IHEP School of Computing 2023

高能物理计算概述

程耀东

chyd@ihep.ac.cn

高能所計算中心

IHEP Computing Center

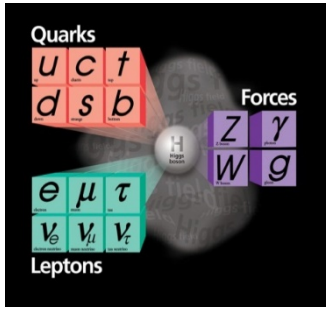
主要内容

- 高能物理数据处理的挑战
- 高能物理科学计算平台
- 计算机体系结构发展变化
- 未来发展趋势

高能物理科学研究

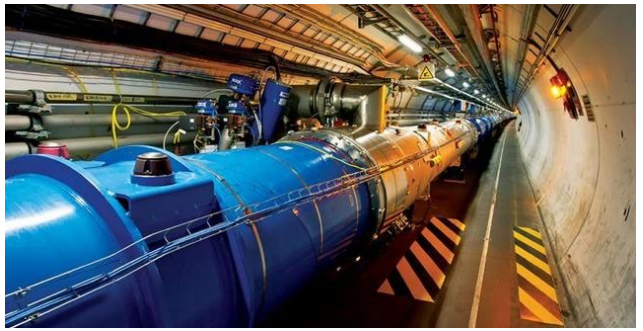
物质结构组成（理论）

- 夸克、轻子、玻色子
- 强力、弱力、电磁力、万有引力



粒子加速器（装置）

- 粒子物理研究的重要手段之一
- BEPCII，LHC，CEPC等等



探测器 （实验）

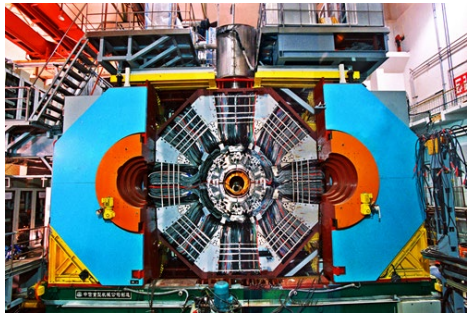
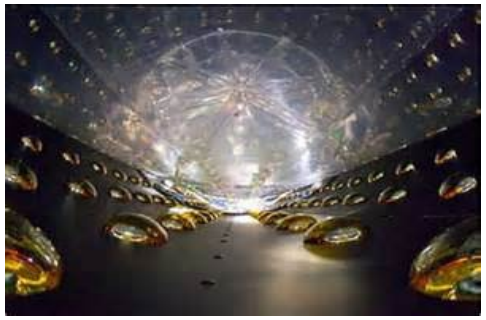
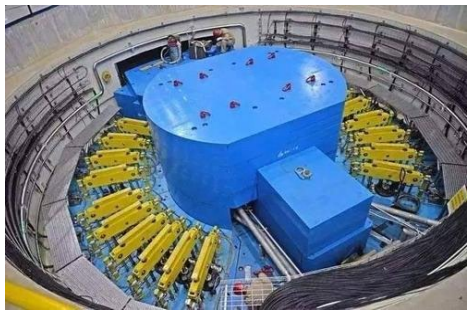
- 探测各类粒子，用于科学研究
- BESIII, JUNO, LHAASO, ATLAS, CMS …

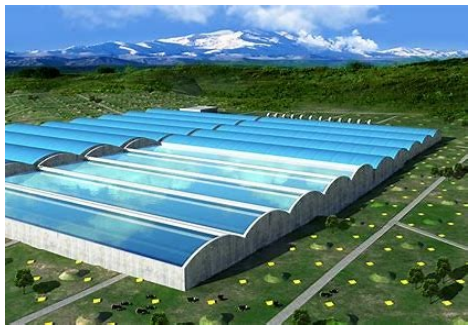
数据分析（科学发现）

- 暗物质/暗能量
- 宇宙起源、…

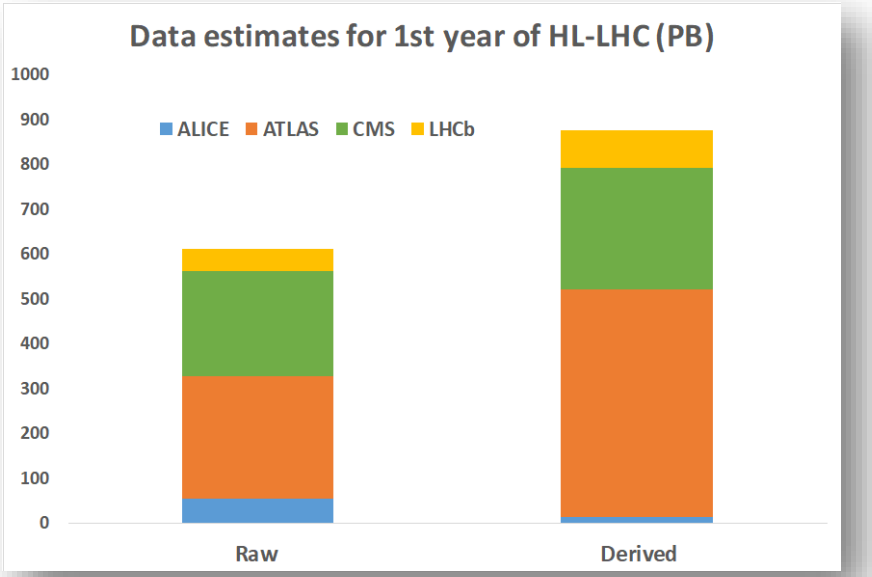
高能物理实验数据挑战

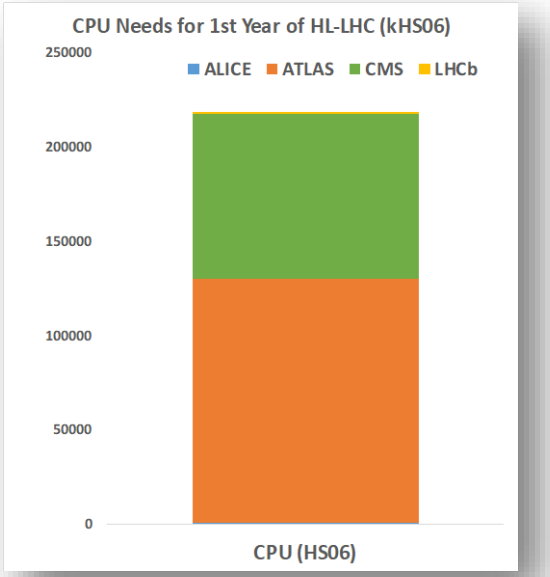
- 北京正负电子对撞机BECPII/BESIII
  - 每年~1PB raw data，已经积累20PB+
- 中微子实验
  - 大亚湾中微子实验已经积累>2PB原始数据
  - 江门中微子实验每年将产生3PB数据
- 高海拔宇宙线实验LHAASO
  - 位于四川稻城海子山，海拔4400米
  - 目前已经运行，每年产生10PB以上的数据
- 高能同步辐射光源HEPS
  - Avg 290PB/year原始数据，
    单个实验数据产生速率> 400Gbps
- 高能空间天文实验HXMT/HERD/eXTP/GeCAM
  等全部运行后预计每年将产生10PB数据
- 阿里原初引力波、中国散裂中子源、南方光源等

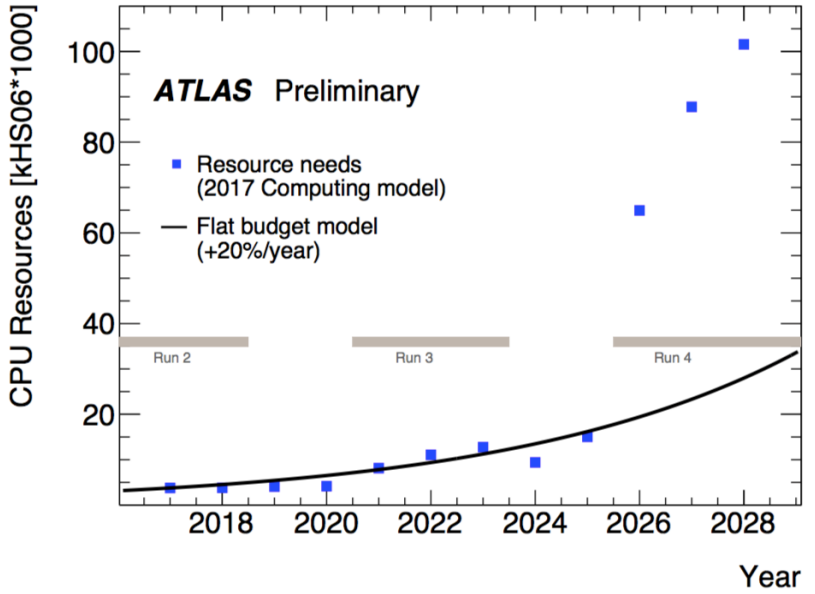







HL-LHC的计算需求





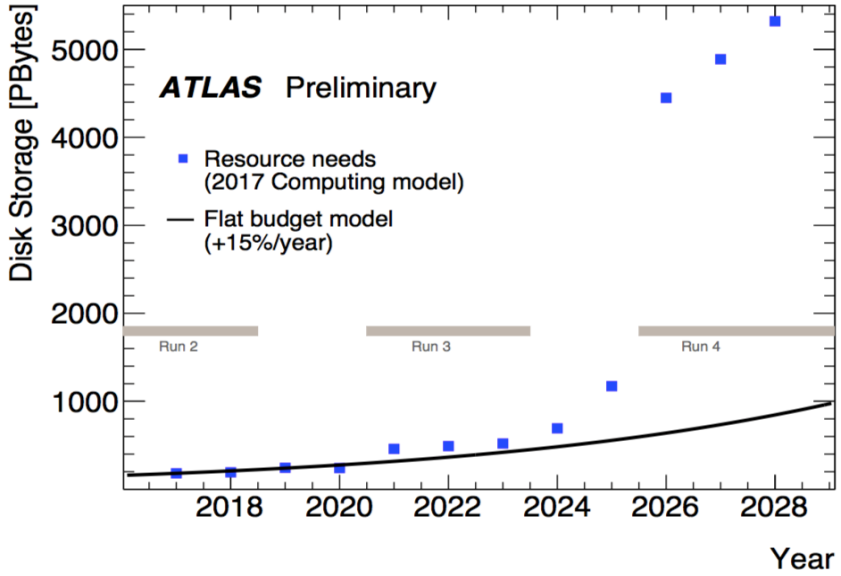


数据与计算挑战:

原始数据 2016: 50 PB → 2027: 600 PB

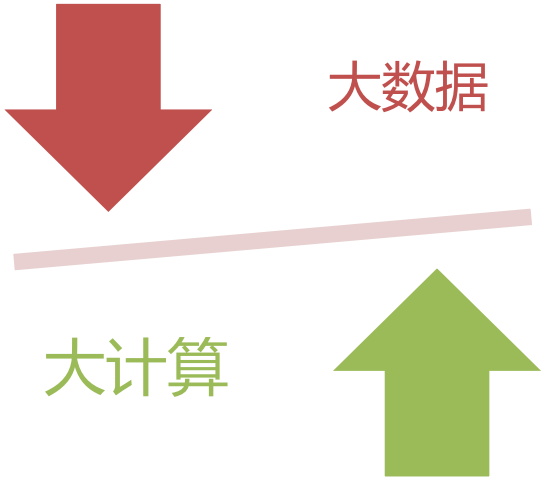
处理后的数据 (1 copy): 2016: 80 PB → 2027: 900 PB

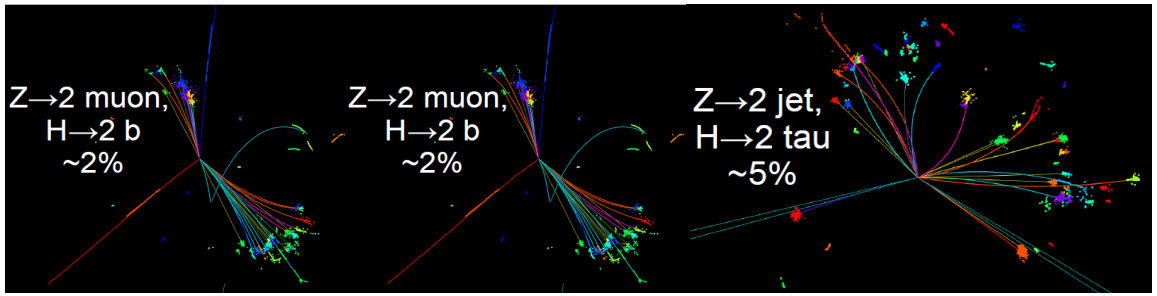
- 10倍以上的存储需求
- 60倍以上的计算需求
  - 假设未来计算机技术平均每年有20%的提升，10年将有6倍的技术提升，仍需要在计算资源方面增加越10倍左右的投入



高能物理实验计算

- 高能物理科学研究能否成功依赖于计算技术的发展
- 实验采集到的数据需要强大的计算系统对其进行分析处理
- 物理模拟及理论计算需要强大的高性能计算支撑
- 不同的数据处理任务采用不同的计算模式
  - 粒子加速器和探测器的计算机模拟设计：计算密集型
  - 粒子探测器观测到的海量科学数据的分析处理：数据密集型
  - 高能物理理论研究中的高强度的科学计算：计算密集型
    - 例如格点量子色动力学（格点QCD)和计算宇宙学

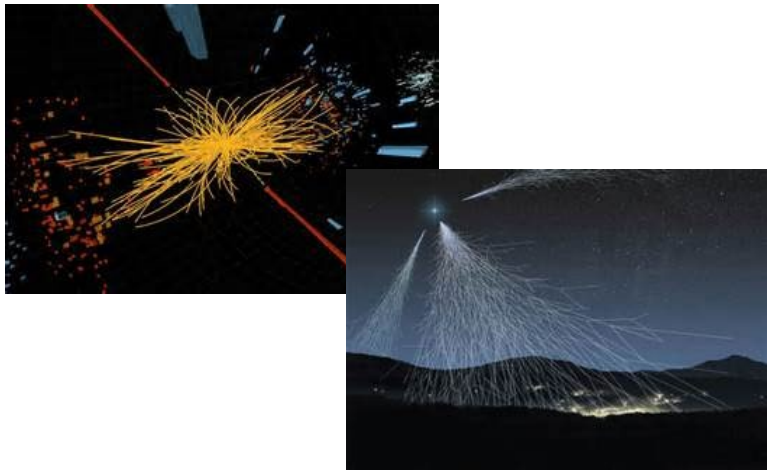


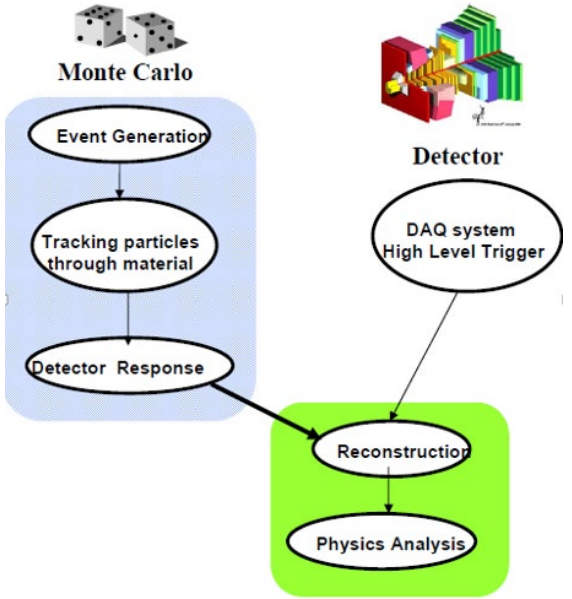


高能物理领域已经步入EB级的大数据时代

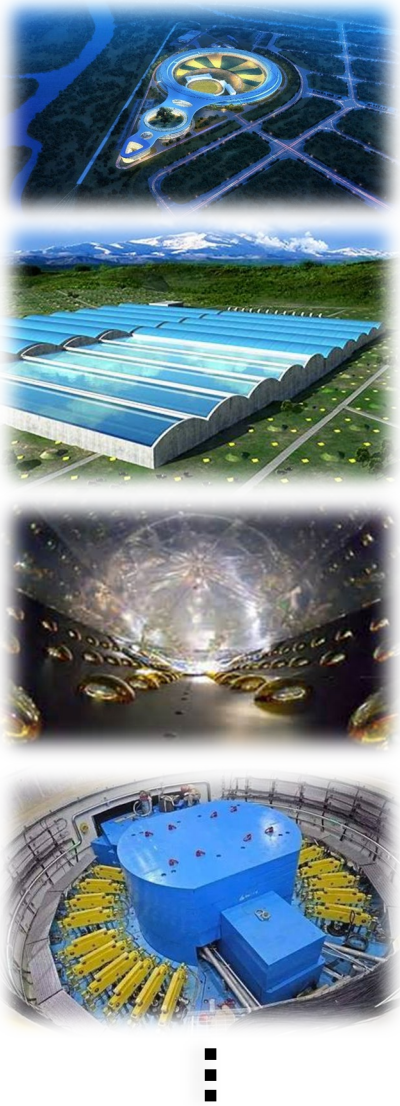
数据处理过程

- **事例：**一次粒子对撞或者一次粒子间的相互作用
  - 粒子物理研究的基本对象
- 探测器记录事例，产生原始数据
  - 以二进制格式记录的探测器信号
  - 由计算机产生模拟实验的蒙特卡罗模拟数据，数字化
- 事例重建
  - 读出Raw/MC Raw数据，处理后产生相关物理信息，如动量、对撞顶点等；
- 数据分析
  - 由上千个属性组成的Event文件，提供物理学家进行分析，并最后产生物理结果





典型的高能物理计算平台



分布式计算系统

前端登录集群

本地计算集群

计算子系统

防火墙

高速核心网络

磁盘存储

磁带存储

数据备份

实验站

专线链路

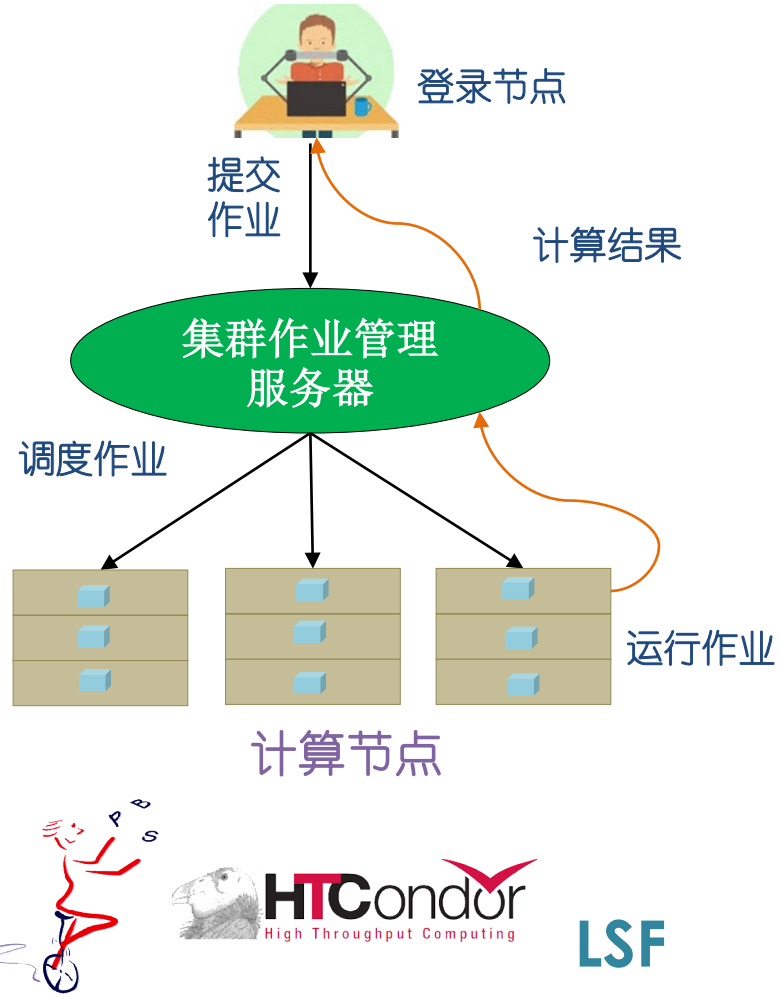
数据传输

网络子系统

存储子系统

本地计算集群

- 管理计算节点，调度作业
- 提供用户提交作业接口
- PBS
  - 开源，简单，历史悠久
  - OpenPBS，PBS Pro，Torque
  - IHEP在2016年以前的调度系统
- HTCondor
  - HTC：High Throughput Computing
  - 开源，更好的性能，更多的功能，调度算法更为公平
  - **IHEP现有调度系统**
- SLURM：高性能计算调度
  - HPC：High Performance Computing
  - GPU、MPI等作业调度
- LSF：商业调度软件

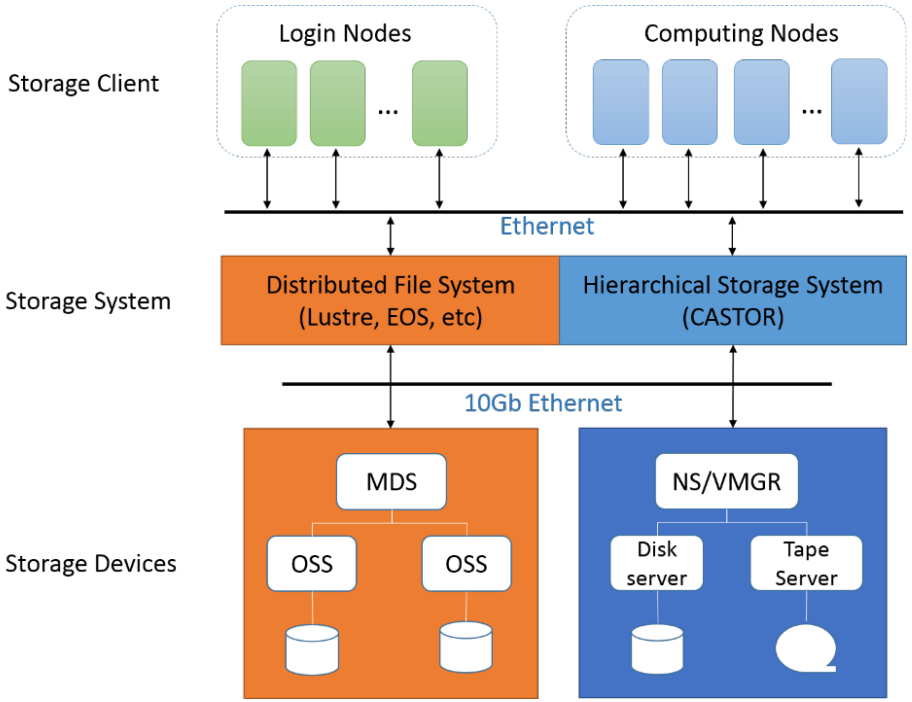


存储系统

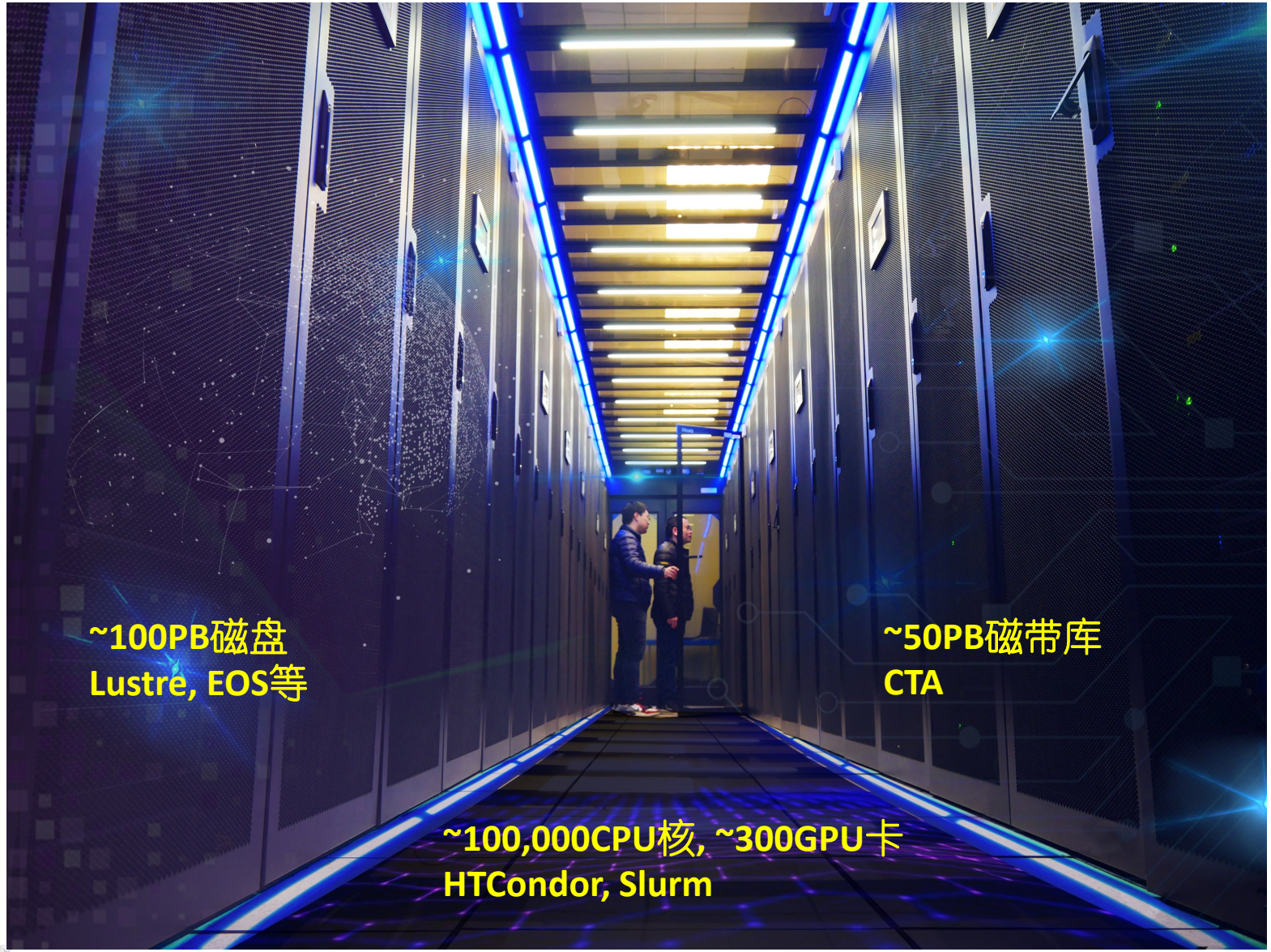
- 磁带存储系统
  - 将顺序设备映射成类似于存储系统的树形目录
  - CERN CASTOR/CTA，enstore等开源软件
  - TSM/HPSS等商业软件
- 磁盘存储系统
  - 分布式文件系统

    Lustre、**EOS**、BeeGFS，GPFS、…
  - 应用层存储系统

    dCache、HDFS、EOS、…
- 用户目录
  - AFS
- 软件库共享系统
  - AFS，CVMFS

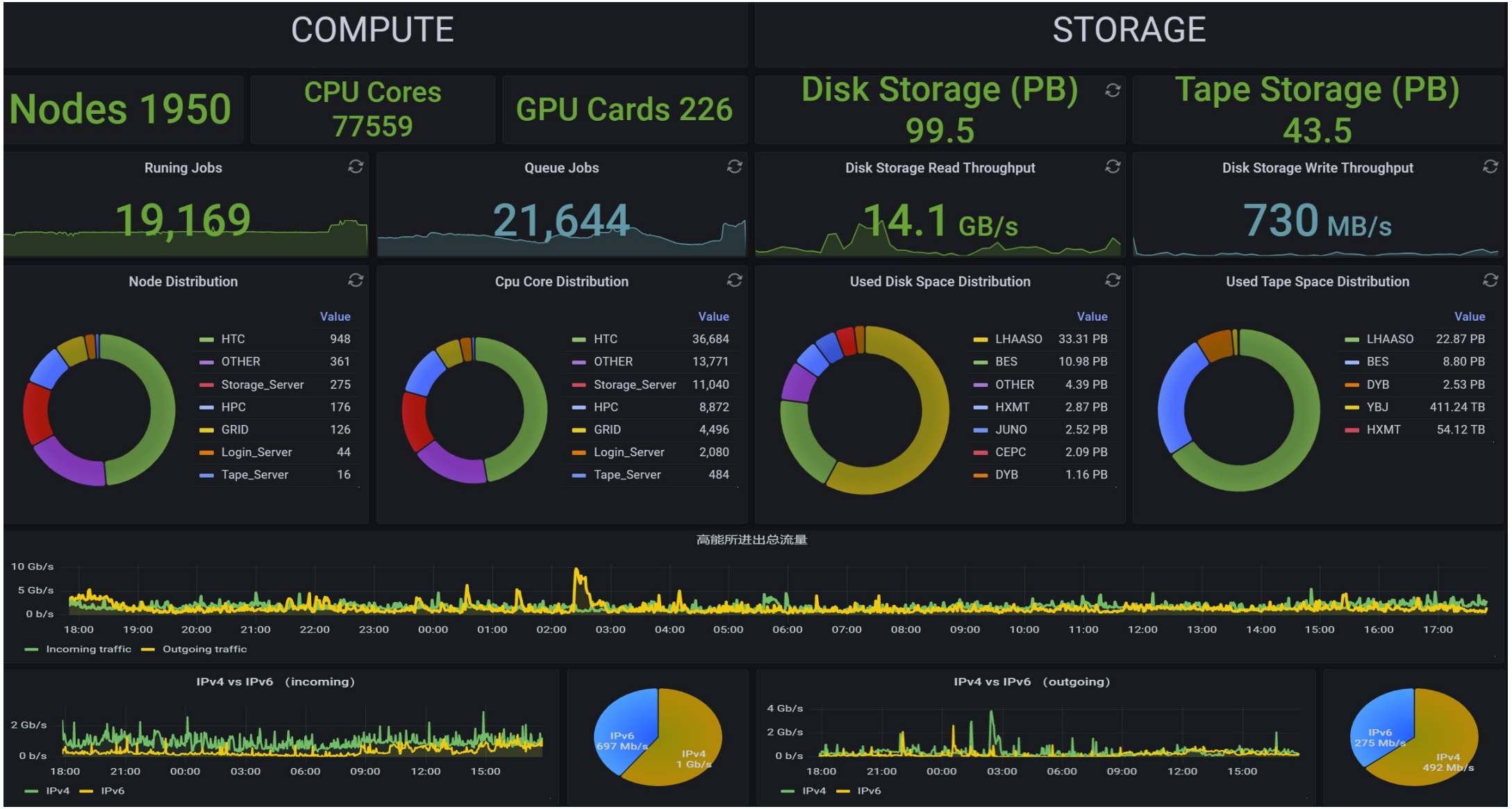


高能所计算集群


~100PB磁盘
Lustre, EOS等
~50PB磁带库
CTA
~100,000CPU核, ~300GPU卡
HTCondor, Slurm

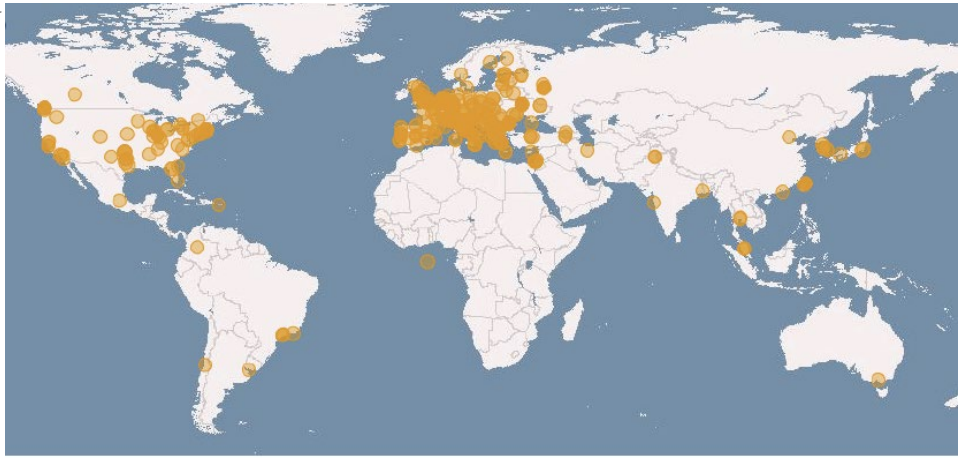

2Tbps计算网络
40Gbps国际带宽
IPV4/V6双链路

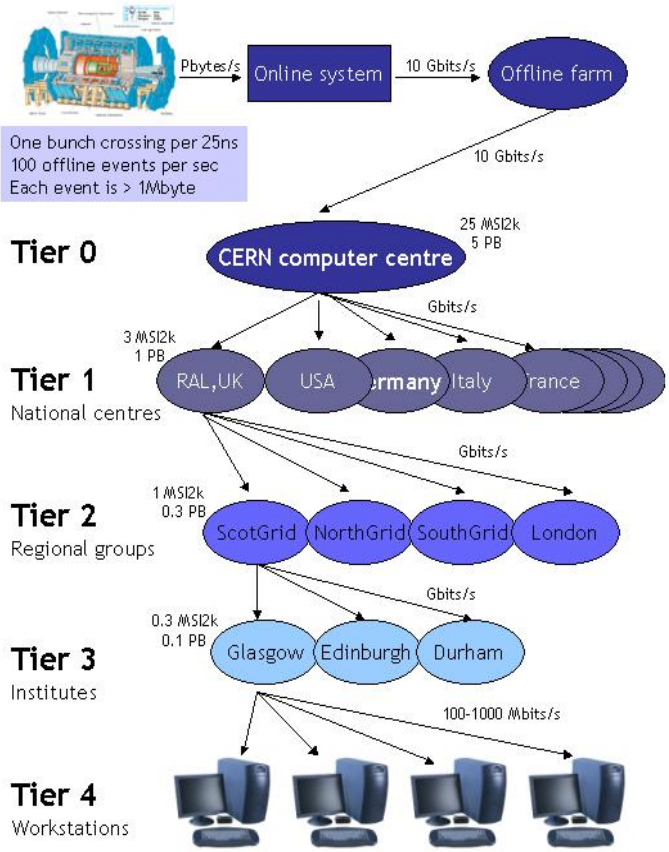
运行监视


COMPUTE
STORAGE
Nodes 1950
CPU Cores 77559
GPU Cards 226
Disk Storage (PB) 99.5
Tape Storage (PB) 43.5
Runing Jobs
19,169
Queue Jobs
21,644
Disk Storage Read Throughput
14.1 GB/s
Disk Storage Write Throughput
730 MB/s
Node Distribution
Value
HTC 948
OTHER 361
Storage_Server 275
HPC 176
GRID 126
Login_Server 44
Tape_Server 16
Cpu Core Distribution
Value
HTC 36,684
OTHER 13,771
Storage_Server 11,040
HPC 8,872
GRID 4,496
Login_Server 2,080
Tape_Server 484
Used Disk Space Distribution
Value
LHAASO 33.31 PB
BES 10.98 PB
OTHER 4.39 PB
HXMT 2.87 PB
JUNO 2.52 PB
CEPC 2.09 PB
DYB 1.16 PB
Used Tape Space Distribution
Value
LHAASO 22.87 PB
BES 8.80 PB
DYB 2.53 PB
YBJ 411.24 TB
HXMT 54.12 TB
高能所进出总流量
10 Gb/s
5 Gb/s
0 b/s
18:00 19:00 20:00 21:00 22:00 23:00 00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00
Incoming traffic
Outgoing traffic
IPv4 vs IPv6 （incoming）
2 Gb/s
0 b/s
18:00 21:00 00:00 03:00 06:00 09:00 12:00 15:00
IPv4
IPv6
IPv6 697 Mb/s
IPv4 1 Gb/s
IPv4 vs IPv6 （outgoing）
4 Gb/s
2 Gb/s
0 b/s
18:00 21:00 00:00 03:00 06:00 09:00 12:00 15:00
IPv4
IPv6
IPv6 275 Mb/s
IPv4 492 Mb/s

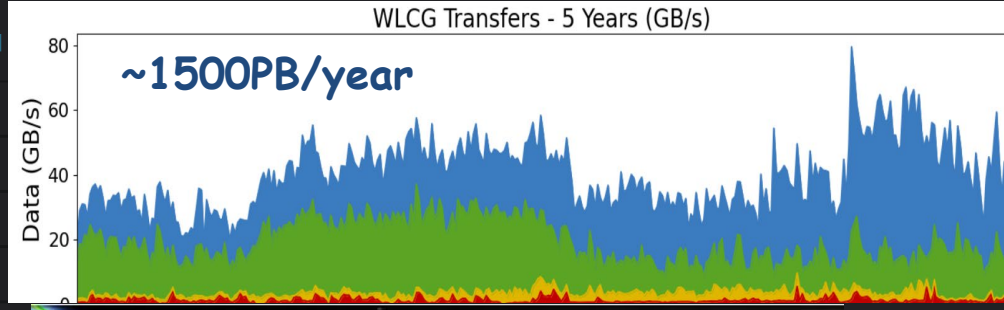
WLCG网格

- WLCG: WorldWide LHC Computing Grid
- Tier 0: CERN
  - 接收原始数据，保存在磁带系统中，并进行第一遍数据重建
  - 向Tier1分发数据
- Tier1: 15个
  - 提供原始数据备份
  - 执行数据重建、分析等任务
  - 提供数据分发等网格服务
- Tier2: 145个
  - 执行模拟、数据分析等任务
- Tier3：本地系统

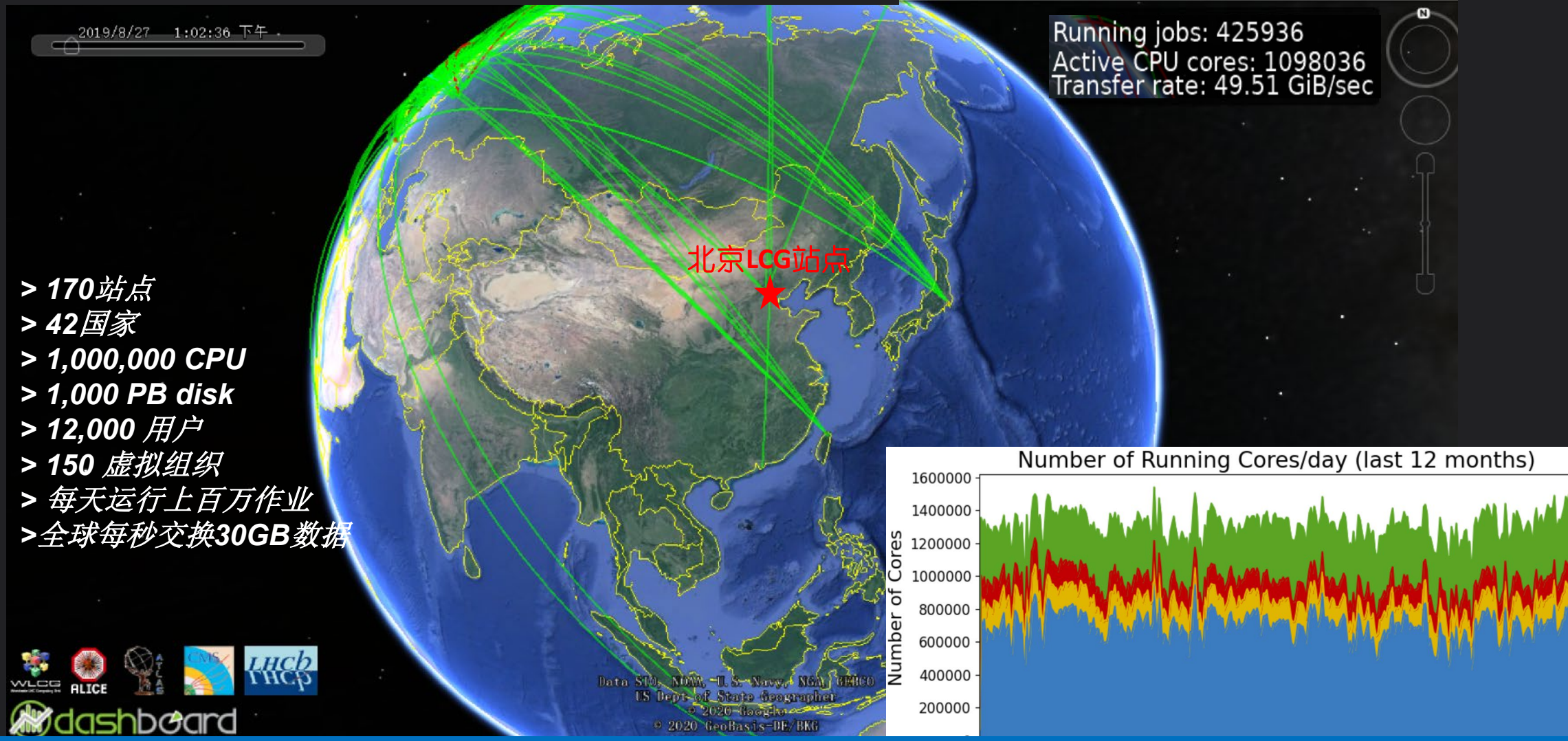




vo	Used	Free	Total
ALICE	151.9 PB	28.8 PB	180.7 PB
ATLAS	488.4 PB	64.7 PB	553.1 PB
CMS	235.1 PB	30.9 PB	266.0 PB
LHCb	102.8 PB	22.5 PB	125.3 PB

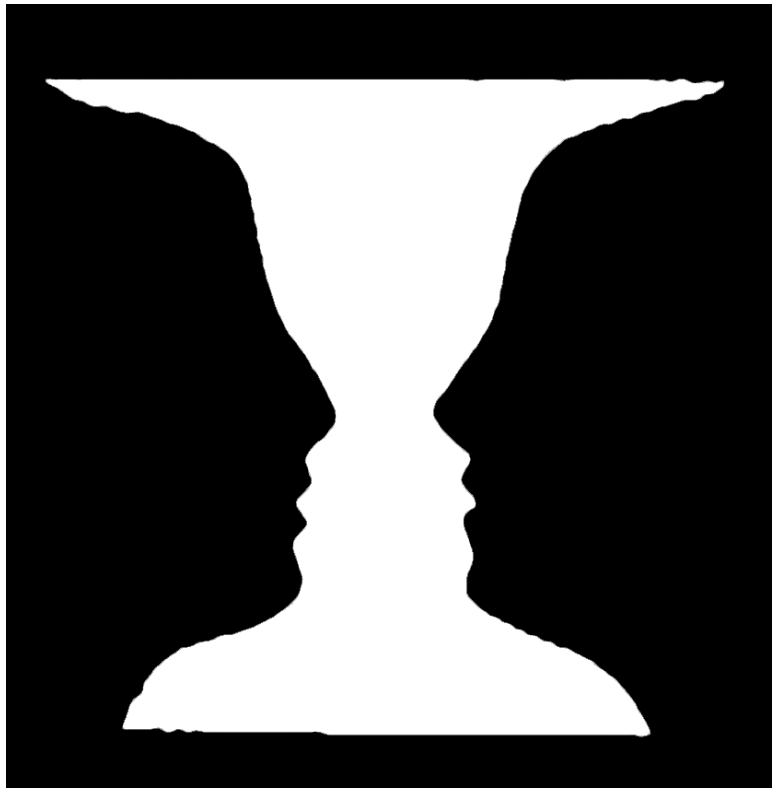


WLCG网格站点



- *> 170站点*
- *> 42国家*
- *> 1,000,000 CPU*
- *> 1,000 PB disk*
- *> 12,000 用户*
- *> 150 虚拟组织*
- *> 每天运行上百万作业*
- *>全球每秒交换30GB数据*

换个角度来看？
——计算机体系结构



摩尔定律

- 二十一世纪以来，摩尔定律发展放缓，甚至面临失效

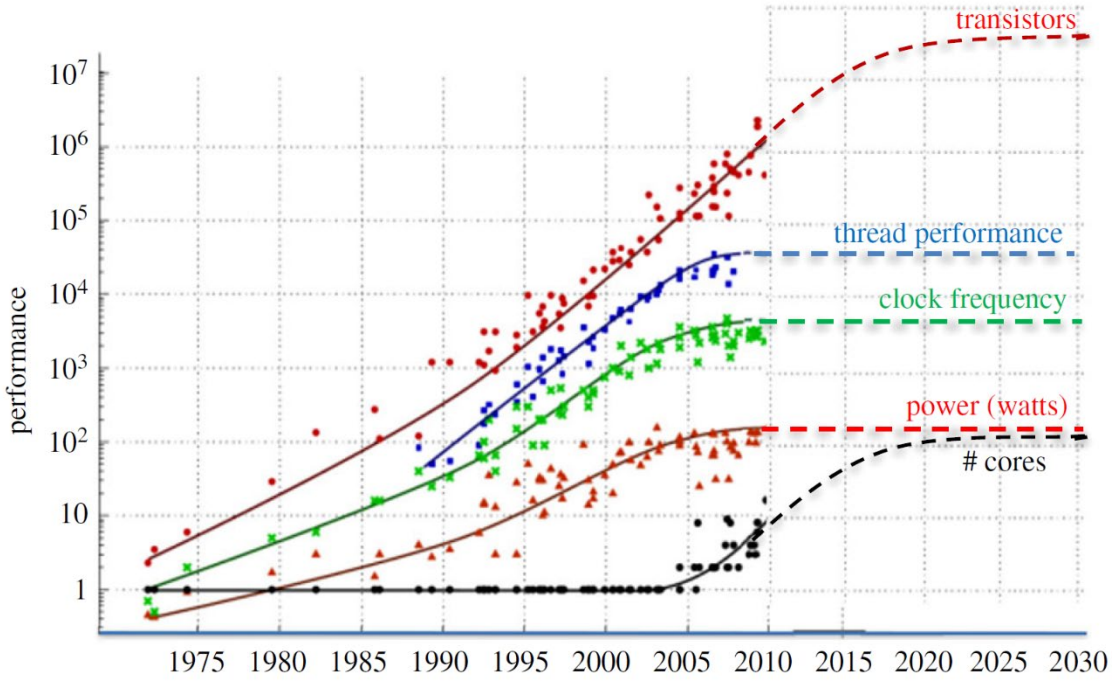
1）晶体管尺寸因不断逼近物理极限而减缓微缩，特别是进入10纳米以后更加趋缓

2）因芯片过热而不可无限提升主频，现大多控制在4GHz以内

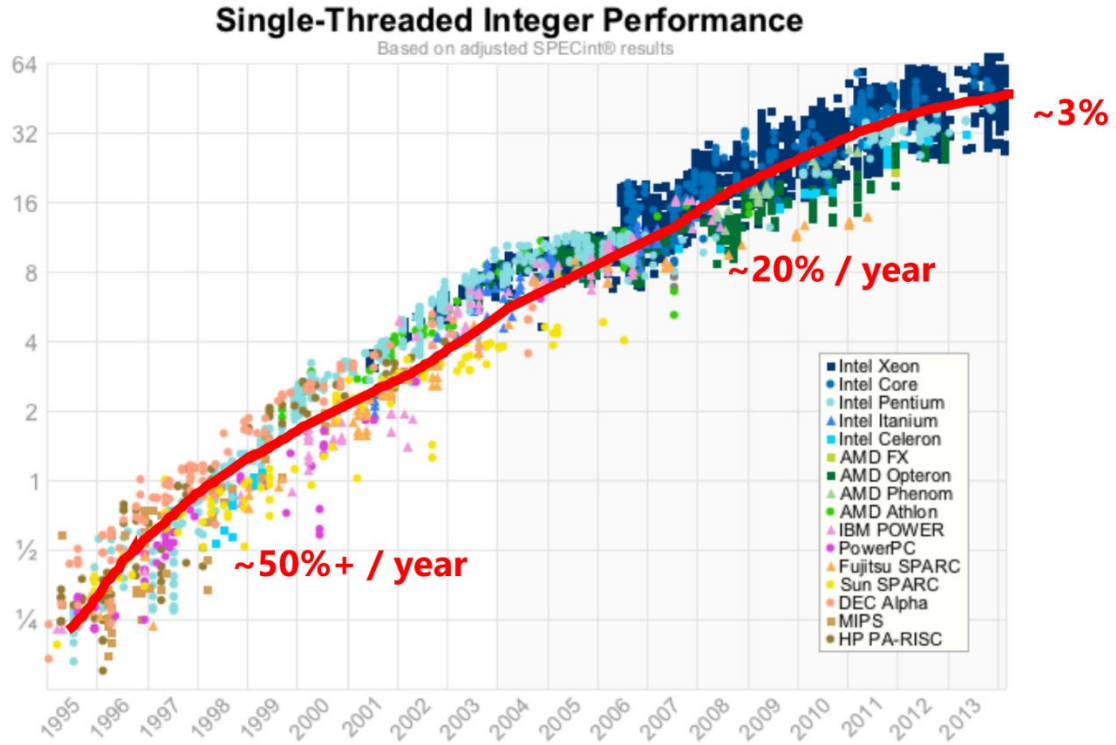
3）处理器自遭遇主频

升级瓶颈后，转向多核架构

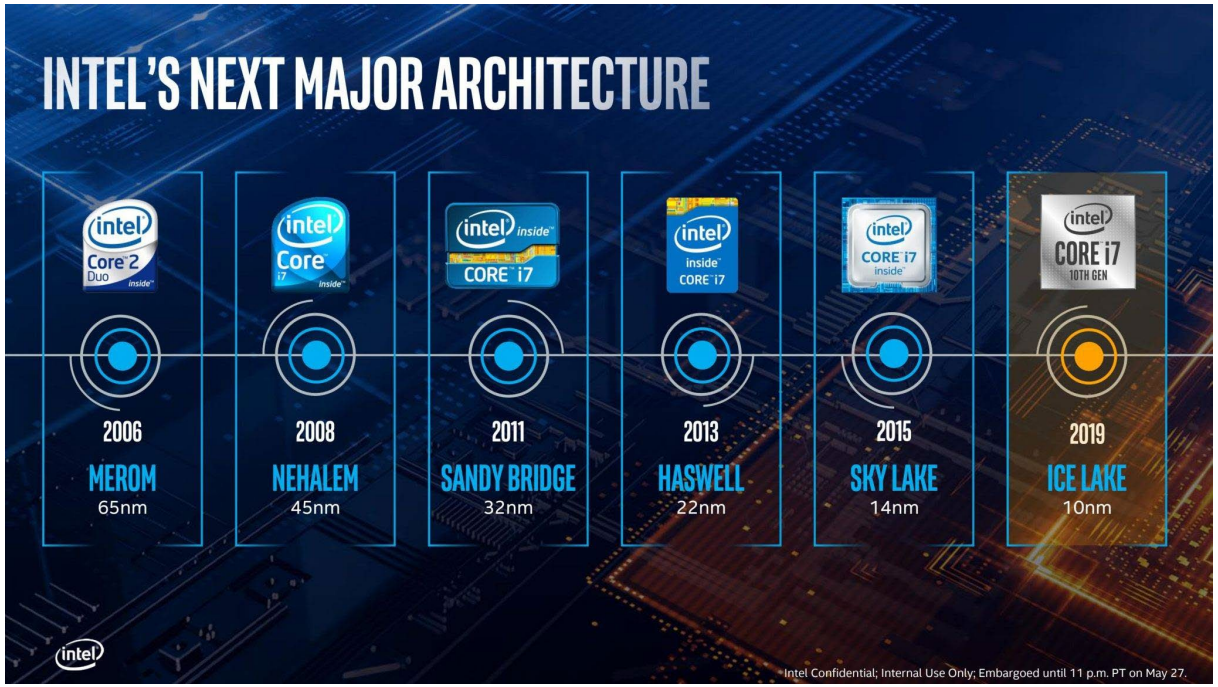
狭义的摩尔定律指每18到24个月，芯片上晶体管集成的密度会翻一番或者价格下降一半。普遍所讨论的是摩尔定律其实包含"摩尔定律"、"登纳德缩放"和"波拉克法则"三个重要法则。



CPU核面临瓶颈



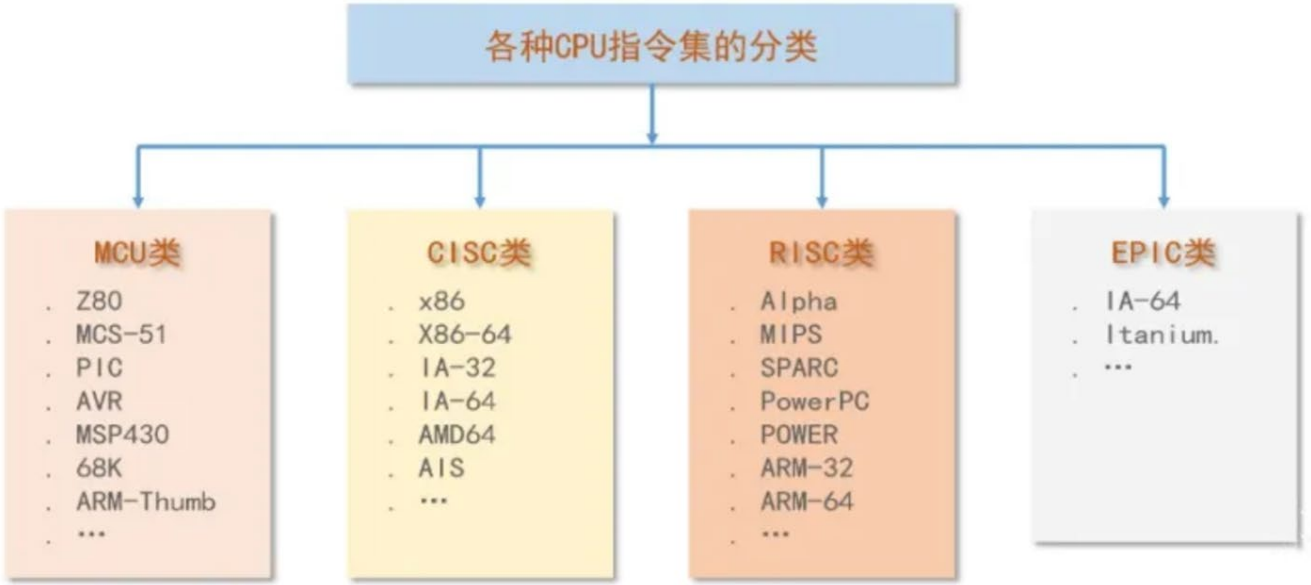
https://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/



制程（Fabrication Process）
CPU 制造厂商也都有各自的路线图来实现更小的制程，例如台积电、三星等在 2022 和 2023 年实现 3nm 和 2nm 的制造工艺

CPU指令集

- 计算机指令是计算机硬件直接能识别的命令
- 指令集架构是指一种类型CPU中用来计算和控制计算机系统的一套指令的集合
- 三种指令集
  - 复杂指令集CISC（Complex Instruction Set Computer）
  - 精简指令集RISC（Reduced Instruction Set Computing）
  - 精确并行指令集EPIC（Explicitly Parallel Instruction Computers）

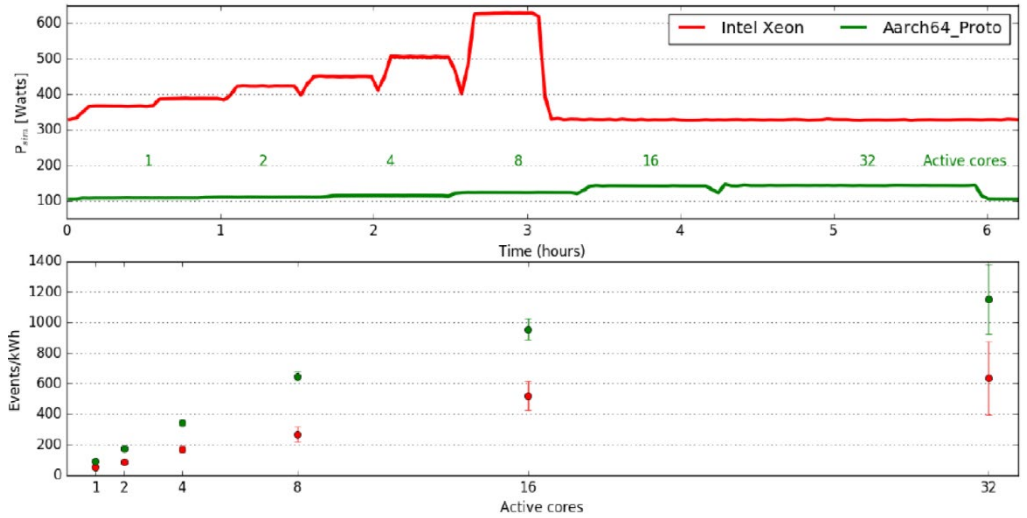


指令集路线	国产 CPU 厂商
MIPS	龙芯
x86	海光
	兆芯
ARM	飞腾
	鲲鹏
Alpha	申威

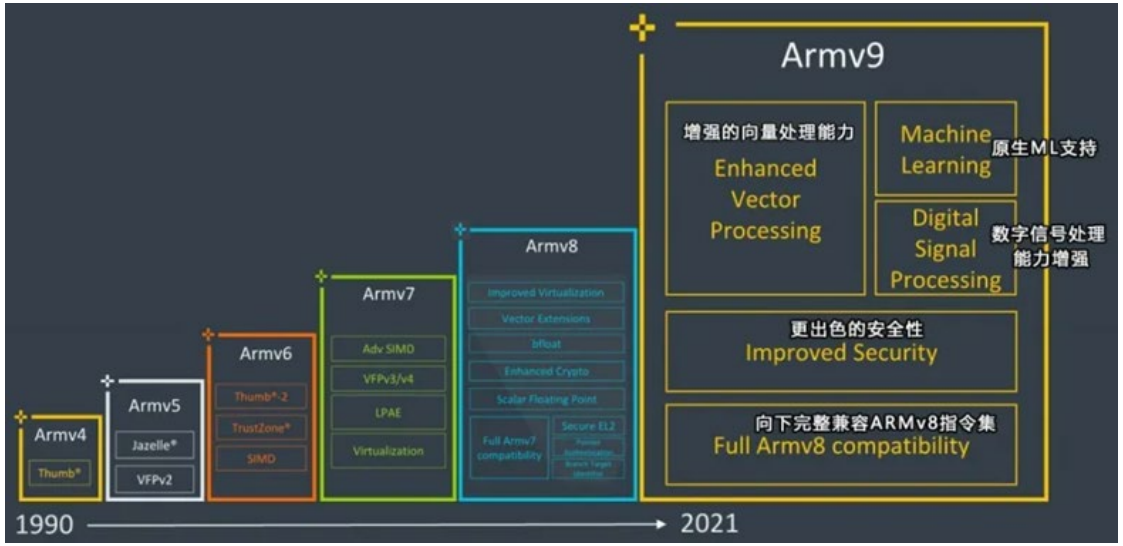
ARM指令集及发展

- 在移动设备等领域占有率高
  - ARM 应用于智能手机＞99%、车载信息设备＞95%、可穿戴设备＞90%。
- ARM通过三种授权模式，应用生态完善
  - 使用层级、内核层级、架构/指令集层级

	x86	ARM	
扩展性	重核多核多线程 高主频	轻核、众核	具有更好的并发性能
指令集	CISC, 通用指令集	RISC, 根据负载优化	匹配业务特征能耗比更佳
供应商	只有两家CPU供应商 Intel处于垄断地位	开放的授权策略 众多供应商	更加灵活丰富的选择
产业链	成熟	完善中	业界热点 快速发展

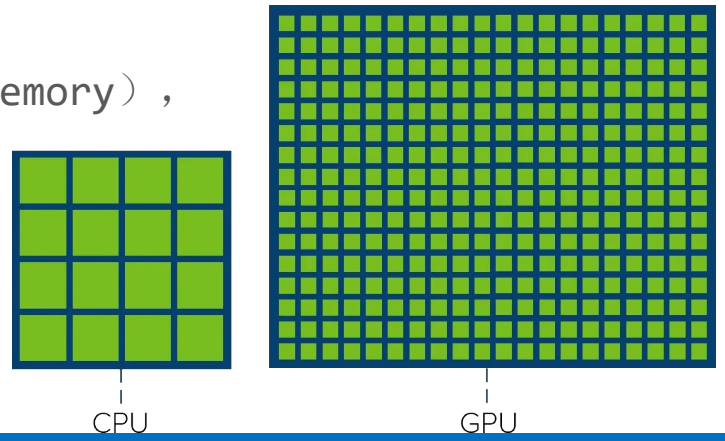


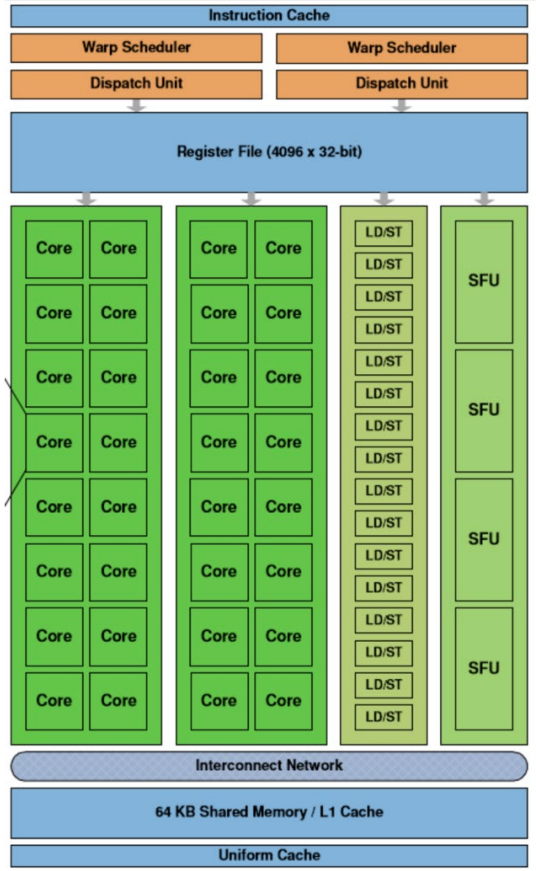
ATLAS测试结果



GPU

- 图形处理单元：Graphics Processing Unit，GPU
- 其设计与 CPU 完全不同，主要提高系统的吞吐量，在同一时间竭尽全力处理更多的任务
- 最初加速图片创建与渲染，由于其高度并行性，在AI及大数据处理方面得到广泛应用
- 流式多处理器（Streaming Multiprocessor，SM）是 GPU 的基本单元，包含以下部分：
  - 线程调度器（Warp Scheduler）：线程束（Warp）是最基本的单元，包含 32 或更多的并行的线程
  - 访问存储单元（Load/Store Queues）：在核心和内存之间快速传输数据
  - 核心（Core）：GPU 最基本的处理单元，也被称作流处理器
  - SM 中还包含特殊函数的计算单元（Special Functions Unit、SFU）以及用于存储和缓存数据的寄存器文件（Register File）、共享内存（Shared Memory），一级缓存和通用缓存等

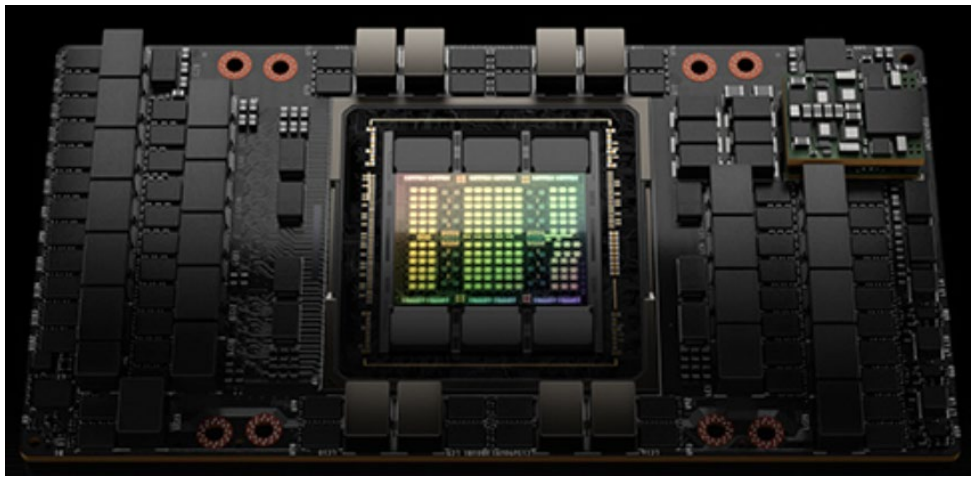




Nvidia Hopper H100架构

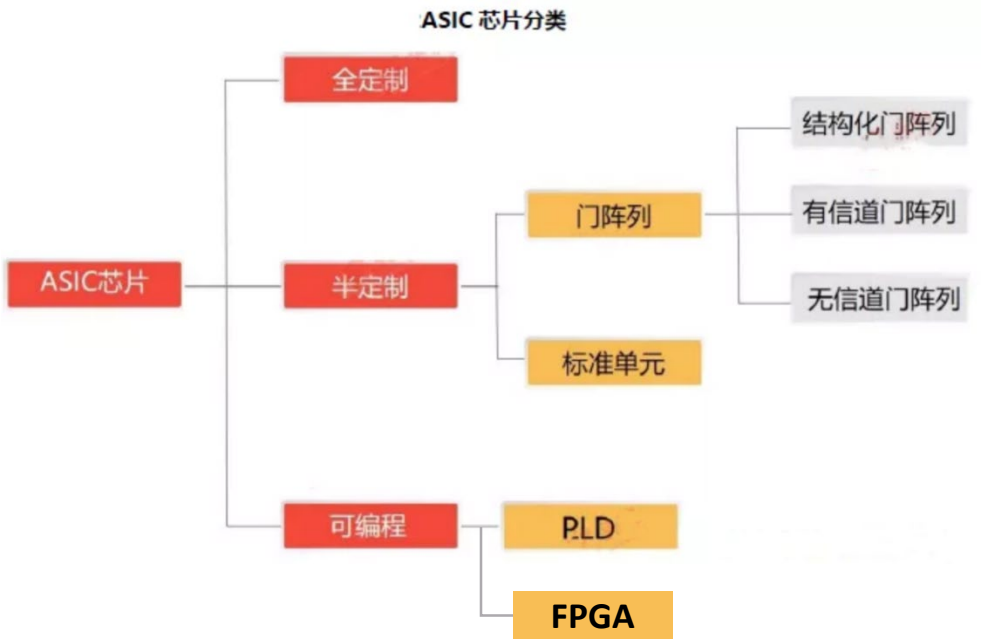
- 144个流式多处理器SM，每个SM有128个FP32内核、64个FP64内核和64个INT32，以及4个Tensor Core；800亿个晶体管，4nm工艺
- H100 GPU在FP16、FP32和FP64计算方面比A100快三倍，在8位浮点数学运算方面快六倍
- 专用核心以支持机器学习等特定应用
  - 比如：张量核心（Tensor Core）和光线追踪核心（Ray-Tracing Core）
  - **cuQuantum**加速量子计算模拟
- 功耗达到惊人的700W！

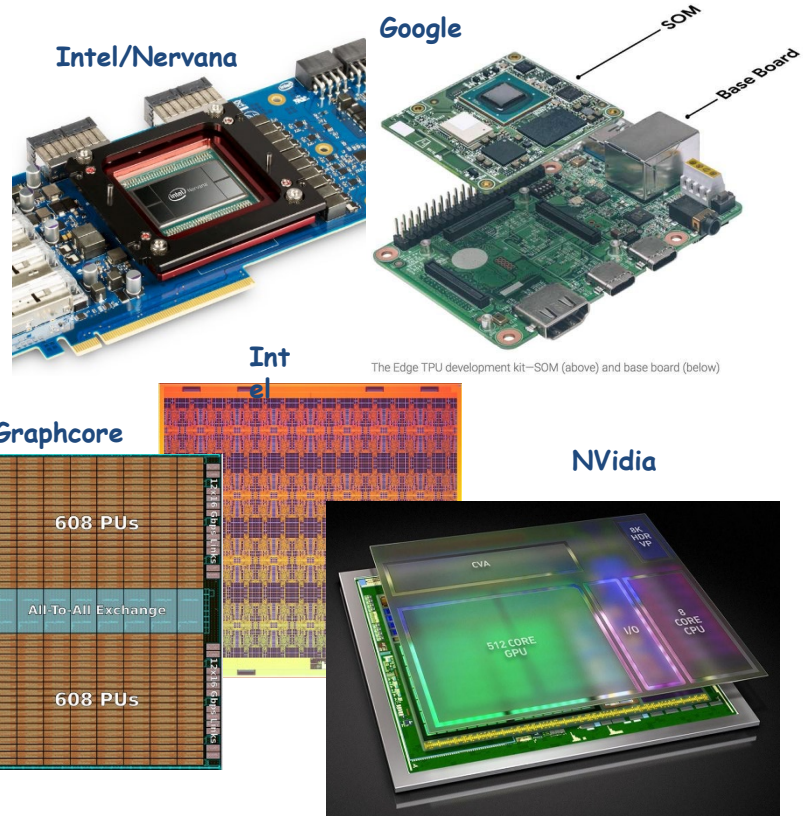
	H100 SXM	H100 PCIe
FP64	30 TFLOPS	24 TFLOPS
FP64 Tensor Core	60 TFLOPS	48 TFLOPS
FP32	60 TFLOPS	48 TFLOPS
TF32 Tensor Core	1000 TFLOPS*	800 TFLOPS*
BFLOAT16 Tensor Core	2000 TFLOPS*	1600 TFLOPS*
FP16 Tensor Core	2000 TFLOPS*	1600 TFLOPS*
FP8 Tensor Core	4000 TFLOPS*	3200 TFLOPS*
INT8 Tensor Core	4000 TOPS*	3200 TOPS*
GPU 显存	80GB	80GB
GPU 显存带宽	3TB/s	2TB/s
解码器	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
最大热设计功耗 (TDP)	700 瓦	350 瓦



ASIC

- ASIC(Application Specific Integrated Circuit):专用集成电路芯片
  - 针对用户对特定电子系统的需求，固定算法最优化设计
  - ASIC 芯片模块可广泛应用于人工智能设备、军事国防设备等智慧终端
  - 比如：NPU (Neural Networks Process Units), TPU(Tensor Processing Units), …
- 优点：面积小、能耗低、集成度高、价格低
- 缺点：设计周期长、更新频繁、市场风险高

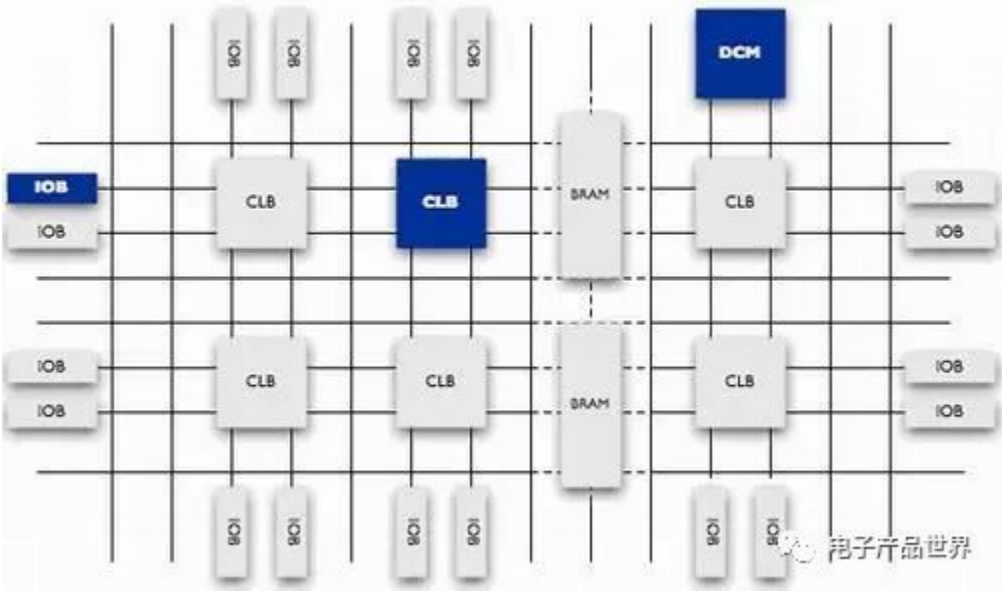




Source	NPU
Amazon	AWS Inferentia
Alibaba	Ali-NPU
Baidu	Kunlun
Bitmain	Sophon
Cambricon	MLU
Google	TPU
Graphcore	IPU
Intel	NNP, Myriad, EyeQ
Nvidia	NVDLA
Huawei	Ascend
Apple	Neural Engine
Samsung	NPU

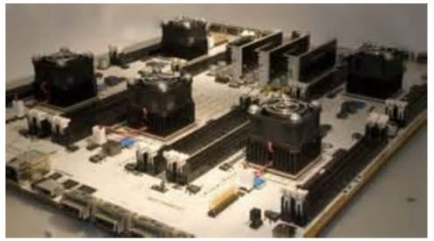
FPGA

- FPGA (Field Programmable Gate Array)：现场可编程门阵列
  - 先购买再设计的“万能”芯片
  - FPGA流片即可以成为ASIC
- 特点
  - 特定的FPGA硬件比通用的CPU更快更高效
  - 与CPU/GPU相比，更节能
  - 与ASIC相比，随时可以更改硬件配置：100ms-1s
  - 用量小时，不需要ASIC流片，成本低
- 内部结构
  - 可编程输入输出单元（IOB）
  - 可配置逻辑块（CLB）
  - 数字时钟管理模块（DCM）
  - 嵌入式块RAM（BRAM）
  - 丰富的布线资源
  - 底层内嵌功能单元

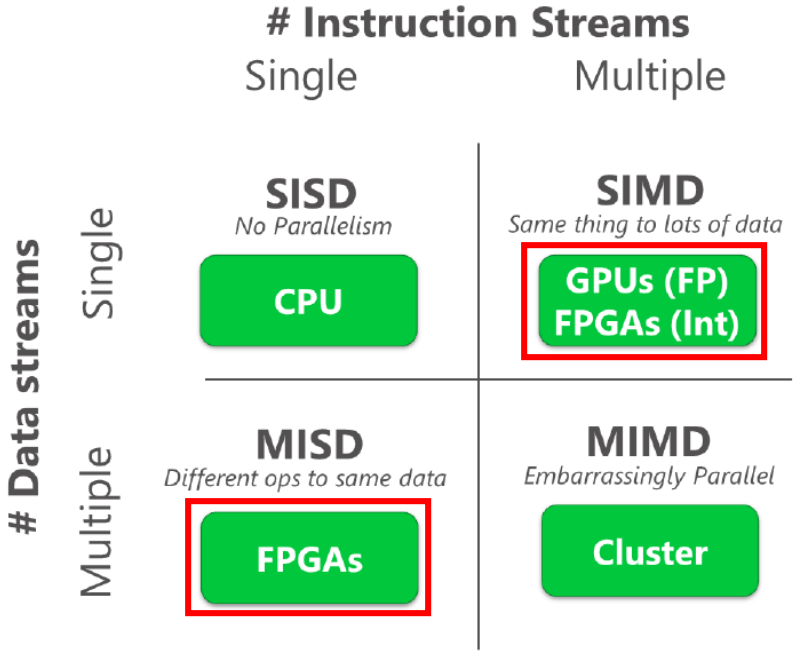


FPGA的并行性

- FPGA支持SIMD,MISD等多种形式的并行化
  - 对于MISD类的应用，即单一数据需要用许多条指令并行处理，FPGA更有优势
- 通过调整硬件资源，在特定的算法上FPGA的性能会超过GPU，因为GPU的硬件资源不能改动
- FPGA的功耗（~10W）远远低于GPU卡（~200W）
- 目前，FPGA在科研领域及大数据等方面应用较多

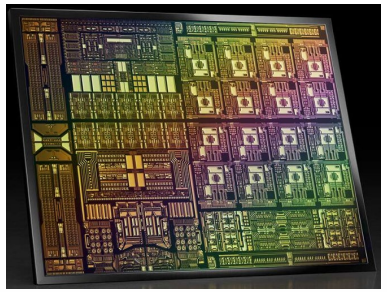


SETI



DPU

- 数据处理单元（Data Processing Unit，DPU)将成为继CPU、GPU的第三块主力芯片
- DPU主要应用场景
  - 算力卸载：作为通用算力补充，提供可编程和开放性
  - 存算结合：通过小型化、算力融合实现单位空间、单位功耗下最优处理效率
  - 以数据为中心：提升数据处理性能，以算力换空间
- DPU算力在数据密集型计算中的比例逐渐提升
- DPU主要实现模式



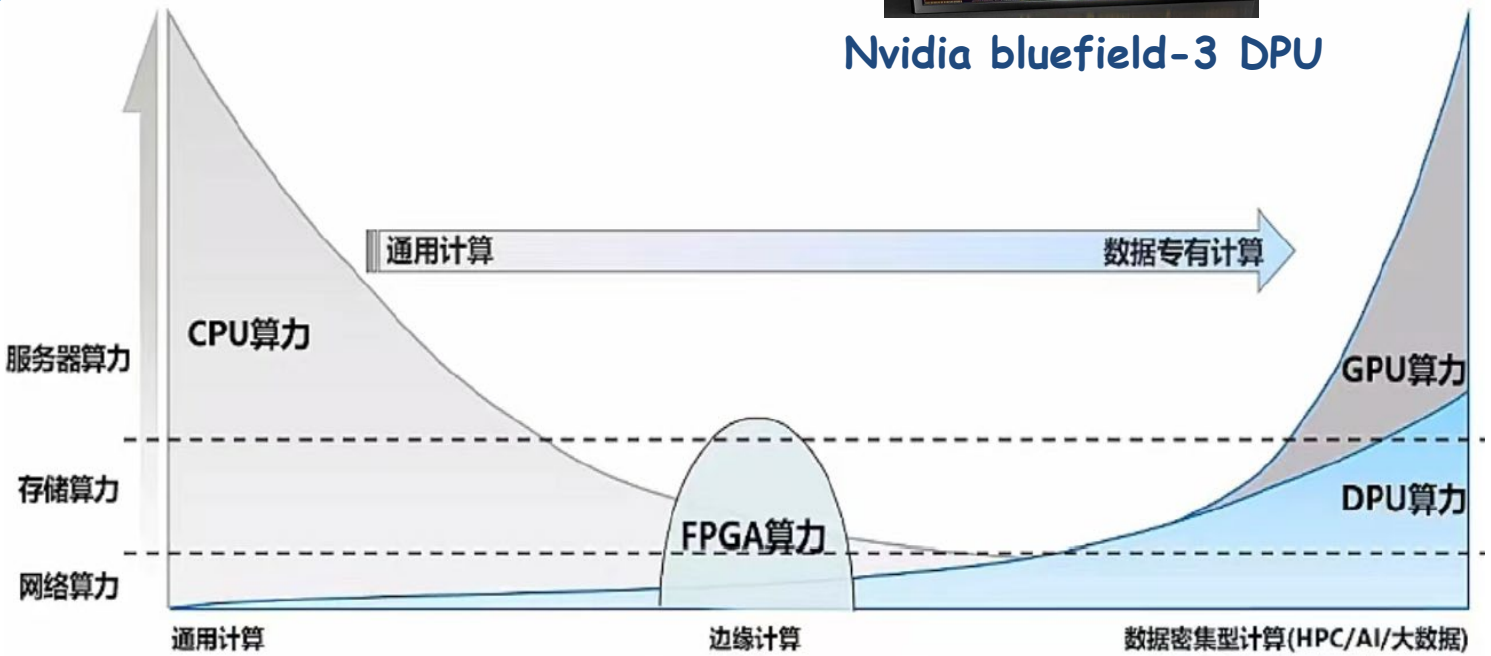
Nvidia bluefield-3 DPU

基于FPGA的DPU

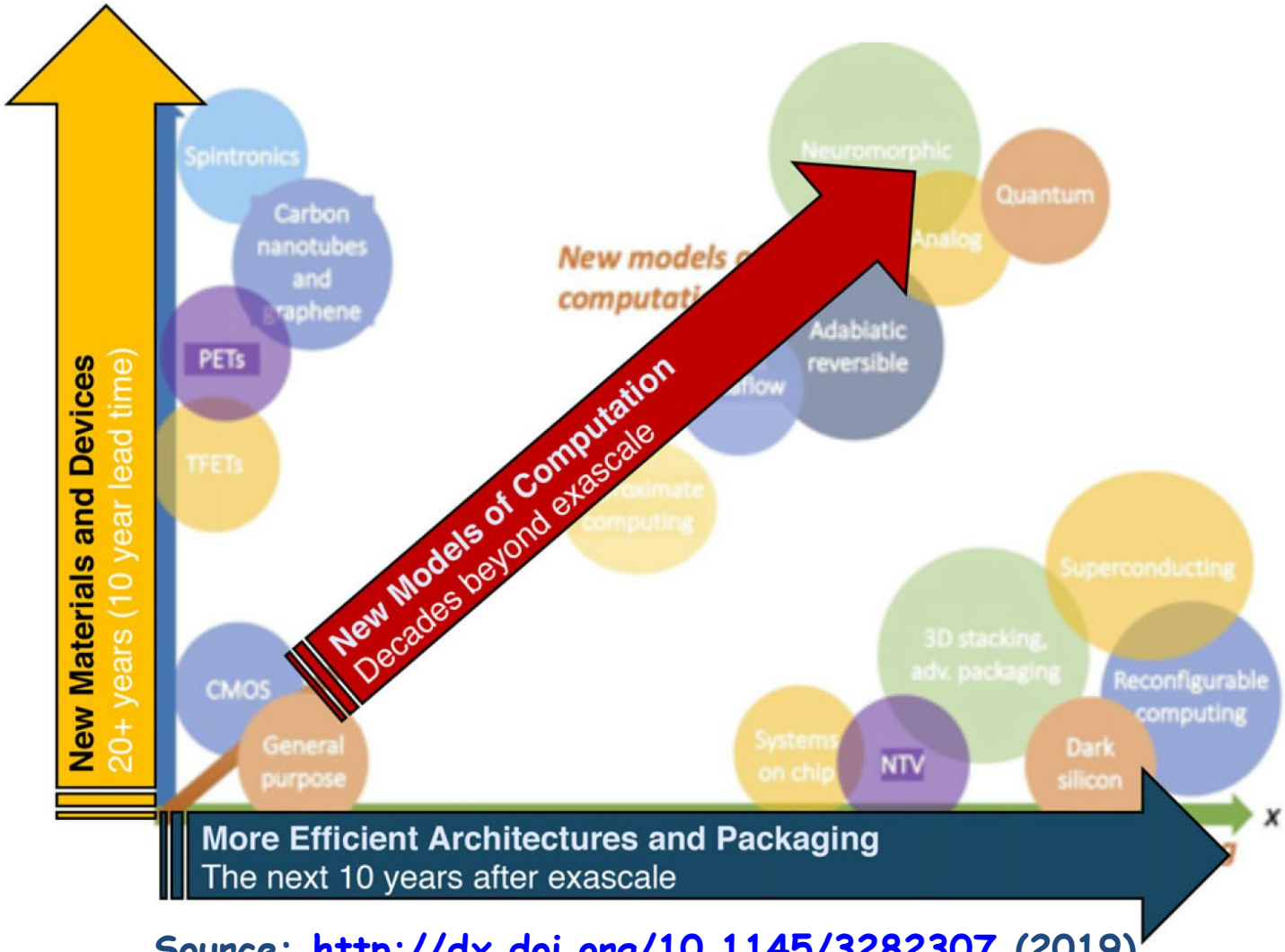
基于ASIC的DPU

基于多核芯片MP的DPU

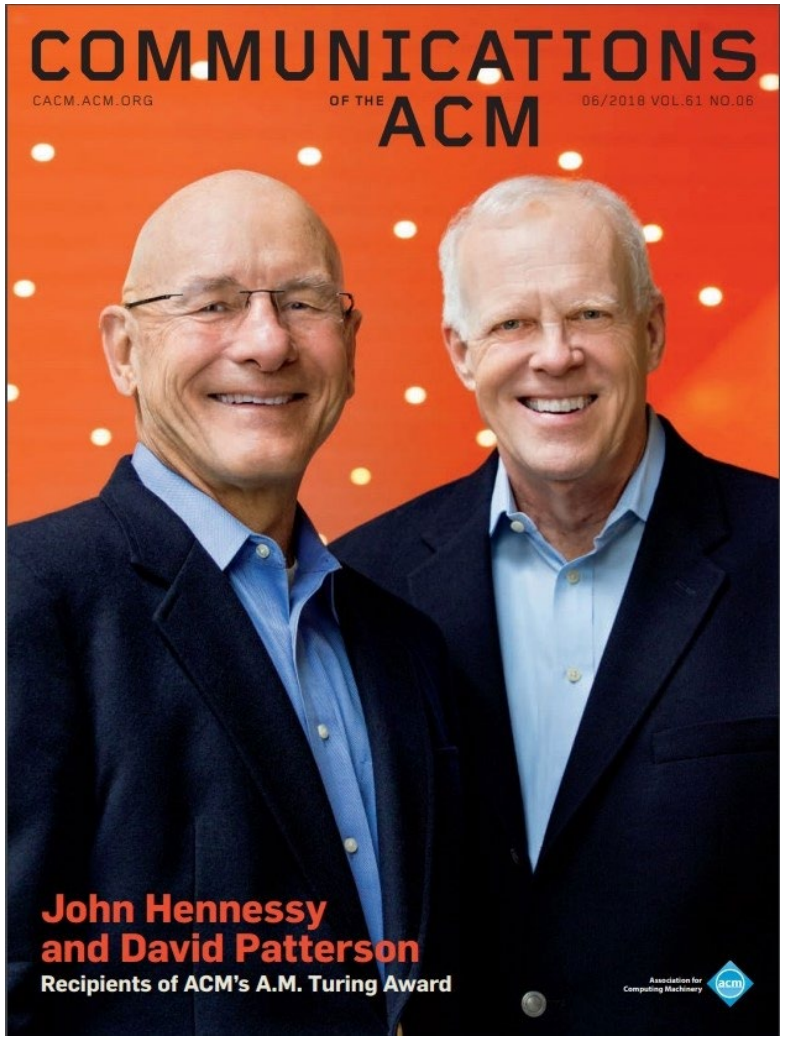
SoC-GP | SoC-NP



A New Golden Age for Computer Architecture?



Source: http://dx.doi.org/10.1145/3282307 (2019)

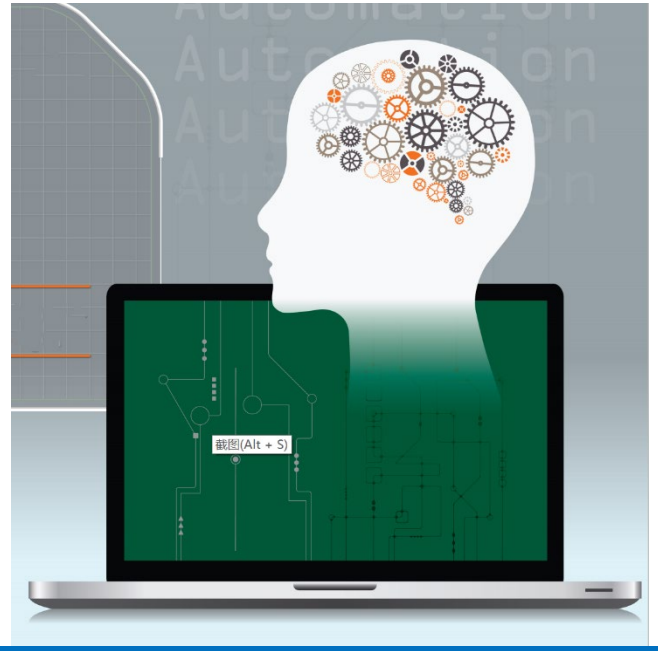


现阶段一些发展方向

- 异构计算渐成主流
  - CPU、GPU、DPU、FPGA、ASIC是目前通用计算领域的主流计算芯片
  - CPU芯片兼顾控制和计算，是构成笔记本、智能终端及服务器计算硬件主体
  - GPU芯片适合通用并行处理，应用领域由早期图像处理拓展至通用加速
  - FPGA芯片具备可重构特性，适合于需要定制的航空航天、车载、工业、科学计算等领域
  - AI ASIC芯片现已成为专用计算加速芯片创新的典型代表（TPU、寒武纪、…）
- 异构计算软件作用日益凸显
  - 从硬件开发转移到深化应用、软硬件融合创新阶段，软件对异构计算的支撑作用越来越明显
- 数据中心异构体系开启由GPU到DPU/FPGA的新变革
  - 将“CPU处理效率低下、GPU处理不了”的负载卸载到DPU/FPGA等专用芯片
- 高性能计算（超算、HPC）已经成为集群计算的重要应用领域
  - 2021年全世界开始步入E级计算时代
  - 格点QCD、宇宙学模拟、粒子物理模拟等是HPC重点应用

异构计算部件

再回到高能物理数据处理
——软件的挑战与发展



计算硬件的变迁

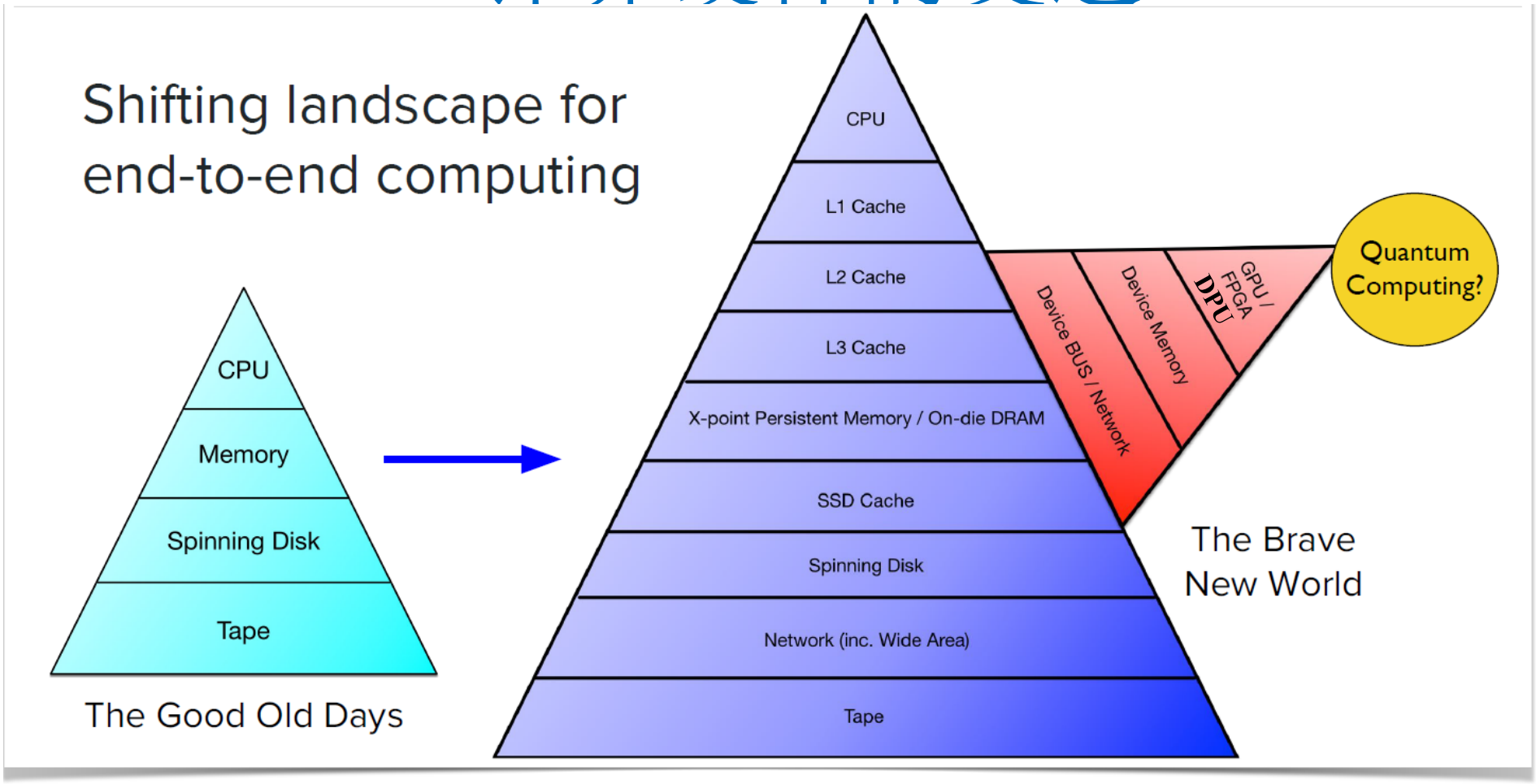


Image Credit: G. Stewart

- 计算加速部件是异构的，软硬协同，编程是不统一的
- 如何整合到高能物理数据处理？给高能物理软件带来了巨大的挑战！

并行处理

- 现代处理器硬件的一个核心特征是“并行处理”
  - SIMD – Single Instruction Multiple Data (vectorisation)
    - 在多个数据对象上执行相同的操作
  - MIMD – Multiple Instruction Multiple Data (multi-threading or multi-processing)
    - 同时在多个对象上执行多个操作
- 由于事例独立性使得高能物理数据处理具有天然并行性，通常采用粗粒度的并行化
  - 将一个任务分成多个作业，同时运行（作业并行）
  - 在一个作业内部同时处理多个事例（事例间并行）
  - 在一个事例上同时执行多个操作（事例内并行）
- 现代计算机硬件需要更加细粒度的并行，比如超级计算、数千个GPU卡并行
  - 比如“太湖之光”的主从核的众核架构等
  - 需要根据物理问题，重新思考新的架构与算法

异构编程

- 目前，已有多种并行硬件架构
- CPU并行指令集各不相同
  - X86: SSE4.2 (Streaming SIMD Extensions), AVX (Intel Advanced Vector Extensions), AVX2, AVX512
  - ARM: NEON, SVE (Scalable Vector Extension)
  - Others: VMX(Altivec), VSX (power7)
- GPU/FPGA架构
  - Nvidia, AMD, Intel, Sugon, Xilinx, Altera(Intel), …
- 其它超越冯诺依曼架构的体系，比如Intel CSA
- 各种编程框架
  - CUDA, TBB, OpenACC, OpenMP, OpenCL (→Vulcan), Kokkos, …
  - HIP (sugon DCU), Verilog, vivada/vitis, …
  - Hadoop, Spark, STORM, …

开源软件

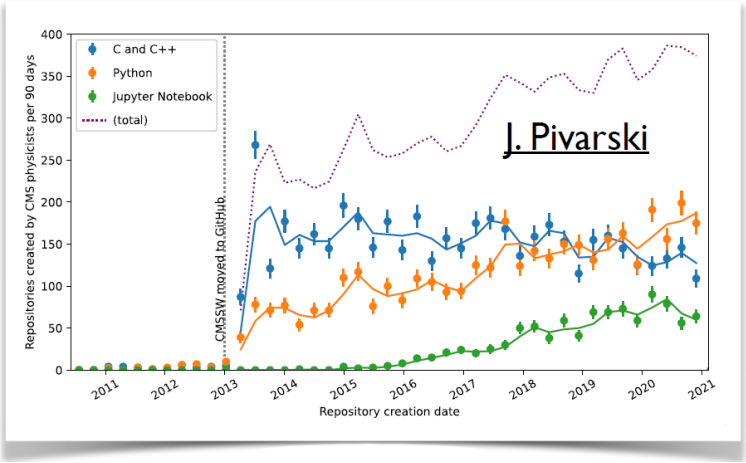
- 高能物理贡献了很多开源软件
  - 存储、计算、网络、数据分析等
  - GEANT4, ROOT, RUCIO, CVMFS, EOS, dCache, CASTOR, …
- 使用大量的业界开源系统
  - Openstack, CEPH, Hadoop, Spark, …
- 与业界开源软件互相融合，取长补短
  - 业界的开源社区活跃，文档齐全
- 挑战：如何整合开源软件，长期发展

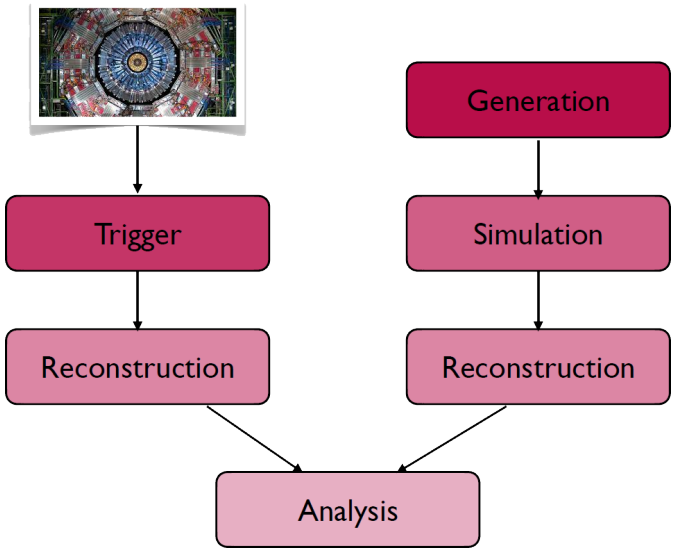
SWIFT-HEP
ExCALIBUR-HEP

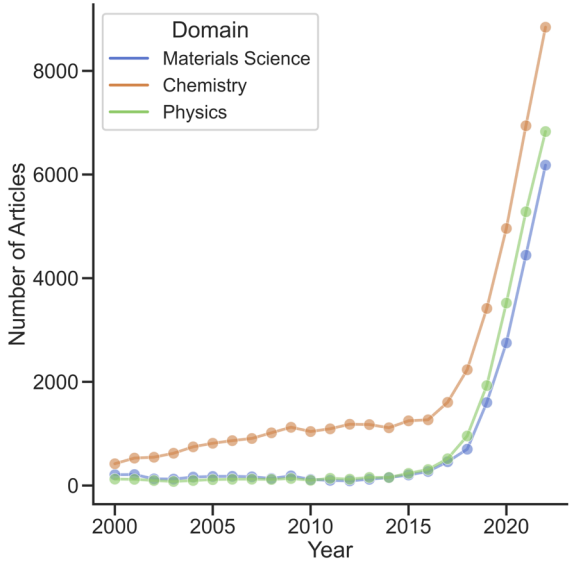

ALFA
HEP-CCE

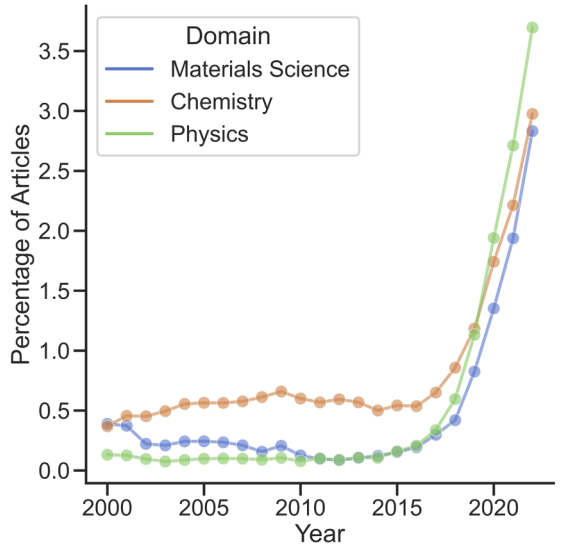

数据科学与机器学习

- 今年来，数据科学与机器学习蓬勃发展
  - 算法、算力、数据组成计算智能
- Python已经变成了数据科学的基本语言
  - 社区活跃，文档丰富
  - Numpy, matplotlib, pytorch, tensorflow, etc
- 上世纪90年代，高能物理领域开始机器学习技术
- 目前，机器学习已广泛应用于高能物理









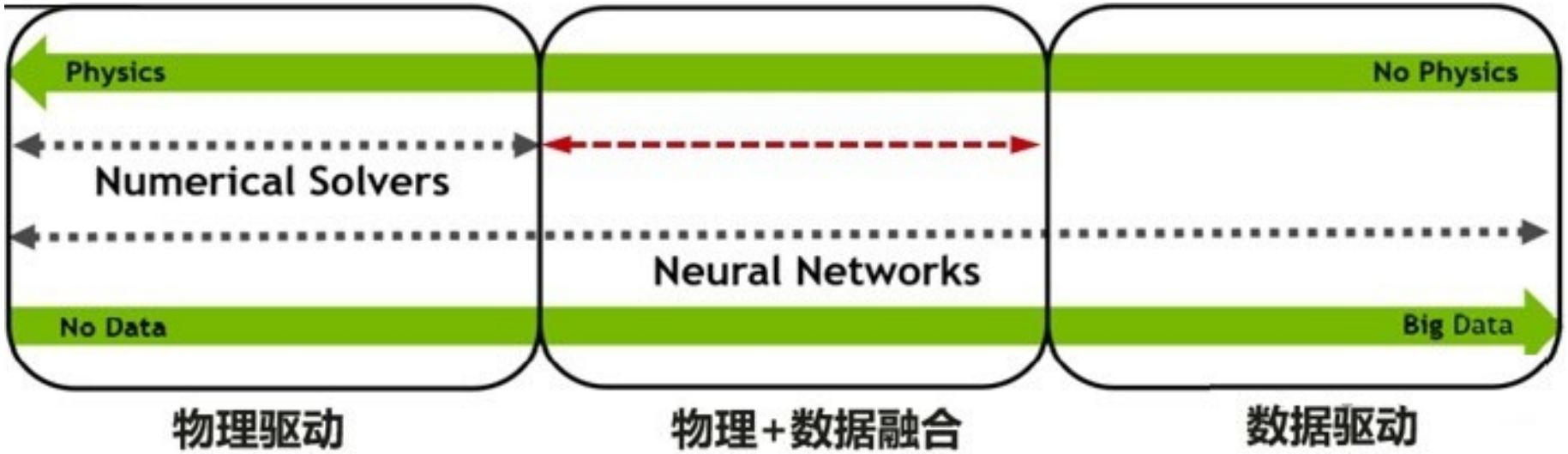
Ben Blaiszik, "2021 AI/ML Publication Statistics and Charts". Zenodo, Sep. 07, 2022. doi: 10.5281/zenodo.7057437.

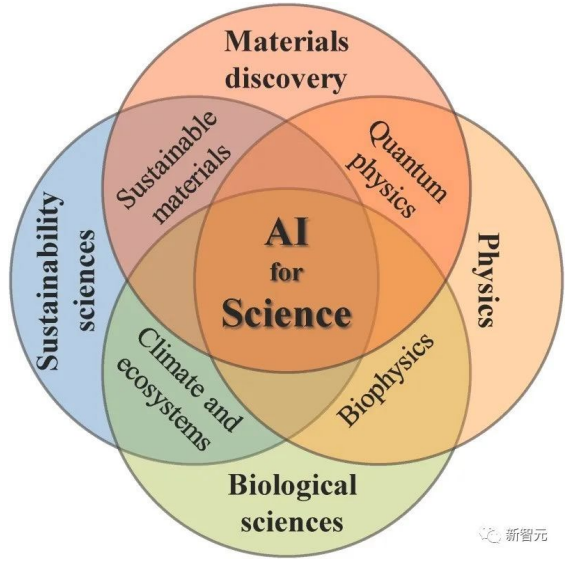
AI4Science

- AI4Science顾名思义：将大数据、机器学习、深度学习等人工智能技术应用到科学研究中
- AI处理多维、多模态的海量数据，不仅将加速科研流程，还将帮助发现新的科学规律
- AI4Science不仅仅是将AI应用到Science，而是两者相互影响与耦合，共同发展
- DeepMind最新研究登Nature，揭示AI时代科研新范式，开拓未知领域，孕育科技革命，其中包括数据采集、数据选择、数据标注、数据生成以及粒子物理应用等

Review | Published: 02 August 2023

Scientific discovery in the age of artificial intelligence https://www.nature.com/articles/s41586-023-06221-2





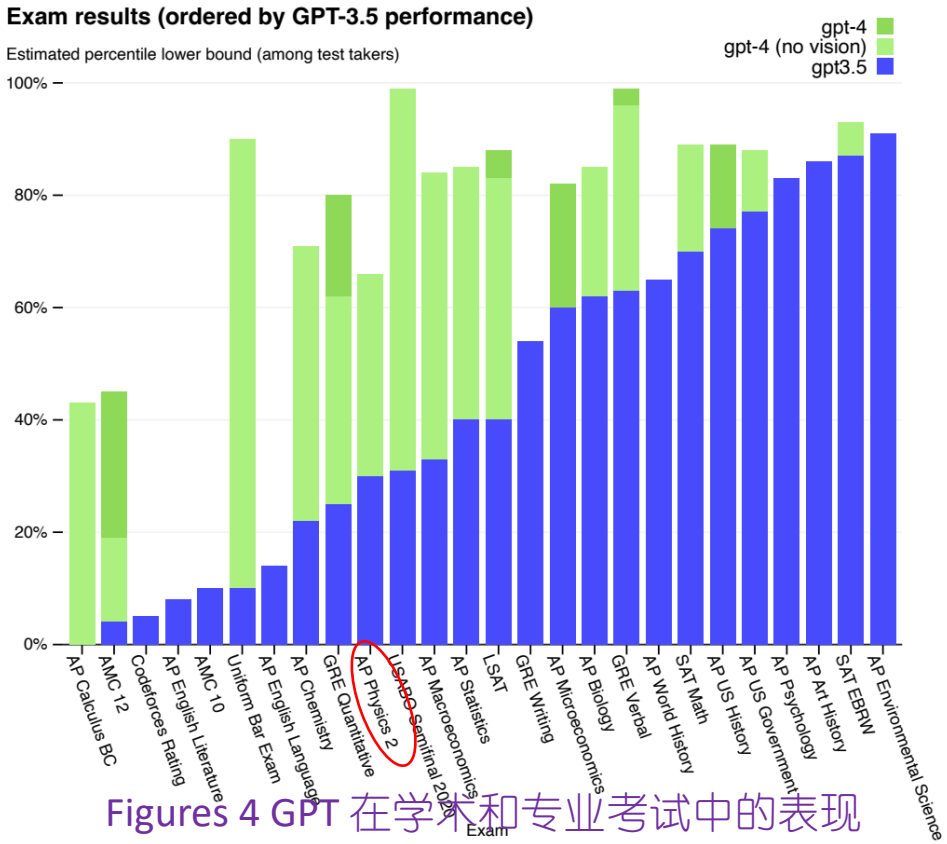
ChatGPT

如果问最近最火爆的科技热点是什么？非ChatGPT莫属。

ChatGPT是由OpenAI发布的人工智能对话生成系统。

聊天、写代码、修改bug、做表格、发论文、写作业、做翻译，无所不能...

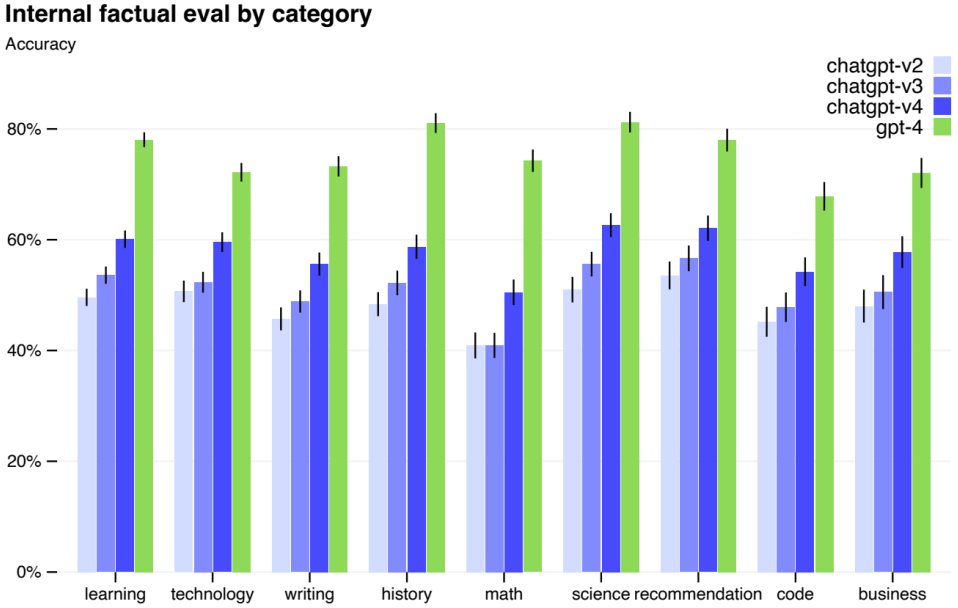
Chat GPT可以做物理吗？



Figures 4 GPT 在学术和专业考试中的表现

“尽管 GPT-4 具有多种功能，但它与早期的 GPT 模型具有类似的局限性。 最重要的是，它仍然不完全可靠”

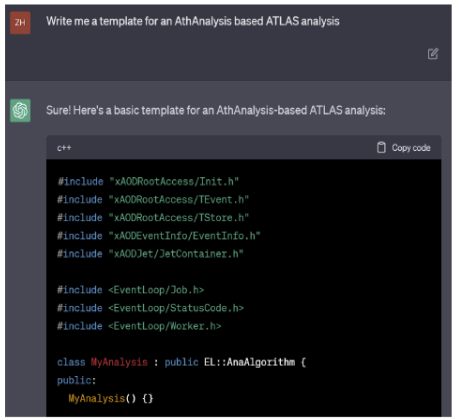
跨学科准确率约为 **75%**（图 **6**）



https://arxiv.org/pdf/2303.08774.pdf

Already useful in HEP IV

Generate ATLAS analysis template



OpenAI 2023年论文

NISQ量子计算时代

- 含噪声的中型量子时代（Noisy Intermediate-Scale Quantum, NISQ）
  - 50-100 Qubit，高噪声
  - 在高能物理领域，在Tracking, Analysis等方面已有应用

Article | Open Access | Published: 14 June 2023

Evidence for the utility of quantum computing before fault tolerance

Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout van den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme & Abhinav Kandala

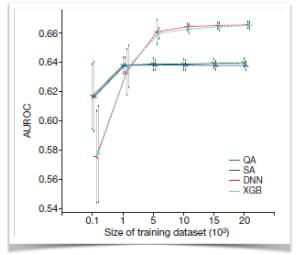
Nature **618**, 500–505 (2023) | Cite this article

新智元

2023年6月，IBM首次验证100+量子比特，无需纠错，依然可取得精确结果，甚至超越经典计算机

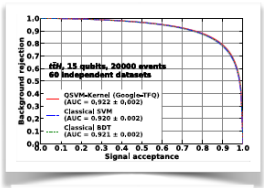
https://www.nature.com/articles/s41586-023-06096-3

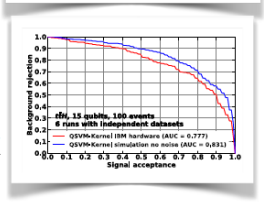
Analysis with ML



Mott, et al, doi:10.1038/nature24047

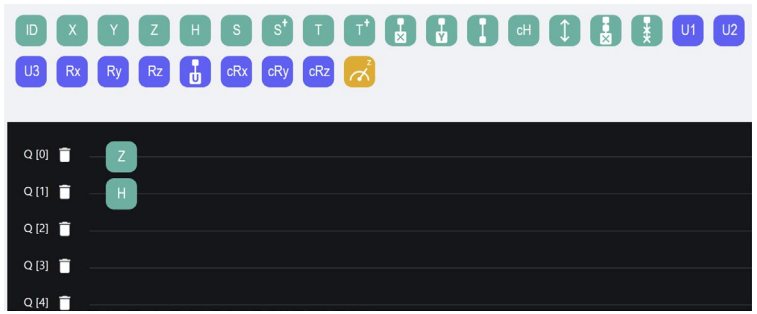
Zlokapa et al, arXiv: 1908.04480





Material courtesy of S.L.Wu, publication coming soon

- 动手编写量子程序：https://qc.ihep.ac.cn/




```
OPENQASM 2.0;
include "qelib1.inc";
qreg q[5];
creg c[5];
z q[0];
h q[1];
```

Qiskit Cirq

https://qc.ihep.ac.cn/login

高能所量子计算平台

总结

- 高能物理计算涉及到计算机体系结构、计算机软件与理论、科学数据处理等多个学科方向
- 计算、存储、网络是基础设施，数据是核心，软件框架是中间件，数据分析是顶层应用，科学发现是最终目标
- 本次暑期培训包括各类实验数据处理（粒子物理、天文实验、光源实验、宇宙线等），各种计算技术（高性能计算、数据存储与管理、网络与安全、软件框架、CUDA/HPC、异构计算、AI以及大模型等），**内容丰富**
- 采用先进的技术，进行规范的使用，以提高数据处理的效率，促进科学发现


我们的征途是星辰大海
仰望星空，脚踏实地
2023.8 拍摄于川西高原