

1

超聚变FusionPoD for AI液冷解决方案

2

G8600E V7解耦方案

3

智算中心节点配置及组网方案

4

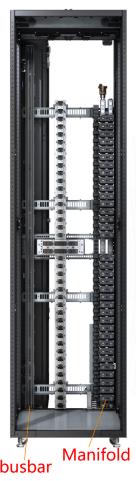
超聚变液冷改造及服务方案

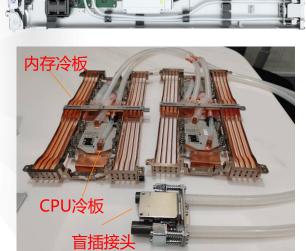


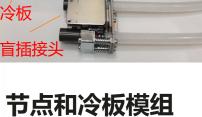
FusionPoD是液冷原生的整机柜液冷服务器,领先产业一代



FusionPoD前视图









液冷门 最高28℃供水

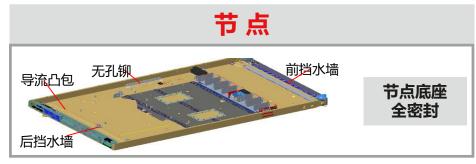
FusionPoD后视图

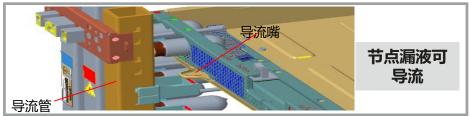
三总线盲插

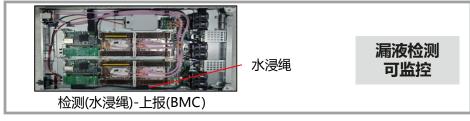
在超聚变的引领下,国内友商已经开始跟随三总线架构,随后会推出类似的盲插整机柜产品。

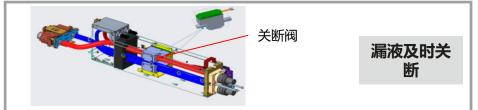


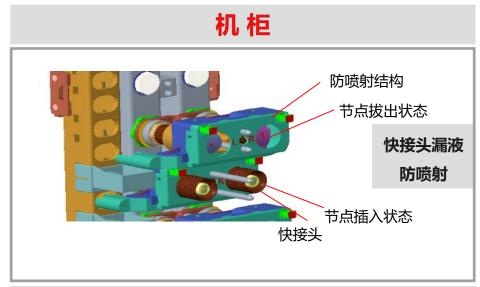
液冷原生: 节点防漏液、机柜防喷射等液冷原生设计, 提供液冷可靠性

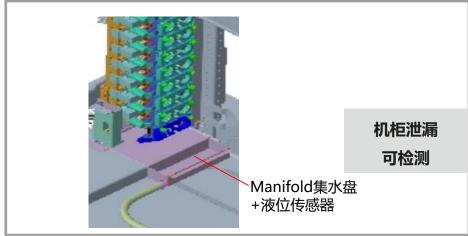








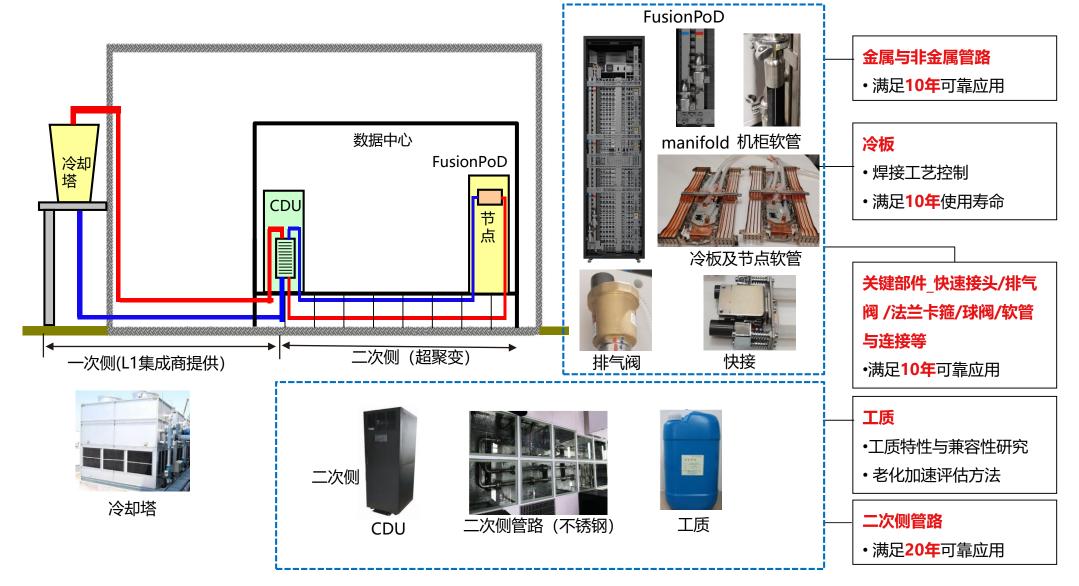




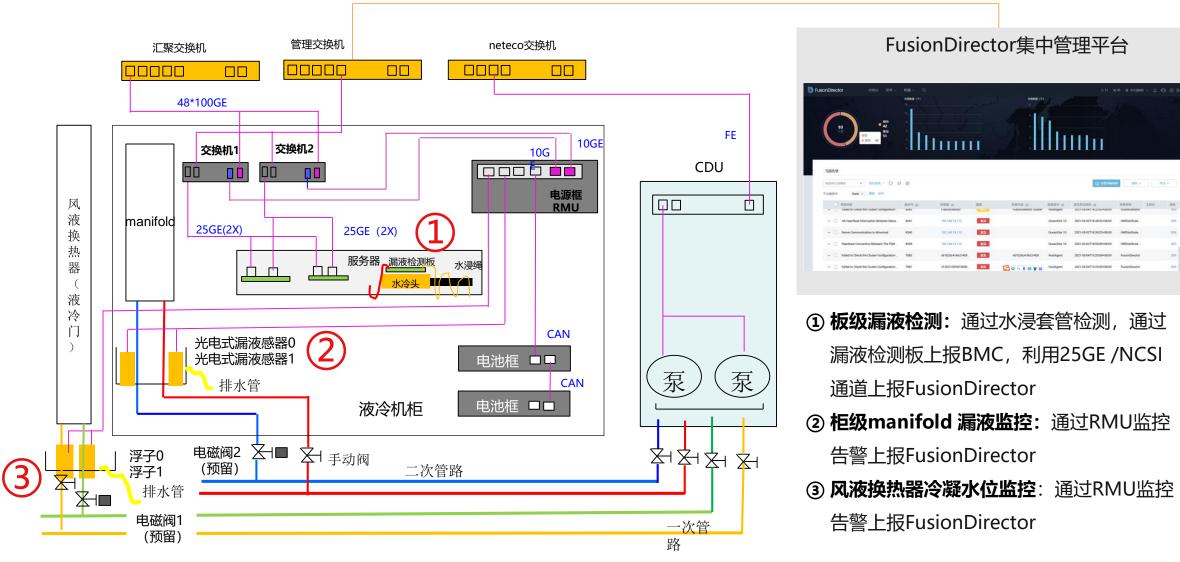
风冷改液冷: "一漏漏一柜" VS 液冷原生: "一漏漏一台"



液冷系统关键部件采用高可靠器件,满足10+年使用寿命

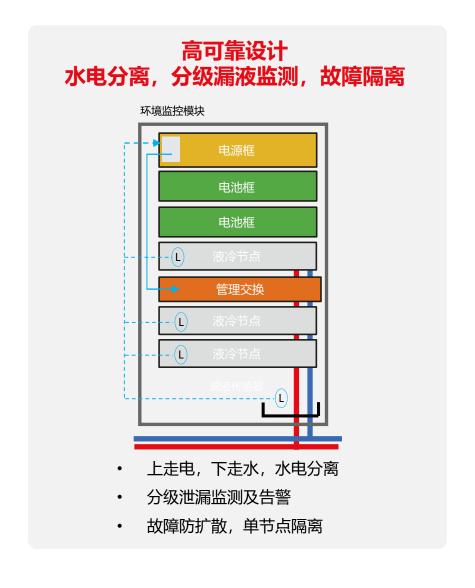


整机柜三级漏液检测能力,漏液精细化告警管理





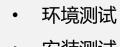
十年可靠性积累,170项测试,全球规模商用



测试保障 170项测试确保系统安全可靠性



- 泄漏测试
- 堵塞测试
- 运输测试



- 安装测试
- 安规测试
- 腐蚀测试
- 老化测试
-





FusionPoD for AI: 单柜部署8台GPU节点64张GPU卡,密度业界最高



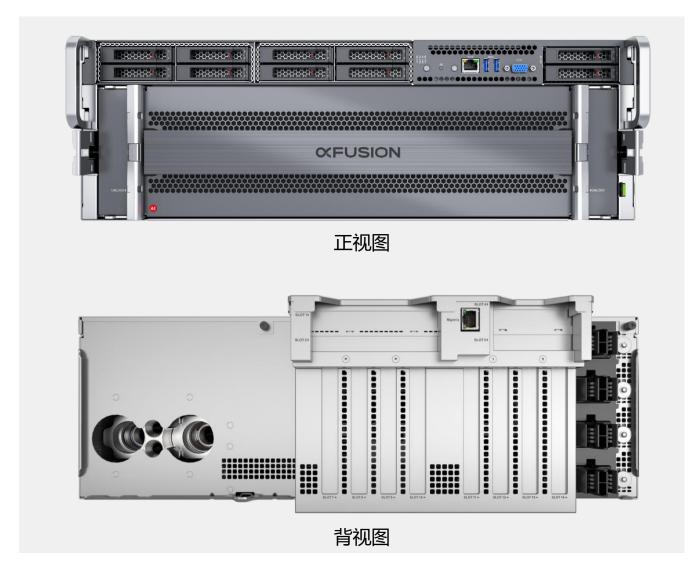
GN560E V7 4U 2*SPR/EMR+32DIMM+8*SXM GPU (**A800/H800/H20**)



模块	FusionPoD for AI整机柜规格
机柜尺寸	高度:2200mm(不含脚轮)/2250mm (含脚轮) 宽度:600mm宽 深度:1200mm(不含液冷门)/1400mm (含液冷门)
服务器节点	8pcs 4U 2P+HGX 8GPU节点
交换机配置	支持5个交换机槽位,支持通用交换机和定 制交换机
供电电源配置	最大 <mark>105kW</mark> 供电; 支持12路3相380V输入,2N供电系统 支持3kW电源模块,支持N+1冗余;
管理模块	适配RM230,提供机柜各部件管理功能,包括资产管理、电源管理,功率封顶、液冷监控、温湿度、二次管路漏液检测等功能
GPU独立维护	GPU模组可以独立插拔维护,不用将服务器 拔出机柜



GN560E V7服务器节点规格配置,CPU/GPU子框独家支持热插拔维护



部件	AI 液冷服务器规格
节点机箱	4U高度,宽度21inch,深度900mm
CPU	每个节点支持2个SPR/EMR处理器,最大功耗支持到 350W
冷板	CPU+内存+VRD+GPU支持液冷散热
内存	支持32根DDR5内存条
存储	支持12*2.5inch NVMe/SAS/SATA SSD,面板热插拔;支 持Raid卡
AI模组	支持OAM类 8卡模组(当前已经适配NV HGX A800/H800
PCle卡	支持8*PCle5.0 x16 FHHL/HHHL 槽位+3*HHHL/2*FHHL
前部IO及指 示灯	支持1个PowerButton按键/指示灯,1个Health指示灯,1个UID按键/指示灯,USB及VGA接口
ВМС	主板内置1711,1个GE管理接口
漏液检测	支持漏液检测



原生液冷服务器易运维,可靠性高,部署周期短

	液冷原生整机柜液冷服务器方案	风冷改液冷服务器方案
施工难易度	管路盲插,工厂预制,即插即用,即拔即断,部 署效率高,对现场工程人员要求低。	部署的节点多,管路手插工作量大,部署效率较 低,对现场工程人员要求高,验收过程中可能无 法察觉隐患
运维难易度	线缆密度:单柜8节点,电源线、管路 盲插,布 线简单,运维人员操作 难度低 ,故障定位 速度提 升50%; 防喷溅:盲插快接头采用防喷溅设计,有效预防 液体喷溅到服务器和PDU内部; 漏液隔离和导流:原生液冷设计,节点底座全密 封,并有挡水墙和导流结构,漏液不扩散;	线缆密度 : 单柜4节点, 机柜背部至少有10个液冷接头, 40根电源线, 布管走线密度大且复杂, 导致运维人员操作难度较大, 故障定位困难; 防喷溅 : 手插快接头没有防喷溅设计, 管路中水压大, 水易喷射至服务器、PDU内部, 导致服务器宕机; 漏液隔离和导流: 管路中液体泄漏后,漏液易扩散,导致下层服务器损坏;
机柜密度(按照 8GPU模组配置)	节点4U高度,单机柜部署密度高,机柜数量减少, <mark>部署和运维工作量降低95%</mark>	风冷改液冷服务器高度8U,单机柜部署密度低, 机柜数量增多,部署和运维工作量大
可靠性	整机柜电源集中在机柜顶部,管路集中在机柜中下部,下走水上走电,水电分离,更可靠	非原生液冷设计,单节点发生漏液易扩散,影响 业务正常运行。水管和PDU纵向混布,存在漏液 短路风险



总结:液冷原生服务器全面领先风冷改液冷服务器,是冷板式液冷服务器首选



液冷原生整机柜服务器 FusionPoD for AI-GN560E V7



手插液冷机柜+风冷改液冷服务器

	节点内部 无孔铆设 计	节点内漏液 导流凸包	节点内部署 前后挡水墙	节点U高 ≤4U	总线盲 插	柜内少线缆	机柜集中供电	机柜水电分离
原生液冷整机 柜服务器	V	√	V	√	√	√	\checkmark	√
风冷改液冷机 柜、服务器	×	×	×	×	×	×	×	×



1

超聚变FusionPoD for AI液冷解决方案

2

G8600E V7解耦方案

3

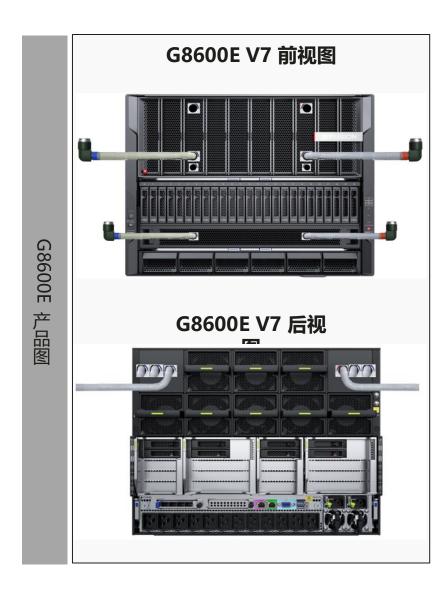
智算中心节点配置及组网方案

4

超聚变液冷改造及服务方案



G8600E V7规格简介



G8600				
应用场景	超大规模AI训练			
类别	8U2P GPU服务器			
GPU	NVLink模组: 8GPU模组			
СРИ	2个第四代英特尔®至强®可扩展处理器(Sapphire Rapids),支持单处理器最大 350W功率			
硬盘	支持多种不同的硬盘配置,硬盘支持热插拔 •最多可配置25个2.5英寸SAS/SATA硬盘 •最多可配置8个NVMe SSD盘 •2个M.2 SSD			
RAID	可选配支持RAID0、1、10、5、50、6、60等,支持Cache超级电容保护			
Ю	高性能拓扑:8*PCIe标卡,2*OCP3.0卡,8*NVMe 5.0,其中8张NVMe盘可以替换成8张标卡 均衡拓扑:12*PCIe标卡,2*OCP3.0卡,8*NVMe 5.0,其中8张NVMe盘可以替换成8张标卡			
电源	6个54V双输入热插拔电源模块,2个12V热插拔电源模块,均支持N+N冗余			
风扇	GPU区域: 10个54V风扇,支持N+1冗余 CPU区域: 5个12V风扇,支持N+1冗余			
温度	5°C~35°C			
尺寸	356mm×447 mm×898 mm (高x宽x深)			



G8600E V7柜级部署

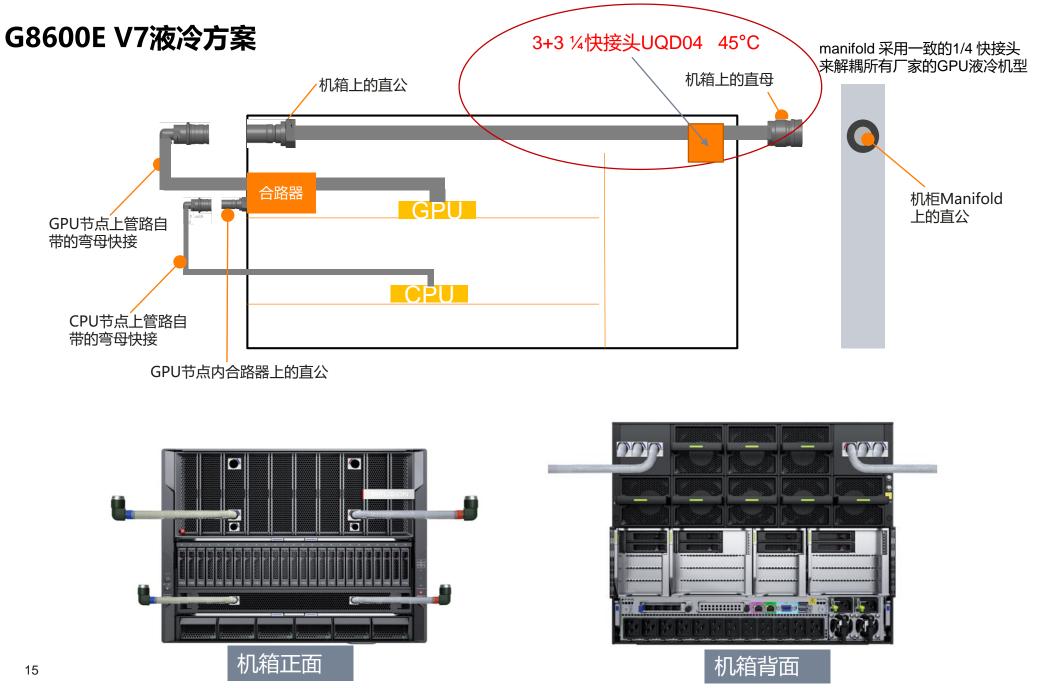
- 1. 考虑到整机柜功耗约50kW,建议PDU 3+3部署在机柜后部;
- 2. 为了将服务器快接到机柜manifold的软管做到最短(流阻小),且冷热水管阻力差异小(冷热水管长度一致),将manifold分置在机柜后部两侧;
- 3. 服务器上自带的快接及其软管长度约400mm;
- 4. 节点内合路,四分之一快接头三进三出。

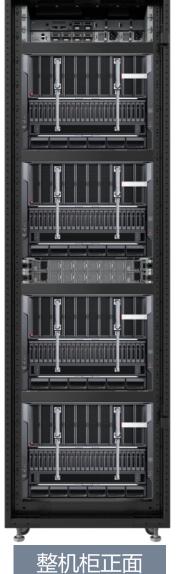


G8600E V7背面

47U高600宽1275深机柜 假面板	$\overline{}$
SW	
sw	+
ŚŴ	+
<u> </u>	
非线型	1
<u> </u>	
	++
IKIH1X	快接
	快接
	快接
G8600_server4	
00000 00.10.1	
走线型	_
AEXON.	快接
	快接
	快接
G8600 server3	
走线架	
	\vdash
	+-+
	++
IEXIENTIX	快接
	
	快接
	快接
G8600 server2	
22222 3011012	
走线架	
72.75	快接
	快接
	快接
	IV JX
G8600 server1	\vdash
	\vdash
管路/走线架	
	SW 走线架 假面板 G8600 server4 - 走线架 假面板 假面板 假面板 (Ba600 server2







XFUSION

1 超聚变FusionPoD for Al液冷解决方案

2 G8600E V7解耦方案

3 智算中心节点配置及组网方案

4 超聚变液冷改造及服务方案



训练服务器配置-GN560E V7

部件	推荐参数
CPU	2 * Intel Sapphire Rapids 8462Y+ (32 核心, 2.8GHz)
内存	2TB (32 * 64GB) DDR5,4800Mhz,RDIMM
硬盘	2 * 960GB SATA SSD(系统盘); 5 * 1.92TB NVMe SSD (本地缓存盘);
以太网卡	2 * 25GE双端口以太网卡;
参数面IB网卡	8 * 400GE单端口NDR IB网卡
存储面网卡	2 * 200GE单端口HDR IB网卡

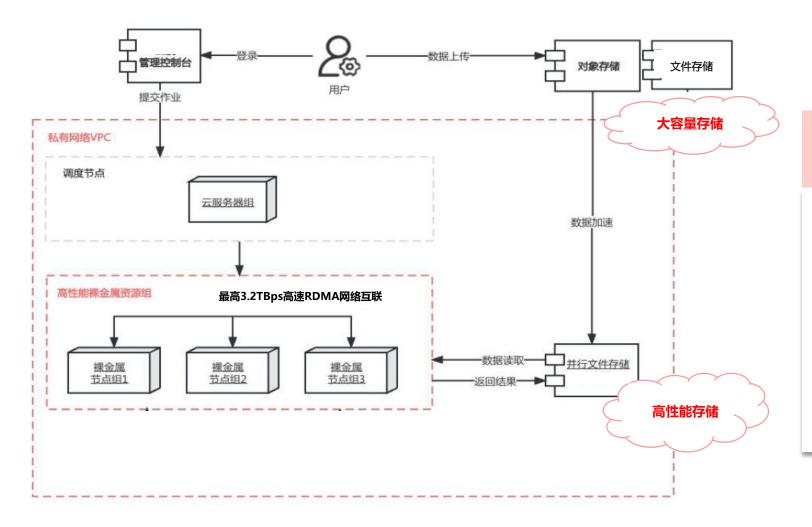


系统节点配置-2288E V6

部件	推荐参数
СРИ	2 * 英特尔至强8352Y处理器 (32核, 2.2 GHz)
内存	16 * 32GB 内存条, DDR4 RDIMM, 3200MHz
硬盘	2 * 960G SATA SSD,5 * 1.92T NVMe SSD;
raid	1 * SAS RAID card,NO Cache raid 0,1,10
以太网卡	2 * 25GE双端口以太网卡;
参数面IB网卡	2 * 200GE 单端口HDR IB网卡
电源	2 * 1500W 白金交流电源,1+1冗余



AI集群存储建设:分级资源配置,统一管理



存储区: 满足海量数据存储以及高IO存储需求, 分级分类实现数据调用

- ✓ **高性能全闪存储**: 用于存储智算区高I/0数据数据,要求使用全闪阵列;海量客户端接入,单文件系统支持数干客户端并行接入;高性能RDMA网络,百GB的吞吐、百万的IOPS能力的同时保证毫秒级的时延;
- ✓ 大容量对象存储: 多层负载均衡架构,从接入层到存储层横向无限扩展;支持分级存储,数据生命周期管理;支持权限控制、数据加密、合规保护、版本控制等多种安全能力;



存储设计: 高性能全闪存储满足训练业务需求

> 整体规划

1. 可用容量3PB

2. 全局存储节点双端口200G IB冗余,接到不同交换机上;

> 性能需求

• *NVMe全闪存储池*:

可用容量: 3.0 PB, EC: 8+2, 任意2个节点同时故障, 数据不丢失

聚合带宽:读 500 GB/s,写 450GB/s; **IOPS (4K)**:写6000W,读4000W

> 硬件配置

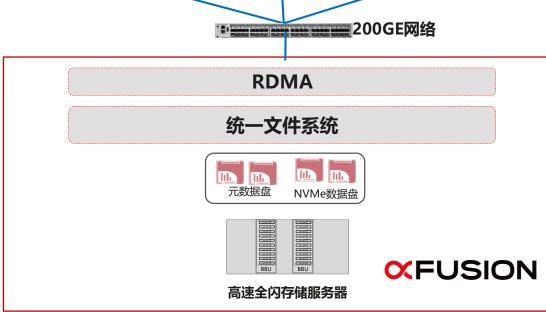
类型	数据节点		
数量	22		
节点配置	 CPU: 2 * Intel SPR 5418Y(24核, 2.0GHz) 内存: 256GB 硬盘1: 2 * 480GB SATA SSD; 硬盘2: 24 * 7.68TB NVMe SSD 网卡1: 2 * 200GE单端口IB HDR网卡; 网卡2: 1 * 25GE双端口以太网卡 电源: 2 * 1500W, 电源冗余 		
20			

> 主要特性

- 对称式, 元数据/存储数据统一部署架构
- · 协议互通:提供私有客户端/NFS/CIFS多种协议原生访问同一文件
- Linux、windows客户端兼容访问
- 支持配额、文件快照、克隆、数据隔离、客户端访问权限、快速数据恢复、等
- 图形化存储管理监控、分析、故障告警等

> 组网拓扑

客户端 (Linux、Windows)
文件协议(NFS、CIFS、FTP等)
计算集群
业务管理节点
200GE网络



对象存储: Fusionone Storage统一存储,支持多种存储协议,满足业务需求

> 整体规划

- 1. 可用容量3PB;
- 2. 全局存储节点2块双端口25GE 冗余,接到不同交换机上,防止单点 故障。

> 容量需求

· HDD混闪存储池:

可用容量: 3.0 PB, EC: 6+2, 支持最多2个节点同时故障, 数据不丢失。

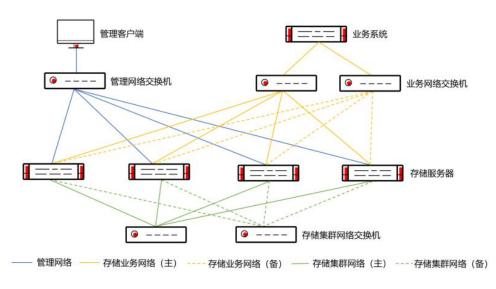
> 硬件配置

类型	存储节点		
数量	8		
	• CPU: 2 * Intel Xeon 4316(20核, 2.3GHz)		
	• 内存: 256GB		
	• 系统盘: 2 * 480GB SATA SSD		
	• 缓存盘: 4 * 7.68TB NVMe SSD		
节点配置	• 数据盘: 36*16TB SATA HDD		
	• 网卡1: 1* GE 四端口以太网网卡;		
	• 网卡2&3: 12* 25GE双端口以太网网卡		
	• 电源: 2 * 2000W,电源冗余		

> 主要特性

- · 可线性扩展至4096节点;分布式Cache,扩展至TB级;
- 高速数据并行重建1TB小于20分钟;
- 统一存储,支持iscsi、rbd、cifs、nfs、s3 等存储协议;
- 支持配额、快照、克隆、监控检测、用户访问权限设置、快速数据恢复等;
- 统一运维,智能化管理,极简运维体验

≻组网拓扑

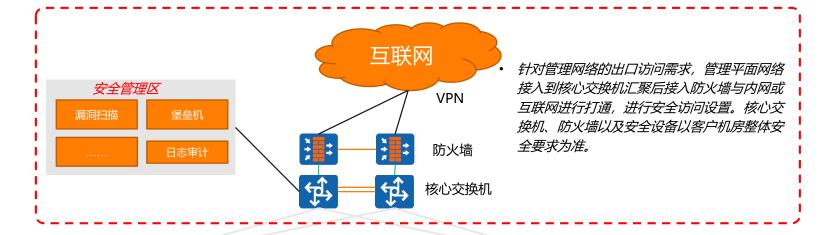


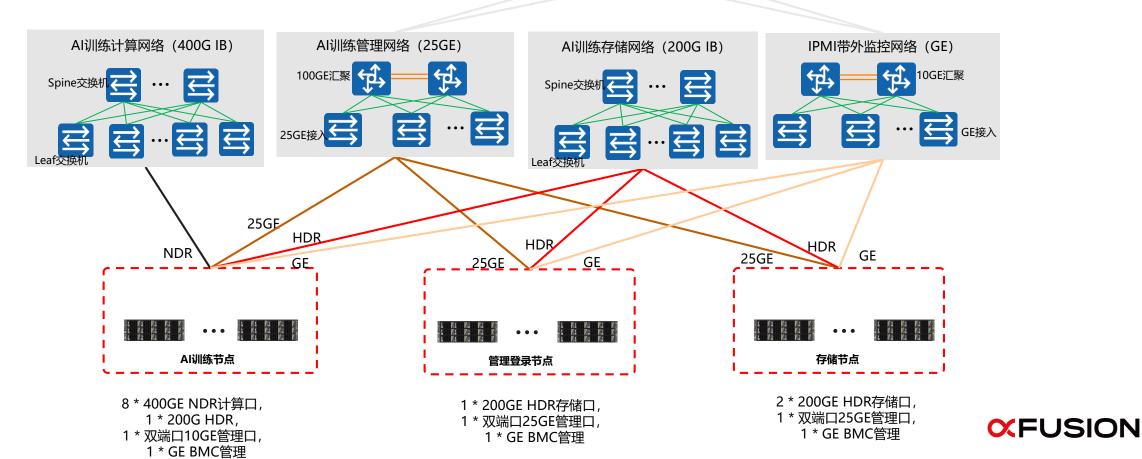
备注:存储设备数量较少时,存储业务网络和存储集群网络可共用交换机,无需分开部署。



AI集群整体网络设计

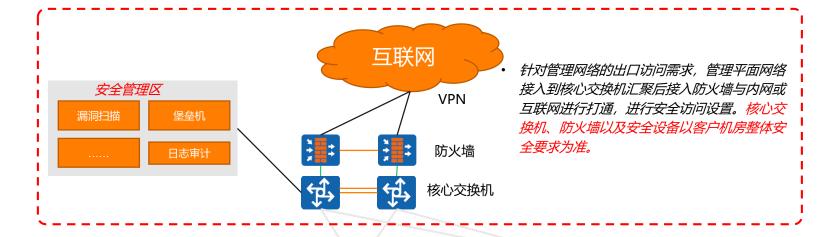
- AI训练计算网络采用400Gb/s NDR网络互联;
- AI训练存储网络采用200Gb/s HDR网络互联;
- AI训练管理网络采用25Gb/s 以太网接入;
- IPMI带外网络采用1Gb/s以太网接入

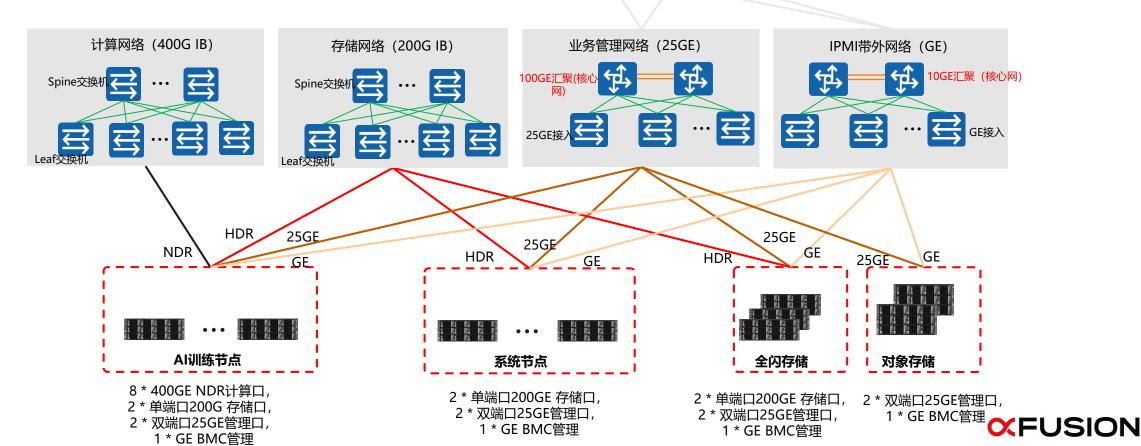




AI集群整体网络设计

- 计算网络采用400Gb/s NDR网络互联;
- 存储网络采用200Gb/s HDR网络互联;
- 业务管理网络采用25Gb/s 以太网接入;
- IPMI带外网络采用1Gb/s以太网接入





网络设计: 计算网络设计可分为两种方案

需求:集群总体规模128台,分两期建设,一期64台,二期64台;按照建设节奏有两种方式建设:

方案	内容	优势	劣势
方案一	正常建设方式,先按照64台规模建设, 后续通过线路改造按照128台规模组网。	不需要提前设Spine交换 机完全建设,节约第一期 的投资成本	后续扩展线缆需要进行改造(主要是Leaf到Spine交换机侧线缆),改造期间需中断一期业务
方案二	平滑扩容方式,先按照128台规模建设 Spine交换侧,Leaf侧连接64台规模, 后续再扩展二期Leaf交换机	二期扩容时,不需要动一期的线缆,建设快,一期业务不中断	一期网络建设成本增加

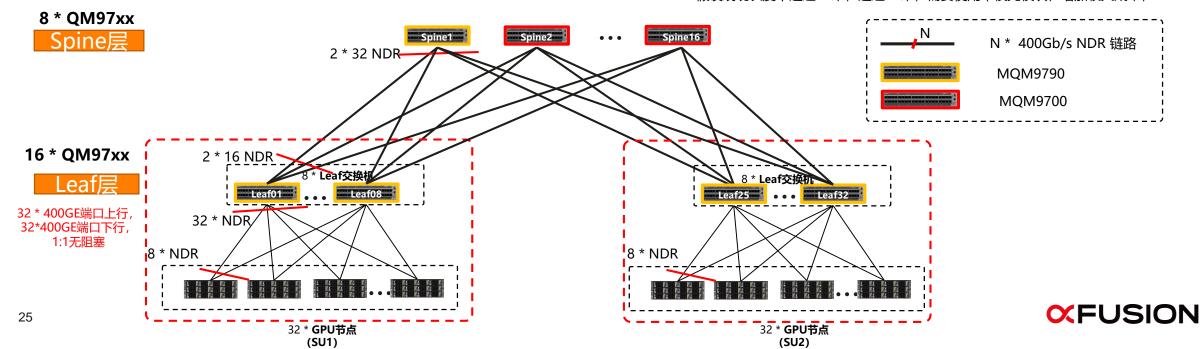


计算网络设计:方案1,按照正常建设方式64台服务器规模建设

- 每台GPU节点8张400G IB网卡,采用QM97xx交换机,胖树双层组网,全局1:1阻塞比;
- 400G IB交换机型号为MQM97xx,规格为64*400GE;
- 64台GPU节点,每台GPU节点配置8个400G网卡;每台GPU节点的8个网卡采用rail的方式分别连接到8台Leaf交换机中,32个8卡GPU节点组成1个SU;64台节点划分成2个SU。
- Leaf层配置16台QM9790; Spine层配置8台QM97xx; 共需24台交换机; 其中包含2台QM9700 (带管理), 22台QM9790 (不带管理);

	类别	描述	单位	数量
交换机	NDR交换机	InfiniBand MQM9790, NDR InfiniBand Switch, 64 个400Gb/s NDR ports, no managed	台	22
		InfiniBand MQM9700, NDR InfiniBand Switch, 64 个400Gb/s NDR ports, managed	台	2
Leaf交换机 to 网卡	NDR光模块	800G光模块,多模,小于50米,NDR Switch	个	256
	NDR儿保坏	400G光模块,多模,小于50m,HCA网卡	\uparrow	512
	光纤跳线	400G to 400G,MPO to MPO,多模, <mark>20m/50m</mark>	根	512
Leaf to Spine 交换机	NDR光模块	800G光模块,多模,小于50米,NDR Switch	根	512
	光纤跳线	400G to 400G,MPO to MPO,多模,20m	个	512

- 以光模块方式部署计算,实际线缆长度以及线缆类型需根据机房实际情况确定;
 - 假设线缆长度不超过50米,超过50米,需要使用单模光模块,增加较大成本;

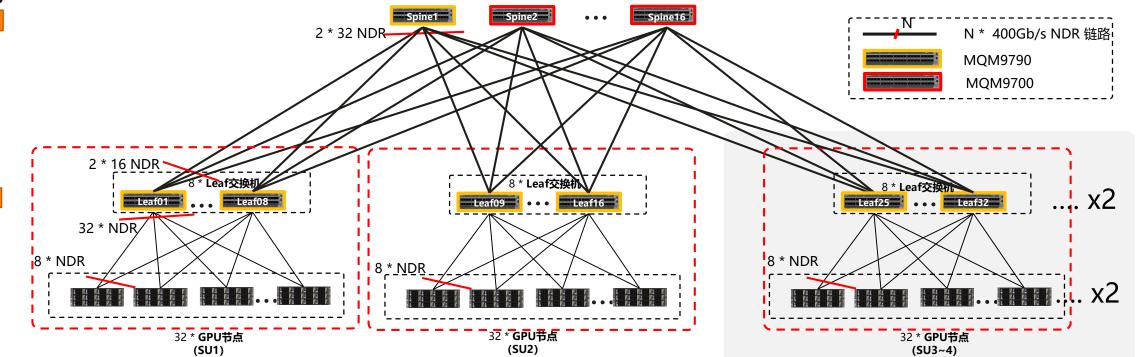


计算网络设计:方案2,按照平滑扩容方式64台服务器规模建设

- 每台GPU节点8张400G IB网卡,采用QM97xx交换机,胖树双层组网,全局1:1阻塞比;
- 400G IB交换机型号为MQM97xx, 规格为64*400GE;
- 64台GPU节点,每台GPU节点配置8个400G网卡;每台GPU节点的8个网卡采用rail的方式分别连接到8台Leaf交换机中,32个8卡GPU节点组成1个SU;64台节点划分成2个SU。二期64台节点也划分成2个SU。
- Leaf层配置16台QM9790; Spine层按128台GPU规格,配置16台QM97xx;共需32台交换机;其中包含2台QM9700(带管理),22台QM9790(不带管理);
- 后续二期扩展时,只需要再增加16台Leaf交换机及相应的光模块等;

		类别	描述		数量	
	交换机	NDR交换机	InfiniBand MQM9790, NDR InfiniBand Switch, 64个400Gb/s NDR ports, no managed		30	
			InfiniBand MQM9700, NDR InfiniBand Switch, 64个400Gb/s NDR ports, managed	台	2	
	Leaf交换机 to 网卡	NDR光模块	800G光模块,多模,小于50米,NDR Switch		个	256
			400G光模块,多模,小于50m,HCA网卡	个	512	
		光纤跳线	400G to 400G,MPO to MPO,多模, <mark>20m/50m</mark>	根	512	
	Leaf to Spine交换	NDR光模块	800G光模块,多模,小于50米,NDR Switch	根	512	
	机	光纤跳线	400G to 400G,MPO to MPO,多模,20m	个	512	

- 以光模块方式部署计算,实际线缆长度以及线缆类型需根据机房实际情况确定
- 假设线缆长度不超过50米,超过50米,需要使用单模光模块,增加较大成本;



16 * QM9790 Spine层

16 * QM9790

Leaf层

32 * 400GE端口上行, 32*400GE端口下行, 1:1无阻塞

26

一期: 64 * GPU节点

二期: 64 * GPU节点

计算网络设计结论: 推荐按照方案二进行计算网络的建设

• 理由:方案1后续扩容需要重新部署Leaf侧到Spine侧的线缆,会导致业务暂停;方案2扩容不需要改动一期部署的线缆,业务不暂停;

设备	配置		单位	方案1: 一期按照正常 建设	方案2: 一期按照平滑 升级	128台整体
交换机	NDR IB交换 机	InfiniBand MQM9790 ,NDR InfiniBand Switch, 64个400Gb/s NDR ports,no managed	台	22	30	46
交换机		InfiniBand MQM9700 ,NDR InfiniBand Switch, 64个400Gb/s NDR ports, managed	台	2	2	2
	NDR光模块	•800G光模块,多模,最大50米,NDR Switch	个	768	768	1536
光模块		•400G光模块,多模,最大50m,HCA网卡	个	512	512	1024
	光纤跳线 400G to 400G	•400G to 400G,多模,20m/50m	根	1024	1024	2048

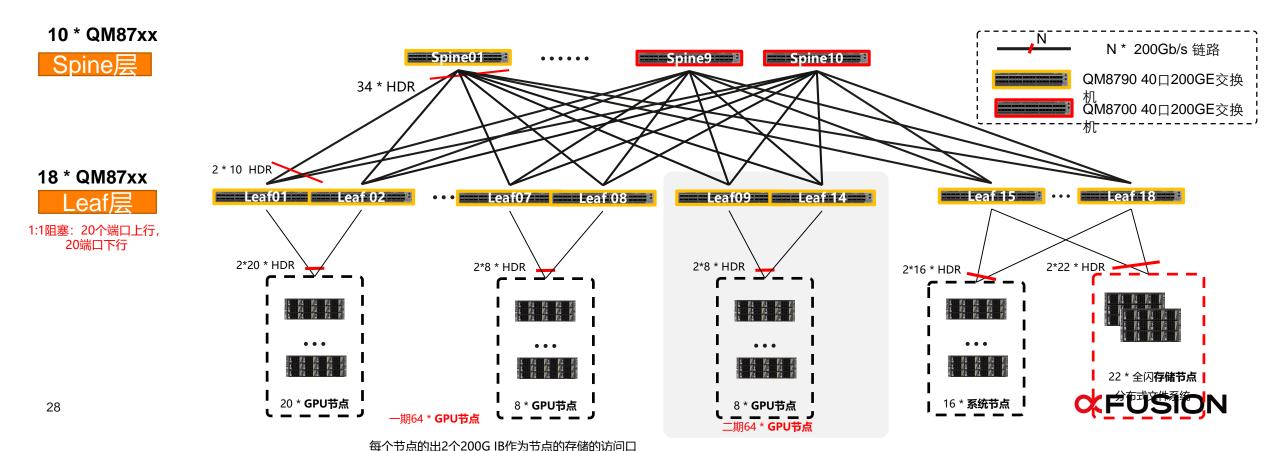


存储网络设计

- 数据访问网络采用200Gb/s HDR网络互联;每台GPU节点和全闪存储节点配置的2个200G IB端口作为存储访问端口,两个IB端口连接到不同的交换机上;
- 整体网络使用两层胖树结构, 达成全局1:1阻塞比;
- 按平滑升级方式64台GPU规格, Leaf层配置12台QM87xx; Spine层配置10台QM87xx; 共需22台交换机; 其中包含2台QM8700(带管理), 20台QM8790(不带管理);
- 后续二期扩展时(推荐平滑升级方式),只需要再增加6台Leaf交换机及相应的光模块等;
- 对象存储通过25GE以太网络接到业务网络中访问;

	类别	描述	单位	64台 数量	128台 数量
交换机	HDR交换机	InfiniBand MQM8790 40个200Gb/s QSFP56,双向带宽HDR端口,no managed	台	20	26
23/1/1		InfiniBand MQM8700 40个200Gb/s QSFP56,双向带宽HDR端口, managed	台	2	2
线缆		AOC有源光缆, 200Gb/s to 200Gb/s, QSFP56, 20m/50m	个	424	672

- 以AOC线缆方式部署计算,实际线缆长度以及线缆类型需根据机房实际情况确定;
- 假设线缆长度不超过50米,超过50米,需要使用单模光模块,增加较大成本;

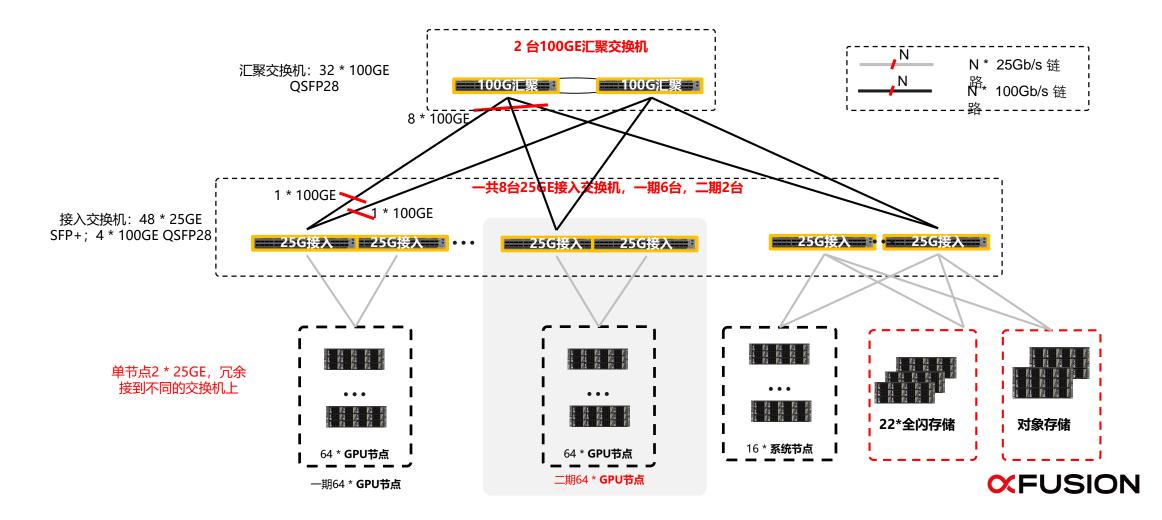


业务管理网络设计

- 业务管理网络采用以太网络互联,每个节点采用双口25GE冗余接入;
- 接入交换机: **8** * 25GE交换机, 48 * 25GE SFP28, 4 * 100GE QSFP28; 一期**6**台, 二期 **2**台;

业务如需接入外网,接入到核心网与外网互联

• 汇聚交换机: 2 * 100GE交换机, 32 *100GE QSFP28;

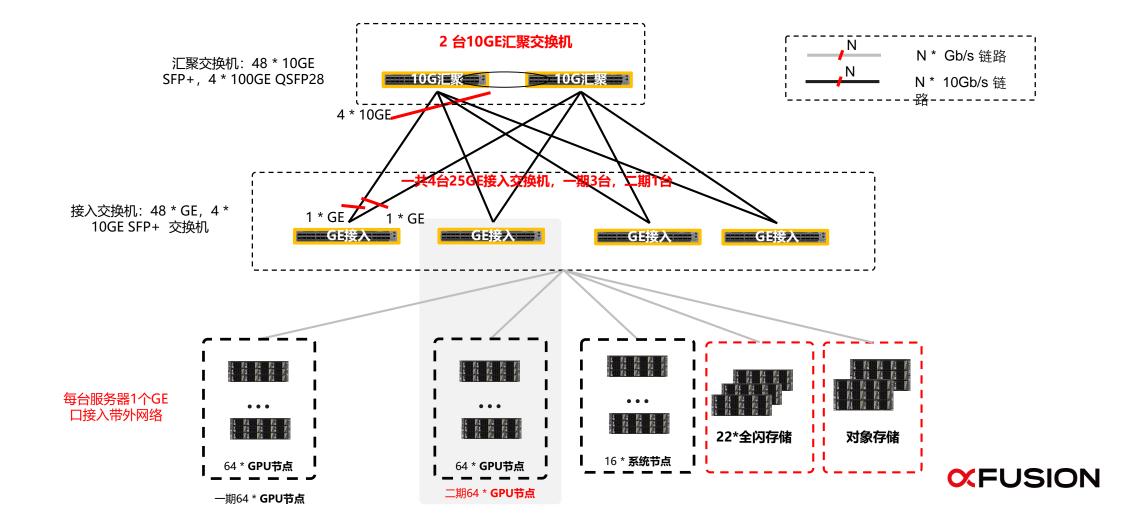


IPMI带外管理网络设计

- 带外管理网络采用以太网络互联,每个节点采用GE接入;
- 接入交换机: **4** * GE交换机, 48 *GE, 4 * 10GE SFP+; 一期**3**台, 二期**1**台;

• 汇聚交换机: 2 * 10GE交换机, 48 *10GE, 4 * 100GE QSFP28;

如需外网在线运维,接入到核心网与外网互联



1 超聚变FusionPoD for Al液冷解决方案

2 G8600E V7解耦方案

3 智算中心节点配置及组网方案

4 超聚变液冷实施服务及维保方案



超聚变液冷集成服务-项目管理

项目交付通过进度管理、集成管理、质量管理、问题与风险管理、变更管理、人员管理、沟通管理等主要管理方法,全方位对各项交付活动进行管控,确保满足客户对进度和质量的要求

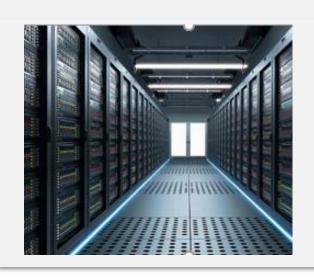
项目准备

- 现场工程勘测
- 规划设计



项目实施

- 由专业的服务团队实施交付工作
- 专家现场保障项目进度和质量



项目验收

由超聚变工程师与客户共同完成项目 验收。签属服务报告,完成服务交付





超聚变液冷集成服务-液冷集成维保及增值服务

基础维保服务外,还提供主动支持服务,让客户放心用



变被动为主动

· VIP备件

专属备件支持,快速响应备件需求,针对 客户极致SLA要求,节省客户备件投资

• 巡检服务

设备检查、故障预测、液冷系统维护、保持设备运行正常,并输出《巡检报告》

• 重大事件保障

关键业务期,专业的保障方案,专家现场保障



超聚变液冷基础维保服务

• 面向客户的技术支持服务,满足客户的多种需求

Super-Care

TAC支持

- 7*24 技术支持服务
- 专家支持团队
- 基于 ITIL的IT平台及流程

在线服务

- 知识库 (案例, 文档)
- 专家论坛及自助服务工具
- 软件更新授权

硬件支持

- 硬件预更换
- 硬件返修

现场支持

- 现场硬件更换
- 现场问题处理

服务内容	Super- Care 标准+	Super- Care 金牌+	Super- Care 白金+
备件先行	7x10xND	7x24x4H	7x24x2H
现场硬件 更换	7x10xND	7x24x4H	7x24x2H
现场问题 处理	7x10xND	7x24x4H	7x24x2H



超聚变新疆本地服务能力介绍

- ・本地设立独立办事处,为新疆客户提供服务支持;
- 具有552万+设备服务经验,可提供优质规划设计与实施服务,培训服务,数据中心咨询服务等;
- 全国48个自有库房备件中心,本地设立乌鲁木齐&克拉玛依两大备件库,覆盖全疆备件需求;





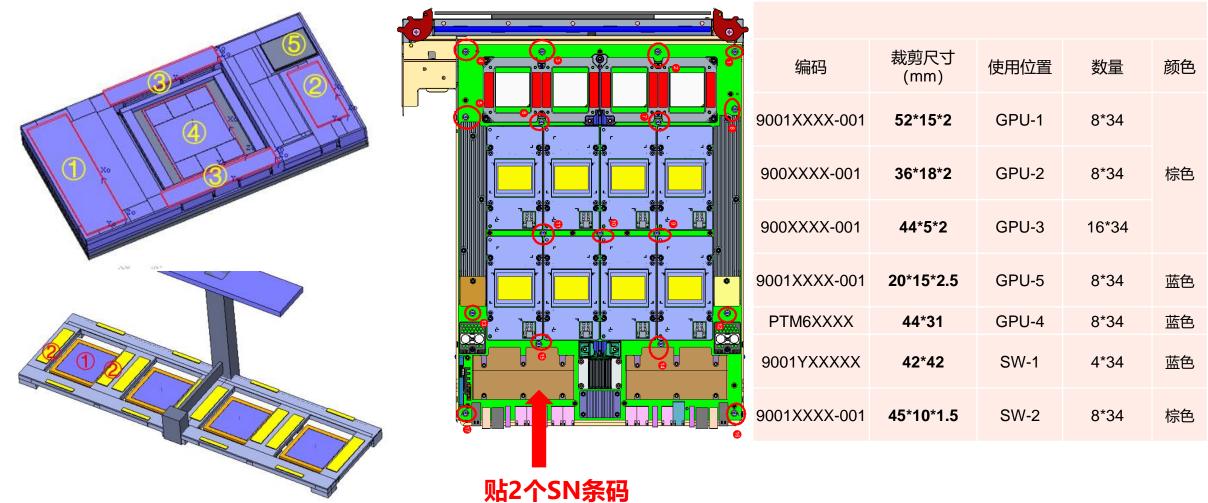






GPU模组组装:

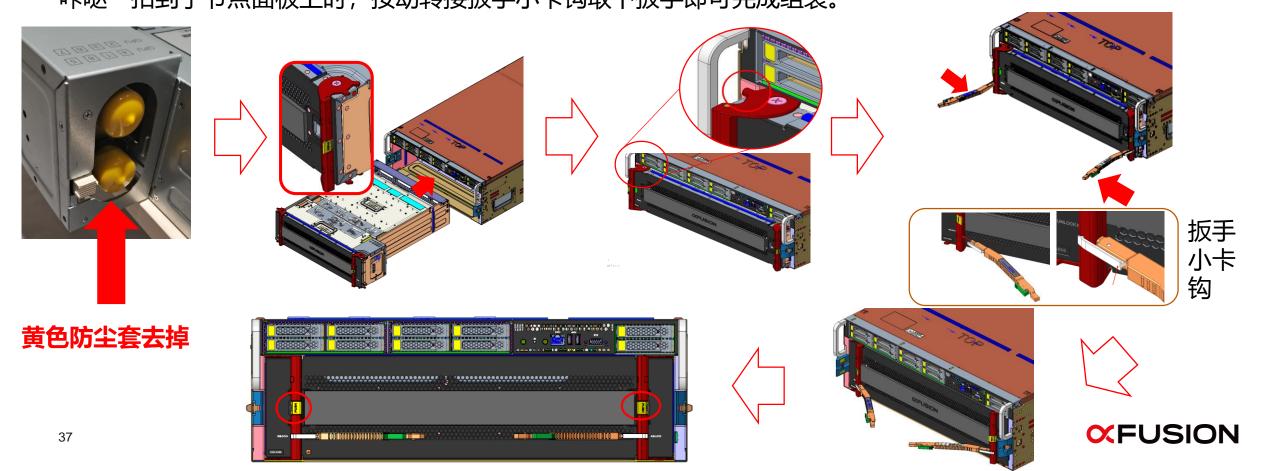
粘贴导热垫和SN条码,SN条码粘贴完成后,用扫码枪扫描记录条码





整机组装

- **1、GPU子节点组装**:将子节点两侧扳手打开(即PUSH按键没有钩住的状态),将GPU子节点按照如图方向组水平推入到机箱内,至扳手卡钩和机箱端面搭接;**注意拔掉冷板进/出水口的黄色保护套。**
- **2、GPU子节点插入**:按照图示将转接扳手插到子节点扳手卡口内,按照红色箭头方向旋转扳手,当"PUSH"按键"咔哒"扣到子节点面板上时,按动转接扳手小卡钩取下扳手即可完成组装。



Thank you.

让数字世界无限可能 Fusion X, Digital Infinity

Copyright©2022 xFusion Digital Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. xFusion may change the information at any time without notice.

