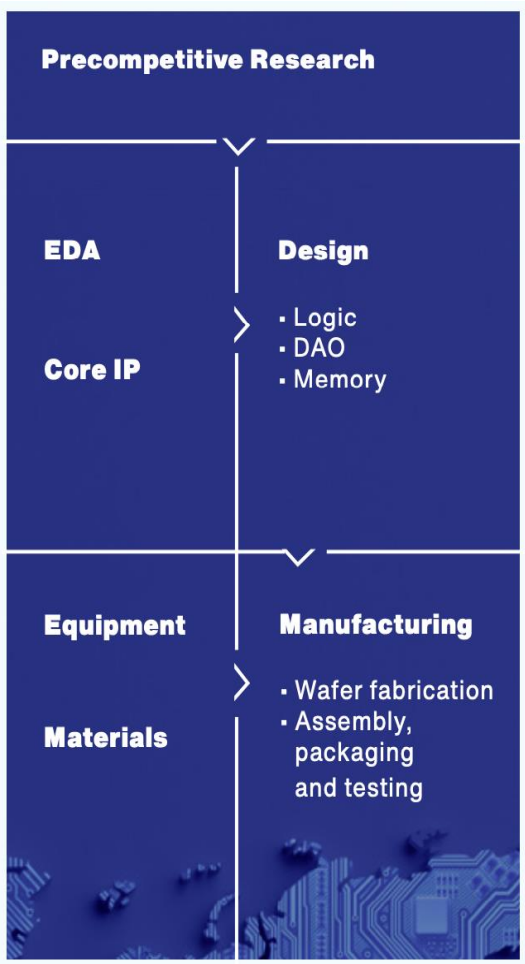



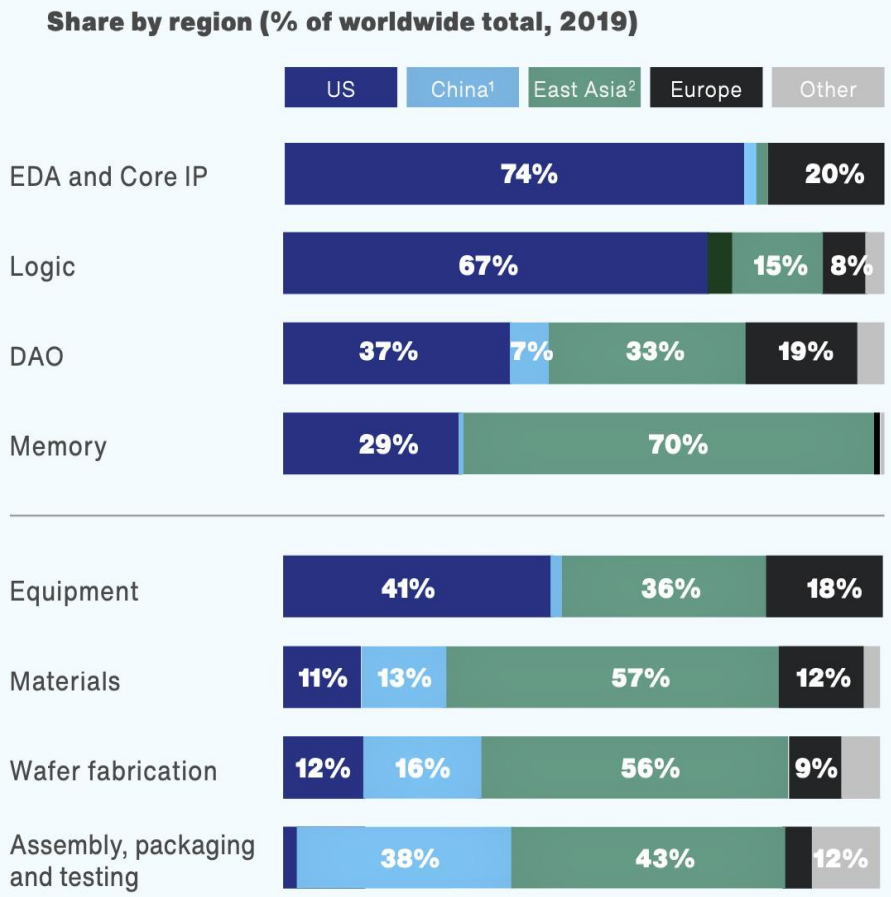

*The Wall Street Journal*：

Supercomputing is essential for the development of nuclear weapons, encryption, missile defense and more…

包括BCG在内的很多智库认为：“技术脱钩在很大程度会带来不确定性，并最终损害美国在半导体行业的国际领导地位!” 两败俱伤?

# Semiconductor Supply Chain




**Cost savings vs. fully localized "self-sufficient" supply chains:**

**$0.9-1.2T**
avoided upfront investment

**$45-125B**
annual cost efficiencies

**35-65%**
enabled reduction in semiconductor prices

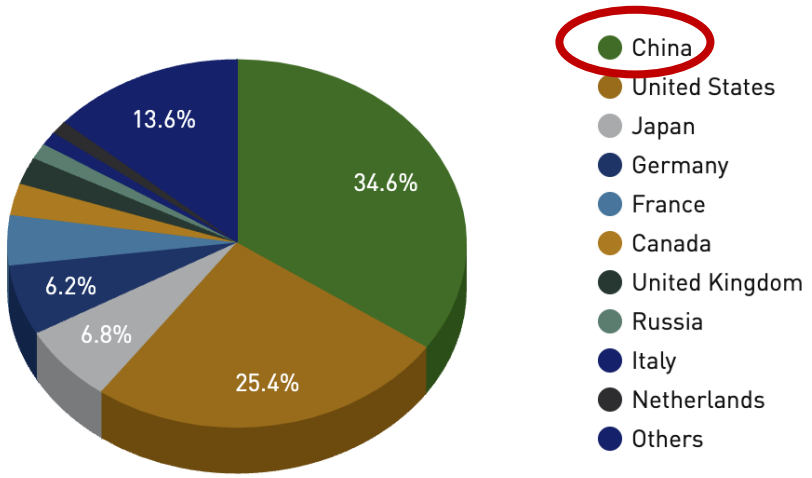
- Key components: Electronic design automation (EDA), core intellectual property (IP), chip design, advanced manufacturing equipment, …
- China is a leader in assembly, packaging, and testing, which is relatively less skill- and capital-intensive, and is investing aggressively to expand throughout the value chain

Source: BCG + SIA, "Strengthening the Global Semiconductor Supply Chain in an Uncertain Era", 2021

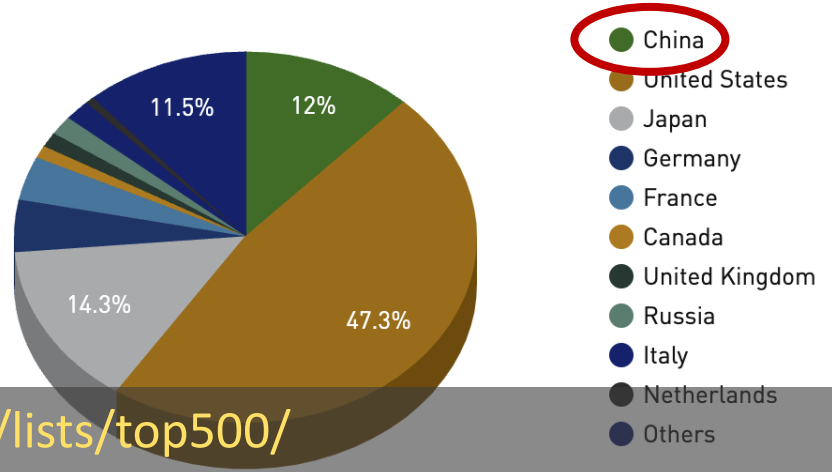
# HPC Top500 List 2022.06

| Rank | System | Cores | Rmax (PFlop/s) | Rpeak (PFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 8,730,112 | 1,102.00 | 1,685.65 | 21,100 |
| 2 | **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>Japan | 7,630,848 | 442.01 | 537.21 | 29,899 |
| 3 | **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>EuroHPC/CSC<br>Finland | 1,110,144 | 151.90 | 214.35 | 2,942 |
| 4 | **Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 2,414,592 | 148.60 | 200.79 | 10,096 |
| 5 | **Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox<br>DOE/NNSA/LLNL<br>United States | 1,572,480 | 94.64 | 125.71 | 7,438 |

**Countries System Share**




**Countries Performance Share**




Source: https://www.top500.org/lists/top500/

# HPCG and Green500

| Rank | Site | Computer | Cores | HPL Rmax (Pflop/s) | TOP500 Rank | HPCG (Pflop/s) | Fraction of Peak |
|---|---|---|---|---|---|---|---|
| 1 | RIKEN Center for Computational Science **Japan** | **Supercomputer Fugaku** — A64FX 48C 2.2GHz, Tofu interconnect D | 7,630,848 | 442.01 | 2 | 16.00 | 3.0% |
| 2 | DOE/SC/Oak Ridge National Laboratory **United States** | **Summit** — IBM POWER9 22C 3.07GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 | 2,414,592 | 148.60 | 4 | 2.926 | 1.5% |
| 3 | EuroHPC/CSC **Finland** | **LUMI** — AMD Optimized 3rd Generation EPYC 64C 2GHz, Slingshot-11, AMD Instinct MI250X | 1,110,144 | 151.90 | 3 | 1.936 | 0.9% |
| 4 | DOE/SC/LBNL/NERSC **United States** | **Perlmutter** — AMD EPYC 7763 64C 2.45GHz, Slingshot-10, NVIDIA A100 SXM4 40 GB | 761,856 | 70.87 | 7 | 1.905 | 2.0% |
| 5 | DOE/NNSA/LLNL **United States** | **Sierra** — IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 | 1,572,480 | 94.64 | 5 | 1.796 | 1.4% |
| 6 | NVIDIA Corporation **United States** | **Selene** — AMD EPYC 7742 64C 2.25GHz, Mellanox HDR Infiniband, NVIDIA A100 | 555,520 | 63.46 | 8 | 1.623 | 2.0% |
| 7 | Forschungszentrum Juelich (FZJ) **Germany** | **JUWELS Booster Module** — AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, NVIDIA A100 | 449,280 | 44.12 | 11 | 1.275 | 1.8% |
| 8 | Saudi Aramco **Saudi Arabia** | **Dammam-7** — Xeon Gold 6248 20C 2.5GHz, InfiniBand HDR 100, NVIDIA Tesla V100 SXM2 | 672,520 | 22.40 | 18 | 0.881 | 1.6% |
| 9 | Eni S.p.A. **Italy** | **HPC5** — Xeon Gold 6252 24C 2.1GHz, Mellanox HDR Infiniband, NVIDIA Tesla V100 | 669,760 | 35.45 | 12 | 0.860 | 1.7% |
| 10 | Information Technology Center, University of Tokyo **Japan** | **Wisteria/BDEC-01 (Odyssey)** — A64FX 48C 2.2GHz, Tofu interconnect D | 368,640 | 22.12 | 20 | 0.818 | 3.2% |

http://hpcg-benchmark.org/

| Rank | TOP500 Rank | System | Cores | Rmax (PFlop/s) | Power (kW) | Energy Efficiency (GFlops/watts) |
|---|---|---|---|---|---|---|
| 1 | 29 | **Frontier TDS** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States | 120,832 | 19.20 | 309 | 62.684 |
| 2 | 1 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States | 8,730,112 | 1,102.00 | 21,100 | 52.227 |
| 3 | 3 | **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland | 1,110,144 | 151.90 | 2,942 | 51.629 |
| 4 | 10 | **Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France | 319,072 | 46.10 | 921 | 50.028 |
| 5 | 326 | **MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect [illegible] Preferred Networks Japan | 1,664 | 2.18 | 53 | 40.901 |

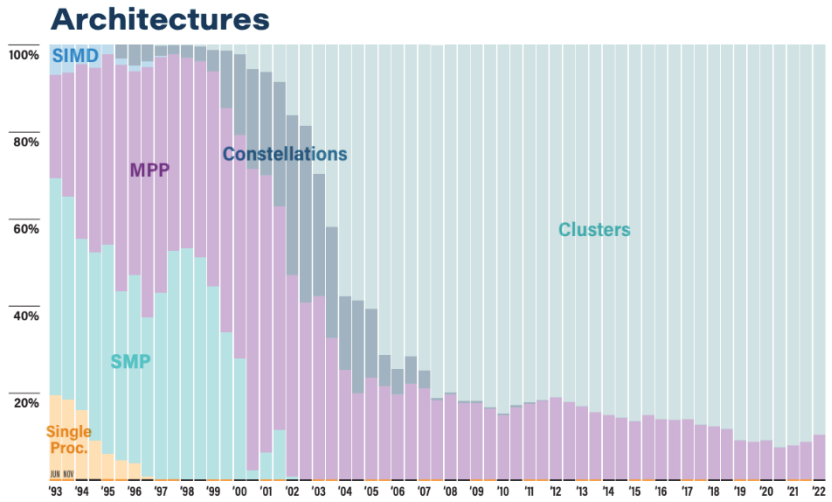
https://www.top500.org/lists/green500/

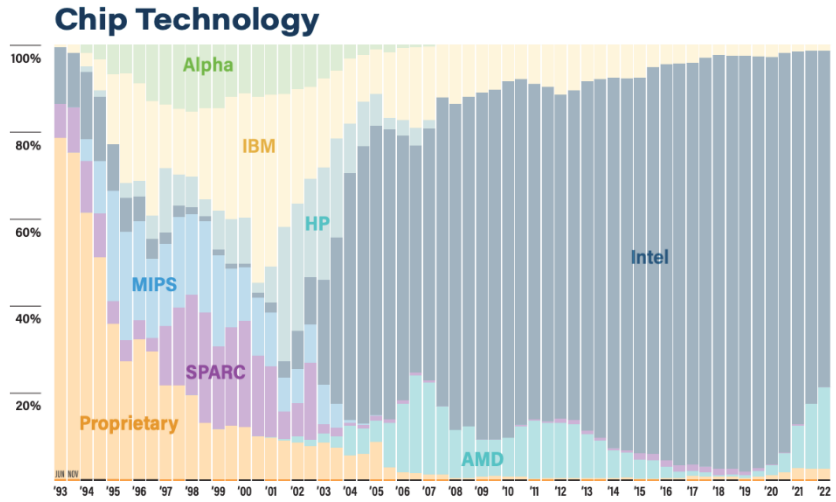
# HPL-AI

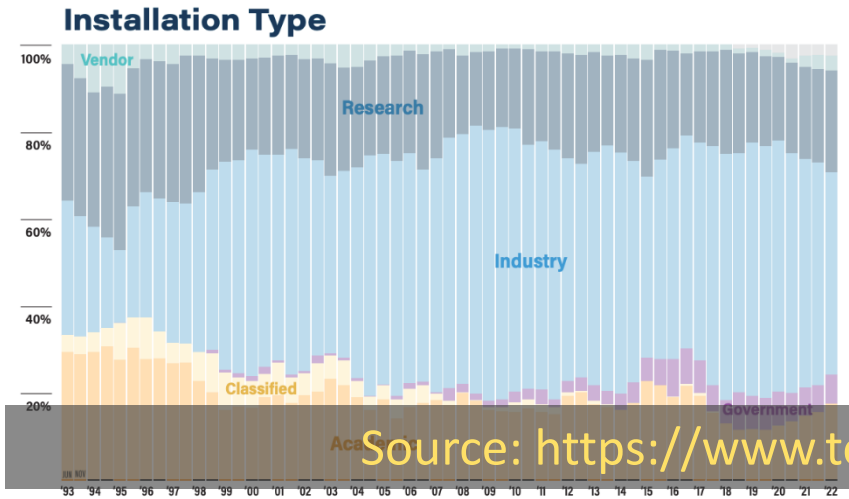
| Rank | Site | Computer | Cores | HPL-AI (Eflop/s) | TOP500 Rank | HPL Rmax (Eflop/s) | Speedup |
|---|---|---|---|---|---|---|---|
| 1 | DOE/SC/ORNL, USA | Frontier | 8,730,112 | 6.861 | 1 | 1.102 | 6.2 |
| 2 | RIKEN, Japan | Fugaku | 7,630,848 | 2.000 | 2 | 0.4420 | 4.5 |
| 3 | DOE/SC/ORNL, USA | Summit | 2,414,592 | 1.411 | 4 | 0.1486 | 9.5 |
| 4 | NVIDIA, USA | Selene | 555,520 | 0.630 | 8 | 0.0630 | 9.9 |
| 5 | DOE/SC/LBNL, USA | Perlmutter | 761,856 | 0.590 | 7 | 0.0709 | 8.3 |
| 6 | FZJ, Germany | JUWELS BM | 449,280 | 0.470 | 11 | 0.0440 | 10.0 |
| 7 | University of Florida, USA | HiPerGator | 138,880 | 0.170 | 34 | 0.0170 | 9.9 |
| 8 | SberCloud, Russia | Christofari Neo | 98,208 | 0.123 | 47 | 0.0120 | 10.3 |
| 9 | DOE/SC/ANL, USA | Polaris | 259,840 | 0.114 | 14 | 0.0238 | 4.8 |
| 10 | ITC, Japan | Wisteria | 368,640 | 0.100 | 20 | 0.0220 | 4.5 |
| 11 | NSC, Sweden | Berzelius | 59,520 | 0.052 | 103 | 0.0053 | 9.5 |
| 12 | Cyfronet, Poland | Athena | 47,616 | 0.050 | 103 | 0.0051 | 10.1 |
| 13 | Nagoya, Japan | Flow Type I | 110,592 | 0.030 | 82 | 0.0066 | 4.5 |
| 14 | NVIDIA, USA | Tethys | 19,840 | 0.024 | 320 | 0.0023 | 10.8 |
| 15 | NVIDIA, USA | DGX Saturn V | 87,040 | 0.022 | 128 | 0.0040 | 5.5 |
| 16 | CloudMTS, Russia | MTS GROM | 19,840 | 0.015 | 319 | 0.0023 | 6.6 |
| 17 | Calcul Quebec/Compute Canada | Narval | 76,320 | 0.014 | 93 | 0.0059 | 2.4 |
| 18 | DOE/SC/ANL, USA | ThetaGPU | 280,320 | 0.012 | 79 | 0.0069 | 1.7 |
| 19 | Indiana University, USA | Big Red 200 GPU | 31,744 | 0.006 | 160 | 0.0026 | 2.4 |
| 20 | Texas A&M University, USA | Grace GPU | 26,400 | 0.004 | 366 | 0.0021 | 1.7 |

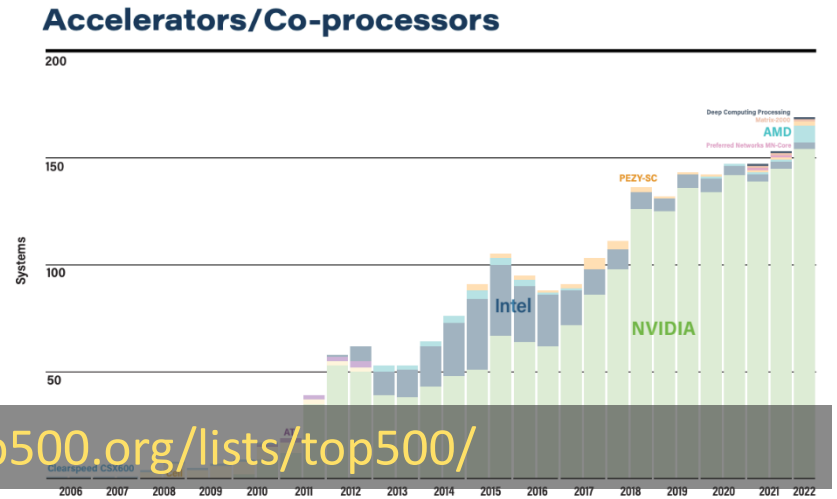
https://hpl-ai.org/

# Some Statistics on Architectures



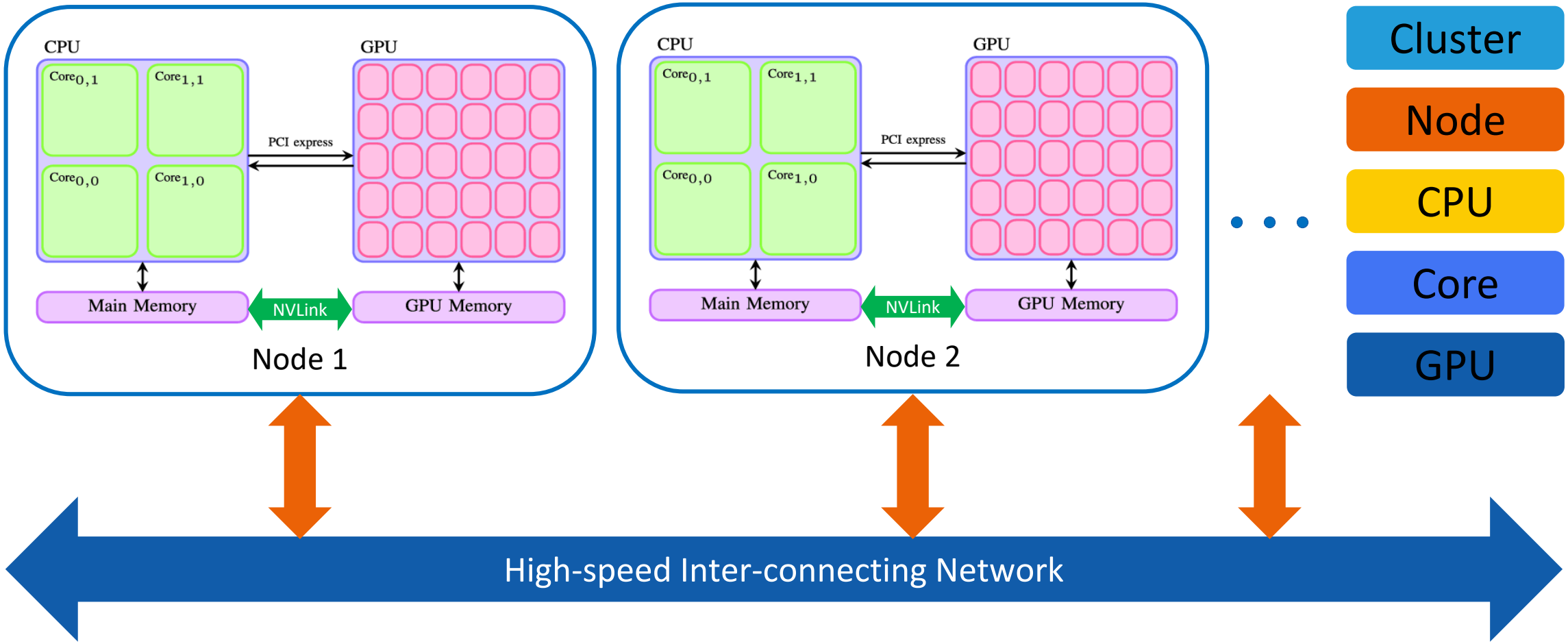




Source: https://www.top500.org/lists/top500/

- GPUs and other accelerators are more and more popular
- nVidia® dominates the HPC accelerator market
- Reasons:
  - Hardware
  - Software
  - Community

# Heterogenous Cluster Architecture




算法的复杂度和可扩展性、对多核CPU的利用率、对众核加速卡的利用率、数据传输时间……

# Fast Algebraic Solvers

How to develop faster and/or better solvers?

# Computational Complexity

Linearization ➔ time marching ➔ spatial discretization ➔ linear solver ➔ parallelization

Q: What prevents us from achieving better performance?

Number of nonlinear iterations

Number of time steps

Solver complexity

**Complexity** $\sim N_{model} \times N_{nl} \times N_{time} \times C_d (N_{space}^d)^k \div P$

How can we get this?

Q: Complexity is not wall-time! What else should we consider for real performance?

Number of simulation runs

Number of spatial DoF

Number of processors

Adaptivity + Fast Linear & Nonlinear Solvers + Parallelization

# Make Good Use of Supercomputers

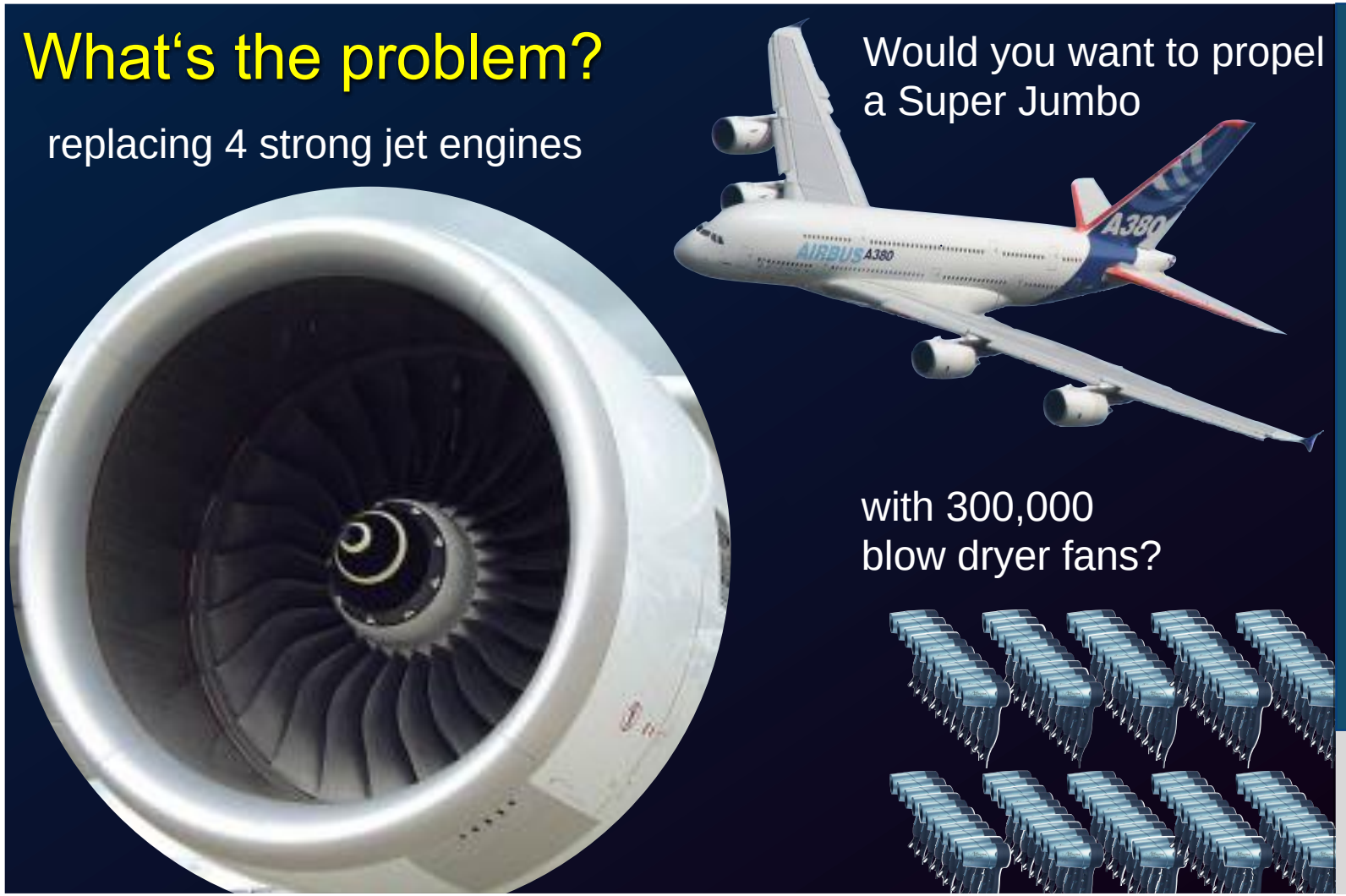
## SUNWAY TAIHULIGHT - SUNWAY MPP, SUNWAY SW26010 260C 1.45GHZ, SUNWAY

| | |
|---|---|
| **Site:** | National Supercomputing Center in Wuxi |
| **Manufacturer:** | NRCPC |
| **Cores:** | 10,649,600 |
| **Memory:** | 1,310,720 GB |
| **Processor:** | Sunway SW26010 260C 1.45GHz |
| **Interconnect:** | Sunway |
| **Performance** | |
| **Linpack Performance (Rmax)** | 93,014.6 TFlop/s |
| **Theoretical Peak (Rpeak)** | 125,436 TFlop/s |
| **Nmax** | 12,288,000 |
| **HPCG [TFlop/s]** | 480.848 |
| **Power Consumption** | |
| **Power:** | 15,371.00 kW (Submitted) |
| **Power Measurement Level:** | 2 |
| **Operating System:** | Sunway RaiseOS 2.0.5 |

Source: https://www.top500.org/lists/top500/


- Power assumption = 15MW
- Peak performance Rpeak = 125PFlops
- HPL performance Rmax = 74%
- HPCG performance = 0.38%

# An Illustration of Parallel Computing




空客A380发动机的推力约为300千牛，最大速度300米/秒，功率为90000千瓦，这大致相当于300,000台普通家用吹风机的功率，但你能用吹风机驱动A380吗？这可行吗？如果可行，带来什么变化？现在，想象一下你有1000万台不同功率的吹风机！

Ulrich Rüde, Friedrich-Alexander-University of Erlangen-Nürnberg, Erlangen, Germany

# Sequential vs Parallel Algorithms

**管理**

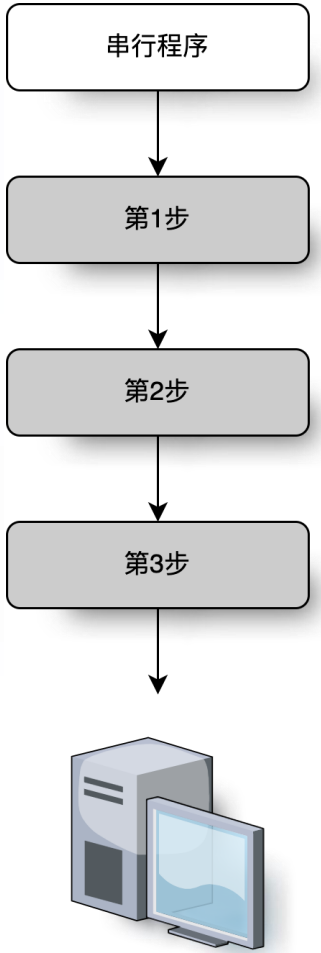
把任务分配给每个进程，让大家都有事干

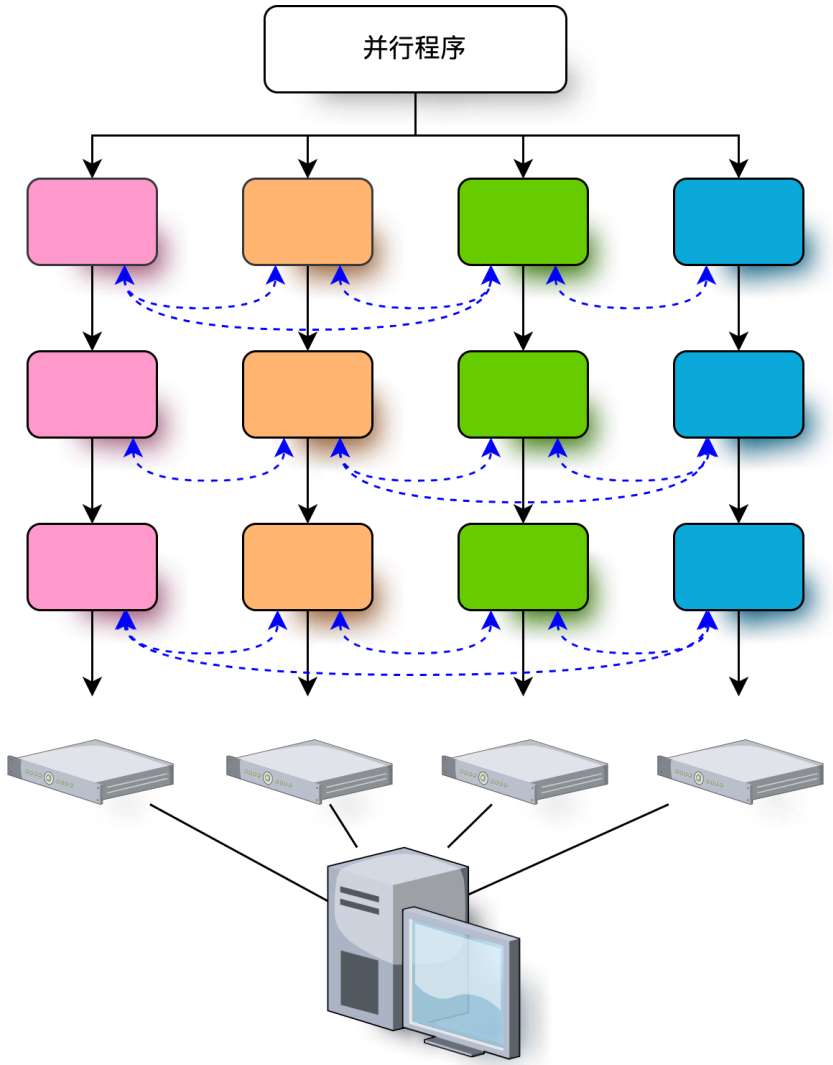
**人事**

减少进程之间交流的次数和开销

**流程**

减少数据之间的依赖关系和相互等待




为了实现整体的效率和资源利用率的最大化，可能会牺牲个人甚至每个人的利益。对应的串行算法不一定快！甚至不收敛，例如： Jacobi和GS方法。

# How to Measure Parallel Efficiency

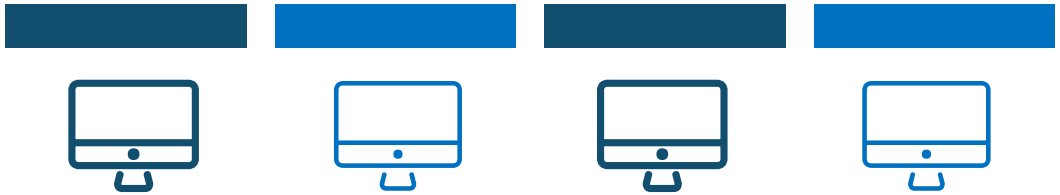
- 强可扩展性

$$\frac{W_1}{kp_1} \qquad S = \frac{T_1}{T_p} \approx k?$$

若需要解决一个任务，把计算资源增大到原来的 $k$ 倍，能加速 $k$ 倍吗?

定义并行效率 $E = S/k$，我们希望是100%!

- 弱可扩展性



$$\frac{W_k}{kp_1} \qquad T_p \approx T_1?$$

若需要解决 $k$ 倍大的任务，计算资源同比例增加，能保持相同耗时吗?

为了效率保持不变， 需要多大的问题规模?

# Three Mountains on Parallel Simulation

Amdahl's Law 1967

如果串行部分占总时间的10%，那并行加速比不可能超过10倍

对于大规模系统来说，很多应用程序的强可扩展性是难以实现的！

Gustafson-Barsis's Law 1988

对于很多应用来说，更需要的是计算越来越大规模的问题！需要是弱可扩展性，而不是强可扩展性

当问题规模与计算资源同比例增大S倍时，加速比最高就可达0.9S+0.1

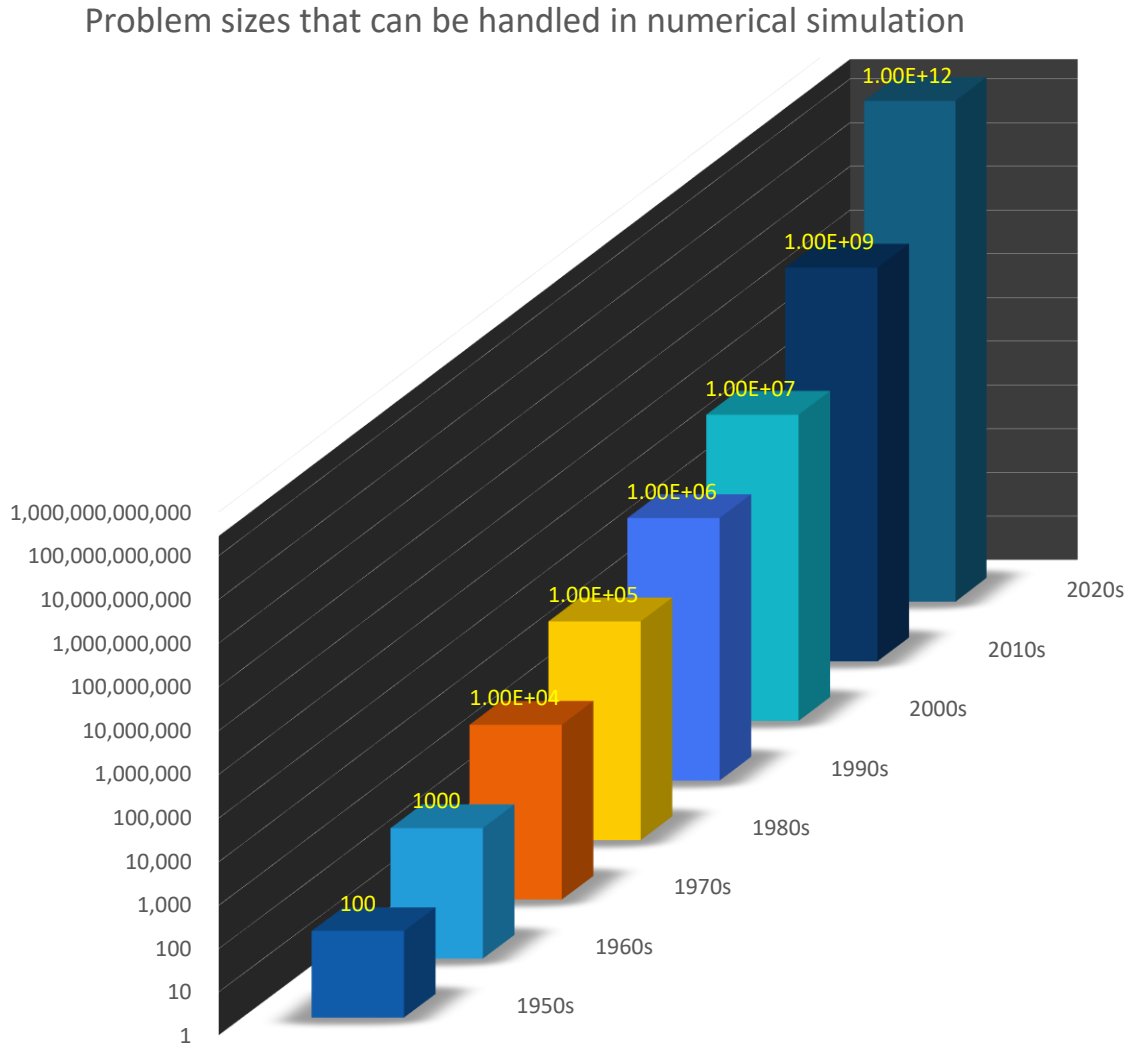
Gabriel Wittum: HPC Paradox

当购买了大10倍的硬件系统，希望能更快地求解大10倍的问题；但现实很残酷，必须有最优算法才可以！

最优算法对于充分发挥HPC效率及实现弱可扩展性至关重要！

硬件投资 >>> 并行效率 >>> 最优算法 >>> 软件投资

# Size of Simulation



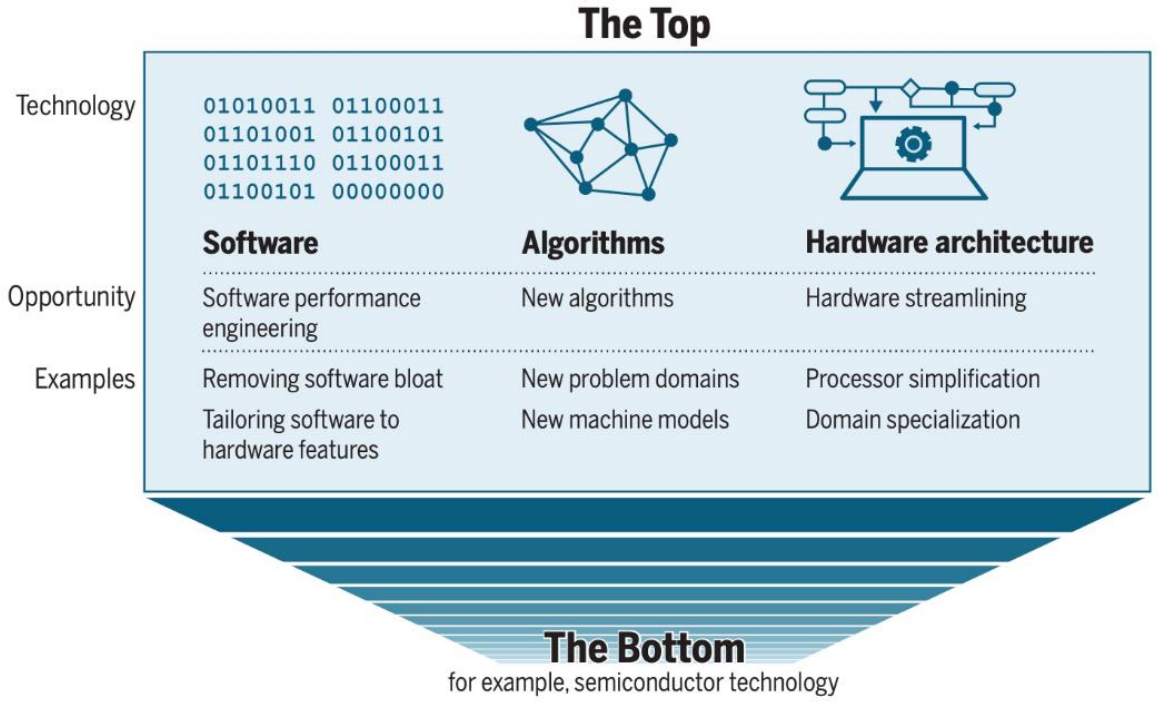

Ref: Manolis Papadrakakis, National Technical University, Greece, “Mastering the Computational Challenges for Solving Large-Scale Problems in Simulation-Based Science and Engineering”, CACM talk, 2022

## Driving Forces

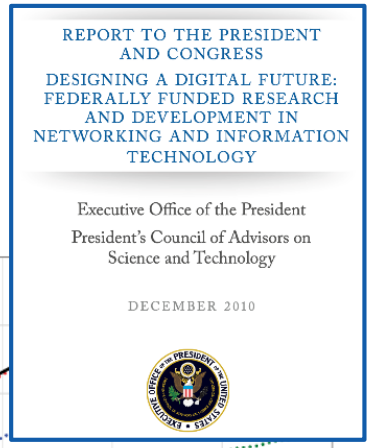
- Hardware improvement
- Algorithm development
- Demand of scientific discovery

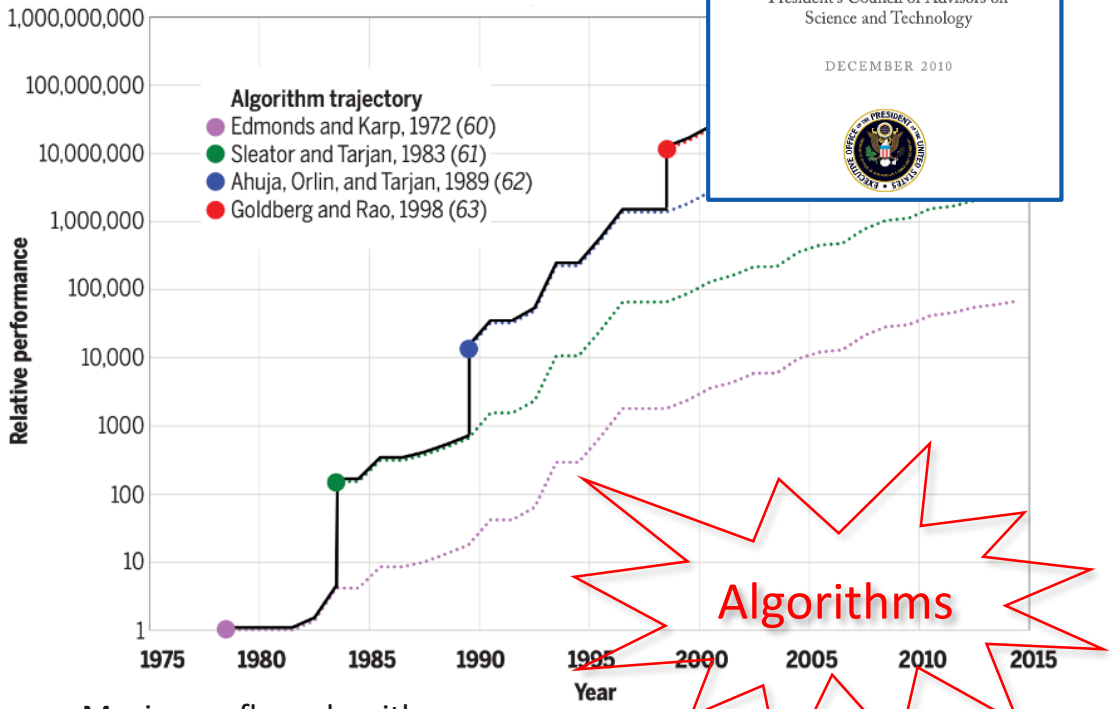
# Room for Improvement

- ~~*"There's plenty of room at the bottom"*~~, R. Feynmann, Nobel prize-winner, 1959
- *"There's plenty of room at the top"*, Leiserson, et al., Science 368, 2020




**Performance gains after Moore's law ends.** In the post-Moore era, improvements in computing power will increasingly come from technologies at the "Top" of the computing stack, not from those at the "Bottom", reversing the historical trend.




Maximum-flow algorithms

arXiv:2203.00671v2, 2022

# Implementation of Algorithms

**Table 1. Speedups from performance engineering a program that multiplies two 4096-by-4096 matrices.** Each version represents a successive refinement of the original Python code. "Running time" is the running time of the version. "GFLOPS" is the billions of 64-bit floating-point operations per second that the version executes. "Absolute speedup" is time relative to Python, and "relative speedup," which we show with an additional digit of precision, is time relative to the preceding line. "Fraction of peak" is GFLOPS relative to the computer's peak 835 GFLOPS. See Methods for more details.

| Version | Implementation | Running time (s) | GFLOPS | Absolute speedup | Relative speedup | Fraction of peak (%) |
|---|---|---|---|---|---|---|
| 1 | Python | 25,552.48 | 0.005 | 1 | — | 0.00 |
| 2 | Java | 2,372.68 | 0.058 | 11 | 10.8 | 0.01 |
| 3 | C | 542.67 | 0.253 | 47 | 4.4 | 0.03 |
| 4 | Parallel loops | 69.80 | 1.969 | 366 | 7.8 | 0.24 |
| 5 | Parallel divide and conquer | 3.80 | 36.180 | 6,727 | 18.4 | 4.33 |
| 6 | plus vectorization | 1.10 | 124.914 | 23,224 | 3.5 | 14.96 |
| 7 | plus AVX intrinsics | 0.41 | 337.812 | 62,806 | 2.7 | 40.45 |

366x

174x

MKL

- Compiler optimization; loop ordering; parallel loops
- Tiling; cache-oblivious divide-and-conquer;
- Vectorization; AVX intrinsic
- Leiserson & Shun, MIT Open Course 6.172

**Source: There's plenty of room at the Top: What will drive computer performance after Moore's law?**
*Science* **368** (6495), June, 2020
http://science.sciencemag.org/content/368/6495/eaam9744

# Scale A Practical Application

自由度数从“亿”扩展到“百亿” 意味着什么？　并不仅规模是大了 **100倍** 而已!

Bill Gropp, NCSA Director, UIUC, President of IEEE Computer Society

… helping industry break through barriers in massively parallel computing …

**2010**

**GigaPowers**
3600 CPU核，油藏模拟

**2015**

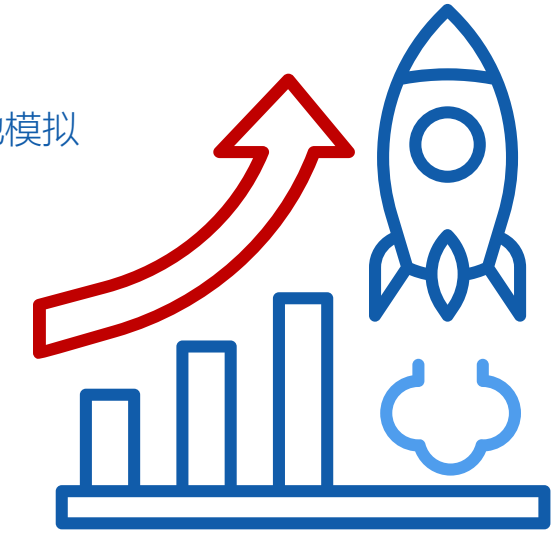
**我们的模拟器**
10000 CPU核，油藏模拟

**2016**

**TeraPowers**
150000 CPU核，盆地模拟

**2017**

**EMPowers**
716800 CPU核，油藏模拟

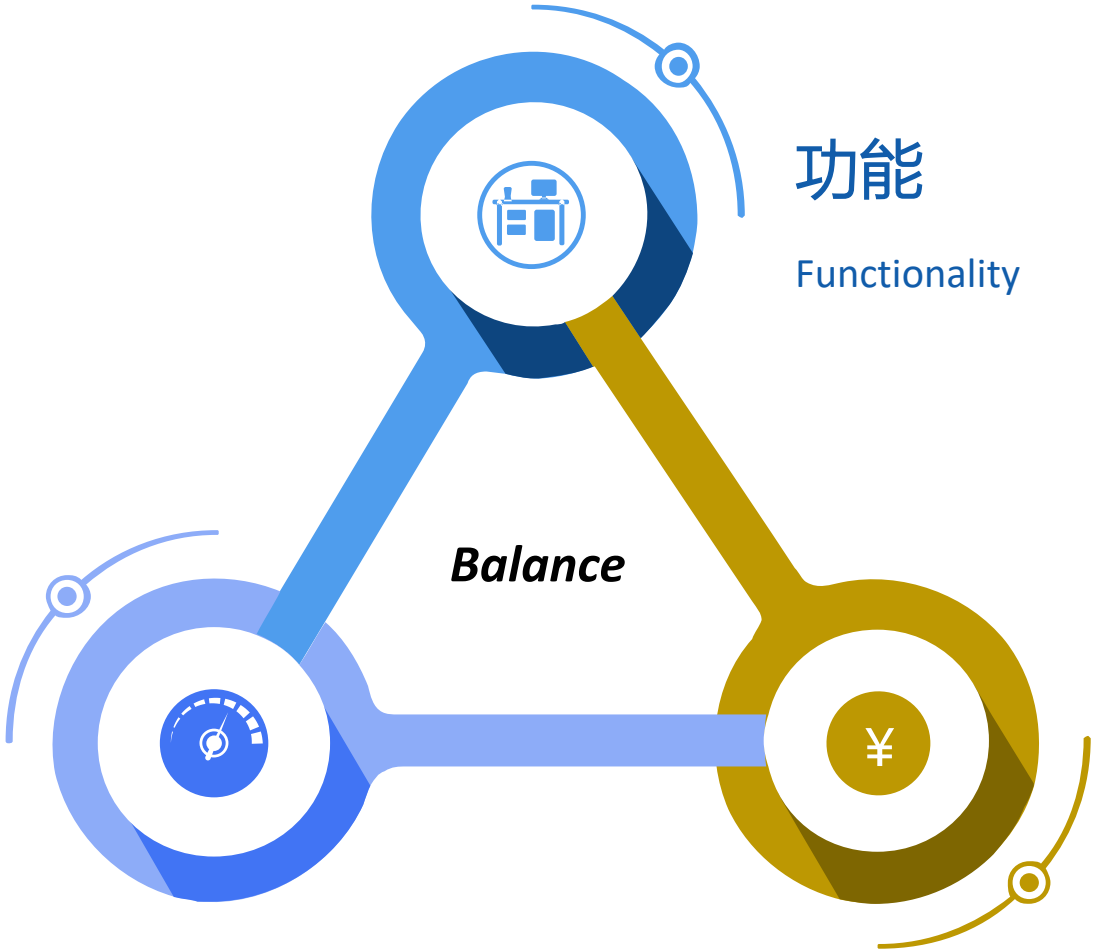


大规模油藏数值模拟器简史

- 更复杂的物理
- 更强的非线性
- 更强的耦合性
- 更小的时间步
- 更多的牛顿步
- 更多次的计算
- 更复杂的机器

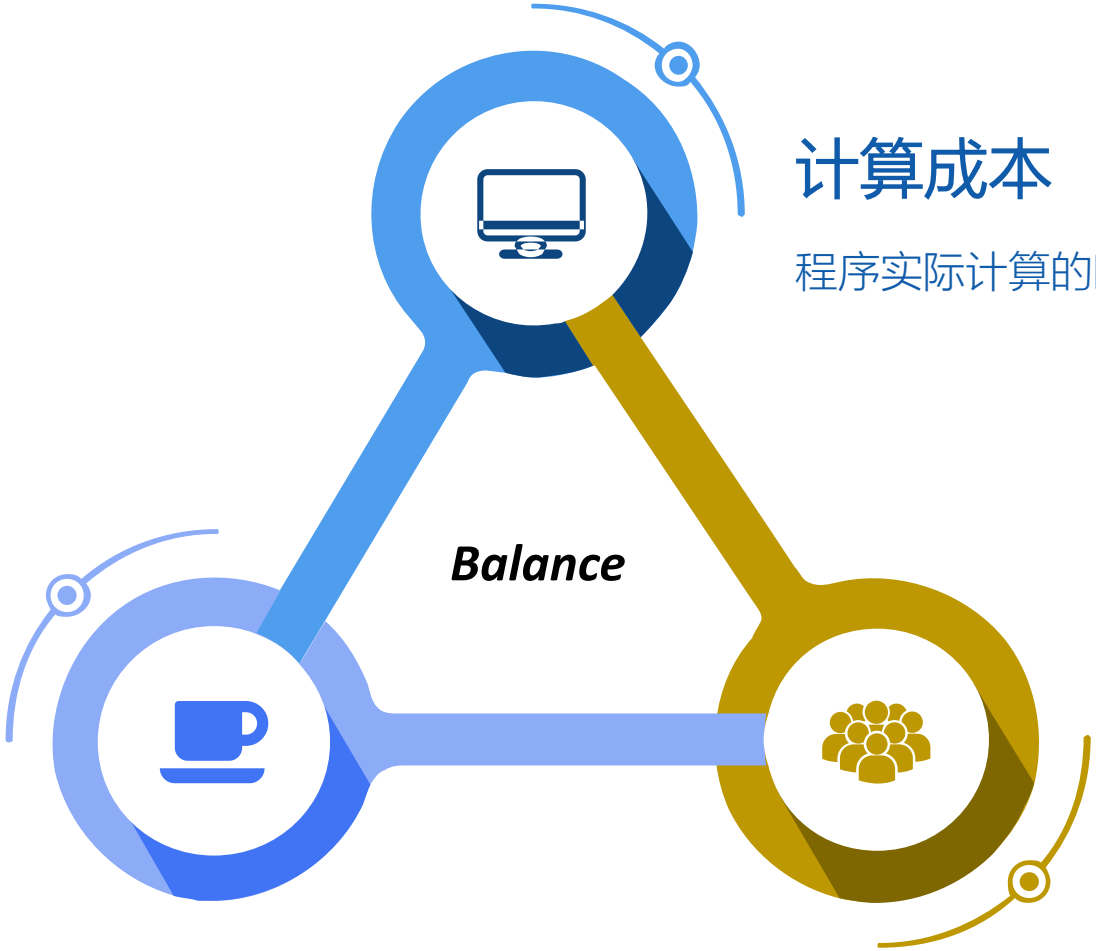
# How Much Do We Expect

三个主要"目标"



# How Much Effort To Put In

三个主要"成本"




## 计算成本

程序实际计算的时间成本

## 开发成本

程序开发与优化的人工成本

## 使用成本

用户使用软件的学习成本

# Main Focus: Algebraic Solvers


- Direct and iterative solvers for sparse linear systems
- Several methods for non-symmetric problems
- Several methods for nonlinear problems
- Algorithm precision and mixed-precision algorithms
- Communication hiding and avoiding algorithms
- Fault resilience and reliability of iterative solvers
- Robustness and adaptivity of iterative solvers

***Gordon Moore:***

**WHATEVER HAS BEEN DONE, CAN BE OUTDONE**

# Reading and Thinking

REVIEW

COMPUTER SCIENCE

## There's plenty of room at the Top: What will drive computer performance after Moore's law?

Charles E. Leiserson[1], Neil C. Thompson[1,2]*, Joel S. Emer[1,3], Bradley C. Kuszmaul[1]†, Butler W. Lampson[1,4], Daniel Sanchez[1], Tao B. Schardl[1]

The miniaturization of semiconductor transistors has driven the growth in computer performance for more than 50 years. As miniaturization approaches its limits, bringing an end to Moore's law, performance gains will need to come from software, algorithms, and hardware. We refer to these technologies as the "Top" of the computing stack to distinguish them from the traditional technologies at the "Bottom": semiconductor physics and silicon-fabrication technology. In the post-Moore era, the Top will provide substantial performance gains, but these gains will be opportunistic, uneven, and sporadic, and they will suffer from the law of diminishing returns. Big system components offer a promising context for tackling the challenges of working at the Top.

- What's your application area?
- Do you do numerical simulation?
- Do you see room for improvement in your simulation? How?
- How much time are you willing to invest on algorithm/implementation?
- Do you think "computational science" is science? Why?
- Do you think "computer science" is science? Why?

# Contact Me

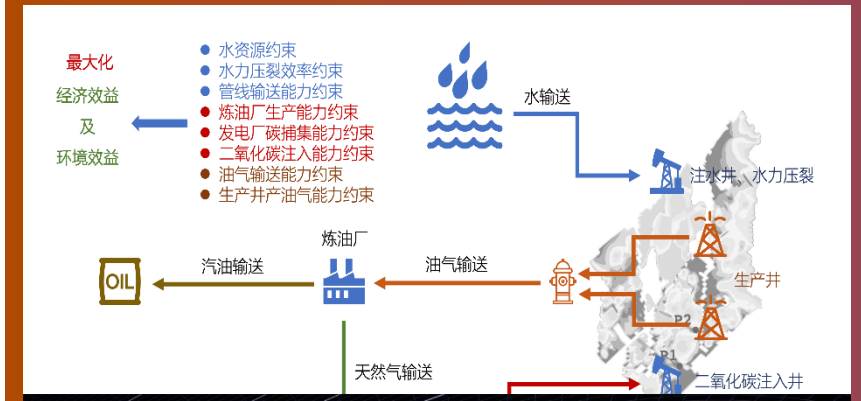
- Office hours: Mon 14:00—15:00
- Walk-in or online with appointment
- zhangcs@lsec.cc.ac.cn
- http://lsec.cc.ac.cn/~zhangcs

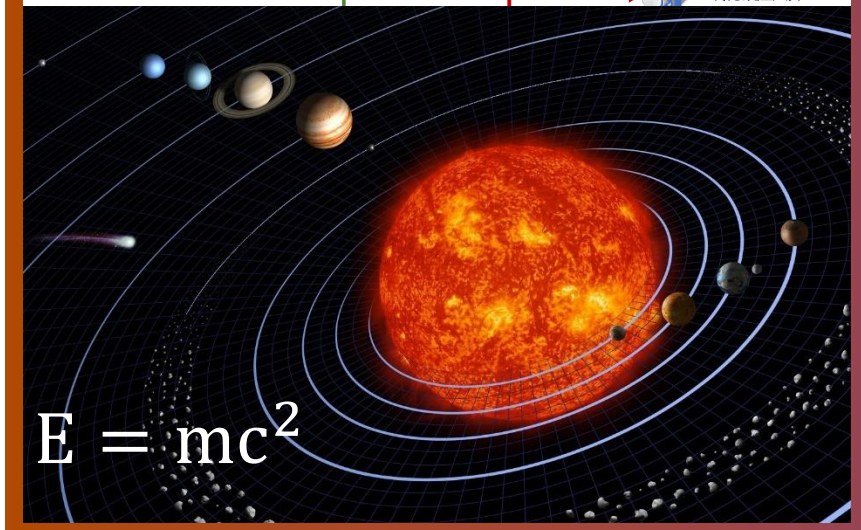
My sincere gratitude to:

Xiaowen Xu, Tao Cui, Bin Dai, Shizhe Li

# Review

## Applications of large-scale simulation




Science & Engineering

Applied Mathematics

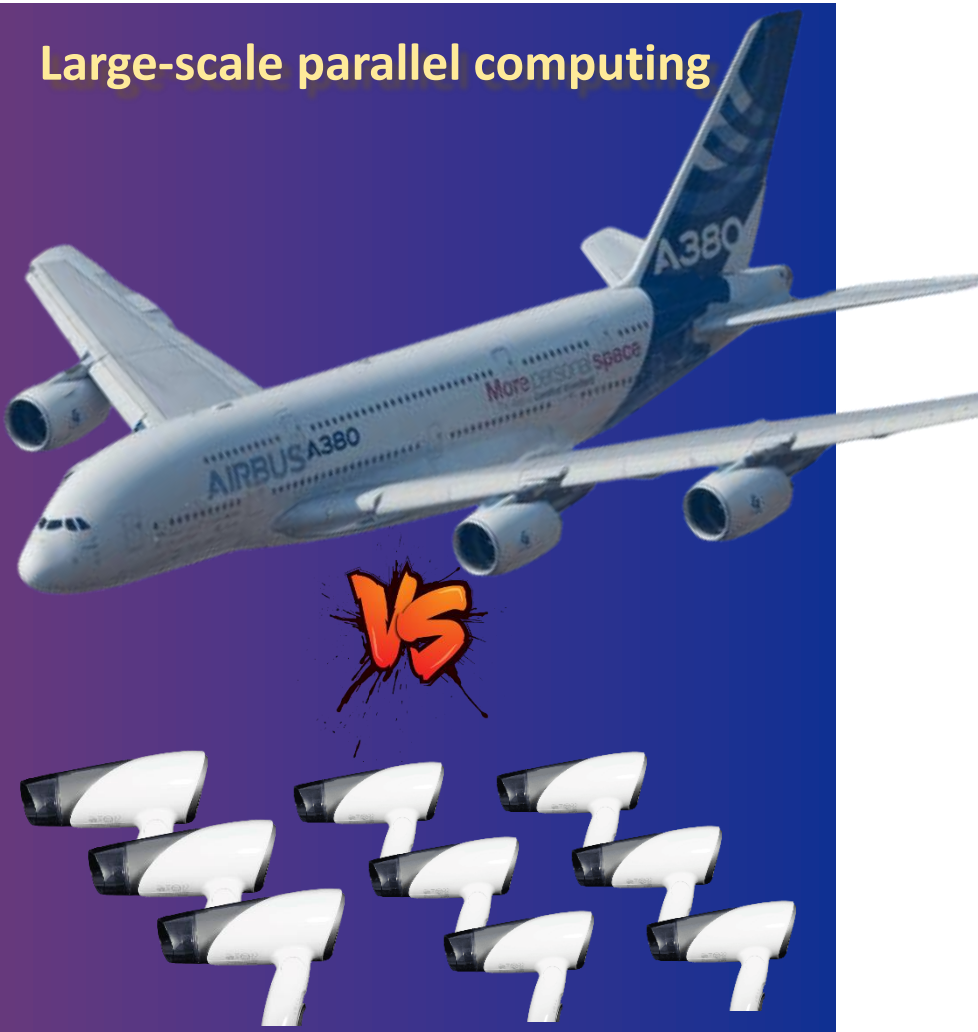
Algorithm Design

Software Engineering

Performance Engineering

Hardware Architecture Design

## Large-scale parallel computing



中国科学院大学
夏季强化课程
2022

# Fast Solvers for Large Algebraic Systems


Chensong Zhang, AMSS

http://lsec.cc.ac.cn/~zhangcs

Release version 2022.06.22