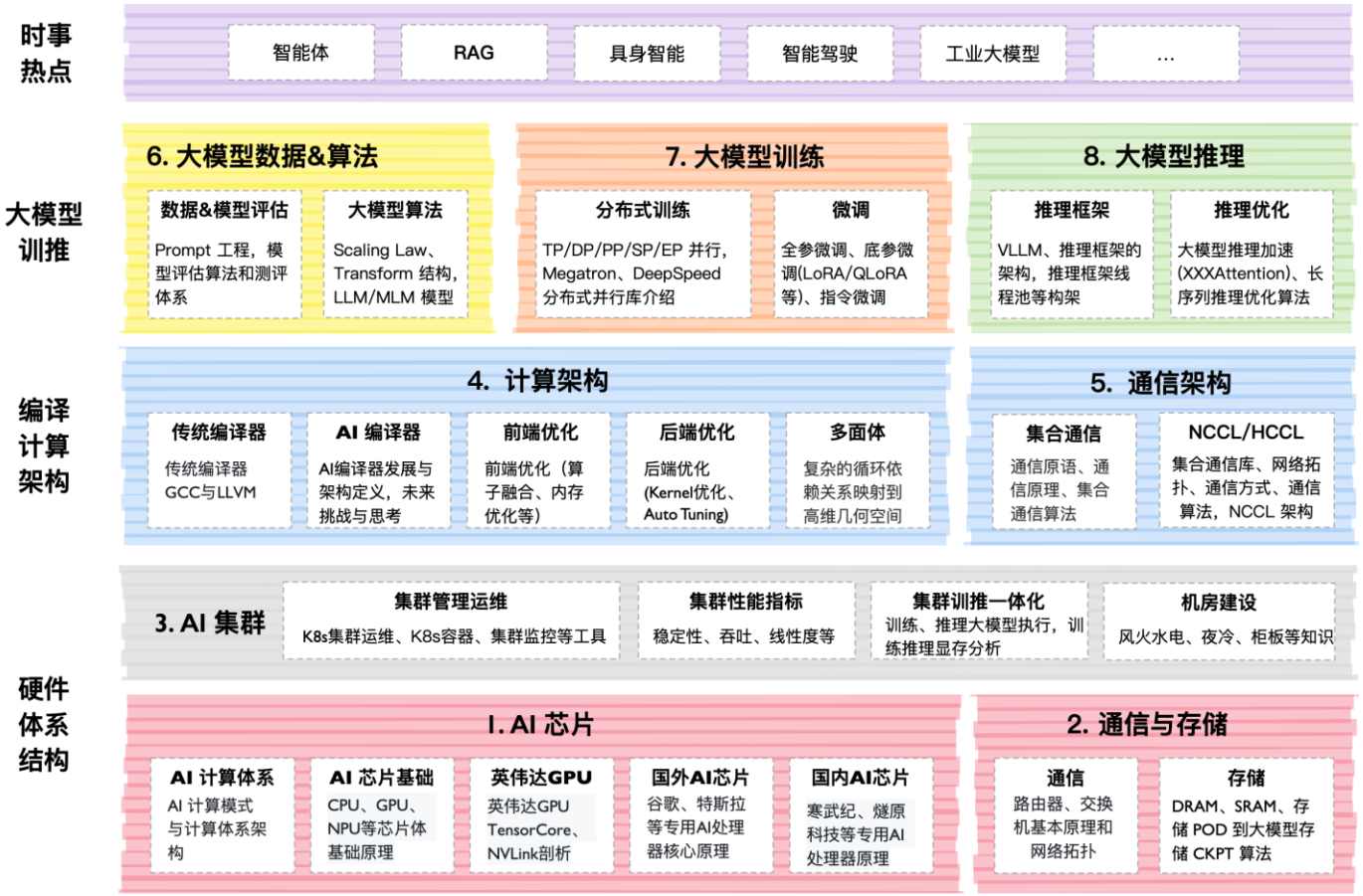

ZOMI
高性能计算HPC
定义
豆包AI生成

Content


计算集群之路
高性能计算定义
算力、分类、集群
高性能计算发展趋势
硬件架构、软件与编程
集群计算：发展历程
整体介绍、初代、二代、三代、未来
集群计算：系统挑战
总览、挑战、架构发展
集群计算：系统架构
L01/L1机房，L2基础，L3 使能

Content


AI 系统 + 大模型全栈架构图
大模型 & AI 应用
加速库 & 三方框架
AI 训练& 推理框架
AI 编译& 计算架构
AI 硬件& 体系结构
RAG
具身智能
自动驾驶
生成式推荐
...
07. 模型应用 & 热点
06. 模型算法与数据
多模态大模型 MLM
语言大模型 LLM
视觉大模型 VLM
视频生成大模型 SoRA (ViT)
文生图大模型 Stable Diffusion
Megatron-LM
DeepSpeed
Colossal-AI
HuggingFace
LLamaIndex
AutoChain
RAGFlow
Milvus
LangChain
PyTorch
TensorFLow
MindSpore
JAX
Caffe
04. AI 训练
vLLM
TensorRT
Xinference
MNN
05. AI 推理
CANN
CUDA
Triton
TVM
GLOW
XLA
TC
LLVM
GCC
AI编译器
传统编译器
超节点
Extra POD
Super POD
K8s 容器化部署
ubuntu 等操作系统
03. AI 集群
NPU
GPU
TPU
CPU
RDMA
IB
Ethernet
HBM
存储阵列
01. AI 芯片 (算)
02. 互联 (网)
02. 存储 (存)


时事热点
大模型训推
编译计算架构
硬件体系结构
智能体
RAG
具身智能
智能驾驶
工业大模型
...
6. 大模型数据&算法
数据&模型评估
Prompt 工程，模型评估算法和测评体系
大模型算法
Scaling Law、Transform 结构，LLM/MLM 模型
7. 大模型训练
分布式训练
TP/DP/PP/SP/EP 并行，Megatron、DeepSpeed 分布式并行库介绍
微调
全参微调、底参微调(LoRA/QLoRA 等)、指令微调
8. 大模型推理
推理框架
VLLM、推理框架的架构，推理框架线程池等构架
推理优化
大模型推理加速(XXXAttention)、长序列推理优化算法
4. 计算架构
传统编译器
传统编译器 GCC与LLVM
AI 编译器
AI编译器发展与架构定义，未来挑战与思考
前端优化
前端优化（算子融合、内存优化等）
后端优化
后端优化(Kernel优化、Auto Tuning)
多面体
复杂的循环依赖关系映射到高维几何空间
5. 通信架构
集合通信
通信原语、通信原理、集合通信算法
NCCL/HCCL
集合通信库、网络拓扑、通信方式、通信算法，NCCL 架构
3. AI 集群
集群管理运维
K8s集群运维、K8s容器、集群监控等工具
集群性能指标
稳定性、吞吐、线性度等
集群训推一体化
训练、推理大模型执行，训练推理显存分析
机房建设
风火水电、夜冷、柜板等知识
1. AI 芯片
AI 计算体系
AI 计算模式与计算体系架构
AI 芯片基础
CPU、GPU、NPU等芯片体基础原理
英伟达GPU
英伟达GPU TensorCore、NVLink剖析
国外AI芯片
谷歌、特斯拉等专用AI处理器核心原理
国内AI芯片
寒武纪、燧原科技等专用AI处理器原理
2. 通信与存储
通信
路由器、交换机基本原理和网络拓扑
存储
DRAM、SRAM、存储 POD 到大模型存储 CKPT 算法

新基建，中国数字经济引擎的顶层规划

“数字经济不能建设在沙滩上”

习总书记2018年关于自主创新的重要讲话

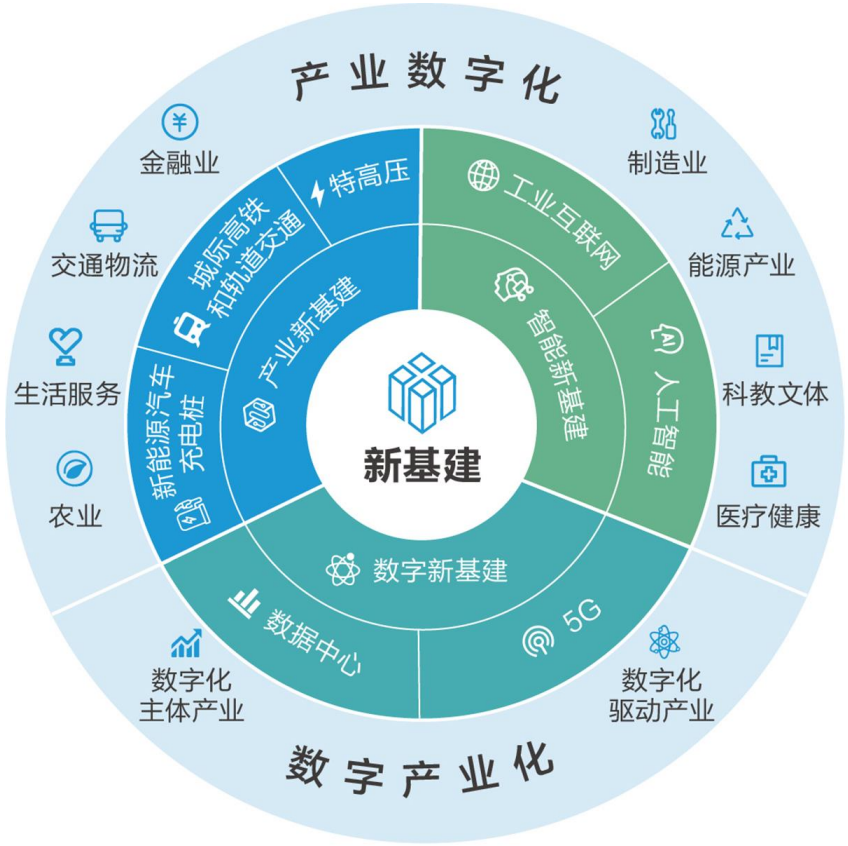
掌握创新和发展主动权，促进中国产业迈向全球价值链高端。

掌握核心技术，加快推进国产自主可控替代

“要适度超前，布局有利于引领产业发展和维护国家安全的基础设施”

习总书记2022年4月26日主持召开中央财经委员会第十一次会议

要加强信息、科技、物流等产业升级基础设施建设，布局建设**新一代超算、云计算、人工智能平台、宽带基础网络**等设施，推进**重大科技基础设施**布局建设



新基建和数字经济

数字新基建（**数据中心**）和智能新基建（**人工智能**）

Content

1. 什么是算力（算力单位、算力空间）
2. 高性能计算（数值分析、并行计算）
3. 集群计算（高性能计算中心、云数据中心、 AI 智算中心）

01

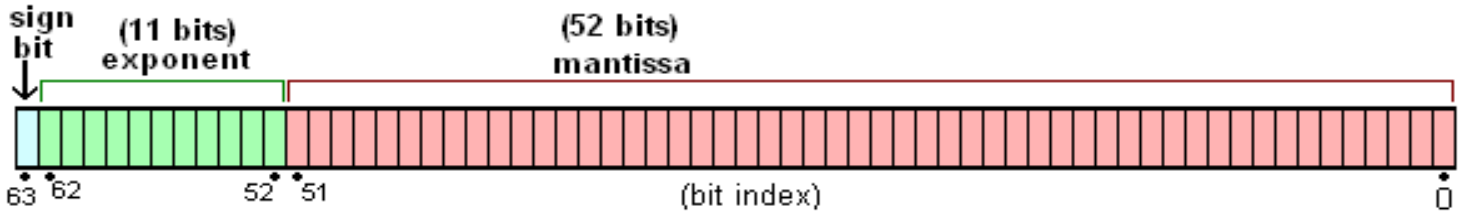
算力

Computational Power

小知识：算力水平是如何衡量的?

- **FLOPS** = **FL**oating point **O**perations **P**er **S**econd，每秒浮点计算次数
- HPC一般数据类型 64位双字长精度（FP64），FLOPS一般采用FP64的算力水平来评估

Processors in Mate60 Pro	FP64 (双精度)	FP32 (单精度)
HiSilicon Kirin 9000	1,166GFLOPS	2,332 GFLOPS



- **IEEE-754 标准** 双精度浮点数 $\boldsymbol{D} = (-1)\boldsymbol{s} \times 1.\boldsymbol{f} \times 2^{\mathbf{e}-1023}$,
- 其中，**s** 是**符号位**（0为正1为负），**f** 是（52位二进制）**尾数**，**e** 是（11位二进制）**指数**

小知识：算力水平是如何衡量的?

Name	Unit（单位）	Value	中文描述	算力水平约等于多少部Mate40 Pro
Kilo FLOPS	KFLOPS	10^3	每秒一千次	0.000000001部
Mega FLOPS	MFLOPS	10^6	每秒一百万次	0.000001部
Giga FLOPS	GFLOPS	10^9	每秒十亿次	0.001部
Tera FLOPS	**TFLOPS**	**10^{12}**	每秒一万亿次	一部Mate40Pro，1/12部XBOX
Peta FLOPS	**PFLOPS**	**10^{15}**	每秒一千万亿次	一千部（武汉超算，算力200P）
Exa FLOPS	**EFLOPS**	**10^{18}**	每秒一百京次	一百万部（深圳超算，算力2.5E）
Zetta FLOPS	ZFLOPS	10^{21}	每秒十万京次	十亿部
Yotta FLOPS	YFLOPS	10^{24}	每秒十亿京次	一万亿部

数字经济时代，算力是国家间的核心竞争力

- 中国人均算力水平当前仅为美国的1/5，提升空间巨大；预计到2026年，中国算力市场将持平甚至超越美国。
- 工信部数据显示，截至2024年三季度末，中国算力总规模达268EFLOPS（每秒百亿亿次浮点运算，以FP32单精度计算），算力应用项目超过1.3万个，在用算力中心机架总规模超过880万架标准机架，算力总规模居世界前列。

6 产经 2025年4月22日 星期二 经济日报

产业聚焦·AI新场景①

人工智能优化算力布局

水库安全管理一刻不能松懈

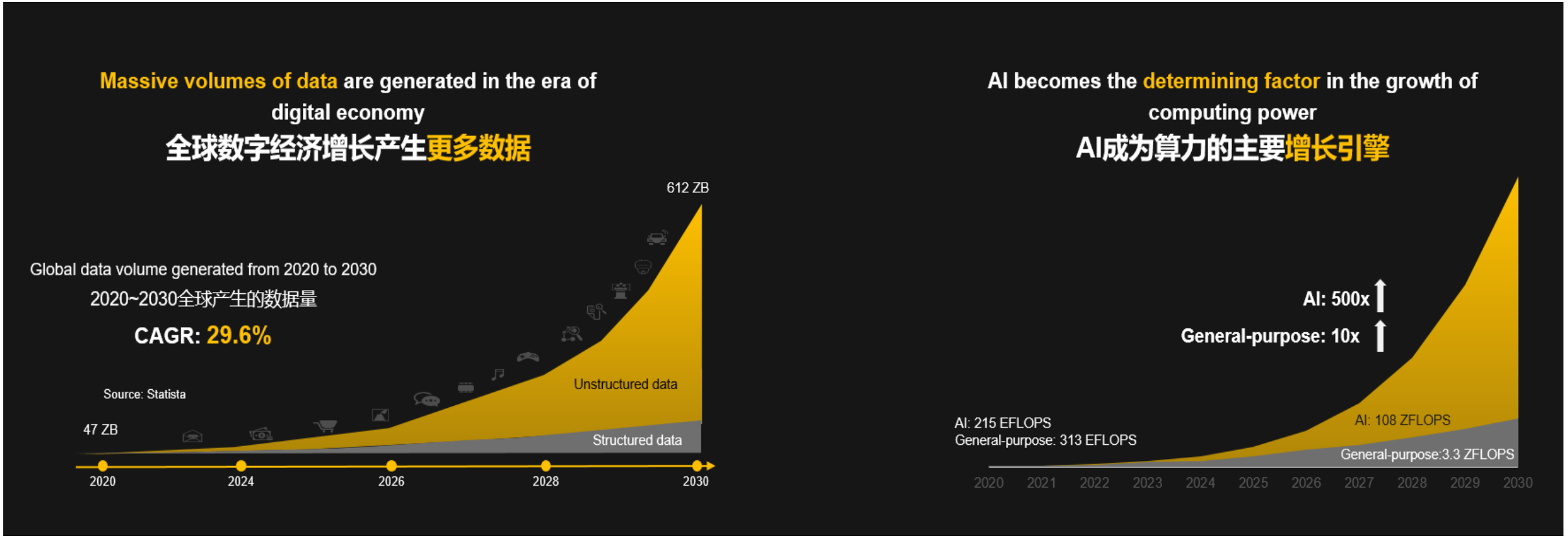
车网互动开启规模化应用

重庆安全技术职业学院

“三位一体”推动高职院校安全教育评价改革

算力规模

- 1961年，AI之父 John McCarthy 提出：“算力服务将成为未来的公共基础设施”，同电话网络一样重要
- 2030年，全球产生数据量年均复合增长29.6%，通用算力需求将增长10倍，AI算力需求将增长500倍



02

高性能计算 HPC

数值分析和并行计算

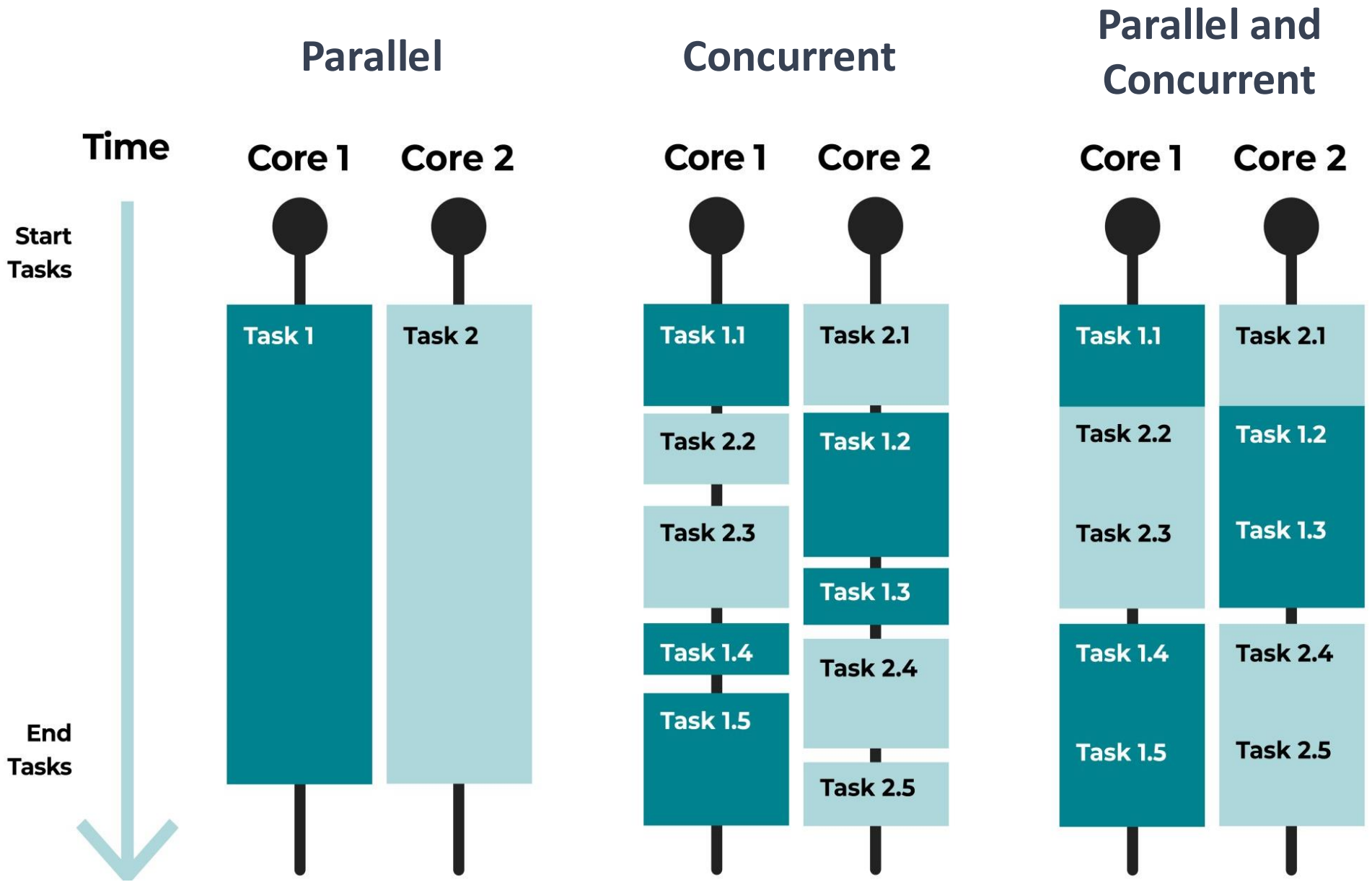
$$\frac{dy}{dx}$$

- 很多人知道韦神 但很多人不知道的是他的研究方向是N-S方程


第一年
第二年
第三年
第四年
第五年

并发计算 vs并行计算

- **并发计算 Concurrent Computing：**
  - 是一种程序计算形式，在系统中，至少有两个以上的计算任务在同时运作，计算结果可能同时发生。并发概念强调的是单个处理器在单位时间内完成多个任务，类似“一个人同时吃三个馒头”
- **并行计算 Parallel Computing：**
  - 一般是指许多指令得以同时进行的计算模式。在同时进行的前提下，可以将计算的过程分解成小部分，之后以并发方式来加以解决。并行概念强调的是多个处理器同时完成多个任务。


Parallel
Concurrent
Parallel and Concurrent
Time
Start Tasks
End Tasks
Core 1
Core 2
Task 1
Task 2
Task 1.1
Task 2.2
Task 2.3
Task 1.4
Task 1.5
Task 2.1
Task 1.2
Task 1.3
Task 2.4
Task 2.5
Task 1.1
Task 2.2
Task 2.3
Task 1.4
Task 1.5
Task 2.1
Task 1.2
Task 1.3
Task 2.4
Task 2.5

03

集群计算

cluster computing

集群计算

- **计算机集群**简称**集群**，是一种计算机系统， 它通过一组松散集成的计算机软件或硬件连接起来高度紧密地协作完成计算工作。

集群计算

- 集群是一个完整系统，而非多个计算机系统，可以被看作是**一整台计算机**（类似于“蚁群”/“蜂群”社会系统），是主要的算力基础设施建设方式。

算力基础设施建设

- 从分层堆叠走向集群全栈优化，达成规模系统最优。

传统DC中服务器部署形态

1. 人工连线耗时长，故障率高，施工维护复杂；风冷为主，系统能效差
2. 计算、散热、电源网络松耦合集成，算力利用率低，算力密度低

服务器主板

服务器节点

Cloud

标准公有云—计算服务器 1U4P

大数据

数据服务器节点 2U2P

高密

HPC服务器节点 2U4P

AI服务器

AI服务器节点 Atlas 800
4CPU+8NPU

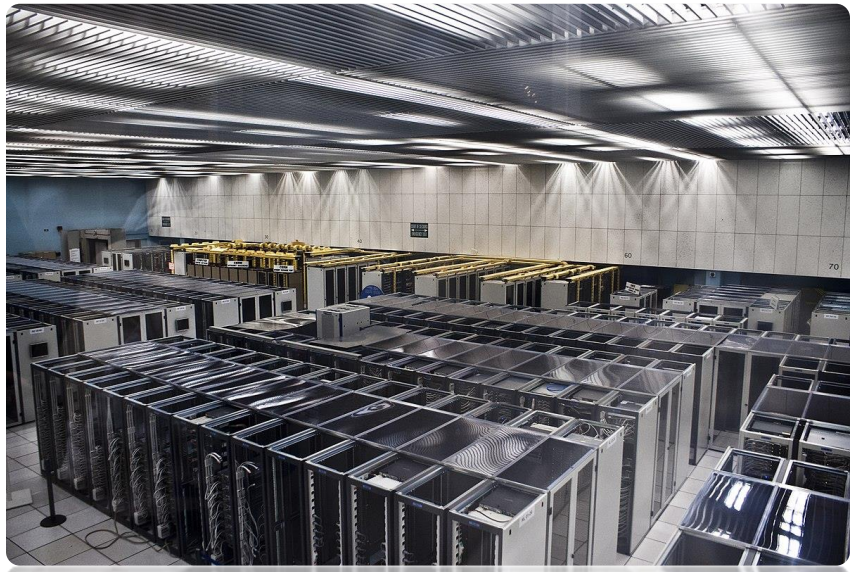
标准服务器机柜内堆叠

集群计算规模应用的场景

高性能计算中心

人工智能计算中心

云数据中心



集群计算规模应用的场景：计算中心

- **专用**：面向特定计算需求的**高性能并行计算**。面向国防、科研等重算力超算场景，以及人工智能大模型训练场景，算力和功耗惊人，体系架构专为特定应用算力摸高定制优化。

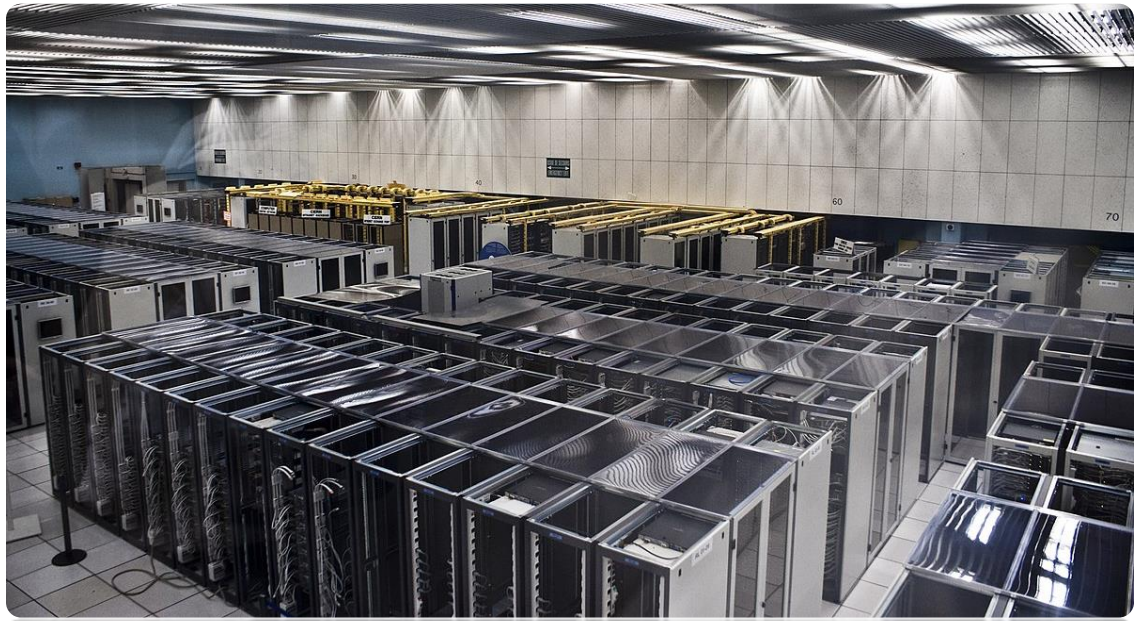
高性能计算中心

人工智能计算中心

集群计算规模应用的场景：数据中心

- **通用**：面向通用计算的**大规模并发处理**。面向大众商用，提供互联网IT服务，并行处理规模惊人，计算相对简单，强调高性价比和高兼容性，对可靠性、可用性、可服务性（RAS）要求高。

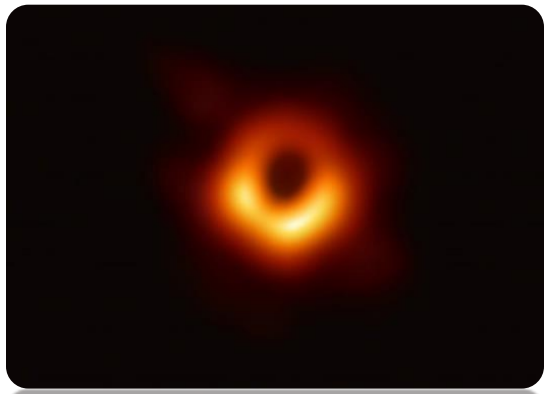
云数据中心



高性能计算

- 衡量和体现国家科技核心竞争力和综合国力重要标志，各国公认支撑科学技术发展战略技术。
- 计算是科研的三大范式之一，科研三大范式：理论、实验、计算。

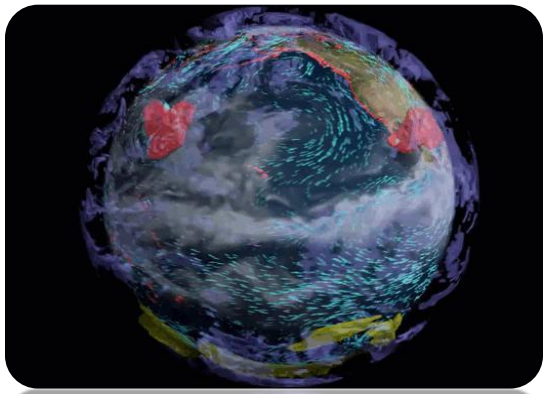
算天



Source: EHT Collaboration

人类第一张黑洞照片

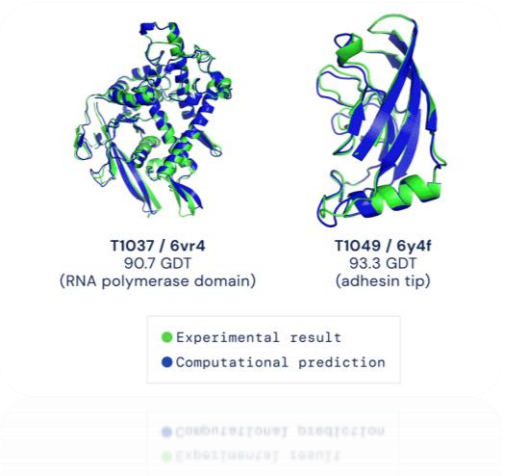
算地



Source: Nvidia

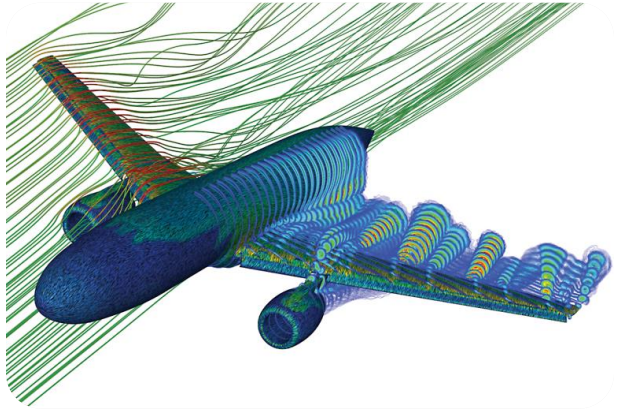
全球水循环变化

算人



生命科学

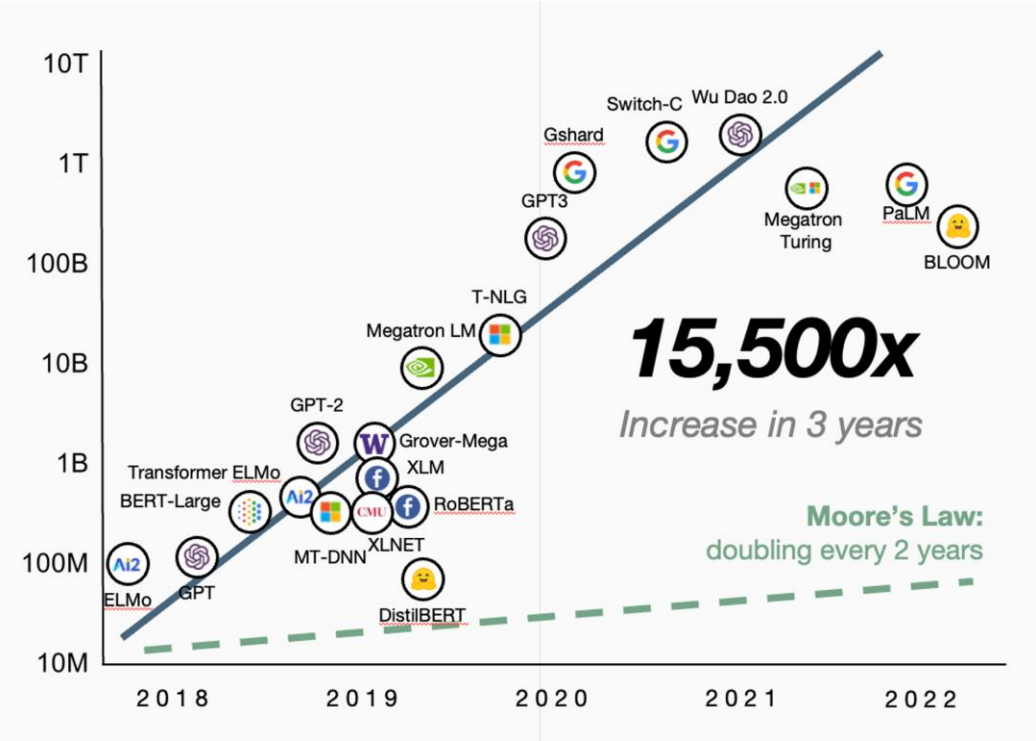
算物

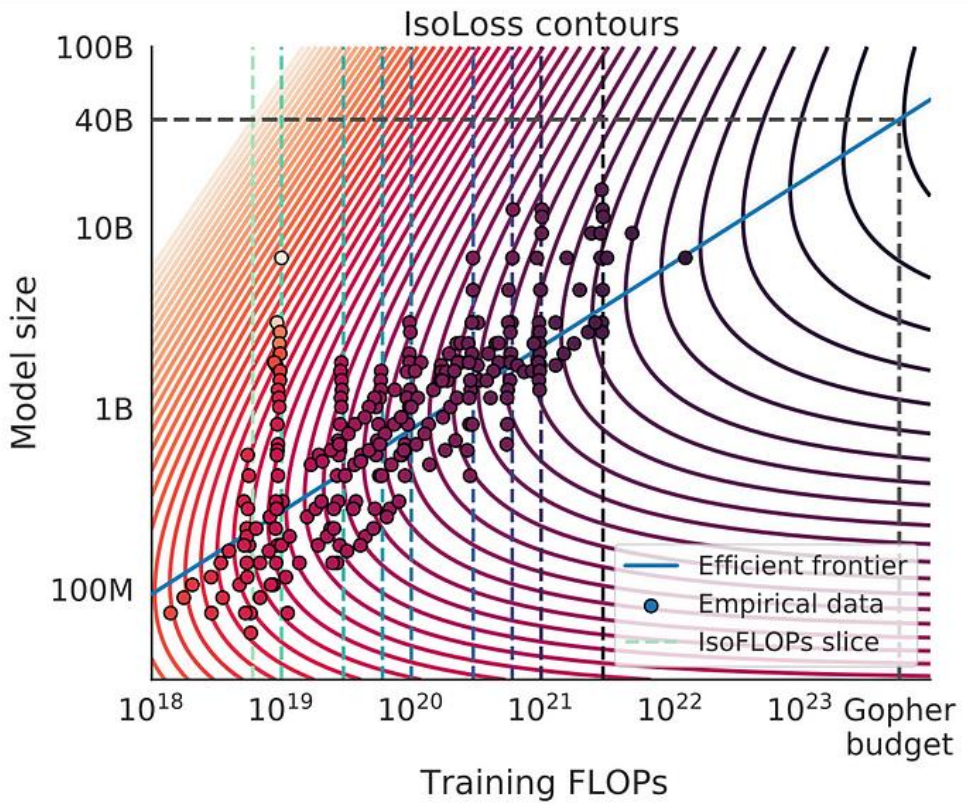


战略装备

人工智能计算

- AI成为新形势下中美科技竞争的新赛道，**大数据+大模型 驱动 训练&推理** 智能算力爆发增长
- 大模型算力需求每2年增长750倍 vs 硬件算力供给按摩尔定律每2年仅增长3倍
- 基于FFN前馈神经网络+Attention注意力机制的Transformer通用大模型架构

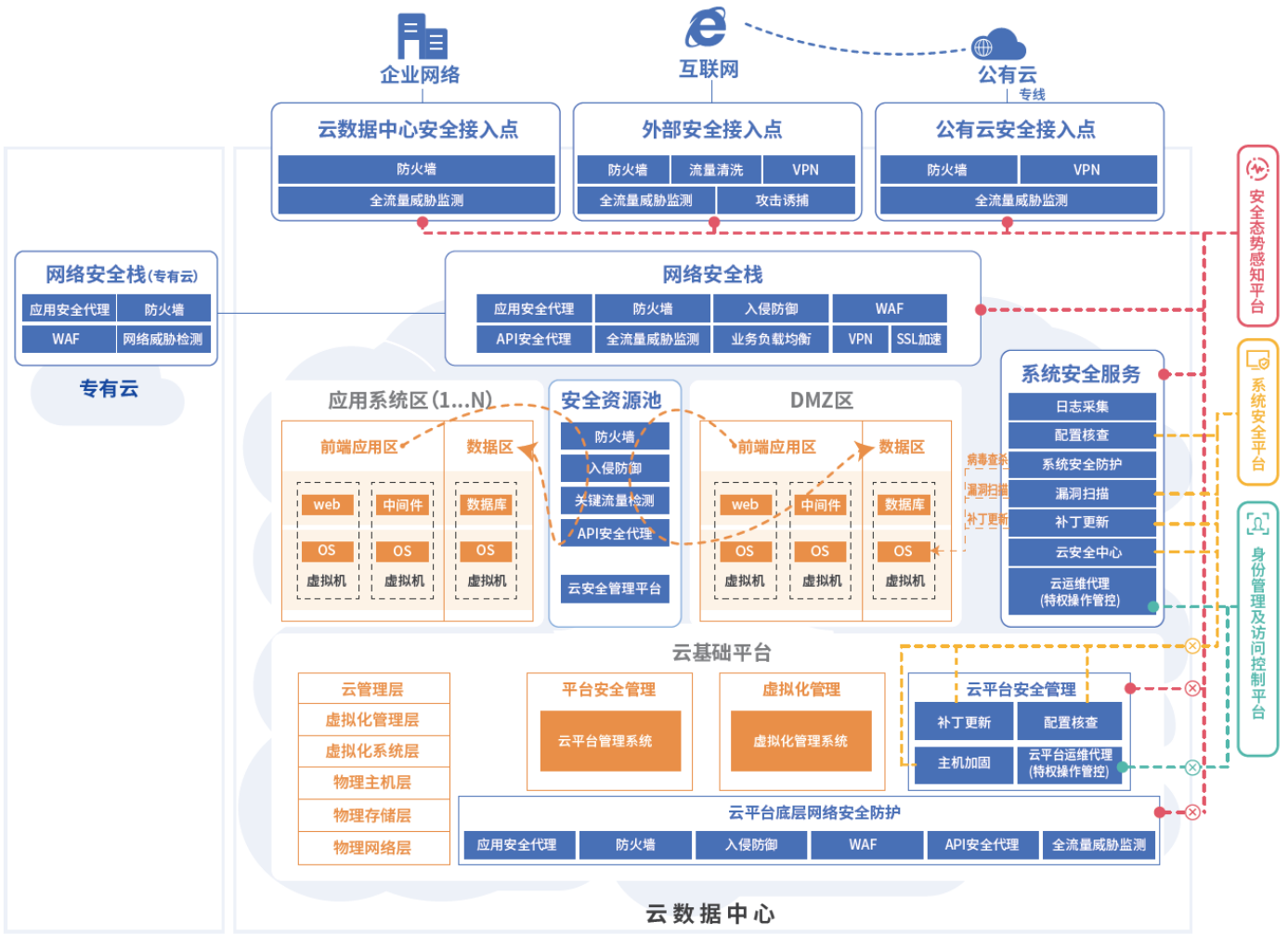




云数据中心

- **特点：高并发通用云计算**

1. **资源按需分配：**资源全面虚拟化，不同粒度按需弹性分配
2. **服务无处不在：**服务通过网络公开，随时随地接入/控制
3. **绿色节能：**降低能耗、提高算效、降低OPEX
4. **集约化建设：**降低TTM和CAPEX



三类计算集群的主要区别

集群类型	应用目的	数据类型	计算特征	网络特征	存储特征
HPC高性能计算中心	国防科研	FP64双精度	重算力，密集计算，高并行	密集通信（每个应用流量特征不同，部分可隐藏）	密集&复杂IO
AI人工智能计算中心	AI训练推理	FP16半精度 BF16/FP8/FP4 FP32全精度		密集通信（部分可隐藏）	密集IO（按节奏迭代）
云数据中心	互联网云计算	INT32整型 FP32全精度	通用计算，高并发	分散通信	分散/密集IO

总结与思考

Summary

- 三类集群计算中心的区别：超算&智算中心为高并行计算，云数据中心为高并发计算
  - 超算中心：国之重器，主要进行FP64双精度的HPC高性能并行计算
  - 智算中心：智算利器，主要进行FP16半精度大模型并行训练和集中推理
  - 云数据中心：互联网云计算基座，主要进行整形和FP32全精度的高并发通用计算

Summary

- 集群计算是主要的算力集成设施建设方式，像水厂、电厂一样
- 算力基础设施建设，从 分层堆叠 走向 集群全栈优化，达成规模系统最优

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIInfra

引用与参考

- https://zhuanlan.zhihu.com/p/683671511

- PPT 开源在：https://github.com/chenzomi12/AIInfra