

数智企业 共创多元算力世界

浪潮信息广东区资深架构师

陈臻淞



数据智能推动数字化转型深入发展

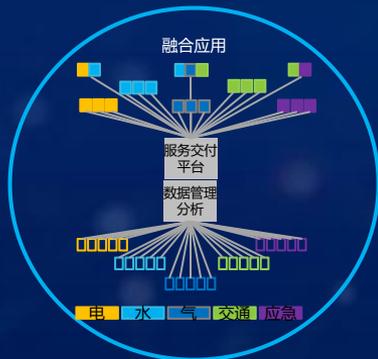
趋势

挑战

路径

数字政府 只跑一次

系统之系统
下一代应用架构与数据共享



融合

数字经济 数字孪生

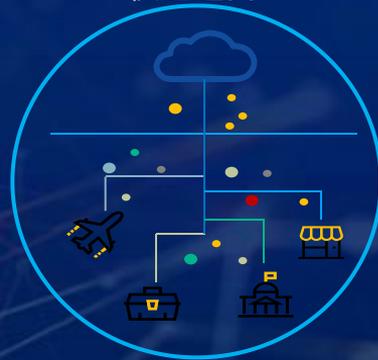
集团管控
工业互联网



智能

数字社会 移动app

互联网+
快速迭代



敏捷



数字科技引领未来

趋势

挑战

路径



语音识别



文字识别



AI质检



自动驾驶



AI for Science



AIGC

模仿、交互

机器学习优化人机协同方式

分析、预测

深度学习提升AI精度和准度

学习、演绎

预训练模型驱动AI自主创新



感知



认知



创造

算力支撑



算力就是生产力 智算力就是创新力

趋势

挑战

路径

模型参数：**2457亿**

| 最大规模 | 中文语言模型 |

训练数据集：**5TB**

| 最大规模 | 中文高质量数据集 |

模型	参数量	数据集	算力当量
源1.0	2457亿	5T	4095PD
GPT-3	1750亿	0.7T	3640PD

大模型

制造

衍生设计
动态模拟

金融

智能客服
欺诈检测

传媒

新闻写作
图文理解

行业模型

古文

对话

技能模型

问答

翻译

2457亿参数
5TB数据集

基础模型

多元算力支持
NVIDIA 燧原 HWJ...



未来算力面临挑战1：性能

趋势

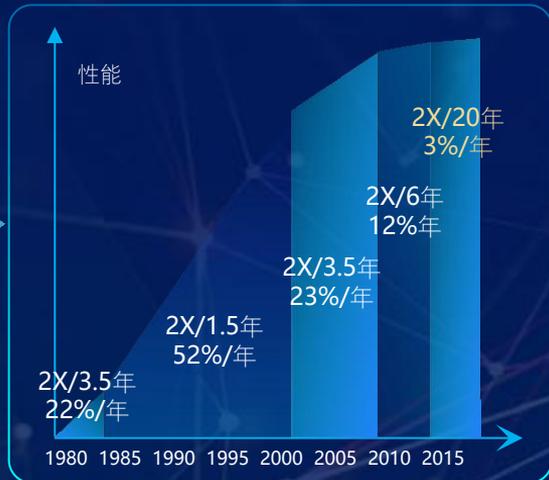
挑战

路径

大模型面临巨量算力挑战

处理器性能提升趋向缓慢

存储墙问题



Operation	Energy (pJ)
计算能耗	
Integer add (32b)	0.1
Integer Multiply (32b)	3.1
Floating Point Add (32b)	0.9
Floating Point Multiply (32b)	3.7
存储访问能耗	
4KB SRAM (64b)	10
1MB SRAM (64b)	100
DRAM	2000

~650倍



未来算力面临挑战2：复杂场景

趋势

挑战

路径





解决之道：系统化设计企业多元算力架构

趋势

挑战

路径



■ 多元算力

- “以应用为导向，以系统设计为中心，建立多元异构算力融合、软硬协同优化”的技术发展路线，实现现代多种算力并存的现代化基础设施

■ 多元算力的内涵

- 多样化：满足多个业务场景
- 专用化：针对特定应用优化
- 异构化：计算单元之间结构设计
- 系统化：统一调度等优化管理
- 绿色化：提升效率，降低能耗，节约成本



多样化：满足现代多业务场景

趋势

挑战

路径

满足关键应用（MES）、通用应用（云）、AI与超算应用（协同设计）、边缘应用（质检）多个业务场景



AI计算

浮点并行计算能力

通用计算

综合算力能力

关键计算

高性能高稳定高安全强一致

边缘计算

性能功耗比，环境适应性



专用化：针对特定领域优化

趋势

挑战

路径

- 形成“CPU+ xPU”的算力格局，应对后摩尔时代性能发展缓慢挑战

X86处理器



- 代表：Intel Xeon、AMD EPYC

- 8086，是英特尔第一款商用16位CPU，被认为是开启x86处理器时代的芯片。后续陆续发布80286/80386/80486等，因此习惯称为X86。主要应用在PC服务器上。

小型机处理器



- 代表：IBM Power、SUN SPARC

- 小型机是指性能和价格介于PC服务器和大型主机之间的一种高性能64位计算机，采用精简指令集处理器。主要配合Unix操作系统环境使用。
- POWER是Performance Optimization With Enhanced RISC的缩写

ARM处理器



- 代表：飞腾KP系列、HW Kunpeng

- ARM处理器是英国Acorn有限公司设计的低功耗成本的第一款RISC微处理器。起初应用在边缘应用场景，目前也开始应用在服务器。
- ARM是Advanced RISC Machine的缩写

DSA-xPU



- 代表：英伟达GPU，谷歌TPU，HWJ等

- 针对特定问题域定制的体系结构，并为该领域提供显著的性能（和能效）增益。
- DSA的例子包括图形处理单元（GPU），用于深度学习的神经网络处理器和用于软件定义网络（SDN）的处理器。



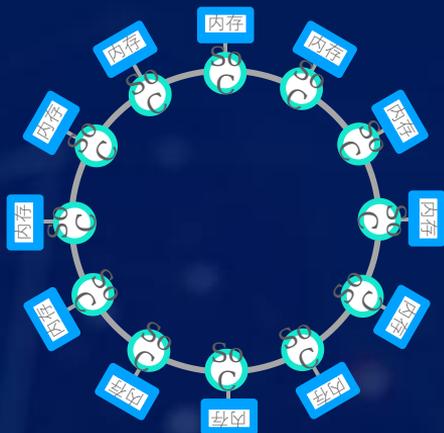
异构化：融合多种算力要素的架构设计

趋势

挑战

路径

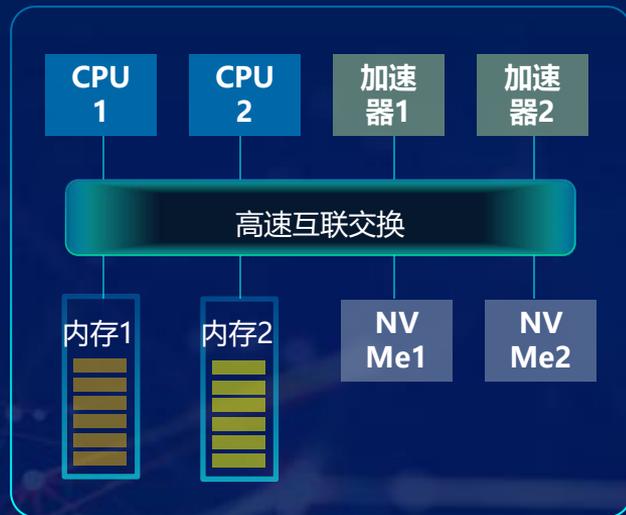
处理器为中心的计算



以数据为中心的计算



面向未来的融合架构



- **问题分析：** AI过程中数据流为硬盘→内存→CPU → 内存→ GPU；CPU 持续读取数据，形成大量并发的IO Stream，以便GPU卡始终保持计算。数据在不同内存区间发生多次数据搬迁；
- **解决思路：** 改变各种处理器（如CPU和AI加速卡）和内存之间的连接架构，构建内存池，多算力共享内存池；



系统化：统一调度管理

趋势

挑战

路径

- 整机系统化设计，企业统一管理调度

智算OS

多元算力

统一纳管GPU和国产AI芯片
算力执行框架消除算力架构间障碍

数据

训练

调优

部署

服务治理

模型开发部署平台

智算中心算力调度平台

多元算力

多种AI芯片接入

并行存储

多文件系统统一管理

高速网络

多IB/RoCE网络适配

灵活调度

万卡超大规模算力毫秒级调度
弹性、独占、切分，资源利用率最大化

模型优化

从数据到训练，再到部署的端到端工具集成
极速构建开发环境，智能适配框架/模型

智能管理

全面监管集群设备状态、运行状况
训练异常和故障的全面检测和自动处理



绿色化：绿色算力设计降低能耗

趋势

挑战

路径

- 统筹考虑PUE基础上，建立对数据中心耗电最大户-算力系统的绿色化设计

智算中心建设面临双碳挑战

2025年数据中心电力消耗

4.05%



远大于超大城市全社会用电总量

2046~2060年

深度脱碳实现碳中和

2031~2045年

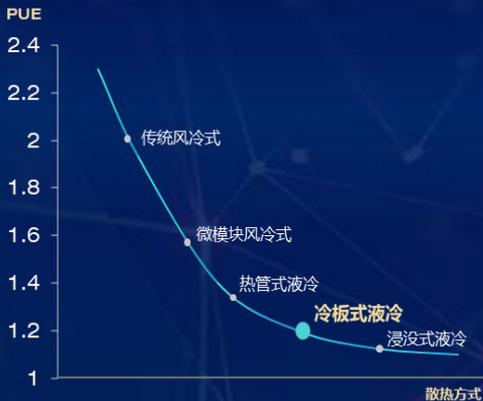
快速降低碳排放

2021~2030年

实现碳排放达峰

北京/上海 新建数据中心PUE < 1.25
上海 PUE > 1.7小老散旧数据中心关停并转

液冷技术支撑绿色降碳



降低数据中心绝对能耗



绿色算力指数

$$\text{绿色算力指数} = \frac{\text{算力输出} / \text{总能耗}}{\text{电力碳排放因子} \times \text{算力输出时长}}$$

建立绿色算力评估指标/指南，引导现代算力的高效部署，降低社会算力总碳排放量



共创现代数智企业的多元算力世界

算力中心

主机：服务器

处理器

内存

存储系统

输入输出IO

CPU

GPU

FPGA

ASIC

DPU

网卡

视频卡

加速卡

数据中心

核心交换机

无线AP

集中存储

分布式存储

边缘侧

边缘服务器

边缘云

终端侧

工控机

传感器

智能终端

摄像头

软件定义存储

软件定义网络

CPU虚拟化

GPU虚拟化

...

本地云

边缘云



智能质检解决方案

应用层

UI交互界面

权限管理

设备管理

数据统计

业务流程

产线管理

数字孪生

AI推理服务

视觉检测系统

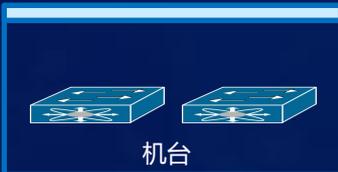
算法镜像仓库

AI算力设备

4G/5G网关 布线简单

以太网关 高稳定性

数据汇聚层



机台



图像与存储管理平台



数据存储

分布式存储

磁带库

汇集

图片压缩

图片存储

整合

管理

协议/以太网

物理接口

感知层



面阵、线阵相机

图像

视频



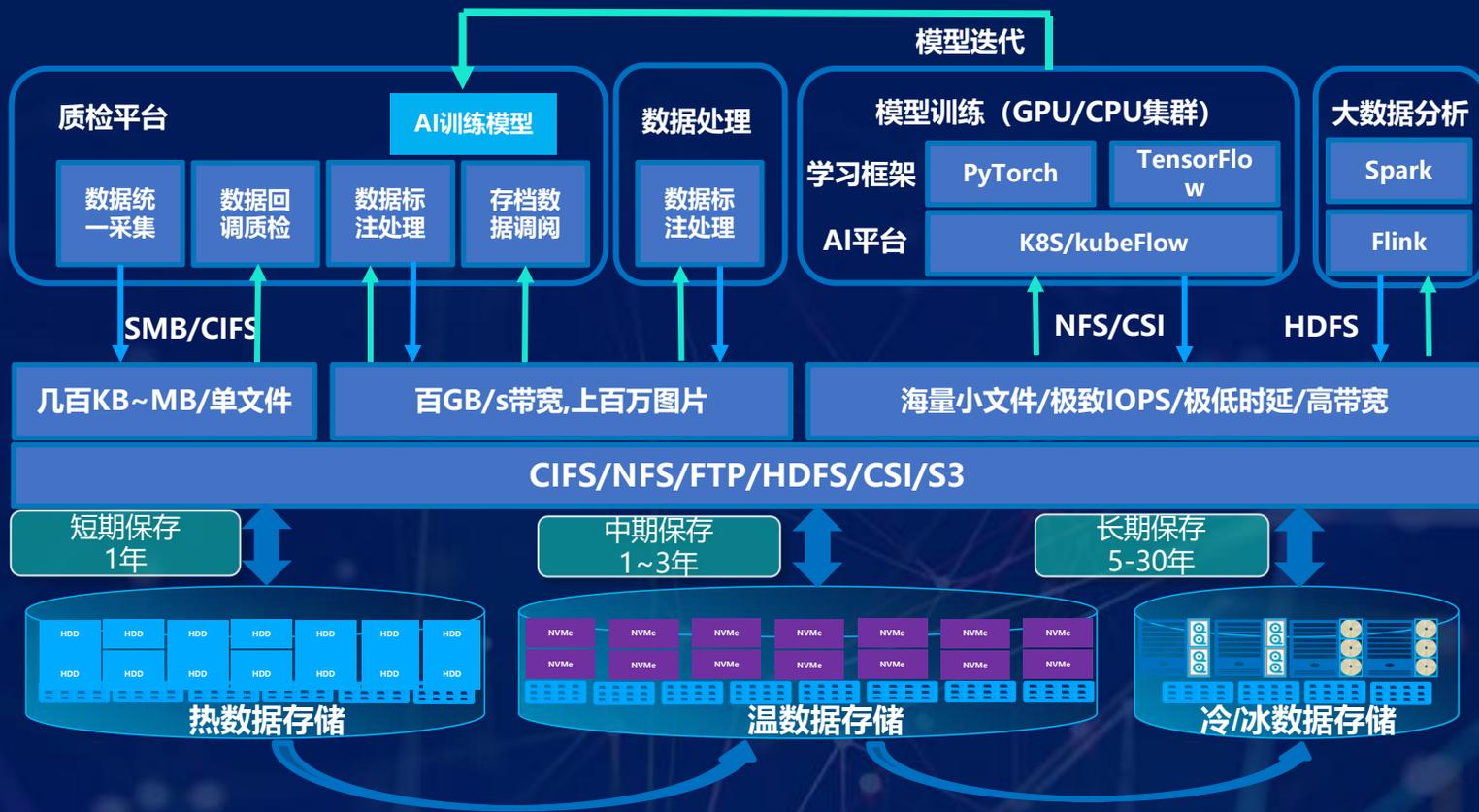
光电、质量传感器

速度、节拍

位置、角度



数据全生命周期管理





机台数据热温冷分级存储





VMware的整套解决方案

浪潮信息是VMware全球七家OEM联盟合作伙伴之一，中国首家合作伙伴，最高合作级别。

The screenshot shows the VMware Partner Locator interface. At the top, there is a navigation bar with the VMware logo and menu items: 应用和云, 网络连接, 工作空间, 安全性, 按行业, 合作伙伴, 资源. Below the navigation bar, there is a breadcrumb trail: 主页 > 合作伙伴 > 合作伙伴查找工具. On the right side, there is a language dropdown menu set to 简体中文. The main content area displays the details for Inspur Electronic Information Industry Co., Ltd. The address and company name are highlighted with a red box: 2 Floor 2, Building C, International Venture Park, No. 2 Xinxl Road, Beijing BEIJING 100085 中国. To the right of the address, it lists the partnership levels: 全球 OEM 联盟 - 核心级 and 技术联盟合作伙伴 - 高级. On the far right, there is contact information: 010-62988886, 发送电子邮件给合作伙伴, 访问合作伙伴网站, and 搜索 VMware Marketplace.

<https://partnerlocator.vmware.com/PartnerView?id=0018000000UG5o3AAD&companyname=INSPUR&usrctr=CHINA&lang=cn>

inMerge1100是基于VMware vSAN架构的自有品牌超融合系统。



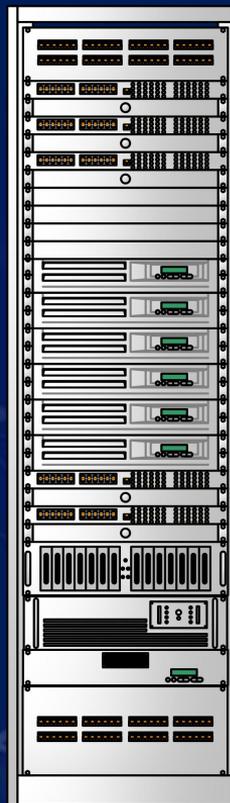
整柜解决方案

SRDC (Smart Rack Data Center)

集成计算、存储、网络、配电、备电、制冷和智能管控于一体的方案产品，点餐式下单，在工厂进行设备集成，可预装客户系统软件，一体化整机交付，具有交付即业务快速上线

关键词

一体化集成 整机柜交付 业务快速上线 无需专用机房



供电模块：
配电单元
PDU

网络模块

安全模块

计算模块

存储模块

备电模块：
UPS
电源包

散热模块

智能触控屏

智能管控平台：
动环监控系统
资源管理系统
视频监控系统

温感烟感器

温湿度传感器

监控摄像头

机柜智能锁



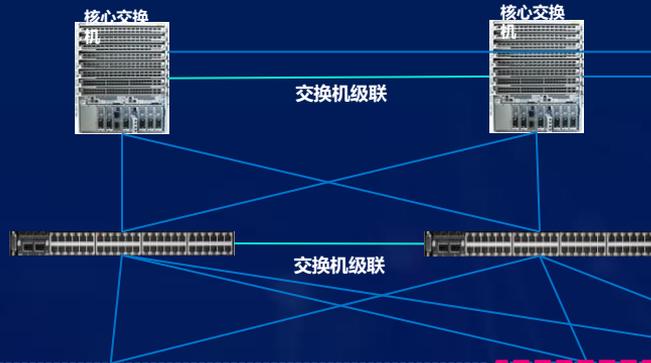


SAP HANA解决方案

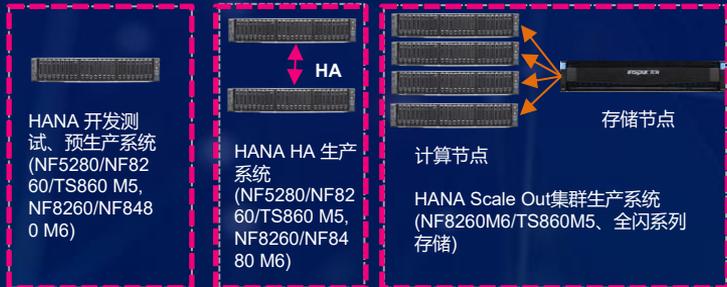
主数据中心

网络设备部分:

- 核心交换机需求;
- 接入层网络设备需求, 包括网络交换机、存储交换机等



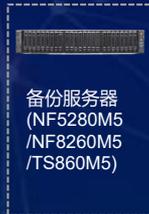
HANA数据库部分



SAP HANA数据库部分:

- 采用经过SAP认证的硬件: 服务器、存储;
- 配置高可用或者集群;
- 测试开发系统服务器需求

备份部分



备份部分:

- HANA数据库服务器、存储

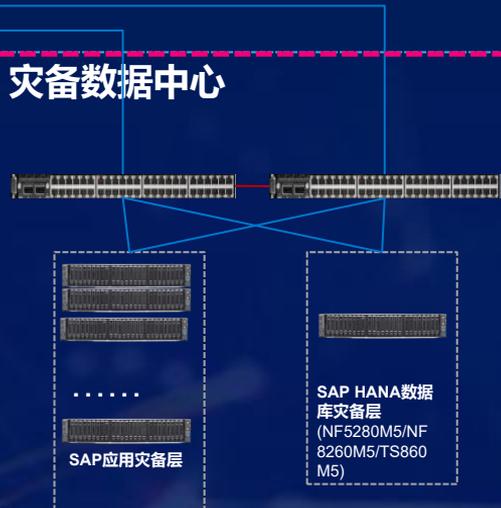
SAP 应用部分



SAP应用层:

- 超融合或虚拟化部署
- 常规部署模式

灾备数据中心



灾备数据库中心部分:

- 应用灾备服务器需求
- 同等配置的数据库灾备服务器需求
- 网络设备需求

开启数字经济新局面

创造算力时代新机遇