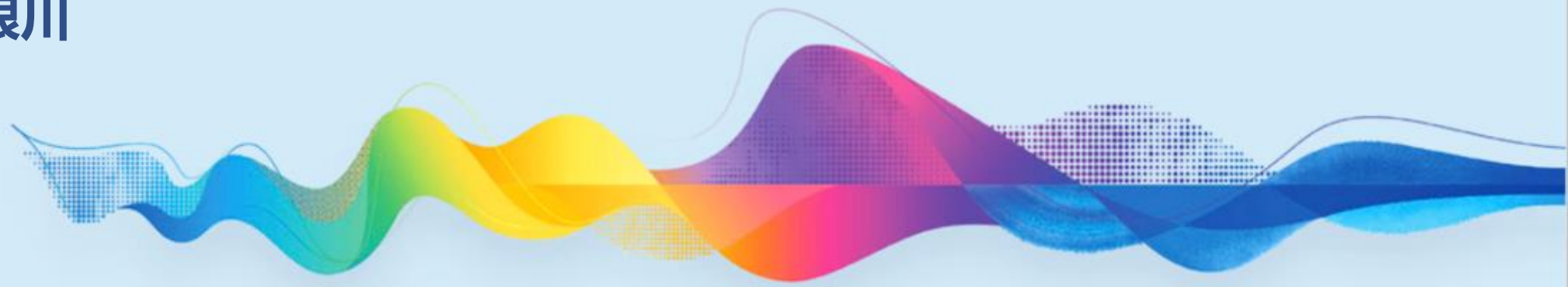
科研创新 算力加速

联想集团 唐珂

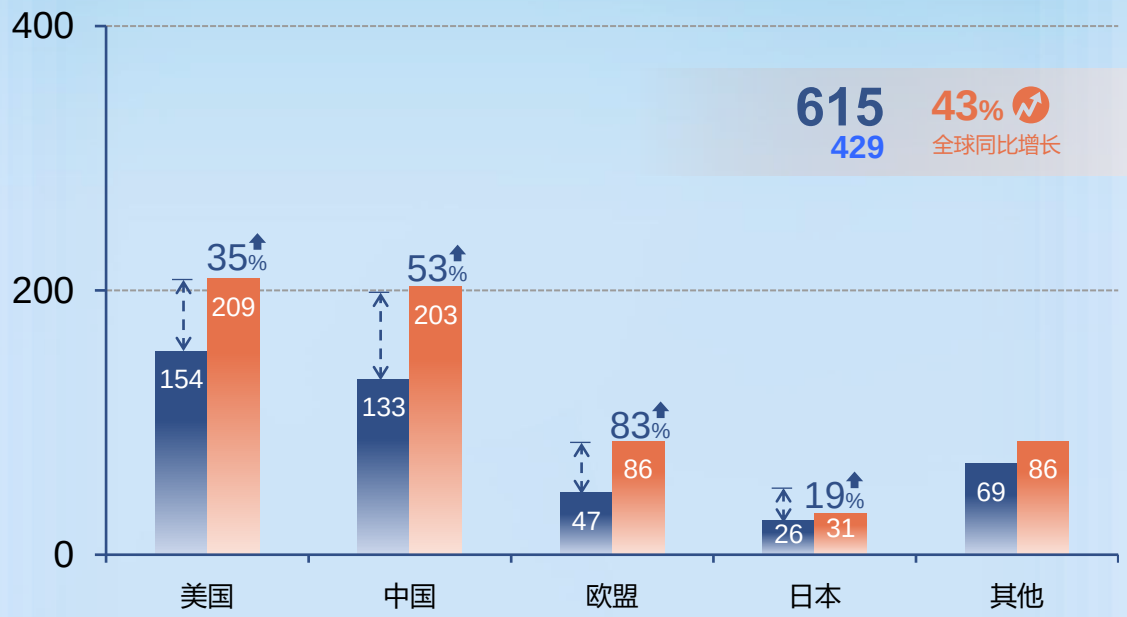
2024.06 宁夏|银川



目录

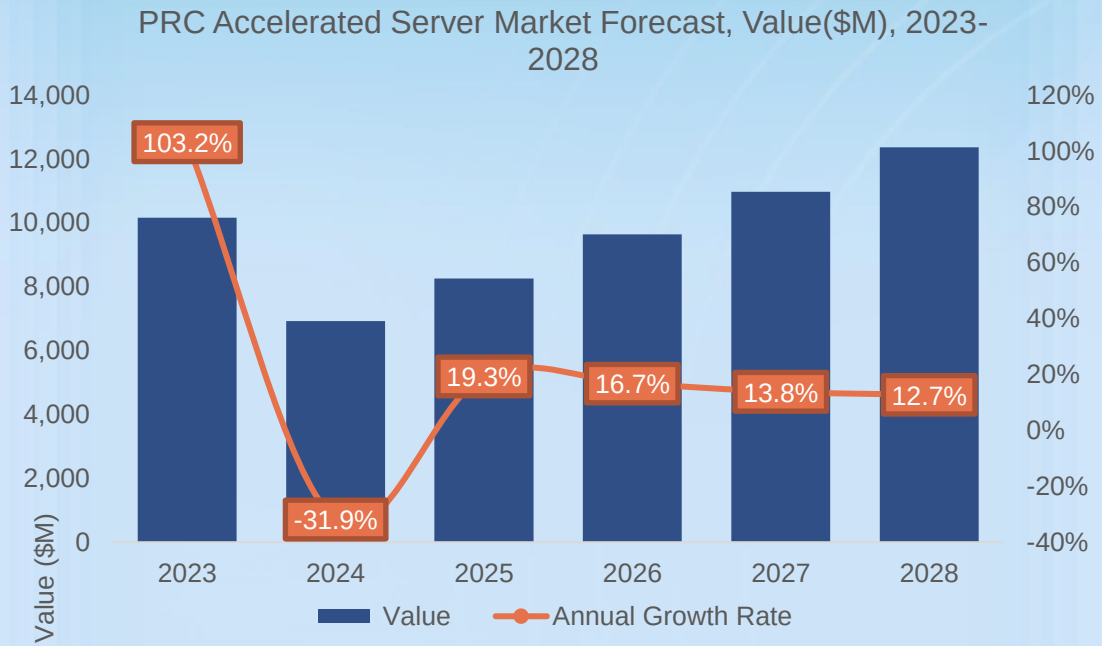
AI算力是算力发展的主要方向

2020-2021全球算力规模分布情况（EFLOPS）



Data Source：毕马威分析

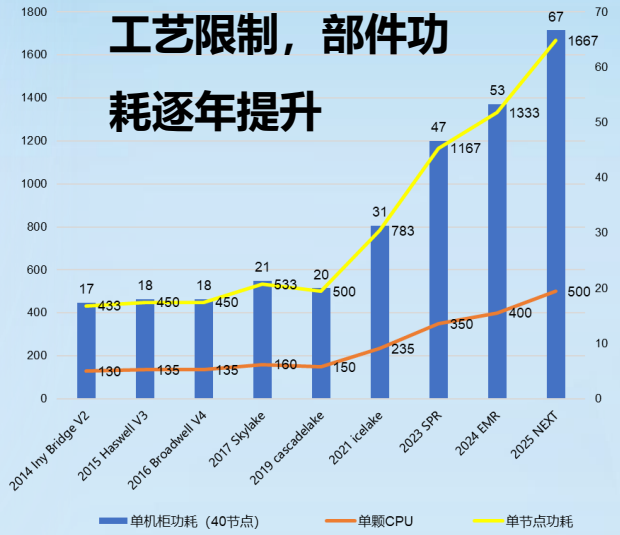
中国AI加速服务器规模及预测 2023-2028



Data Source：IDC

处理信息需要能量成本，在快速提升

部件功耗



焦耳定律

$$Q = I^2Rt$$

电流通过晶体管产生的热量

电流、电阻、通电时间成正比

应用功耗

OpenAI

每天响应约2亿个需求

≈50万度电力

≈1.7万美国家庭1天用电

2024

全美AI数据中心的耗电量

占全美总用电量2.5%

足以点亮整个纽约市

GPT-3：消耗1300MWh

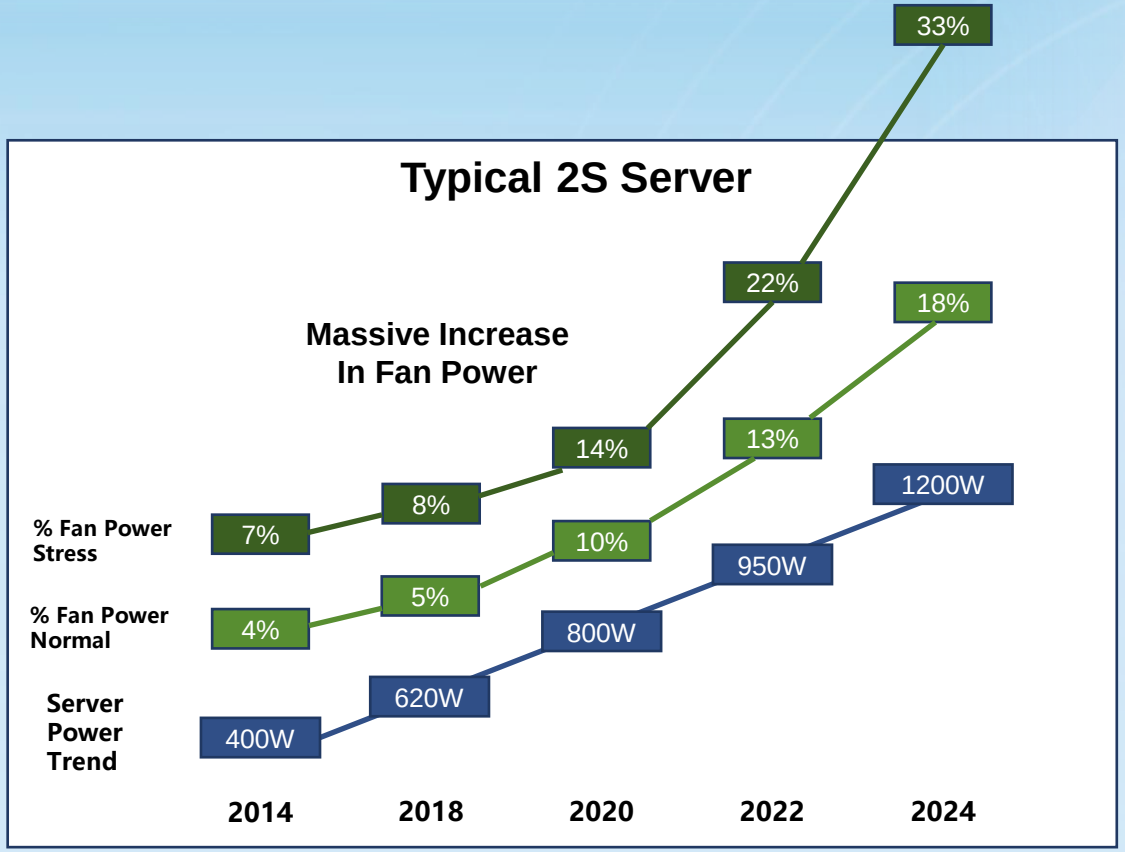
- 相当于162.5W小时=185.5年流媒体播放
- 一张文生图消耗的电量能把一部手机充满
- GPT-6的训练曾让微软的电力崩溃

- 1961年，物理学家兰道尔提出兰道尔定律，指出计算机中存储的信息发生不可逆变化的时候，系统的熵会增加并且伴随着能量的耗散
- 在AI训练阶段，LLM需要收集和预处理大量的文本数据，随着模型的迭代，处理的参数会指数级增长
- scaling law 规模法则依然是万能解药等情况下，耗电量将指数型增长，计算的尽头是能源

传统风冷遇到瓶颈，能源利用率降低

随着功耗增加，非计算部件特别是**散热部件的功耗占比越来越高**，空气作为解热介质已经面临瓶颈

Major Device	Item Power (W)	Qty/System	Device Power(W)
主板			
Processor IceLake	350	2	700
DDR4 Memory	5.4	16	86.4
BMC	5	1	5
Other Motherboard Logic/IC	5	1	5
制冷			
System Fans(8086)	150	4	600
硬盘			
U.2	25	1	25
M.2	5	0	0
扩展			
PCI Express Slots(X16)	40	1	40
OCP 3.0	20	1	20
输入输出			
USB 3.0	4.5	2	9
VGA	2.5	1	2.5
总体			
Total Estimated System Power			1508.23
Total			
Total Estimated System Power*0.8 (For PSU sizing,with 100% Fan Duty)			1326.584



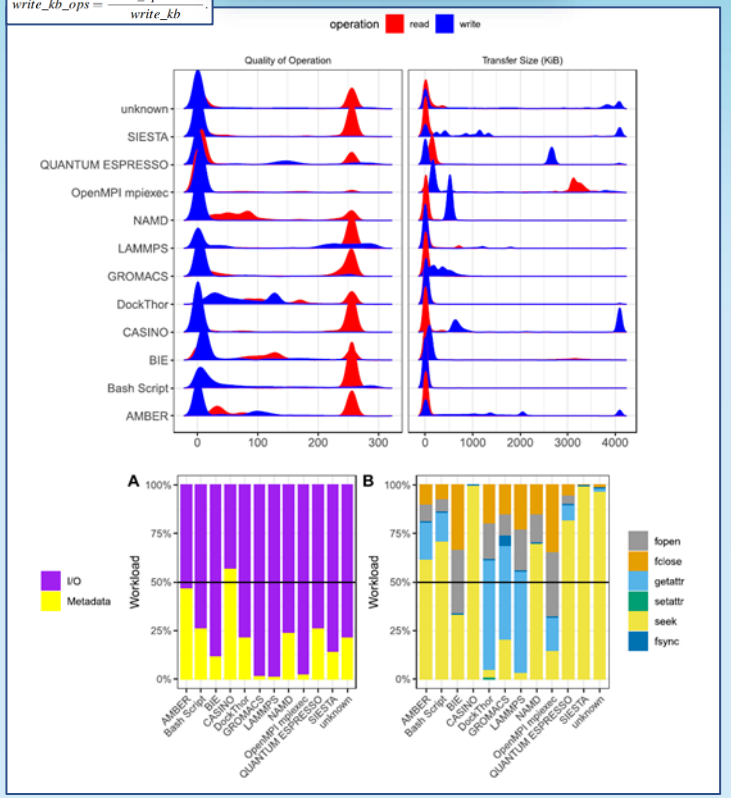
算力增长的同时，存力需求接踵而来

HPC IO

- 大多数应用readQO > 1，表示RPC粒度更小，read更琐碎，影响性能
- 75% 的传输请求 < 100 KiB，SIESTA、QUANTUM ESPRESSO、CASINO 和 AMBER，至少有 50% 的时间内传输大小 > 1 MiB；LAMMPS、DockThor 和 Bash 脚本，它们的操作有 75% 的时间内 < 40 KiB
- 通常的元数据操作负载在10%到25%之间。
- 寻址操作在大多数应用程序中占主导地位，表明存在大量的随机文件访问。

$$read\_kb\_ops = \frac{read\_ops * 1024}{read\_kb}$$

$$write\_kb\_ops = \frac{write\_ops * 1024}{write\_kb}$$



Source: AndréRamosCarneiro, Uncovering I/O demands on HPC platforms: Peeking under the hood of Santos Dumont, 18 August 2023

ML IO

- 80%的ML工作负载的读取和写入调用都小于1MB
- 大量的小的读取和写入请求会过载并行文件系统
- 大的顺序读取和写入可以从GPFS获得更高的性能，而小的读取和写入在BB上更高效

Sc. Domains	Num				
	<1M	1M-10M	10M-100M	100M-1G	>1G
Biology	2.92e+7	922.12	678.61	70.07	2.48
Chemistry	8.63e+5	1135.22	21.2	0	0.02
Comp. Sc.	5.14e+6	421151.22	69558.12	1.45	4.91
Earth Sc.	5.57e+5	24435.34	382.81	0	0
Engineering	4.67e+5	12.99	104971.99	0.74	0.24
Fusion	3.05e+7	80.81	87.57	83.66	0
Mach. Learn.	3.90e+5	28126.52	6484.93	0.59	0
Materials	5.33e+6	7037.46	103.03	0.29	0.16
Physics	1.5e+7	1004.92	6644.35	25.76	31.34

(a) Mean number of read calls.

Sc. Domains	Number of write calls				
	<1M	1M-10M	10M-100M	100M-1G	>1G
Biology	5.41e+7	79.57	11.62	0.29	0.03
Chemistry	1.77e+7	117.08	305.52	0	0
Comp. Sc.	1.98e+6	406.57	5883.53	0.26	0.01
Earth Sc.	6.93e+4	48.97	7.49	0.10	0
Engineering	5.49e+5	1192.63	241313.12	0	0
Fusion	2.05e+3	325.89	959.40	239.85	0.17
Mach. Learn.	1.62e+3	89.91	1.48	0	0
Materials	4.49e+6	2.34	17.58	0.26	0.23
Physics	3.59e+6	959.99	44.69	1.50	0

(b) Mean number of write calls.

问题：

1. NVMe的成本仍是HDD介质的10倍+
2. IO性能需考虑：IO整体时间占比是否值得被优化
3. 存储容量不可忽略，性能和容量选择平衡点
4. 不同应用IO表现不同，对共享存储会有影响
5. 解决随机小IO，分级策略是个好的选择，用户的使用习惯是否养成，临时文件如何处理

Source: Ahmad Maroof Karimi, Arnab K. Paul, and Feiyi Wang, I/O performance analysis of machine learning workloads on leadership scale supercomputer, October 2022

数据中心面临重大挑战

≈1.59

2020年全球大型数据中心平均PUE

数据来源：数据中心标准组织Uptime Institute报告，2023

PUE

Power Usage Effectiveness

电力使用效率

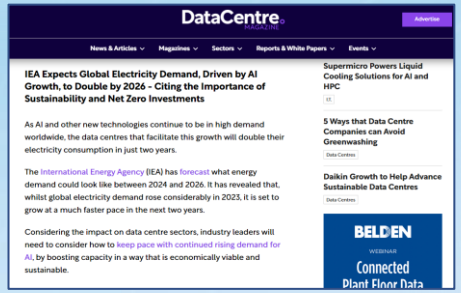
$$\frac{\text{IT设备能耗 + 配套设备能耗}}{\text{IT设备能耗}}$$

越接近1，浪费的能源越少

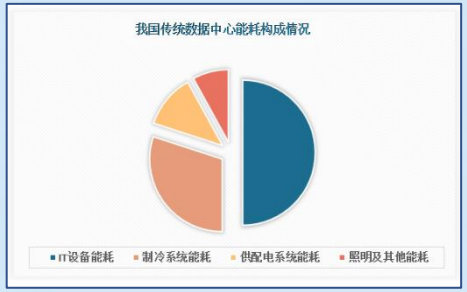
460TWh

人工智能和加密货币的耗电量

占全球能耗2%



数据来源：国际能源署（IEA）



40%

现有冷却技术下冷却系统能耗占比总能耗过高

数据来源：中国数据中心行业现状深度研究与未来前景分析报告（2022-2029年）

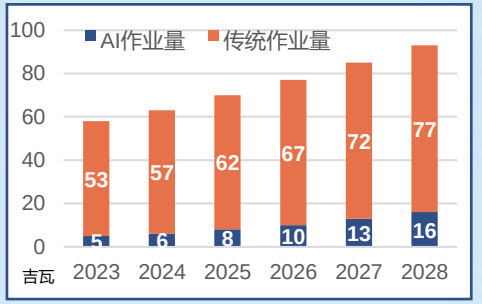
工信部等七部门：2025年新建大型、超大型数据中心

PUE降到1.3以下

25%~33%

年复合增长率

全球数据中心能耗增长快速飙升



数据来源：Bank of America 提交机构报告

目录

联想集团 开启第五个10年

1984-1993

创业时期

- 11位创业者
- 20万元启动资金
- 代理分销业务起步
- 创新起步：联想汉卡

1994-2003

PC品牌时期

- 发布联想自有品牌
- 1997年位列中国第一
- 1999年位列亚太地区第一

2004-2013

全球化时期

- 2004年收购IBM的个人电脑业务
- PC销量5500万台，2013年位列全球PC第一
- 渡过全球金融危机

2014-2023

3S转型时期

- 2014年收购摩托罗拉移动及IBM x86服务器
- 实施3S战略
- 向服务和解决方案提供商转型

2024至今

AI 变革时期

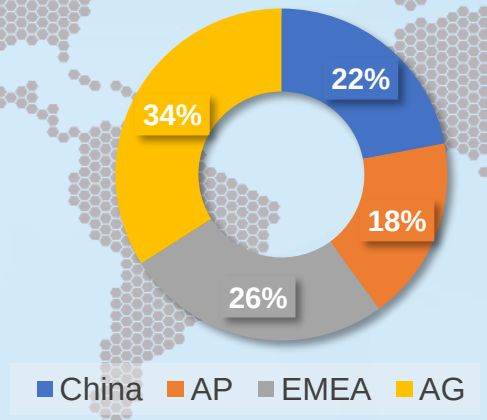
- 混合式人工智能
- 全栈人工智能
- 让世界充满AI

联想集团：植根于中国的全球化企业

2023/2024财年收入

4074

亿人民币



全球各大区营收占比
2023/2024财年

全球研发基地	18
全球工厂	约30个
研发投入	200亿
工程师、研究人员和科学家	18000+
专利及专利申请	33,000+

连续13年入选
位列《财富》
世界500强
第217名

福布斯 Forbes (2023年度)

全栈产品及解决方案，助力高校数字化建设

Lenovo 联想

PC份额全球第一
(IDC 2023)

服务器份额全球第三
(IDC 2023)

HPC全球第一
(TOP500 2023)

IT服务份额中国第二
(IDC2023)

智慧办公 Smart Device


笔记本


会议套装


台式机


工作站


显示器


打印机


选件

智能设备 Smart Device

商用平板

商用智慧屏

智能边缘


工控机


边缘计算网关


云终端

智慧教育设备


智慧黑板


智慧互动大屏


备授课5.0

基础设施 Smart Infrastructure

服务器

存储

软件定义

Microsoft
Red Hat
NUTANIX

方案服务 Smart Service

T1 支持服务

- 尊享服务
- 零碳服务
- 客制化服务
- 整体维保
- 性能升级
- 技术培训

T2 管理服务

- 运维服务
- DaaS（TruScale）

T3 方案服务

- 智慧城市
- 智慧园区
- 智慧零售
- 智慧教育
- 智慧制造
- 混合云
- .....

联想中国 基础设施业务群
收获客户和市场认可

2024第一季度
服务器市场**排名前三**

2023第四季度
服务器市场**增速第一**

面向智能计算、科学计算、通用计算设计

Lenovo 联想

应用

行业方案：AI模型开发　智能制造　智慧医疗　智慧金融　…

应用生态企业：OpenCSG　4Paradigm　诺谛智能　Lenovo 联想新视界　安想 中国平安 Lenovo 联想　科大讯飞 iFLYTEK　商汤 sensetime　阿里云

开发平台

Pytorch/TensorFlow/DeepSpeed等
基于不同GPU算子库实现的算法框架和并行加速库

管理平台

联想万全异构算力管理平台
统一管理、调度和优化智能和科学算力

联想云
统一管理CPU为主的通用算力

硬件平台

智能计算	科学计算	通用计算

服务器：WA7780 G3　WA5480 G3　WR5220 G3

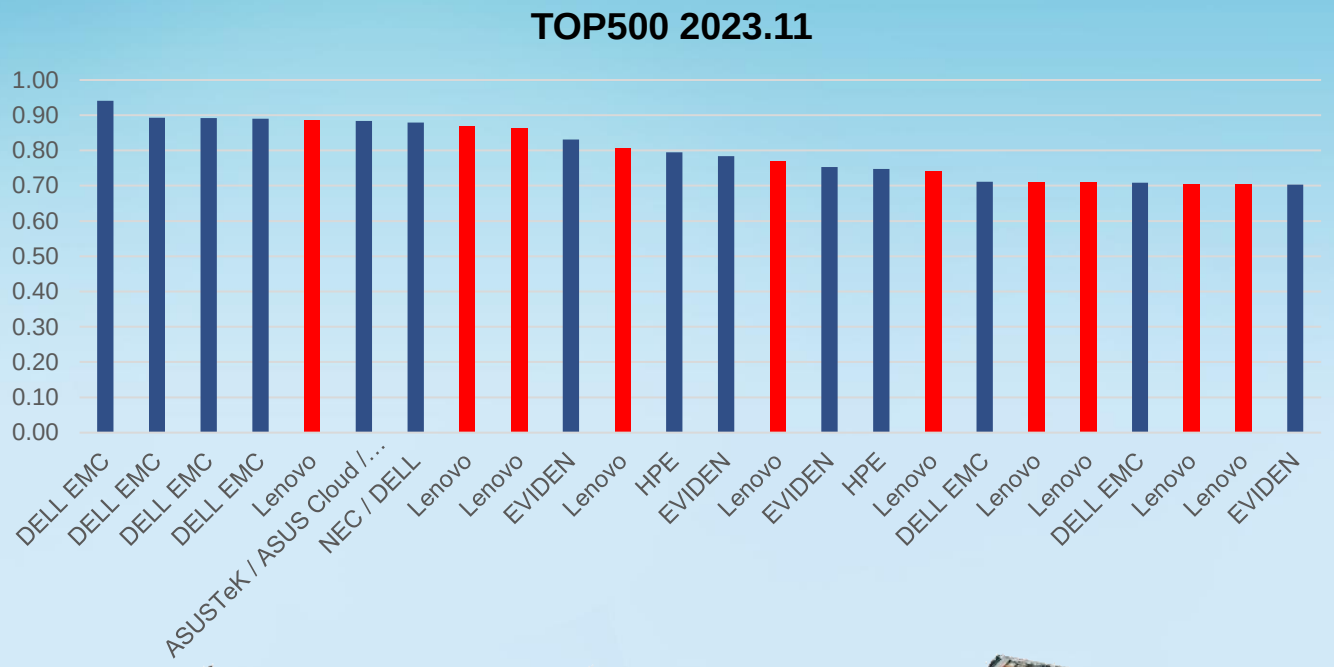
存储：DE-F SAN存储 DM-F 统一存储　EF300 /EF600

交换机：400G　200G

边缘：SE350 V2　SE360 V2

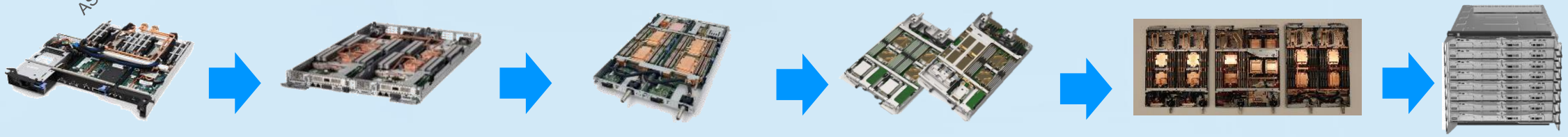
AI 硬件设备 GPU 芯片：intel　天数智芯　Enflame 燧原科技　昆仑芯　Cambricon 寒武纪　摩尔线程 MOORE THREADS　METAX 沐曦　壁仞科技 BIREN TECHNOLOGY

联想液冷服务器历经十六年发展



已部署
70,000
液冷节点

TOP500
42%
液冷装置



2008
第一代液冷高密
INTEL E5-2600 V1/V2

2015
第二代液冷高密
INTEL E5-2600 V3

2017
第三代液冷高密
INTEL 第一/第二代可扩展

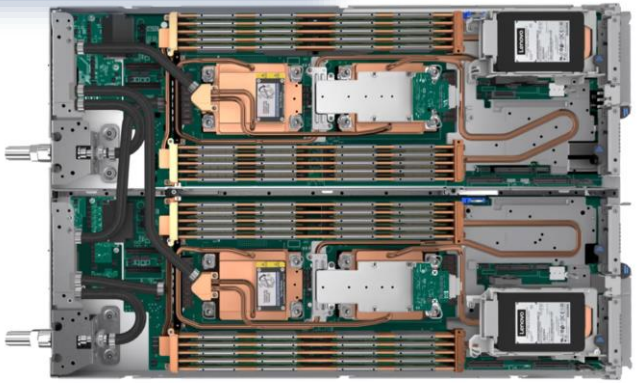
2020
第四代液冷高密
INTEL 第三代可扩展

2022
第五代液冷高密
INTEL第四/五代

2025
第六代液冷高密
INTEL第六代

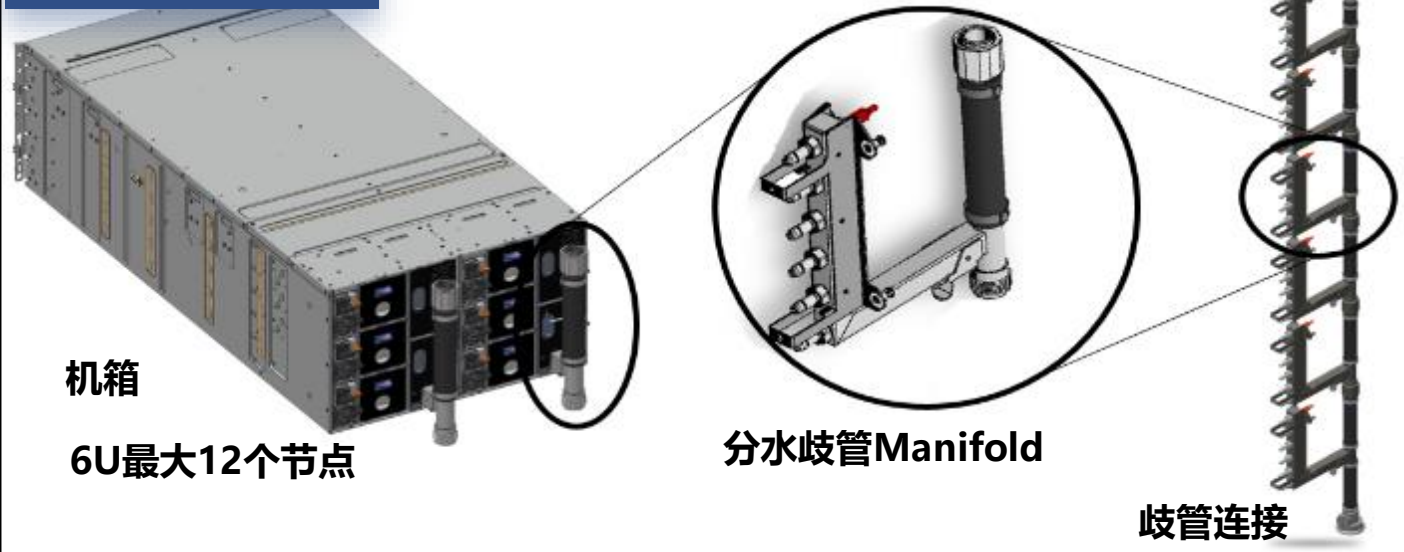
极致的产品工艺设计

节点设计



- **全铜冷板设计**，提升散热效率和可靠性
- **98%液冷覆盖**，PUE≤1.1
- **前后分离**，快速维护
- **CPU并行水路**，保证相同工作温度
- **无风扇设计**，超强静音

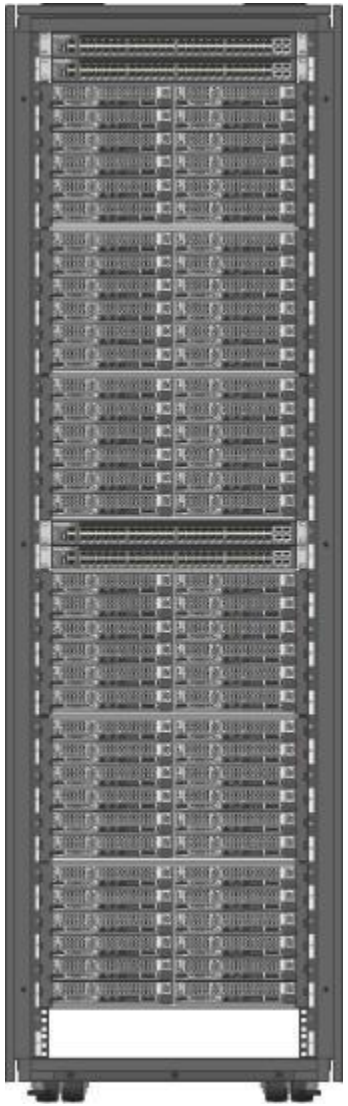
机箱设计



机柜设计

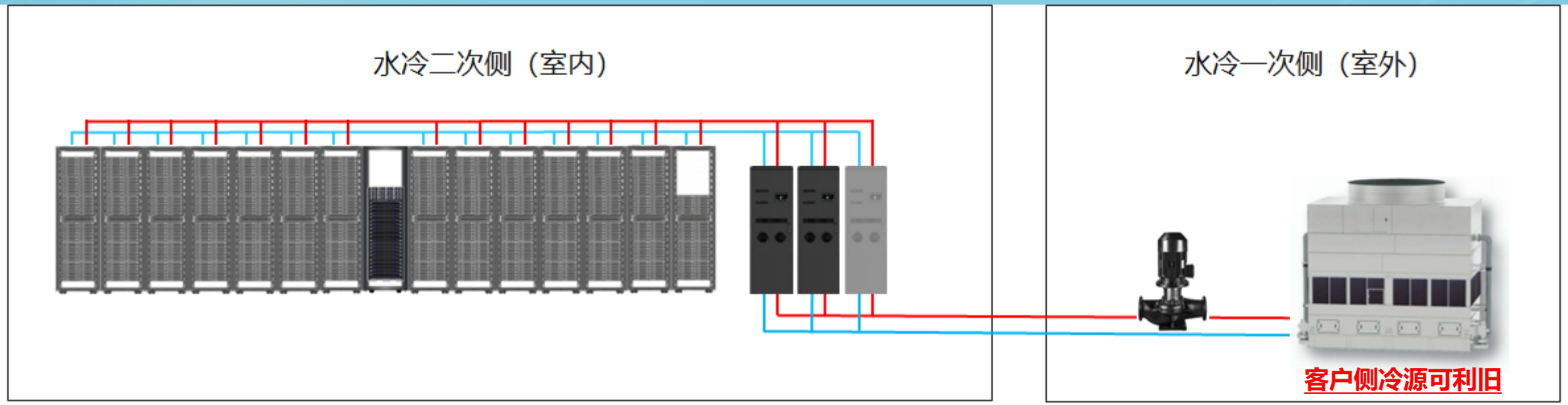
42U标准尺寸机柜

整机柜，侧视图



整机柜，正视图

液冷配套部分



水冷服务器

采用水冷方式的高性能服务器，提高冷却效率，增加单机负载，降低芯片高负载下的运行温度

二次侧回路

连接CDU和水冷服务器，提供安全可靠和清洁的通路条件，环路或主分管方式

CDU 冷量分配控制器

对二次侧回路冷却液的温度、流量和压力进行控制，保证冷却液水质

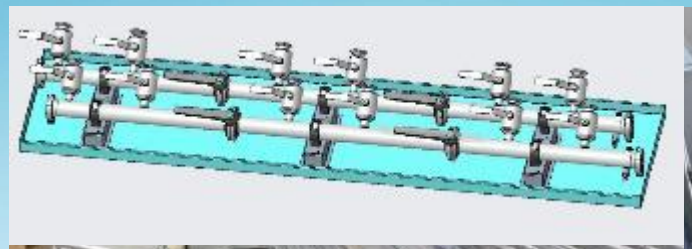
一次侧回路

连接CDU和一次侧冷源，提供一次侧冷却液的通路和水力

一次侧冷源

闭式冷却塔或干冷器，提供水冷系统的冷源，将水冷服务器的热量排出机房

二次侧回路



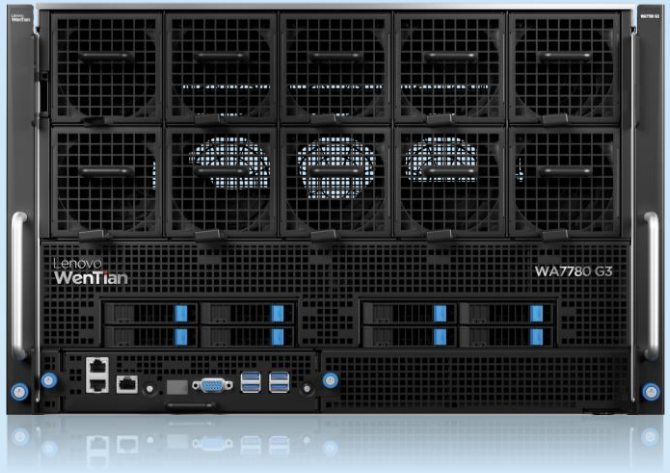
- 二次侧回路采用304不锈钢管路
  - 洁净，可靠
- 工厂预制、测试，现场拼接
  - 精确，现场快速部署，不动火
- 环路方式
  - 提高可维护性，单点故障不停机
- 主分管方式
  - 二次侧管路简单，可靠性高
- 软管连接方式
  - 小规模部署
- 漏水检测绳，接入CDU实时告警

AI产品方案

WA7780 G3
大模型训练AI服务器

1* HGX™ H20 8-GPU

支持训练场景



WA5480 G3
训推一体AI服务器

最多支持10张双宽GPU

兼容国内国际多GPU生态

支持训练，推理等多种场景

基于Intel Xeon 6联想问天服务器

联想问天 WR5220 G5

通算基石

联想问天 WA5480 G5

训推一体

高性能：输出超强算力

225% 单处理器核数增加

2X AI工作负载性能提升

高可靠：减少停机时间

BMC内嵌智能引擎

无中断固件升级方案

高扩展：提供极致IO能力

14% 内存带宽提升、10% PCIe带宽提升

33% NVMe SSD 数量提升

低能耗：降低PUE和TCO

CPU,内存,GPU 高性能风冷+液冷组合

“羊角” EVAC散热器

万全异构智算管理平台

Lenovo 联想

AI数据采集和处理、AI模型库、AI模型服务等

联想异构智算管理平台（HIMP）

异构资源调度

- 高效的调度算法
- 通信优化
- 编译优化

高效

异构算力管理

- 异构算力统一监控管理
- 容错和断点续训
- 内核级虚拟化

稳定

GPU基础软件栈

- GPU设备插件接口
- AI算子库和框架
- 开发工具链与运行时

多元

联想混合云

基础设施集群 (AI、科学、通用计算)

万全异构智算管理平台-创新点

算力匹配魔方

针对场景，全自动规划和调度最佳算法和集群配置

- 数百种集群配置
- 数十种算法

魔方中的交叉点，代表一种场景和与之最匹配的算法和集群配置

GPU内核态虚拟化

从用户态到内核态，虚拟化GPU算力利用率从80%提升到95%

- 在GPU驱动层创建核心虚拟化算法
- 对GPU内核做全局高效的管控和调度

高效断点续训技术

提升训练效率
节省额外支出百万元/月

- 以AI预测AI训练故障，在断点前针对故障特征做优化备份，在断点后极速恢复

AI&HPC融合调度

1小时内自动完成跨集群资源调度和共享

- 可切换“语言”分别指挥AI和HPC调度器
- 可跨集群全局动态管理和调度算力资源

联想全球TOP HPC算力案例

Lenovo 联想

西班牙巴塞罗那超算中心
排名**19/121**位，理论峰值**46.3/10PF**

西班牙巴塞罗那超算中心，TOP500第**19/121**位

德国莱布尼兹计算中心
排名**40/52位**，理论峰值**27/28PF**

韩国国家气象中心，TOP500第**47/48**位

韩国国家气象局
排名**47/48**位，理论峰值**25.49PF**

加拿大公共服务，TOP500第**102/103**位

德国马克斯-普朗克研究所
排名**89**位，理论峰值**16.03PF**

澳大利亚国家超算中心，TOP500第**136**位

德国卡尔斯鲁厄理工学院
排名**97**位，理论峰值**15PF**

奥地利VSC研究中心，TOP500第**162**位

联想中国高校科研算力案例

Lenovo 联想

北京大学 PEKING UNIVERSITY	2018年 “未名一号”
上海交通大学 SHANGHAI JIAO TONG UNIVERSITY	2022年 “思源一号”
华南理工大学 South China University of Technology	2023年 华南理工大学计算中心
浙江大学 ZHEJIANG UNIVERSITY 超重力离心模拟与实验装置国家重大科技基础设施 Centrifugal Hypergravity and Interdisciplinary Experiment Facility（CHIEF）	2024年 超重力实验室
北京市气象局	2021年 冬奥会气象服务器任务

- 中科院网络中心
- 中科院数学所
- 中科院过程所
- 中科院大气物理所
- 中科院地球物理所
- 中科院高能所
- 中科院遥感所
- 中科院空间中心
- 中科院国家天文台
- 中科院上海生命科学院
- 中科院大连化物所
- 国家海洋局
- ……

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

中国科学院数学与系统科学研究院 Academy of Mathematics and Systems Science Chinese Academy of Sciences

中国科学院过程工程研究所 Institute of Process Engineering, Chinese Academy of Sciences

中国科学院高能物理研究所 Institute of High Energy Physics Chinese Academy of Sciences

- 北京大学
- 清华大学
- 复旦大学
- 南京大学
- 南方科技大学
- 厦门大学
- 大连理工大学
- 中国石油大学

……

清华大学 Tsinghua University

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

南京大学 NANJING UNIVERSITY

西湖大学 WESTLAKE UNIVERSITY

全栈AI 助力智能化
转型每一步