COMP4007: 并行处理和体系结构

第一章：并行处理绪论

授课老师：王强、施少怀
助　　教：刘虎成、田超

哈尔滨工业大学（深圳）

课程组信息

- 授课老师

王强，助理教授。

- ✓ 2014年于华南理工大学取得学士学位。
- ✓ 2020年于香港浸会大学计算机系取得博士学位。
- ✓ 主要研究方向为GPU计算、节能计算和分布式并行计算。

施少怀，教授。

- ✓ 2010年于华南理工大学计算取得学士学位。
- ✓ 2013年于哈尔滨工业大学（深圳）取得硕士学位。
- ✓ 2020年于香港浸会大学计算机系取得博士学位。
- ✓ 主要研究方向为分布式机器学习系统和高性能计算。

- 助教

刘虎成（hucheng.lew@qq.com），田超（tianchao_1995@163.com）

- 授课地点

理论课：A304；实验课：T2-109

预期学习成果

- 理解并行计算机体系结构的基本概念和原理
- 掌握常见的并行计算范式以及并行计算性能分析方法
- 掌握现代主流并行编程框架的编程方法以及优化技巧
  - OpenMP
  - CUDA
  - MPI
- 设计开发用于实际科学或工程问题的高性能计算系统

课程计划

1. 并行处理绪论 (1学时) (10月30号，周一)
   - 基本概念、发展趋势、基本原理、重要性

2. 并行计算机体系结构 (3学时) (10月30号，周一；11月1号，周三)
   - 并行计算机结构、分类方法、SIMD

3. 并行计算范式与性能评估 (2学时) (11月6号，周一)
   - 基本编程模型、存储管理、评估分析

4. 基于OpenMP的并行编程 (6学时) (11月8号，周三；11月13号，周一；11月15号，周三)
   - 编程范式与语法
   - parallel for语句
   - 高级特性：变量共享/私有化、依赖与同步、并行方式

5. GPU计算和CUDA并行编程I (4学时) (11月20号，周一；11月22号，周三)
   - GPU硬件架构
   - CUDA编程基础
   - CUDA线程组织

课程计划

6. GPU计算和CUDA并行编程II (4学时) (11月27号，周一；11月29号，周三)
   - GPU内存层次结构
   - GPU访存优化
   - 分支与bank冲突

7. 基于MPI的并行编程I (4学时) (12月4号，周一；12月6号，周三)
   - MPI简介与编程基础
   - 点到点通信（阻塞与非阻塞）

8. 基于MPI的并行编程II (4学时) (12月11号，周一；12月13号，周三)
   - 集体通信
   - 基于流水线的通信优化

9. 并行计算高级主题 (4学时) (12月18号，周一；12月20号，周三)
   - 基于MPI和OpenMP联合的高扩展性并行
   - 多GPU编程
   - 基于MPI和CUDA联合的多层次并行
   - 分布式机器学习训练

实验课程

1. 实验一： OpenMP 编程 (11月10号，周五)
   - 搭建实验环境
   - 编写、调试、编译、运行代码

2. 实验二：利用 CUDA 在 GPU 上编程 (11月24号，周五)
   - 搭建实验环境
   - 编写、调试、编译、运行代码

3. 实验三：MPI 集群编程 (12月8号，周五)
   - 搭建实验环境
   - 编写、调试、编译、运行代码

4. 实验四：结合 OpenMP、CUDA 和 MPI 的编程示例 (12月22号，周五)
   - 结合不同编程框架的代码示例

评分 (Final Exam)

- 平时作业：4次，共10%
- 实验：4次，共40%
- 期末考试（闭卷）：50%

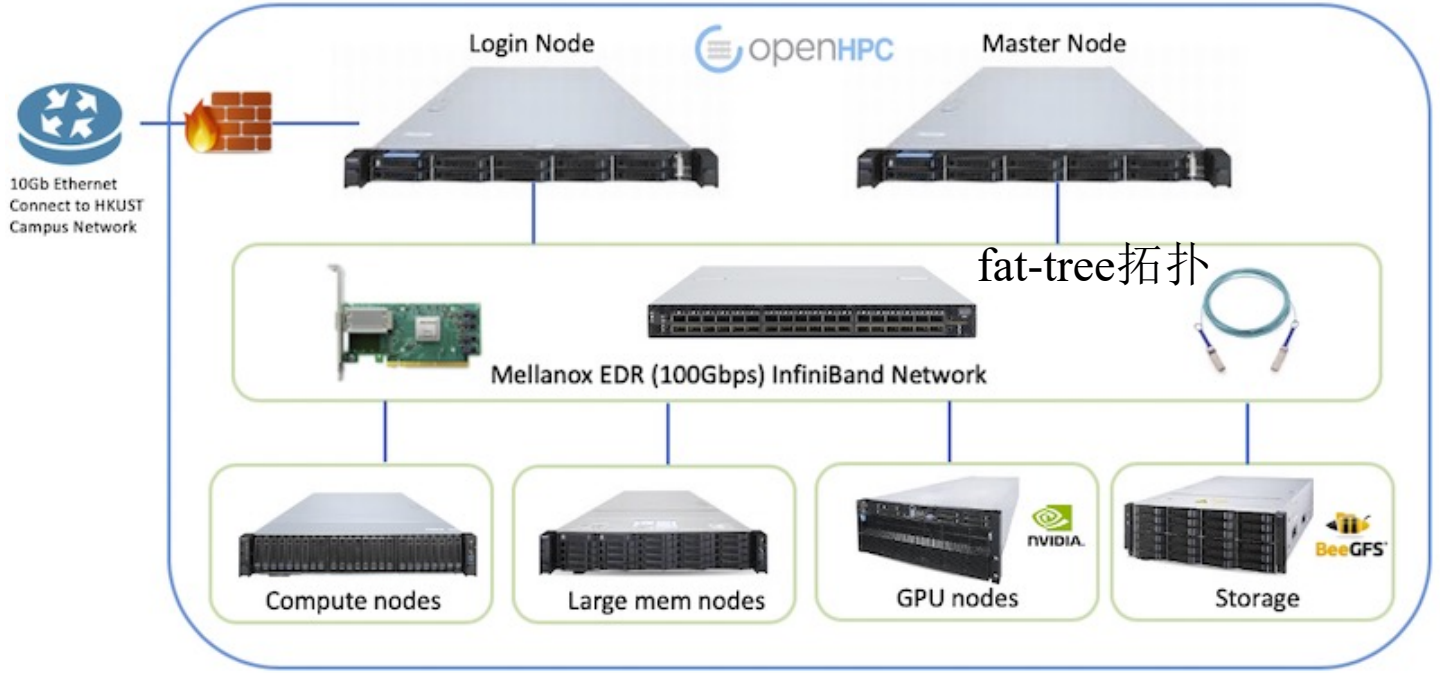
什么是高性能计算

- HPC（高性能计算）通常指的是以聚合**计算能力**的方式提供更高性能（更短耗时）以解决科学和工程中的大型问题。
  - **计算能力：**从多核处理器到超级计算机
  - **大型问题：**可被拆分为小问题，从而并行计算
  - **更高的性能：**尽可能更快的解决问题

- 通常是在讨论两个主要领域：
  - 硬件
    - 从多核处理器到超级计算机
    - Top500榜单：https://www.top500.org/
  - 软件栈
    - 充分利用计算资源的软件架构
    - 低计算复杂度的算法

超级计算机的硬件

- 单个节点
  - 处理器：e.g., CPU, GPU, etc.
  - 内存：e.g., DRAM, HBM
  - 外存：HDD, SSD, etc.
  - 互联：QPI, UPI, PCIe, NVLink, …
- 集群
  - 节点互联:
    - 因特网 或 InfiniBand
  - 交换网络
    - 互联拓扑
    - Fat-tree, BCube, etc.
  - 管理节点
    - 登录，监看，…



性能度量

- FLOP: floating point operation
  - 双精度 (default, 8 bytes) or 单精度 (4 bytes)
  - 用来衡量任务的工作负载

- FLOP/s (or FLOPS): floating point operations per second
  - 用来衡量系统处理目标任务的性能

- 常用单位
  - Kilo: KFLOP/s = 10^3 FLOPS
  - Mega: MFLOP/s = 10^6 FLOPS
  - Giga: GFLOP/s = 10^9 FLOPS
  - Tera: TFLOP/s = 10^12 FLOPS
  - Peta: PFLOP/s = 10^15 FLOPS
  - Exa: EFLOP/s = 10^18 FLOPS
  - Zetta: ZFLOP/s = 10^21 FLOPS

世界超算 No.1 (Supercomputer Fugaku): ~537 PFLOPS
世界超算 No.2 (Summit): ~200 PFLOPS
The top500 list: https://www.top500.org/

超级计算机——Fugaku（富岳）（#1）

- 计算峰值性能
  - 核心数：48
  - SVE 512-bit × 2 向量计算器 / 核心
  - 核心频率：2.2 GHz
  - 双精度：64 bits
  - 512-bit addition or multiply (fma) per cycle
    - 512/64=8 double-precision FLOP per cycle
  - 峰值性能：2.2×10^9 (Frequency) × 8 × 2 × 2 (FLOP per cycle) × 10^{-12} (Tera) × 48 (Cores) = 3.3792 TFLOPS

(Source: https://www.riken.jp/en/news_pubs/news/2020/20201117_2/index.html)

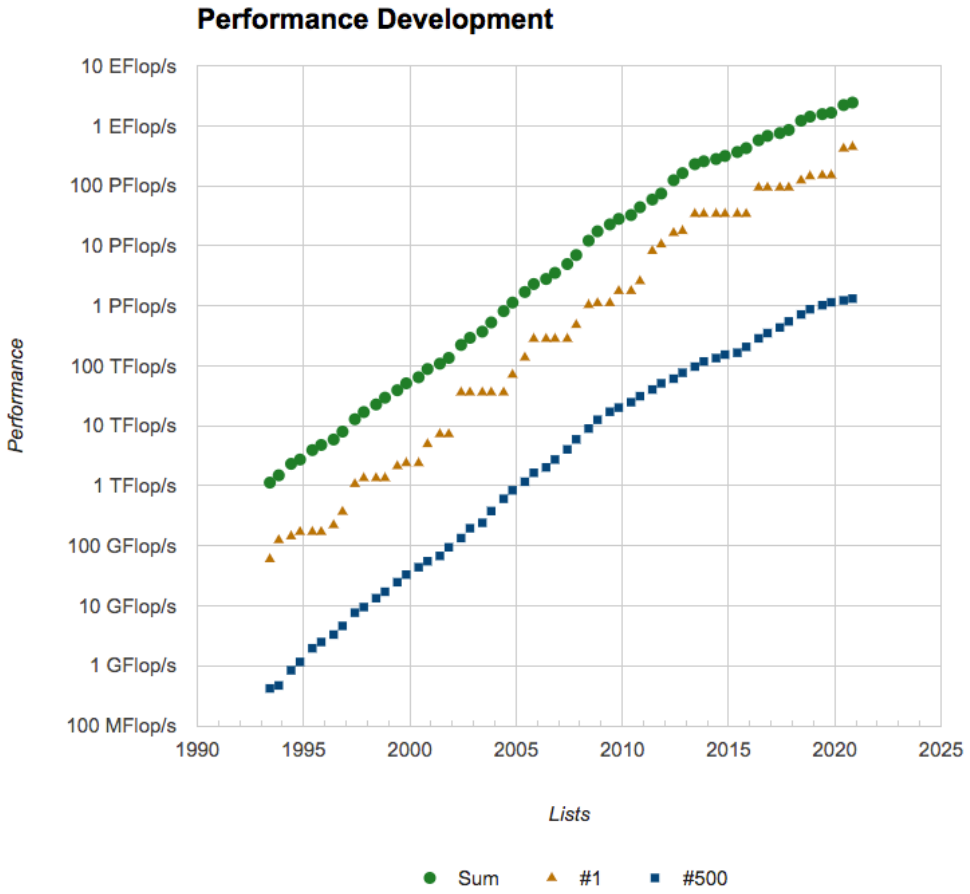
- Supercomputer Fugaku (#1 supercomputer)
  - 峰值性能: ~537 PFLOPS
  - 158,976节点

Architecture		Armv8.2-A SVE 512bit With the following Fujitsu's extensions: Hardware barrier, Sector cache, and Prefetch
Core		48 cores for compute and 2 or 4 cores for OS activities 4 CMGs (NUMA nodes)
Performance	Normal Mode: 2.0 GHz	DP: 3.072 TF, SP: 6.144 TF, HP: 12.288 TF
	Boost Mode: 2.2 GHz	DP: 3.3792 TF, SP: 6.7584 TF, HP: 13.5168 TF
Cache*1 *2		L1D/core: 64 KiB, 4way, 256 GB/s (load), 128 GB/s (store)
		L2/CMG: 8 MiB, 16way L2/node: 4 TB/s (load), 2 TB/s (store) L2/core: 128 GB/s (load), 64 GB/s (store)
Memory		HBM2 32 GiB, 1024 GB/s
Interconnect		Tofu Interconnect D (28 Gbps x 2 lane x 10 port)
I/O		PCIe Gen3 x16
Technology		7nm FinFET

(Source: https://www.r-ccs.riken.jp/en/fugaku/project/outline)

Detailed report at: https://www.icl.utk.edu/files/publications/2020/icl-utk-1379-2020.pdf

TOP500算力榜



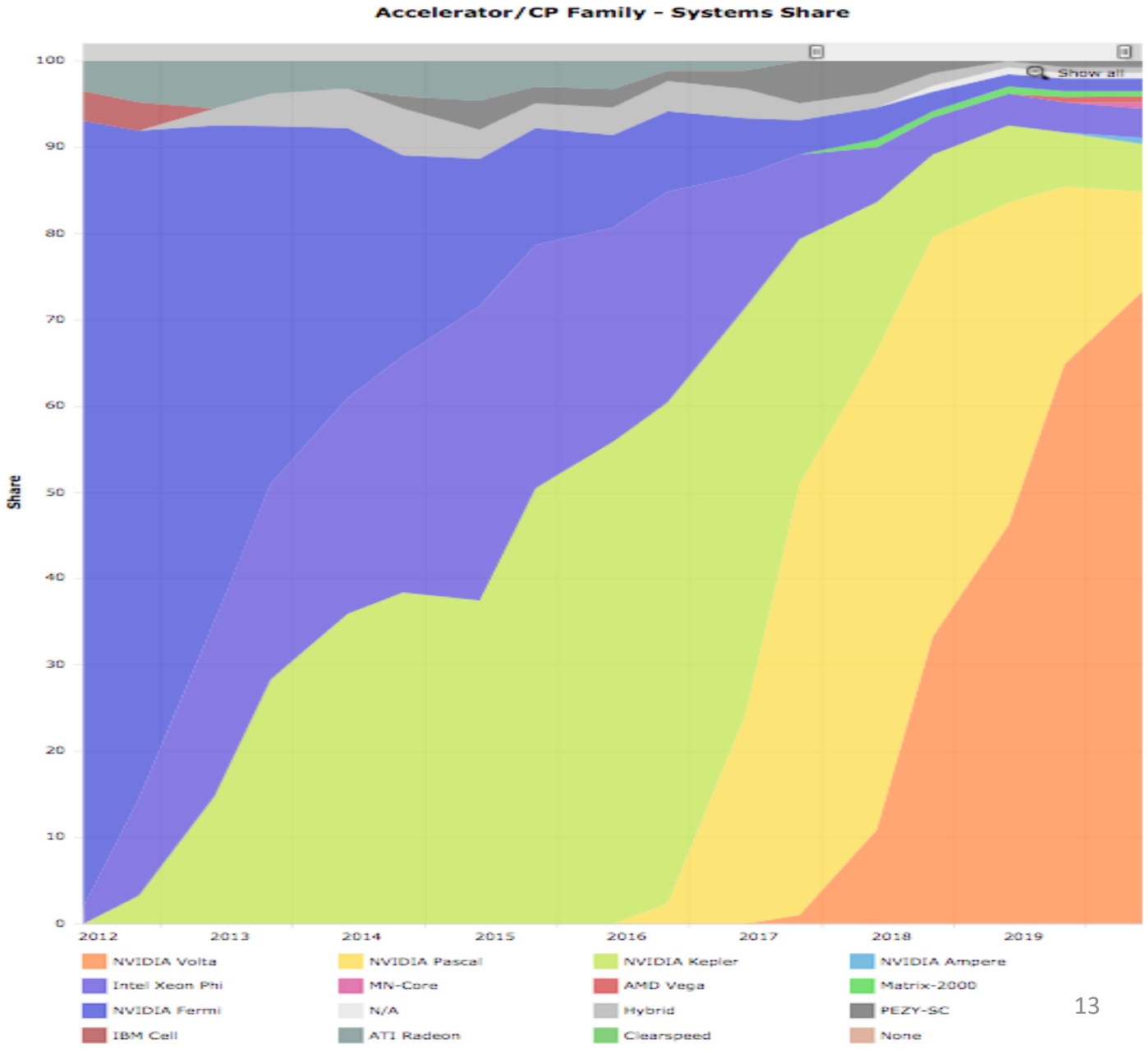
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, 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	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70,870.0	93,750.0	2,589
6	**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

- 图形处理单元 (GPUs) 在超算中扮演重要角色

Source: https://www.top500.org/statistics/perfdevel/

TOP500算力榜

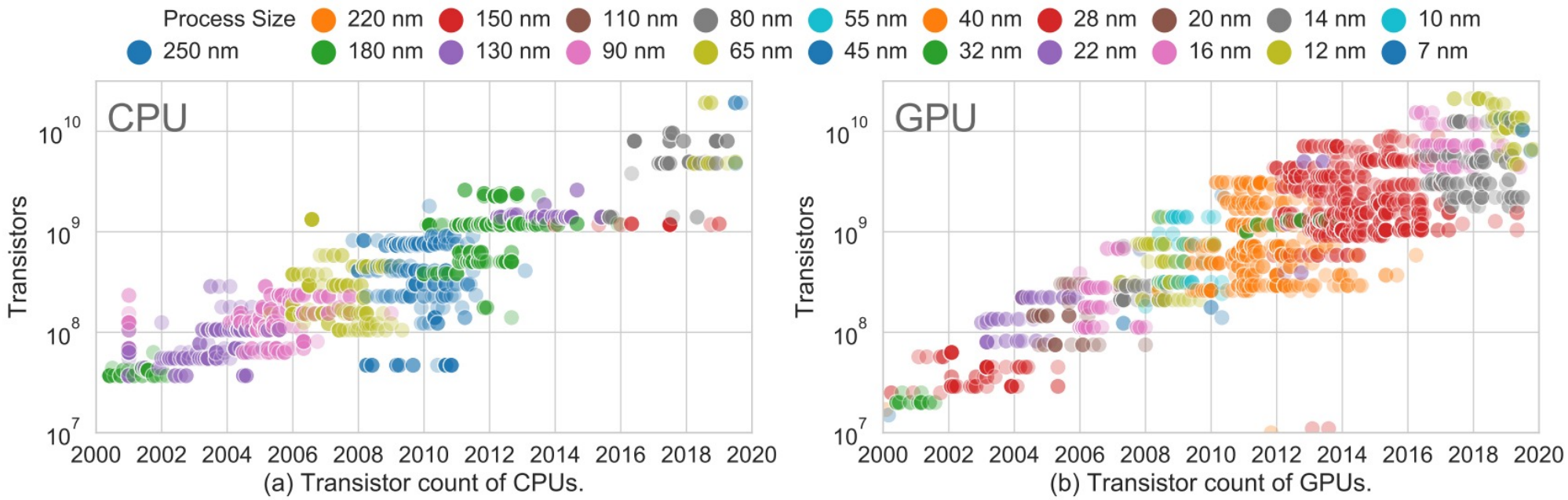
- Top500中约92%的超算系统配备了GPU (June 2020)



Source: https://www.top500.org/statistics/overtime/

处理器的发展历程

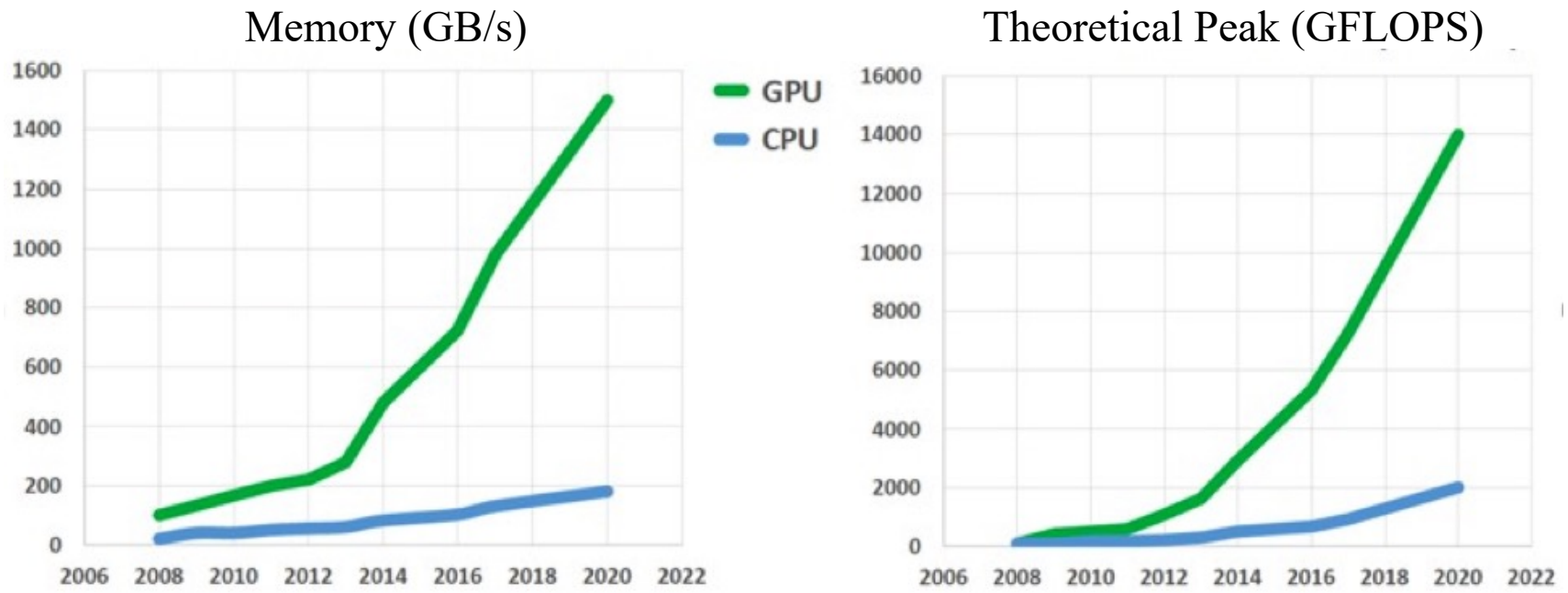
- 摩尔定律，1965 年由 Intel CEO、共同创始人 Gordon Moore 提出
  - 在密集集成电路中的晶体管数量大约每两年翻一番
- 黄氏定律，2018 年由 Nvidia CEO 黄仁勋提出
  - GPU 性能不到两年就会翻一番



晶体管数量对比

Source: https://arxiv.org/pdf/1911.11313.pdf

处理器的发展



性能对比：CPU vs. GPU

Source: https://www.nextplatform.com/2019/07/10/a-decade-of-accelerated-computing-augurs-well-for-gpus/

为什么需要高性能计算

- 大型问题：输入数据-> 处理 -> 输出结果
  - 模型体量：太大；时间需求：太长；数据规模：太多
- 许多来自科学与工程上的高性能需求
  - 工程：如计算流体动力学 (CFD) 仿真
  - 地理
  - 分子动力学
  - 物流学
  - 量子力学
  - 大数据分析
  - …

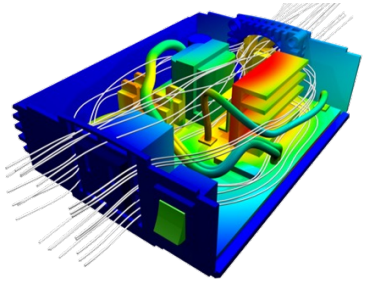
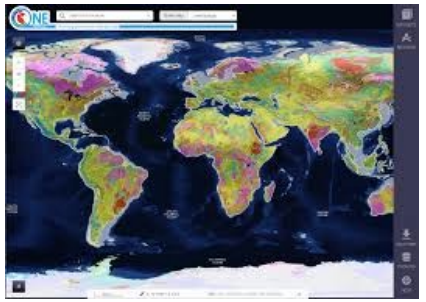
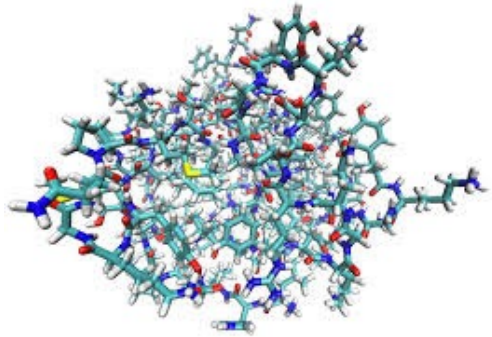
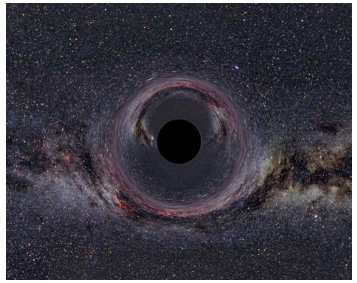
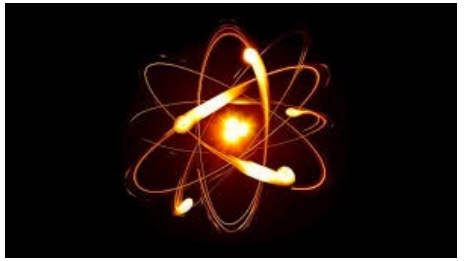
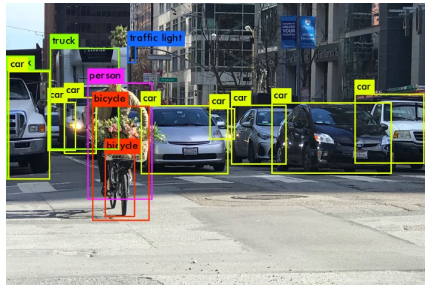
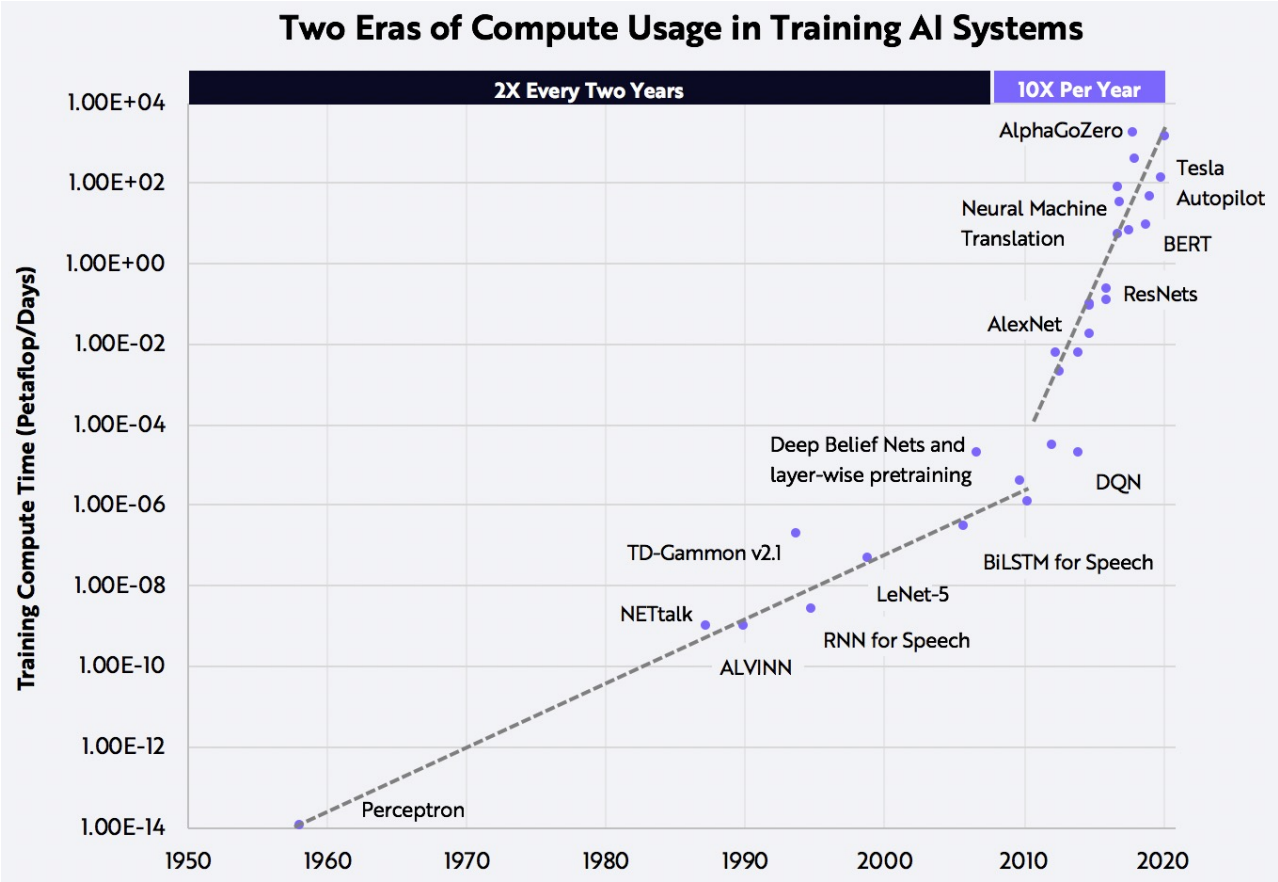







Image credits:
https://www.npd-solutions.com/cfd.html
http://www.onegeology.org/
https://www.epcc.ed.ac.uk/blog/2014/10/06/improving-tinker
http://www.sci-news.com/physics/naked-singularities-saddle-shaped-universe-04886.html
https://online.stanford.edu/courses/ee222-applied-quantum-mechanics-i
https://rishi30-mehta.medium.com/object-detection-with-yolo-giving-eyes-to-ai-7a3076c6977e

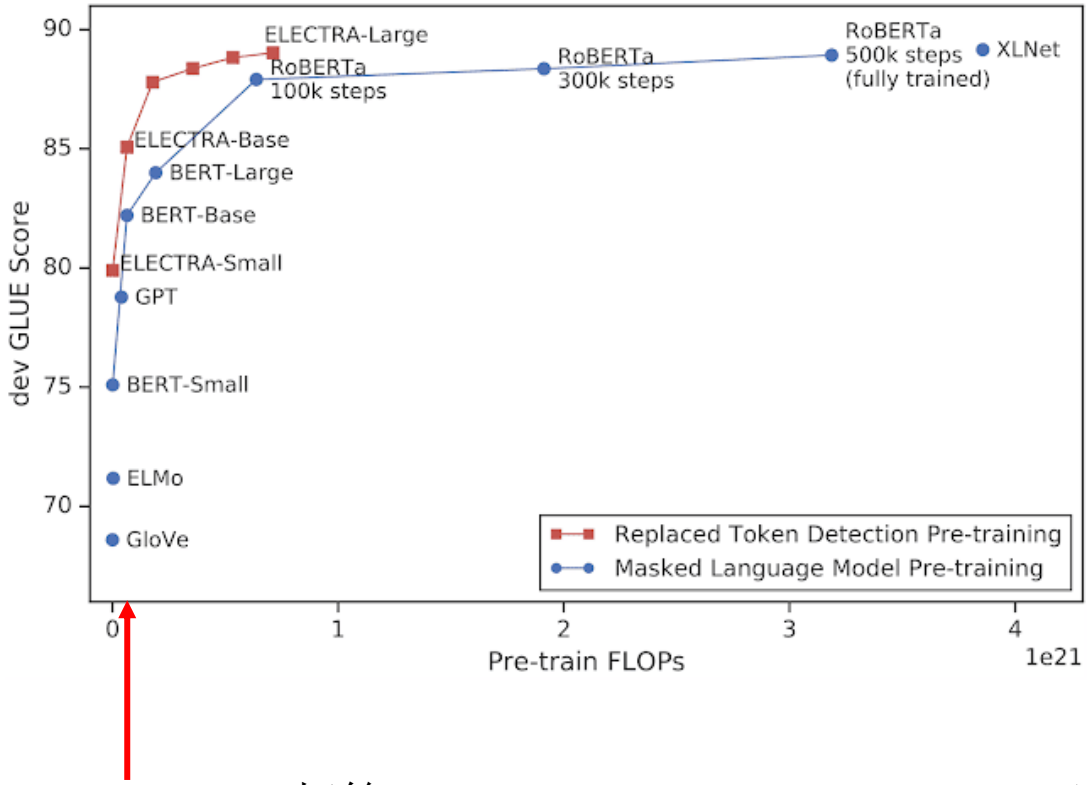
人工智能与高性能计算

- AI模型对算力的需求极高



Source: https://ark-invest.com/articles/analyst-research/ai-training/

Source: https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html



- No.1 超算: 2.146176 EFLOPS（1000T Per B）
  (Half precision for AI)!

人工智能与高性能计算

索 引 号： 000014349/2017-00142　　主题分类： 科技、教育\科技
发文机关： 国务院　　成文日期： 2017年07月08日
标　　题： 国务院关于印发新一代人工智能发展规划的通知
发文字号： 国发〔2017〕35号　　发布日期： 2017年07月20日

国发[2017]35号

国务院关于印发
新一代人工智能发展规划的通知
国发〔2017〕35号

2017年国务院《新一代人智能发展规划》
（一）构建开放协同的人工智能科技创新体系

1. 建立新一代人工智能基础理论体系
2. 建立新一代**人工智能关键共性技术体系**
3. 统筹布局人工智能创新平台
4. 加快**培养聚集人工智能高端人才**

标　　题： 五部门关于印发《国家新一代人工智能标准体系建设指南》的通知　　发文机关： 国家标准化管理委员会 中央网信办 国家发展改革委 科技部 工业和信息化部
发文字号： 国标委联〔2020〕35号　　来　　源： 标准委网站
主题分类： 科技、教育\综合　　公文种类： 通知
成文日期： 2020年07月27日

国标委联 [2020]35号

国家标准化管理委员会 中央网信办 国家发展改革委 科技部 工业和信息化部
关于印发《国家新一代人工智能标准体系建设指南》的通知
国标委联〔2020〕35号

2020年五部门：国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部
《国家新一代人工智能标准体系建设指南》

G. 行业应用：智能医疗、智能教育、智能政务等

……

B. 基础软硬件平台：智能芯片、系统软件、开发框架

基座：硬件（AI算力）+软件（AI框架）

来源：中华人民共和国中央人民政府 www.gov.cn
https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm
https://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm

人工智能与高性能计算

- AI模型对算力的需求极高
- 在单块GPU/TPU设备上训练大型模型非常慢
  - 例如，在一块 Tesla P100 GPU 上完全训练一个 ResNet-50 模型需要 **11天**
  - 用一块 Google 三代 TPU 进行 BERT 预训练任务，花费超过 **1.5 个月**

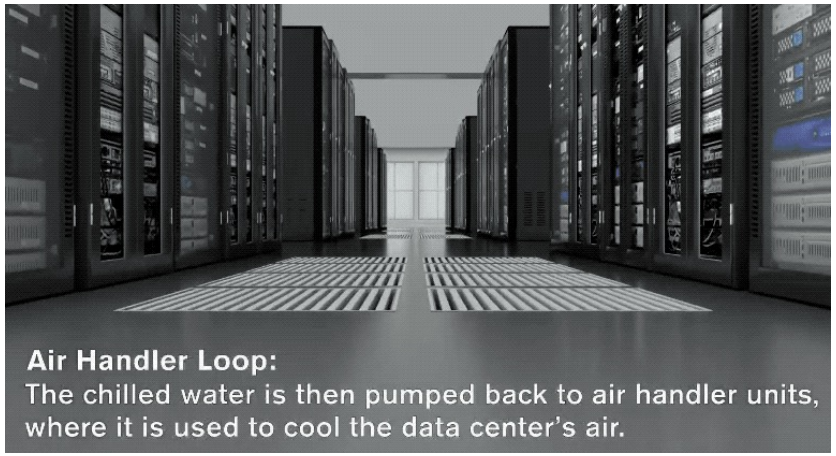
Training ResNet-50 to 75+% accuracy (90 epochs)

1 hour	20 minutes	6.6 minutes	1.2 minutes	45.6 seconds	28.2 seconds
June 2017 Facebook 256 GPUs	September 2017 UC Berkeley 2048 KNLs	July 2018 Tencent & HKBU 2048 P40 GPUs	April 2019 Fujitsu 2048 V100 GPUs	July 2020 Nvidia 1840 A100 GPUs	July 2020 Google 4096 TPUv3

December 2021	ResNet-50 4320 A100 GPUs, NVIDIA: ~20s	BERT 4320 A100 GPUs, NVIDIA: ~14s

人工智能与高性能计算

- 大语言模型的参数量：
  - GPT-3：1.5亿
  - Llama-2：700亿
- 2023年2月，ChatGPT 吸引了全球 16 亿次访问。
- 按照一万张Nvidia A100来算的话，运行一个月就要用掉 585 万度电，大概与我们 8 万人用电相当。
- 散热方面，利用蒸发水来散热，但运行起来需要消耗大量的清水，一个用户与 ChatGPT 进行 25～50 个问题的对话，大概就相当于请 ChatGPT 喝了 500ml 水。

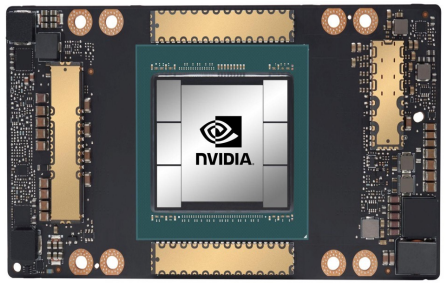


AI 不止费显卡，住恒温的大 house，胃口还出奇的好，大口吃电，大口喝水。

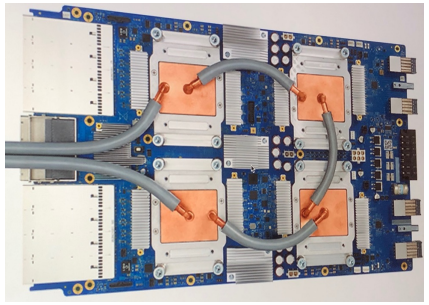
来源：https://mp.weixin.qq.com/s/ZcNcx-fQygB5CswcUptIUQ

人工智能与高性能计算

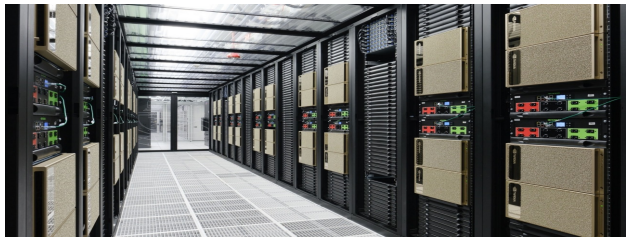
硬件：AI加速器和集群



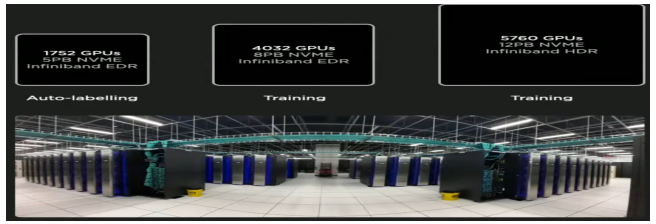
英伟达GPU



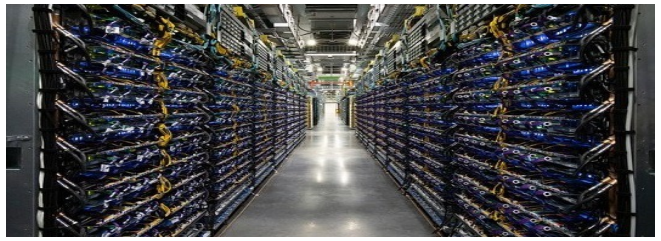
谷歌TPU



Nvidia GH200 (2023):
256 H100 GPUs, 1 EFLOPS



Tesla Dojo (2021):
5760 A100 GPUs, ~3 EFLOPS



Google TPUv4 (2022):
4096 TPUv4, ~1.1 EFLOPS

分布式AI框架：高效易用

AutoDist　Alpa　RAY

deepspeed

NVIDIA Megatron-LM

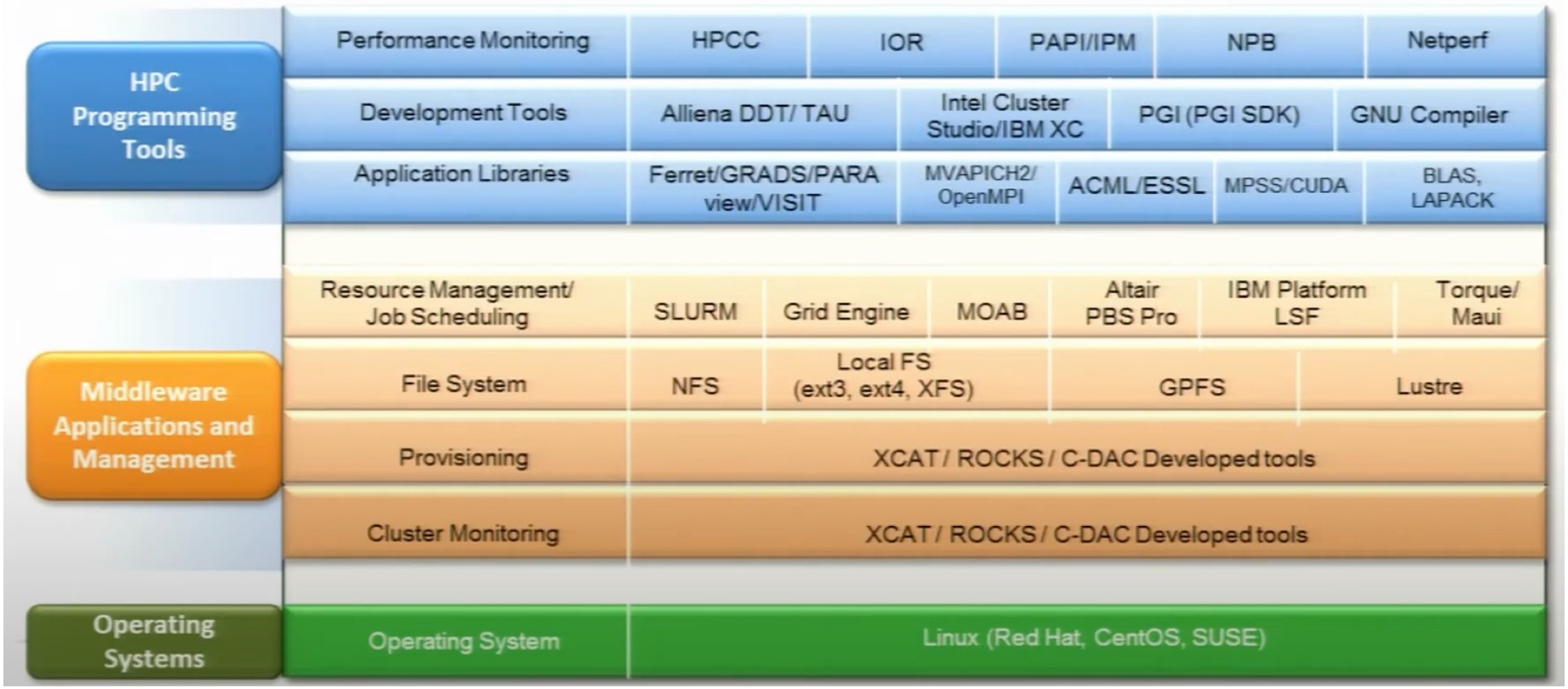
HOROVOD

AI框架：完备的算子库

PyTorch　TensorFlow

AI高性能计算国外研究现状

如何进行高性能计算：软件栈



Source: https://nptel.ac.in/courses/128/106/128106014/

如何进行高性能计算：并行编程

- 并行编程
  - 多线程：共享内存
    - **OpenMP** for CPUs, **CUDA** for GPUs
  - 多进程：分布式内存
    - **Message passing interface (MPI)**, Spark, etc.
  - 存储
    - Hadoop
  - 互联协议
    - TCP/IP, RDMA, etc.

- 设计低复杂度算法
  - FFT：快速傅里叶变换，时间复杂度 O($nlogn$)
  - 矩阵乘法 A × B, 维度为 n × n
    - 标准复杂度 O(n^3)
    - 可以缩减为 O($n^{2.37\underline{548}}$) [1] ~ O($n^{2.372\underline{8596}}$) [2]

[1] Don et al., "Matrix multiplication via arithmetic progressions," Proceedings of the nineteenth annual ACM symposium on Theory of computing, 1987.
[2] Josh et al., "A Refined Laser Method and Faster Matrix Multiplication," SODA 2021