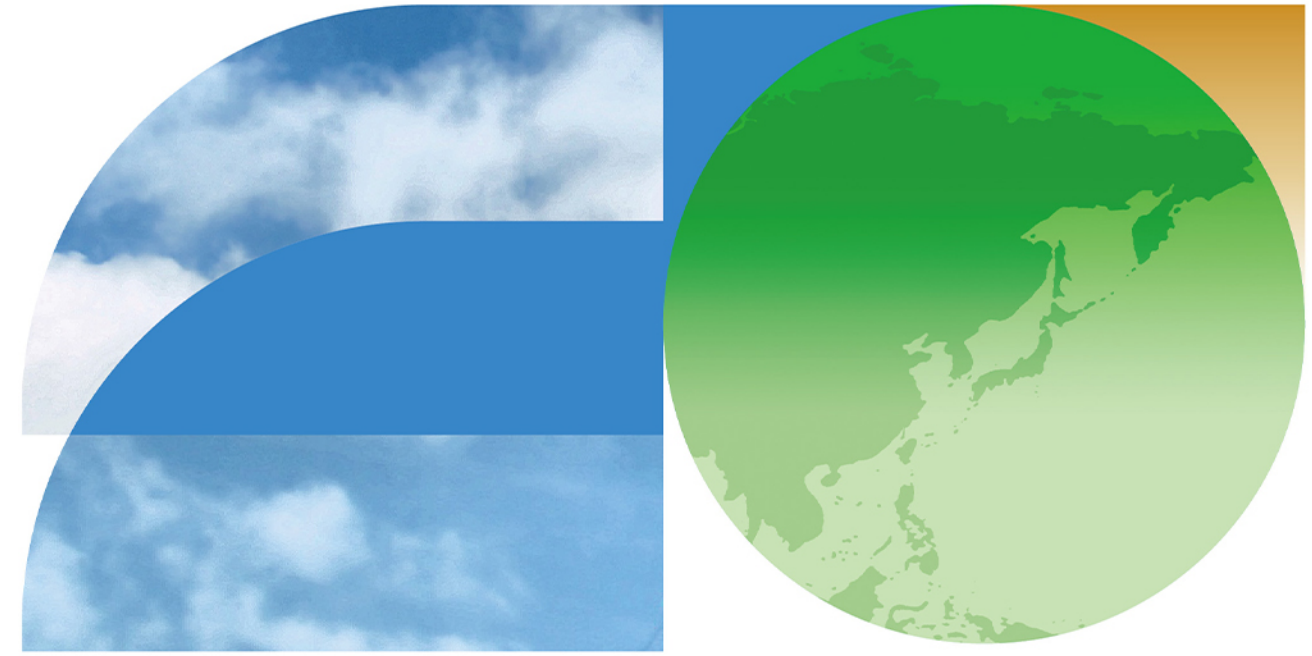


联想高性能计算解决方案白皮书

售前专线 400-819-6776
售后专线 400-106-8888

卷首语

随着以人工智能为代表的新一代信息技术的迅速发展和广泛应用，算力不仅成为助力产业升级、加快科技经济融合发展的重要支撑，也成为各区域科技创新和基础研究的竞争高地。高性能计算在“新基建”下的内涵和外延不断扩大，不只在传统行业应用发挥着重要作用，在新兴领域也产生了新的突破。

同时随着“十四五”规划提出“建设数字中国”未来愿景，高性能计算更进一步成为数字社会发展的基石。千行百业对数据处理的需求越来越高，不仅在石油勘探、气象预报、航天、科学研究等领域处处可见高性能计算的身影，在金融、政府信息化、教育、企业、网络游戏等新兴领域对高性能计算的需求也出现了爆发式增长。这趋势的背后是，政府、企业客户、科研机构等已经将算力视为智能化转型的基础。没有算力支持，就无法释放数据的真正价值，指导业务决策和业务创新的目标更是无从谈起，同时也意味着数字城市失去了高效运行的基础。

联想作为全球最大的高性能计算解决方案提供商，坚持开放、融合的技术路线，并以高效能的高性能计算解决方案不断推动企业智能化转型。联想凭借坚实的技术积累、顶尖的人才储备、完善的生态布局，持续引领行业。2020年11月20日，TOP500组织发布了最新的全球高性能计算机TOP500榜单，联想制造的高性能计算机入围182台，数量远超其他厂商，以36%的总体份额，再次名列全球高性能计算提供商份额第一名。目前，联想在全球部署的高性能计算机正在政府机构、商业公司、学术机构发挥力量，在科学计算、城市管理、人工智能开发等领域助力中国新基建，赋能实体经济，加速构建新发展格局。

任重道远须策马，风正潮平好扬帆。为了与行业伙伴共同推动高性能计算在各行各业的落地实践，联想依托于自身出色的高性能计算优势，整合优质的内外部资源，推出了《联想高性能计算解决方案白皮书》，涉及政府、科研、教育、气象、海洋、石油石化、CAE仿真计算、生命科学、人工智能等各个行业，涵盖高性能计算产品、高性能计算硬件和环境配套、高性能计算软件、高性能计算集群实施服务等，帮助客户、合作伙伴在通过高性能计算解决方案进行行业应用过程中有据可依，进而实现业务价值的提升。

最后，联想希望与客户、合作伙伴共乘时代风帆，不断推动高性能计算与行业应用的深度融合，进而顺利实现企业智能化转型，为赋能新基建，实现经济高质量发展作出贡献。

联想集团高级副总裁
联想企业科技集团总裁

CONTENTS

联想高性能计算

高性能计算多年来一直是科技综合实力竞争的至高点，也在一定程度上反映了各大服务器厂商系统研发方面的实力。作为行业的技术领先者，联想公司在这一领域积累了长达 20 多年的丰厚经验，并在关键技术领域不断创新，取得大量里程碑式的成果。

联想从 2001 年进军高性能服务器领域，是最早针对高等院校和科研院所进行产品开发与市场拓展的厂商，并在市场中一直处于技术领先地位。截止 2020 年，先后为数万个用户成功实施了高性能集群。曾经两次承担了中国科学院网络计算中心主节点的建设任务，并且成功地与威廉姆斯车队进行合作，成为国产品牌中最早将高性能业务拓展到海外的企业。

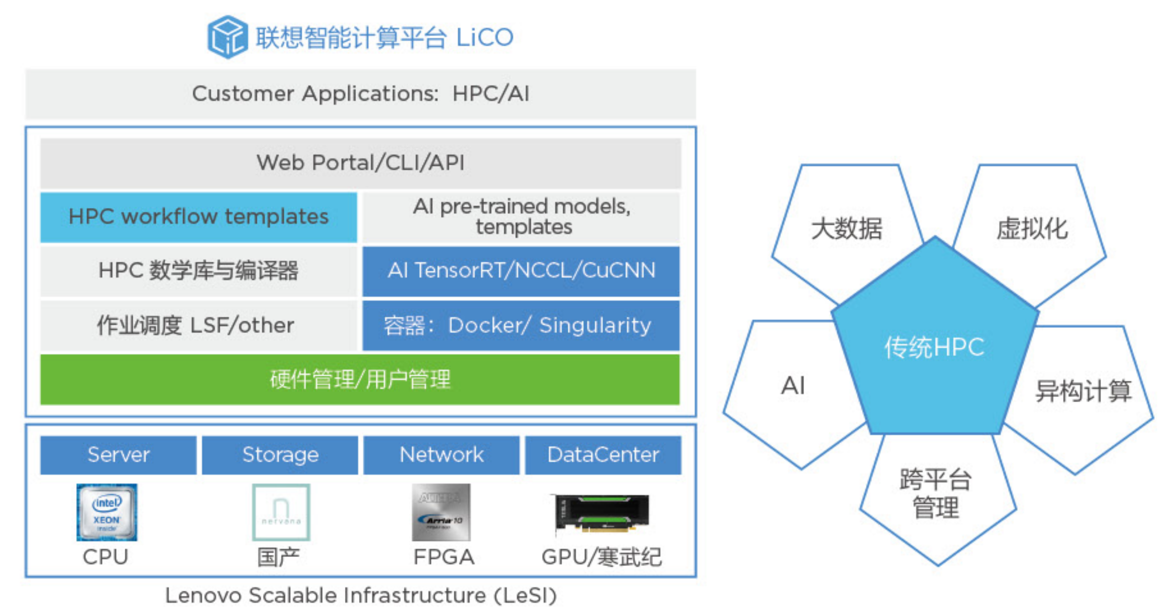
联想顺应国际主流技术发展趋势，以市场需求为驱动，吸收国内外最新技术成果，进行了大量创新性研发，突破包括系统设计与优化、系统基础架构、系统软件等在内的一大批高性能服务器的关键核心技术，开发出一系列可扩展、易管理、好使用、稳定可靠的高性能服务器产品，并配备可满足用户个性化需求的行业解决方案，提供从系统层到应用软件层的全面解决方案和技术服务。

联想高性能计算一贯秉承开放，融合和高效的理念，并以此来服务高性能计算各行各业，成就高性能计算客户。

联想HPC发展愿景

开放

融合



高效

2002 年 7 月，联想研制成功“深腾 1800”万亿次集群系统，安装在中科院数学与系统科学研究院。这是世界上第一个实际速度超过 1 万亿次的大规模集群系统。曾入选新华社 2002 年中国十大新闻及两院院士评选的 2002 年中国十大科技进展，并荣获 2004 年国家科技进步二等奖。2002 年末，另一套深腾 1800 大规模集群系统安装在中科院大气物理所国家重点实验室。2002 年 12 月 30 日，联想深腾 1800 中标大庆油田，使该油田第一次在国内实现三维叠前深度偏移地震资料处理。

2003 年，联想成功研制“国家网格主节点—联想深腾 6800 超级计算机”，安装在中科院计算机网络信息中心。这是当时世界上 Linpack 效率（78.5%）最高的高端通用计算机，其组合查询性能名列当时所有大型服务器的第四位， 其典型应用 MM5 的测试结果在 2004 年 3 月列世界所有超级计算机的第一位。该机荣获 2005 年国家科学技术进步二等奖、2005 年国家重点新产品奖、2004 年信息产业重大技术发明奖。联想深腾 6800 自 2004 年初在网络中心对外服务以来，一直 7 X 24 小时稳定运行，在双星计划、气候模式计算、油藏模拟、材料科学计算、流体力学计算等领域取得了 150 多项重要计算成果。

联想深腾系列高性能计算机成为最早进入世界 TOP500 的一批国产计算机，分列当时世界 TOP500 的第 14、43、98 和 299 名。这是一个历史性的突破，联想深腾系列高性能计算机已成为国际知名国内主流的品牌。联想在推动高性能技术产业化方面取得了突破性进展，联想的高性能计算机广泛应用于许多关键领域，在国民经济和社会发展中发挥重要作用。

目前，集群已成为世界高性能计算机体系结构的主流，联想深腾 1800、深腾 6800 和深腾 7000 为这一趋势的形成做出了重要贡献。2002 年 8 月初，世界上主流并行编程环境 MPI-ch 的发明人、美国阿贡实验室 William Cropps 参观联想深腾 1800 后写道：“We see the future of clustering computing”。

联想在高性能服务器基础技术方面有着长期的积累，有齐全的产品线和严格的质量控制体系，为高性能计算机的研制和生产奠定了坚实的基础。

在产品设计上，联想坚持用户导向的原则，同时结合对新技术的深入理解和消化吸收，始终遵循模块化设计思想，在充分综合考虑各模块精密配合和整机系统合理整合的基础上，先设计出高性价比、稳定的产品方案，然后对方案进行工程计算仿真，同时不断地结合验证性实验，最终才形成可行的开发方案，从而保证为用户在最短的时间里开发出最贴近的具有竞争力的产品。

在研究开发上，联想建立了与国际接轨的两级研发体系，即公司级研发平台和各事业部研发中心。公司级研发平台由联想研究院、软件中心、板卡中心和工业设计中心组成。事业部研发中心隶属于各事业部，直接承担具体的专项技术开发工作。联想在高性能服务器技术上已突破并拥有了核心技术，拥有知识产权的系统设计与优化技术、系统监控技术、系统管理技术、高可用和负载均衡技术以及基础架构技术等关键技术，在高性能计算机系统技术方面已申请国家发明专利 85 项，其中，46 项已获授权。

在工程技术上，联想拥有针对服务器的部件及整机进行专业性测试的全套技术。部件测试包含外观、结构、功能、兼容性、可靠性、安全性、性能和环境 8 个方面的测试，以保证所有部件符合联想服务器技术特性和质量标准的要求，对服务器的核心部件如电源、内存，还建立了专业化的实验室，实现了部件的自动测试。如自动电源测试实验室和自动内存测试实验室，国内功能全面、技术先进的系统测试实验室，以及高温实验室、电磁兼容检测实验室、噪音实验室、湿热实验室等等，所有产品需要在这些实验室中通过一系列的严格检测，只有通过了这一系列的严格检测的服务器产品，才可以顺利出厂，提供给客户。联想始终严格执行国际标准的质量控制体系，是通过 ISO9000 – 2000 版质量认证体系的服务器厂商。

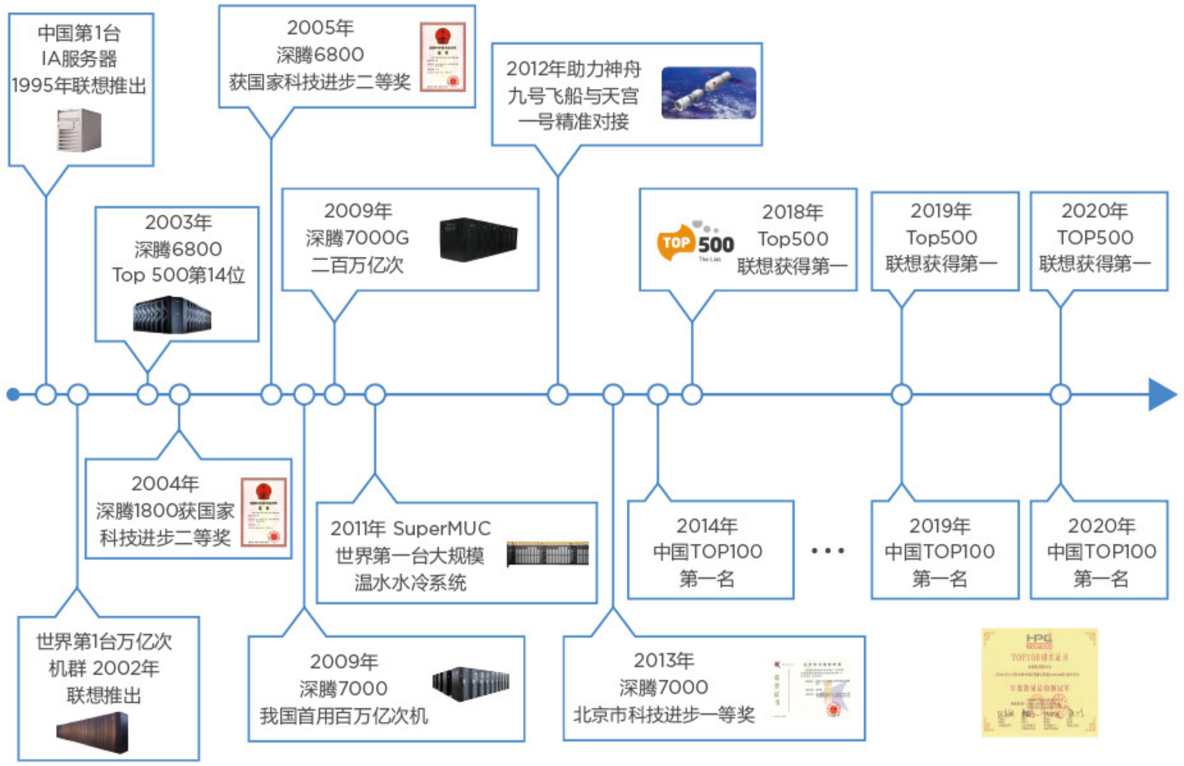
在技术服务与方案上，联想服务器应用方案中心拥有雄厚的技术力量，在硬件平台、操作系统、数据库、软件、网络、存储、集群技术等方面有着多年的技术和经验积累，可以分别从不同的技术层面为用户提供有效的产品应用和方案支持服务。中心拥有先进的实验环境，包括方案集成实验室、性能评测实验室、数据中心、客户实验室四个部分，为用户提供方案开发、测试、方案移植、优化以及培训、咨询等服务，及时、快速、可靠地解决用户系统在使用过程中所遇到的技术问题，使客户的系统可以更加安全稳定地运行，以保障和促进客户业务的顺利开展并取得更大的成功。

2014 年 9 月 29 日， 联想宣布完成对 IBM x86 业务的收购，从此，具有 IBM 的优秀的 X86 超算产品线与 IBM x86 大批 HPC 专家加入了联想。与联想原来的 HPC 团队，形成了全新的联想 HPC 业务团队。

2016 年 7 月 1 日，从 ISC 2016 凯旋归来的联想集团再度吹响集结号，在北京隆重召开了以“开启 E 级计算新篇章”为主题的首届全球高性能计算峰会。本次大会联想正式发布了面向 E 级计算的高性能计算机系统深腾 x8800。

2018 年 6 月 30 日，联想正式升级高性能计算机系统为深腾 x9000，这是联想面向智能高性能计算的统一平台，该平台涵盖传统高性能计算和人工智能技术。这是联想集团 3S 战略的具体方案落地，也是向人工智能方向迈出的一大步。

时光的积淀成就联想HPC的辉煌



2020 年 11 月 20 日，TOP500 组织发布了最新的全球高性能计算机 TOP500 榜单。榜单显示，在全球浮点运算性能最强的 500 台高性能计算机中，联想制造的高性能计算机入围 180 台，数量远超其他厂商，以 36% 的总体份额，再次名列全球高性能计算提供商份额第一名。

目前，联想在全球部署的高性能计算机正在政府机构、商业公司、学术机构发挥力量。在科学计算、城市管理、人工智能开发等领域助力中国新基建，新 IT 赋能实体经济，助力构建新发展格局。

02

联想高性能计算行业解决方案

联想政府科研教育行业 HPC 解决方案

系统总体架构

联想高性能计算系统分为以下 6 个子系统：

1. 计算系统：用于执行用户计算任务的子系统

根据计算系统中的服务器节点的硬件配置及提供的计算服务的不同，计算系统中的服务器节点可分为 CPU 计算节点、GPU 加速节点和大内存多核胖节点。

① CPU 计算节点　② GPU 加速节点

③ 大内存多核胖节点

2. 存储系统：用于存储用户数据或计算任务数据的子系统。根据存储系统的架构设计和构建。

① 存储节点与磁盘阵列

② 传统分布式存储系统

③ DSS-D/DSS-G 高端并行存储系统

3. 服务系统：用于管理和监控集群系统以及用户接入的子系统。根据提供的服务类型不同，服务系统中的功能节点分为登录节点、可视化节点、管理节点和监控节点四类。

① 登录节点：系统使用者通过该节点登录集群进行作业。

② 可视化节点（包括远程可视化节点）：系统使用者远程可视化。

③ 管理节点：系统管理员通过该节点管理和配置整个集群。

④ 监控节点：系统管理员和运维人员通过该节点对整个集群系统的软硬件资源进行监控。

4. 网络系统：用于子系统内连接和子系统间互连。根据网络连接的子系统不同，网络系统分为计算网络、存储网络、集群管理和监控网络。

① 计算网络　　④ 集群监控网络

② 存储网络　　⑤ KVM 系统

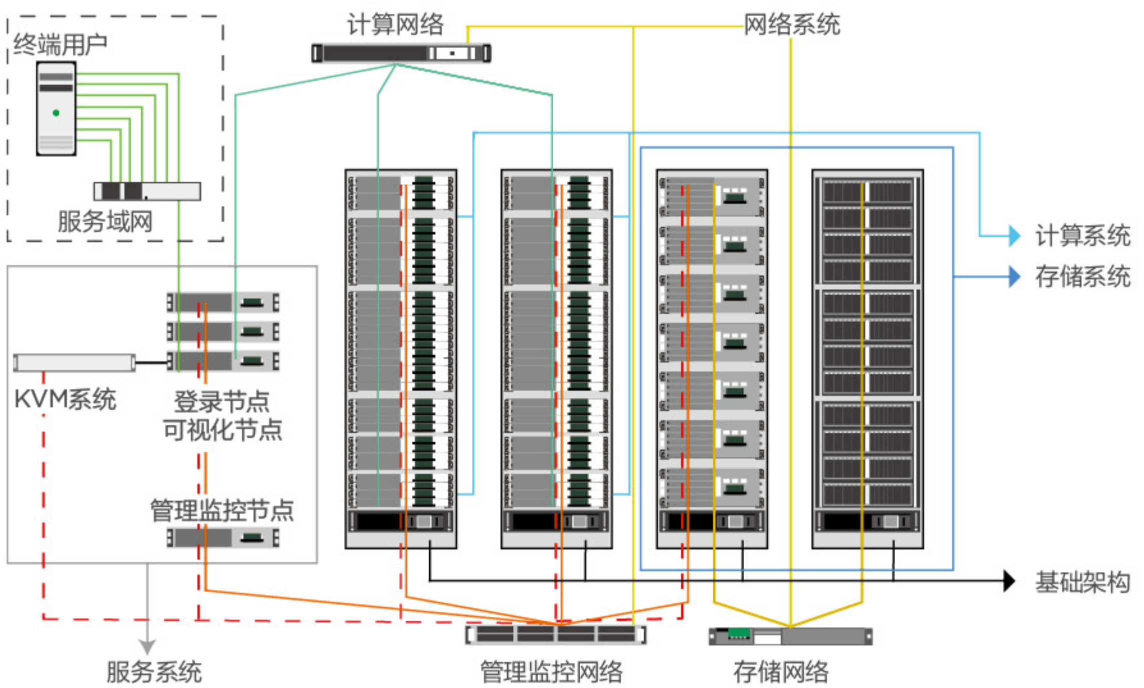
③ 集群管理网络

5. 基础架构：为整个集群系统提供供电、散热、布线等服务。基础架构包括机柜、电源系统、散热系统和布线系统，聚合整个集群系统的基础设施同时为集群提供供电、散热和布线等基础服务。

① 机柜　　③ 散热系统

② 电源系统　　④ 布线系统

6. 软件系统：用于整合管理硬件资源，为用户提供软件平台与服务。联想高性能集群系统提供丰富的软件系统层来支撑集群硬件系统向用户提供服务。联想高性能计算软件系统 —— LiCO 包括快速集群部署、易于集群管理、合理资源调度、实时监控集群状态等多种工具，同时又集成很多不同领域热门用户应用及相应的开发环境。



联想高性能计算高性能集群硬件系统架构图

联想高性能集群系统构成						
子系统	子系统功能模块					
计算系统	CPU 计算节点	GPU 计算节点	大内存节点	N/A	N/A	N/A
存储系统	I/O 节点 + 盘阵	I/O 节点 + 存储节点	N/A	N/A	N/A	N/A
服务系统	管理节点	监控节点	登录节点	可视化节点	KVM 系统	N/A
网络系统	计算网络	存储网络	管理网络	监控网络	N/A	N/A
基础架构	机柜	电源系统	散热系统	布线系统	N/A	N/A
软件系统	集群软件系统	节点操作系统	编程环境	应用软件	文件系统	测试工具

联想高性能计算集群系统构成表

联想高性能集群系统的软件系统层次图

① 节点操作系统：集群中各子系统节点支持多种操作系统，如 Red Hat Enterprise Linux、CentOS、SUSE、Windows Server 等。

② 集群软件系统：集群软件系统提供集群的管理、集群的监控、用户作业和集群资源的调度、集群的优化调试等功能。

③ 文件系统：存储系统的软件系统，用于存储资源的整合管理和满足计算系统的存储需求。

④ 编程环境：用于支撑用户计算程序的编写和运行，以及集群调试优化和测试程序的运行。

⑤ 用户应用软件（应用软件）：用户生产工具。

政府科研和教育常见应用特性分析

联想高性能计算系统集成和支持了多个领域的多种商业和开源的科学计算应用软件，包括但不限于商业应用软件如：

MOLPRO、GAUSSIAN、Q-CHEM、MATLAB、VASP、TURBOMOLE、WIEN2K、SPARTAN、Materials Studio、Linda、LS-DYNA、FLUENT、CRYSTAL、ADF 等；

开源应用软件如 GROMACS、GrADS、ABINIT、GMT、CASTEP、CPMD、DL_POLY、DOCK、GRAPES、MM5、NAMD、AutoDock、VMD、TINKER、OpenMX、VENUS、WRF 等。

以下对这些软件的特点进行简单的分析

1. MOLPRO

MOLPRO 是国际上广泛使用的专业级电子结构量化计算软件，MOLPRO 的特色是高精度计算，通过多参考 CI，耦合簇和有关的方法，广泛处理电子相关问题。直接积分局域电子相关方法使得同分子尺寸正相关的计算量大为减少，从而能够对更大的分子体系进行准确的从头计算。MOLPRO 基于 Linux 操作系统运行，程序主体是用 Fortran 90 语言编写，支持大多数常见的 MPP 超级计算机和计算集群。对于分布式内存系统如集群，由 GlobalArrays 并行工具集支持并行运行；对于共享内存系统，由 OpenMP 的方式实现了部分代码的并行化。需要注意的是，MOLPRO 的并行部分通讯代价较大，一般不建议基于普通以太网运行。

MOLPRO 中并行化了传统积分的 SCF、DFT、MRCI、MP2、LMP2、CCSD(T) 能量和 SCF、DFT 梯度计算，以及直接积分和密度拟合的 SCF、DFT、LMP2 和 LCCSD(T) 计算。在最新版本中还并行化了 DF-HF、DF-KS、DF-LCCSD(T)。

MOLPRO 目前的并行代码包括两种形态，一是运行于分布式内存的多处理器系统上，如计算集群，这种方式需要 GlobalArrays 并行工具的支持；另一种方式是运行在共享内存系统上，这部分代码采用 OpenMP 进行并行化。由于计算模型导致的并行任务间通讯负载较大的问题，MOLPRO 并行版在普通网络上难以表现出良好的加速比，通常建议其在专用高速网上运行。

MOLPRO 通常有三种编译模式：

1. 串行模式： 运行时只有单个实际执行体；

2.MPP 模式： 一个任务被分为多个执行实体，同时运行在多个处理器上；

3.MPPX 模式： 多个任务分配到多个处理器上同时执行，每个任务用一个处理器串行执行，例如，在梯度的有限差分求解中，多个构型计算在多个处理器上同时进行，处理不同的数据，属于 SPMD 的计算模式。就目前而言，这种模式只用于数值求一阶梯度和 Hessian。

MOLPRO 的运行形态在编译时指定为上述三种模式之一，并且会在输出结果中显示。

2. GAUSSIAN

Gaussian 是目前最流行的量子化学综合软件包，用于研究许多化学领域的课题，例如取代基的影响，化学反应机理，势能曲面和激发能等。

Gaussian 用 Fortran 语言编写，从量子力学的基本原理出发，可计算能量、分子结构、分子体系的振动频率以及大量从这些基本计算方法中导出的分子性质。它能用于研究不同条件下的分子和反应，包括稳定的粒子和实验上难以观测的化合物，例如瞬时的反应中间物和过渡结构。

Gaussian 03 软件主要采用两种并行方式，对于 IBM 小型机或者 SGI 的 Origin 系列或者 Altix 系列的高性能机来说，采用基于 OpenMP 的并行方式。OpenMP 模式下的多个线程并发地运行在基于共享内存方式或者基于 NUMA 体系结构的高性能计算机上，以此提升程序地运算速度和运算规模。这种并行方式受到单节点的处理器数的限制。如果服务器的处理器数为 4，则 OpenMP 并行方式最多只能使 4 颗处理器并行工作。

对于分布式内存的集群系统，Gaussian 03 用 Linda 并行库来实现跨节点的并发执行。TCP-Linda 是专门为 Gaussian 03 设计，用以实现其分布式内存系统下的并行。Gaussian 03 在提交任务的时候选择分布式内存方式还是共享内存方式。

Linda 模型定义了一个功能强大的逻辑存储器 (TS) 和在其上的一组核心操作 (in, out, read, eval), 它们能够方便地嵌入到不同的语言 (如，C, Fortran) 中而构成相应的并行语言 (C-Linda 和 Fortran-Linda 等)，该模型同时还支持动态程序设计和两种不同的编程风格 (Master/Slaver 和 Divide-and-Conquer)，为用户开发不同类型的应用程序提供了灵活的手段。

TS 是一个可被多个进程共享的、能同时存放数据和任务的数据箱。进程从 TS 中提取任务或数据进行计算，并将结果或生成的新任务放入 TS 中。并行执行的进程之间通过 TS 进行间接的通信和同步。

3. Q-CHEM

Q-CHEM 是由遍及全球的多个研究机构共同开发的从头计算量子化学软件包，它包含了当前流行的各种最先进量子化学理论方法和工具。Q-CHEM 的特点是与当前最先进的量子化学理论和计算方法结合的非常紧密。Q-CHEM 采用了全新的方法理论、最先进的算法和现代编程技术进行设计，系统地组合各种计算方法和工具，极大地提高了计算速度和准确性，能够以较高精度进行大分子体系的运算。

Q-CHEM 主要用作电子结构从头计算，可以对分子的基态和激发态进行第一定律计算。Q-CHEM 可以采用串行和并行两种方式运行。Q-CHEM 的 HF、DFT 方法及它们的二阶导数计算。采用 MPI 编写，使用分布式模式并行运行。Q-CHEM 对解析和数值积分的计算采用的动态负载均衡的办法，并且在求解耦合扰动 SCF 方程中采用全局内存使用共享的方式，使得它能够利用并行计算机的分布内存进行大结构的频率计算。

4. MATLAB

MATLAB 是建立在向量、数组和矩阵基础上的一种分析和仿真工具软件包，自 1984 年由美国 MathWorks 公司推向市场以来，历经二十多年的发展与竞争，已成为国际公认的最优秀的工程应用开发环境。

MATLAB 包含各种能够进行常规运算的“工具箱”，如常用的矩阵代数运算、数组运算、方程求根、优化计算及函数求导积分符号运算等；同时还提供了编程计算的编程特性，通过编程可以解决一些复杂的工程问题；也可绘制二维、三维图形，输出结果可视化。

MATLAB 在 2004 年拥有了第一代的并行与分布式处理产品分布式工具箱（Distributed Computing Tools）。分布式工具箱将待处理任务分配给多个处理器，使采用不同运算结构的 MATLAB 用户共享同一个处理器池的计算能力。在最新的版本中，MATLAB 能使跨节点运行的任务之间相互通信，做到真正的并行运算，此功能是通过 MPI 实现。分布式计算工具箱可以在多处理器计算环境中使用 MATLAB 解决计算、数据密集型问题。

5. VASP

VASP 是固体物理程序中非常稳定，性能优秀的软件之一，被广泛的用于固体性质的计算，一直是相关领域研究人员的关注热点。VASP 使用赝势和平面波基组，进行量子力学分子动力学计算，它基于 CASTEP 1989 版，由奥地利 Vienna 大学开发。

VASP 中采用基于有限温度下的局域密度近似（用自由能作为变量）的方法，对每一 MD 步骤用有效矩阵对角方案和有效 Pulay 混合求解瞬时电子基态。这些技术可以避免原始的 Car-Parrinello 方法（基于电子、离子运动方程同时积分的方法）存在的一切问题。离子和电子的相互作用采用超缓 Vanderbilt 赝势 (US-PP) 或投影扩充波 (PAW) 方法描述，可以在相当程度上减少过渡金属或第一行元素的每个原子所必需的平面波数量。同时，VASP 可以很容易地进行力与张量的计算，以达到把原子衰减到其瞬时基态的目的。VASP 的并行由 MPI 实现，可以在分布式内存环境上进行并行计算。

6. TURBOMOLE

TURBOMOLE 是一个强大的量子力学软件包，由德国卡尔斯鲁厄（Kalsruhe）大学的 Ahlrichs 教授领导的理论化学研究小组开发而成。TURBOMOLE 经过了十多年的不断改进，已经成为化学、物理和工程人员的有力工具。

与众多同类软件不同，TURBOMOLE 着重解决如何用尽量短的时间和尽量少的内存需求，快速稳定地处理工业应用型的分子。特别是独有的 RI-DFT 方法，可以较其它大多数量子力学程序节省 10 倍的 CPU 时间。由于它使用了辅助函数（单位元分解近似使得 TURBOMOLE 非常适合于计算大分子或重复计算中等分子。TURBOMOLE 可以以串行和并行两种方式运行。并行代码用 MPI 编写。

7. WIEN2K

WIEN2K 采用密度泛函理论来计算固体的电子结构，是基于键结构计算最准确的方案——完全势能（线性）增广平面波 ((L)APW)+ 局域轨道 (lo) 方法。在密度泛函中可以使用局域（自旋）密度近似 (LDA) 或广义梯度近似 (GGA)。

WIEN2K 主要运行在 Unix/Linux 上，有串行和并行方式。

WIEN2K 程序的 I/O 操作较频繁，可能导致处理器的闲置状况，通常解决方案是扩充读写缓冲，减少对磁盘的访问。WIEN2K 中某些代码对处理器资源的消耗极大，其中一些已经用 MPI 方式加以并行化。WIEN2K 的并行模式分 K 点（K point）并行和细粒度（fine grained）并行，二者属于不同层次上的并行划分，在安装时的并行机选项中可以指定具体的并行方式。

8. SPARTAN

SPARTAN 是由美国 Wavefunction 公司研发的多功能化学软件，是化学领域强大的分子力学和量子力学计算程序。SPARTAN 具有优秀的图形显示功能和成熟的算法，为制药与生物技术机构在靶标识别与确立、先导物的选择与优化和工艺优化过程提供关键数据支持。

SPARTAN 采用的理论方法有从头 Hartree-Fork、密度泛函（局域密度近似，以及 BP，BLYP，EDF1，B3LYP）、MPn（MP2，MP3，MP4，LMP2）、精细的相关能计算（CCSD，CCSD(T)，OD，OD(T)，G2，G3）、半经验方法（MNDO，MNDO(d)，AM1，PM3）、蒙特卡罗和分子力学。

SPARTAN 具有强大的图形和用户界面功能：从计算的 IR 数据绘制红外光谱图、从 Hartree-Fock 分子轨道理论绘制 13C NMR 图、分子数据库及搜索、氢键显示、内反应坐标 (IRC) 的产生和追踪、SDF，TGF，MOL 各式的分子坐标文件的输入输出、动画速度的人为控制和键型显示等。

9. Materials Studio

Materials Studio 是 Accelrys 公司专为材料科学领域开发的新一代材料计算软件，用以帮助研究人员解决化学及材料工业中的许多重要问题。Materials Studio 软件采用 Client/Server 结构。Materials Studio 将高质量的材料模拟带入了个人电脑的时代，极大的降低了研究的门槛。

多种先进算法的采用使得 Materials Studio 在性质预测、聚合物建模和 X 射线衍射模拟等方面具有强大的功能。同时，Materials Studio 可以生成的高质量图片、能处理各种来源的图形、文本以及数据表格。

Materials Studio 采用客户端 / 服务端的工作模式，较小计算量的任务直接运行在客户端，而长时间或大运算量的计算则运行到服务器端。

Materials Studio 中许多模块都实现了并行化，其中 Dmol3 模块是世界上最快的从头计算代码之一，业已成为制药业的一个重要的工具。而 CASTEP 模块的并行版本则可以模拟包含数百原子的大系统。

Materials Studio 的所有模块都是通过第三方队列软件支持串行工作提交，比如主流的 OpenPbs 和 LSF。对于 Linux 平台，还支持跨节点的 MPI 并行方式。

10. Linda

Linda 是美国耶鲁大学与科学计算协会共同研制的用于实现并行程序设计的、与机器无关的程序环境。Linda 通过增加一些函数对传统的程序设计语言进行扩充（如 C、Fortran 等），使其能实现并行程序的设计。将 Linda 映射到各计算语言中，就形成了可进行并行计算的程序语言 C-Linda、Fortran-Linda。Linda 可以运行在共享和分布式内存系统上。

Linda 运行在分布存储多处理机上，引入了虚拟共享存储概念，通过在各处理机上实现一个虚拟的共享存储器，将分布存储多处理机模拟成为基于共享存储器的多处理机，从而支持共享存储编程方式，提供比消息传递方式更方便的并行编程模式。

Linda 的虚拟共享存储器称为元组空间（Tuple Space）。元组空间由一组有序的元组（tuple）组成，元组的每个域都包含有实际的数据。元组空间是相联存储器，元组的标识与选择是通过域值匹配，而不是通常采用的地址选择方法。

Linda 提供了以下 4 种对元组的基本操作：

① out：将数据放入元组空间，整个操作是顺序进行的。

② eval：功能同 out，但它是并行执行的。

③ in：从元组空间中选择匹配的数据，并将数据从元组空间中删除。

④ rd：功能同 in，但它不将数据从元组空间中删除。

当两个进程需要交换数据时，不是采用消息传递来直接通信，而是通过读写元组空间来完成（即共享内存系统的并行程序设计）。例如，如果 A 进程需要向 B 进程传递数据，则需要完成以下的操作：A 向 Tuple 空间写一个数据，B 在 Tuple 空间中读取需要的数据。Linda 在编译系统和运行系统上做了很多优化工作，使用 Linda 编写的应用程序在效率上能够接近用传统消息传递方式编写的应用程序。

TCP-Linda 是专门为 Gaussian 03 设计，实现 Gaussian 03 在分布式内存系统上并行的必须程序，使 Gaussian 03 能突破单节点的限制，充分利用可用的计算资源，对于极耗资源的量子化学计算来说，具有相当的实用意义。

11. LS_DYNA

LS_DYNA 最初是由 John O. Hallquist 博士在美国劳伦斯 . 利弗摩尔国家实验室开发。LS_DYNA 最早实践了显式有限元理论，当代所有的显式求解程序都源于 LS_DYNA。

LS_DYNA 以拉格朗日（Lagrange）算法为主，同时有欧拉（Euler）求解技术；以显式求解为主，兼有隐式求解功能；以结构分析为主，兼有热分析、流体 - 结构耦合功能；以非线性动力分析为主，兼有静力分析功能（如动力分析前的预应力计算和薄板冲压成形后的回弹计算）；特别适合求解各种结构的高速碰撞、爆炸和金属成型等高度非线性瞬态动力学问题。

LS_DYNA 主要应用在如下领域：

① 汽车工业：碰撞分析、乘员约束系统及气囊设计、乘客被动安全、部件加工、轮胎在积水路面排水性和动平衡分析；

② 电子工业：跌落冲击分析、包装设计、热分析、电子封装、电子产品抗冲击性设计；

③ 航空航天：鸟撞、叶片包容性设计、异物损伤分析、飞机结构冲击动力分析、碰撞、坠毁、冲击爆炸及动态载荷、特种复合材料设计、制造工艺仿真及优化；

④ 建筑业：地震安全、混凝土结构、爆破拆除、公路桥梁设计；

⑤ 石油工业：液体晃动、完井射孔、管道设计、爆炸切割、事故模拟、海上平台设计。

LS_DYNA 具有良好的并行设计，能够支持分布式及共享内存方式的并行。LS_DYNA 能够支持网格计算 (Grid Computing)，将多台计算节点组合成多个网格点以实现协同计算。

在用户模型定义准确、单元规模和处理器数目匹配的情况下，LS_DYNA 的性能加速比和处理器数目基本呈线性关系。

LS_DYNA 得到了主流作业调度软件 Platform LSF 的有力支持。在共享内存和分布式内存系统上，LS_DYNA 作业都具有检查点 (checkpoint) 和重启功能。

12. FLUENT

FLUENT 软件最初原发于英国谢菲尔德大学 (Sheffield University)，现已成为一款商用 CFD 软件。FLUENT 占有全球商用 CFD 软件领域第一的市场份额，在航空航天、旋转机械、航海、石油化工、汽车、能源、计算机 / 电子、材料、冶金、生物、医药等领域有广泛的应用，美国宇航局 (NASA)、美国国防部 (DOD)、美国能源部 (DOE)、ABB 公司、西屋公司、波音公司、福特公司、三菱公司、IBM 公司、杜邦公司及各大汽车厂商等都是 FLUENT 的用户。

FLUENT 主要用于计算流体流动和传热问题。其提供的非结构网格生成程序，对相对复杂的几何结构网格生成非常有效。FLUENT 具有网格自适应能力，能高效地精确求解大梯度的流场。

FLUENT 采用的是客户端 / 服务器 (Client/Server) 设计模式。用户在客户端进行模拟问题的设定和数据输入，然后将计算任务提交服务器；服务器进行模拟计算并将结果传回客户端。

FLUENT 的并行计算主要是在服务器端进行的。FLUENT 并行运算由一个主机进程和多个节点进程组成。节点进程的数量和系统的处理器数量有关，计算任务均由节点进程完成。FLUENT 能支持共享式和分布式内存系统上的并行计算。

在网格处理技术方面，FLUENT 在并行环境下可以实现网格的自动分区并动态地平衡计算负载。

FLUENT 的并行效率和互连网络有很大关系。核心的线性方程组求解过程产生频繁的信息传递和交换，因而使用如 InfiniBand 等高速网能有效地提高程序的扩展性。ISGI、HP 和 Intel 等多种优化 MPI 接口也对 FLUENT 的并行性能有非常正面的影响。此外，FLUENT 中不同算法在共享式或者分布式系统上也会表现出一定的性能落差。

13. CRYSTAL

CRYSTAL 是研究晶态固体的最流行程序之一，并且是第一个公开发布的程序。CRYSTAL 程序密度泛函或各种混合近似方法，计算周期体系的电子结构。周期体系的 Bloch 函数作为原子中心 Gaussian 函数的线性组合展开。采用强大的屏蔽技术研究实空间区域。

CRYSTAL 可以用于研究分子，多聚物，表面及晶体的物理和电磁性质，使用全电子基组或者价基组进行限制性闭壳层、限制性开壳层或者非限制性计算。可以自动操作空间对称性并对分子提供点群对称性和平移对称性。

CRYSTAL 具有如下功能：

① 哈密顿量计算，求解基于数值求积分方案的数值网格并进行基于数值方案的密度拟合；

② 能量导数计算；

③ 提供单点能计算和自动几何优化等计算类型；

④ 支持多种基组：Gaussian 型基组、用户自定义赝势基组、Hay-Wadt 大核和小核基组等；

⑤ 独有的周期体系；

⑥ 多种波函特性与分析：能带结构、态密度、全电子电荷密度、电场梯度、结构因子、静电势及其导数等；

⑦ CRYSTAL 已经实现了代码的并行化，以 MPI 实现，能够在分布式内存系统上进行并行计算。

14. ADF

ADF 是一个通用密度泛函程序包，基于密度泛函理论 (DFT)，主要应用于量子化学计算。广泛应用于医药化学、材料学等研究及应用领域，尤其是同类和异类催化、无机化学、重元素化学、生物化学及多种光谱学。

ADF 的优势主要体现在计算速度和对过渡金属的处理上面。主要具有以下功能：几何优化、过渡态、反应路径及红外频率的表征及计算；分子性能测试的分析与表征；模拟溶剂、蛋白质及其它物化环境；强大的图形用户界面；进行聚合物、异类催化剂及大量结晶的研究。

ADF 主要有如下特色：

① 精度方面：精确的、可调节的数值积分方案；应用于整个周期表的，最高到四 zeta 的 Slater 型轨道；无需赝势或 ECP 近似；对过渡金属化合物稳定的 SCF 收敛性；现代的交换相关能泛函和势能泛函。

② 独有的 slater 轨道：采用 Slater 型基组 (STO)；正确反映原子轨道的尖峰区和渐进行为；BAND 采用数值的轨道和 STO 轨道。

ADF 以单程序多数据 (SPMD) 的模式实现并行计算，在处理器中分配格点和原子对，然后对其执行相同的计算。ADF 的并行用消息传递机制实现，基于 MPI 和 PVM 开发了专用的并行库，极大地提高了通讯的效率。通常情况下，高达 90% 以上的 CPU 时间都被用来进行数值积分。

15. GROMACS

GROMACS (http://www.gromacs.org/download/index.php) 是由 Groningen 大学开发一个分子动力学模拟包，在基于 Fortran77 的 GROMOS 的基础上，加入了 C 语言和大量的优化代码。GROMACS 支持所有的当代分子动力学应用算法，其一个著名的用途是进行快速准确的蛋白质模拟。

GROMACS 具有如下特点:

① 性能远高于同类软件：将最内层循环的针对不同体系结构进行优化。

② 交互界面友好：提供清晰的拓扑和参数文件；进行及时的进度提示；能用 gzip 读取压缩文件；提供简单实用的命令行控制方式。

③ 数据文件独立于硬件体系，无须区分大小端 (big/little endian)；版本间数据文件格式兼容。

④ 能够以并行方式运行，并行代码用 MPI 编写而成。

⑤ 精巧的算法保证对性能调节不会影响计算过程和结果的准确性。

⑥ 能够自动的构建蛋白质分子拓扑。

⑦ 正在向量子化学和生物信息学进行扩展。

GROMACS 代码中对部分算法进行了并行化。例如对 MD (Molecular Dynamics) 算法，对数据进行一维的分割，各计算点只和相邻点发生数据通讯，这样可以使并行代码编写简单化，而且还能将琐碎的通讯合成大数据量的通讯，由此减少通讯的额外开销。除 MD 算法外，GROMACS 还对 FFT 和排序算法进行了并行化。

16. GrADS

GrADS (http://www.iges.org/grads/) 是用于气象数据处理和显示的交互式工具，通过其提供的集成环境可以容易地对数据进行读取、加工、图形化显示和打印输出，同时具有分析功能强，坐标丰富，显示速度快等优点，因此已经成为国内外气象数据显示方面通用的标准图形环境之一。

GrADS 既可以处理站点数据，也可以处理格点数据。GrADS 对数据的处理使用四维数据环境，包括了三维空间和时间。在 GrADS 中，原始数据和元数据分别存储在不同的文件中。原始数据文件保持的是纯数据，没有空间和时间标识。元数据文件称为数据描述文件，包含了对原始数据文件的描述以及如何获取等信息。所有数据集通过数据描述文件被放置到四维空间中。GrADS 可以识别二进制的数据格式，也可以采用 GRIB，NetCDF 或 HDF-SDS 等通用的数据交换格式。

用户可以在交换环境下，通过在命令行输入类似 Fortran 语句的表达式进行操作。可以将键入的命令记录到一个文本文件中，通过 exec 命令批处理执行。同时 GrADS 提供了一种脚本语言作为编程界面，解释器就是 GrADS 本身，由 run 命令控制执行。该语言提供了变量、流控制、输入输出等高级语言功能，用户可以容易地以编程方式处理复杂的分析和图形显示。除了能够批量方式运行 GrADS 处理冗长的数据外，使用脚本还可以让用户显示按钮或下拉菜单，并根据用户选择执行相应的动作。

GrADS 提供了丰富的内建函数，用户也可以借助任何编程语言添加自己的函数。GrADS 支持多种图形化显示方式，并可以 postscript 或其他图形格式输出。

GrADS 目前版本为 1.8s11 和 1.9b4，支持包括 Linux，Mac，IRIX，AIX，Windows 在内的大多数主流操作系统。

17. ABINIT

ABINIT (http://www.abinit.org/) 是由 Xavier Gonze 所领导、世界上众多电子结构计算研究人员集体开发的平面赝势法电子结构计算程序包。

程序包基于密度泛函理论 (Density Functional Theory, DFT)，采用赝势和平面波基矢的方法来处理由电子和核所组成的体系，它可以计算体系的总能、电荷密度以及电子结构。

ABINIT 也可以按 H-F 力和压力来优化体系的几何结构，或根据这些力进行分子动力学模拟或计算得到动力学矩阵、Born 有效电荷及介电张量。在密度泛函理论的框架下可以计算分子体系的激发态，或基于多体微扰论 (GW 近似) 来处理激发态。除了主要计算模块外，程序包也提供了一些不同的工具模块用来处理计算结果和数据。

ABINIT 具有如下计算功能:

① 通常的电子结构计算 (晶格参数、结合能、体弹性模量、能带结构、电子态密度和电荷密度分布)；

② 分子动力学模拟；

③ 晶体磁性和磁结构的计算；

④ 基于线性响应或密度泛函微扰理论计算声子色散曲线、声子态密度和热力学量，介电性质、Bohr 有效电荷、压电系数和弹性常数、电 - 声耦合常数；

⑤ 计算线性和非线性的光学性质；

⑥ 采用 GW 近似对半导体的 LDA 带隙进行修正；

⑦ 基于含时密度泛含理论计算体系的激发态的性质。

18. GMT

GMT (http://www.soest.hawaii.edu/gmt/) 是大约 60 多个 UNIX/LINUX 工具的集合。它使得用户可以处理二维及三维的数据集 (包括 filtering, trend fitting, gridding, 投影等) 。GMT 支持 30 多种常用的投影及线性、对数和指数拉伸，并且支持诸如海岸线、河流及行政界线等。

GMT 系统最早由哥伦比亚大学的 Wessel 和 Smith 开发，全球大约有 6000 多位的使用者。早期是在 UNIX 系统上开发，可以和其它许多 Unix 的工具一起使用。GMT 主要是在指令模式下执行。GMT 是免费软件，数据输出的主要格式是 PS(postscript file)，因此几乎在各种平台上都可支持。

19. BLAST

BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) 是基于局部序列排比的常用数据库搜索工具，建立在严格的统计学的基础之上的，它集中于发现具有较高的相似性的局部比对。

BLAST 具有如下特点:

① 使用方便、功能齐全;

② 速度快、结果可信度高;

③ 配套数据库不断更新;

④ 提供 web 上的免费服务;

⑤ BLAST 算法本身具有良好的并行性，其并行版基于 MPI。

20. CASTEP

CASTEP (http://www.tcm.phy.cam.ac.uk/castep/index.html) 是由剑桥大学凝聚态理论研究组开发的一套先进的量子力学程序，可以进行化学和材料科学方面的研究。CASTEP 一个基于密度泛函方法的从头算量子力学程序，可以模拟固体、界面和表面的性质，适用于多种材料体系，包括陶瓷、半导体和金属等。

基于总能量赝势方法，CASTEP 根据系统中原子的类型和数目来进行各种性质的预测。CASTEP 使用的平面波赝势技术已经过实践的考验，每年发表的数百篇科学文献充分说明了其在许多领域中的成功应用。

CASTEP 主要具有如下功能:

① 基于总能量赝势方法，预测出包括晶格常数、几何密度、弹性常数、能带、态密度、电荷密度、波函数以及光学性质在内的各种性质；

② 电子的交换和相关效应采用局域密度近似 (LDA) 和广义密度近似 (GGA) ;

③ 电子波函数用平面波基组扩展 (基组数由 Ecut-off 确定) ;

④ 分子轨道波函数采用原子轨道的线性组合 (LCAO) 构成。

21. CPMD

CPMD (http://www.cpmd.org/) 是 以 CPMD (Car-Parrinello Molecular Dynamics) 方法为基本思想所编写，是 1993 年 IBM 瑞士研究所的一个计算物理研究计划的产物。

CPMD 是一种量子化学计算方法，将分子动力学和密度泛函有机结合，最大限度的发挥了密度泛函的优势。CPMD 是从头计算法的延伸，在精确模拟和简化模型上有独特的优势。CPMD 主要应用于固体和液体，静力学和动力学，固态物理和固态化学。

CPMD 主要有如下特点和功能:

① 孤立体系和周期边界体系的计算;

② 对分子和晶体使用对称性;

③ 波函优化;

④ 几何优化与过渡态寻找;

⑤ 恒定能量，恒温，恒压的分子动力学;

⑥ 路径积分分子动力学;

⑦ 线性响应函数计算，NMR，Raman 和 IR;

⑧ TDDFT 计算激发态;

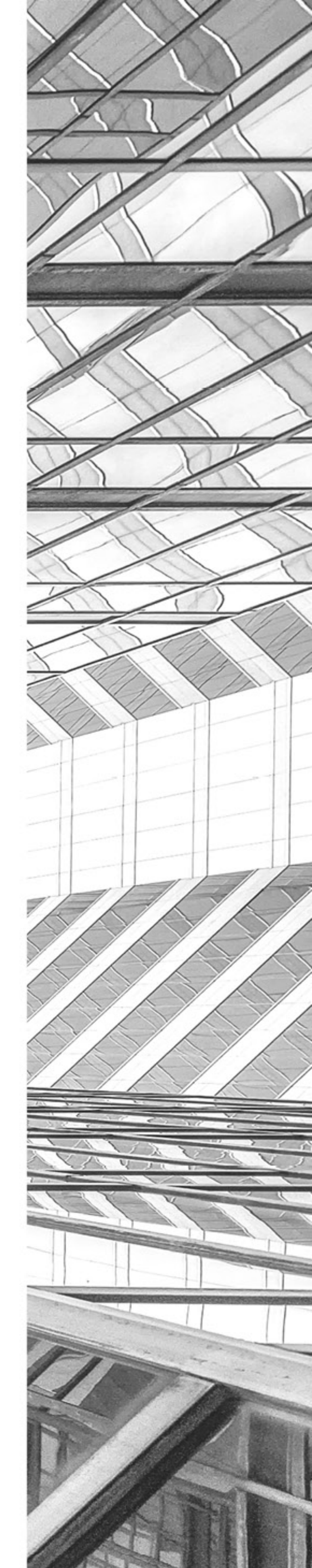
⑨ 多电子特性;

⑩ 周期体系的偶极矩;

⑪ 激发态的分子动力学;

⑫ QM/MM 方法;

⑬ OpenMP 并行化。



22. DL_POLY

DL_POLY (http://www.cse.scitech.ac.uk/ccg/software/DL_POLY/) 是串行和并行分子动力学模拟软件包，由 Daresbury 实验室的 W.Smith 和 T.R.Forester 开发。

DL_POLY 可以串行和并行两种方式运行。对于并行方式，既能以 MPP 模式并行，也能在分布式内存系统上运行。

DL_POLY 可以有两种并行的方式：

① 复制数据的方法，适合在数十量级的处理器上模拟十万以下个数的原子；

② 区域分解的方法，能够在 8 至 1024 个处理器上，模拟百万量级个数的原子。

对于一个 DL_POLY 许可，同时提供两个版本。DL_POLY 还提供基于 JAVA 语言的图形用户界面。

23. DOCK

DOCK (http://dock.compbio.ucsf.edu/) 是 Kuntz 研究小组发展的分子对接程序，可能是目前应用最为广泛的分子对接程序之一。它能自动地模拟配体分子在受体活性位点的作用情况，并把理论预测最佳的方式记录下来。而且该方法能够对配体的三维结构数据库进行自动搜索，因此被广泛应用于基于受体结构的数据库搜索的药物设计中。

用 DOCK 进行药物设计以及数据库的搜索基本上可以分为下面几个步骤：

配体和受体相互作用位点的确定，评分系统的生成，DOCK 计算及 DOCK 结果的处理与分析。活性位点的确定和表达是 DOCK 最重要的特点之一。活性位点特征的确定对于 DOCK 研究是非常重要的，因为配体分子和受体相互作用过程的模拟主要就是参考几何位点的几何特征进行的。在 DOCK 中，活性位点的确定通过 sphgen 程序来完成。DOCK 软件包中 sphgen 程序生成受体表面所有的凹陷的负像，并对这些负像进行聚类分析。在 DOCK 程序中，表面点采用了 Richards 提出的模型。在这些表面点的基础上，采用 sphgen 程序生成了负像，它实际上由一些与分子表面点相切的圆球叠加而成。

在生成负像的基础上，就可以进行配体分子和活性口袋之间的匹配。在这里，配体也采用一组球集来表示，和负像不同的是，配体所用的球集表示配体所占的空间区域。如果配体分子能和活性口袋形成比较好的匹配，那么配体的球集一定能和活性口袋中的负像形成好的叠合。配体分子和负像之间的匹配原则是基于配体和受体之间球集的内坐标的比较。

按照匹配原则得到了配体和受体之间的匹配情况之后。就要通过合理的得分函数来选择最优的结果。DOCK 提供了多种得分函数来评价配体和受体之间的结合情况，包括原子接触得分以及能量得分。

DOCK 进行分子对接时，配体分子可以是柔性的。对于柔性的分子，其键长和键角保持不变，但可旋转二面角是可以发生变化的。在 DOCK 中，柔性分子的构象变化通过下面的操作实现：首先是刚性片断的确定，然后是构象搜索。构象搜索采用两种方法：一种是锚优先搜索 (anchor-first search)，第二种方法是同时搜索 (simultaneous search)。

24. GRAPES

GRAPES 全球区域同化预报系统是中国气象局负责研究开发的、研究与业务通用的数值天气预报系统，属于“十五”国家科技攻关计划。

GRAPES 是集常规与非常规变分同化、静力平衡与非静力平衡、全球与区域模式、科研与业务应用、串行与并行计算、标准化与模块化程序、理想实验与实际预报等为一体，中小尺度与大尺度通用的先进数值预报系统。

GRAPES 作为我国研制的天气预报模型，具有下列创新点：

① 可以把空中和地面的资料全部用到数值预报中，改变了国家花费几十亿元发展的气象卫星和多普勒天气雷达闲置的情况。

② 性能优于已有的国外模式，对于暴雨、强对流天气等可以有更好的预报结果。

③ 较高的预报准确率，在可用预报时效和预报精细程度方面有很好的表现。

GRAPES 应用非常广泛，目前国内很多气象部门已经在对 GRAPES 进行研究和应用。

25. MM5

MM5 (http://www.mmm.ucar.edu/mm5/) 是美国宾州州立大学和美国国家大气研究中心 (NCAR) 在 20 世纪 90 年代初研制发展的新一代中尺度天气模型。中尺度强对流天气系统尺度小，具有突发性和多发性特点，且天气变化激烈，造成的灾害范围广。MM5 的出现对于区域中尺度数值天气预报的研究有着重大的意义。

MM5 是建立在多重嵌套网格上的格点差分模式。MM5 模式是一种双向嵌套有限差分预报模式，其计算域按照模式预报的不同要求和分辨率分成母域和嵌套域，嵌套域层数可以达到 9 层。MM5 模式计算涉及到粗网格 (母域) 和细网格 (嵌套域) 两个计算区域，数据关系复杂。MM5 有限区域差分模式的所有物理量的计算都是在网格空间上进行的。每一个物理量的模式积分在每个水平经纬网格点上是相互耦合的，但在垂直层上无数据相关性。

大气运动的物理特征非常适合于并行处理，如大气中的水平和垂直运动、各种非绝热物理过程等，在一定程度上都可以看作相互独立的。

MM5 基于消息传递机制实现并行计算，低层依赖 RSL（RuntimeSystemLibrary）库实现。RSL 库是基于消息传递标准 MPI 之上的并行计算应用程序接口，主要针对有限差分格式网格计算。

MM5 采用水平二维数据划分实现分布式并行计算。算法的基本思想是：（1）将处理机划分为 Nx × Ny 的逻辑网格单元，分别对母域和嵌套域的计算网格点进行经向（x 方向）和纬向（y 方向）二维剖分，将各个网格层上的计算划分为 Nx × Ny 个逻辑计算单元，分布到各处理机上进行；（2）采用 RSL 库特殊的数据结构，计算出每个逻辑计算单元之间存在数据相关性的网格点，实现网格层内部及粗 / 细网格层之间的通讯。

MM5 的数据通讯包括域内通讯和域间通讯两种。域内通讯是指计算域（母域或嵌套域）内部由于数据划分，使计算分布在各个处理结点上进行，相邻计算子域之间需要调用 RSL 库显式地通过消息传递来进行数据的通讯。域间通讯是母域和嵌套域之间由于“动力”和“反馈”的相互作用而进行的数据交换。

MM5 采用一种基于不规则区域的数据划分策略，使用该方法能够基本保证负载平衡。理想情况下，采用不规则区域数据划分方法，基本能将网格点均匀的分配到各个处理器上。

26. NAMD

NAMD（http://www.ks.uiuc.edu/Research/namd/）是一款并行的分子动力学计算软件，曾获得 2002 年度的金铃奖（Gordon Bell Award），用于在大规模并行计算机上快速模拟大分子体系。NAMD 用经验力场，如 Amber，CHARMM 和 Dreiding，通过数值求解运动方程计算原子轨迹。用于预测生物分子的动力学行为和重要性质，如弥散因子和内聚能等等。

NAMD 具有如下基本功能：

① 以两种任务类型提交运算：几何优化和分子动力学；

② 有效共轭梯度最小化；

③ 周期和非周期分子动力学模拟；

④ 速率重新标度、球边界条件和谐振原子限制；

⑤ 恒温和恒压的分子动力学模拟、热平衡分子动力学模拟；

⑥ 多时间步计算；

⑦ 化学和构象自由能计算；

⑧ 基于 Tcl 的脚本；

⑨ 两种计算特性：均方位移和内能分解。

NAMD 可以以串行和并行两种方式运行。NAMD 的并行机制基于 Charm++ 对象，可以在高端的并行机上进行上千个处理器以上规模的并行计算。对于使用普通以太网的分布式内存系统，可以进行数百个处理器规模的运算。

27. WRF

WRF（http://www.wrf-model.org/index.php）模式系统是由许多美国研究部门及大学的科学家共同参与进行开发研究的新一代中尺度预报模式和同化系统。

WRF 模式是一个完全可压非静力模式，控制方程组都写为通量形式。网格形式采用 Arakawa C 格点，有利于在高分辨率模拟中提高准确性。模式的动力框架有三个不同的方案。前两个方案都采用时间分裂显示方案来解动力方程组，即模式中垂直高频波的求解采用隐式方案，其他的波动则采用显示方案。第三种模式框架方案是采用半隐式半拉格朗日方案来求解动力方程组。这种方案的优点是能采用比前两种模式框架方案更大的时间步长。

WRF 的动力框架具有一定的优越性，在前处理和所选物理过程相同的情况下，WRF 对大部分中尺度天气系统的高度场、风场、散度场、水汽通量场以及垂直速度场等物理量的模拟效果要好于 MM5。

WRF 模式系统具有可移植、易维护、可扩充、高效率、方便等诸多特性，使新的科研成果运用于业务预报模式更为便捷。WRF 有着广泛的应用前景，包括在天气预报、大气化学、区域气候、纯粹的模拟研究等方面的应用，它将有助于开展针对不同类型、不同地域天气过程的高分辨率数值模拟。

28. AUTODOCK

AUTODOCK（http://autodock.scripps.edu/）是由 David S. Goodsell 博士在 Olson 实验室撰写完成。之后的版本又加入了新的科学理解策略，计算过程更稳健和快速。AUTODOCK 采用模拟退火和遗传算法来寻找受体和配体最佳的结合位置，用半经验的自由能计算方法来评价受体和配体之间的匹配情况。

在 AUTODOCK 中，配体和受体之间结合能力采用能量匹配来评价。最初版本中，能量匹配得分采用简单的基于 AMBER 力场的非键相互作用能。非键相互作用来自于三部分的贡献：范得华相互作用、氢键相互作用以及静电相互作用。在 3.0 版中，AUTODOCK 提供了半经验的自由能计算方法来评价配体和受体之间的能量匹配。

在 3.0 版本中，AUTODOCK 采用了一种改良的遗传算法，即拉马克遗传算法 (LGA)。测试结果表明，LGA 比传统的遗传算法比模拟退火具有更高的效率。LGA 方法把遗传算法和局部搜索 (local search) 结合在一起，遗传算法用于全局搜索，而局部搜索用于能量优化。在 AUTODOCK 中，局部搜索方法是自适应的，它可以根据当前的能量调节步长大小。LGA 算法引入了拉马克的遗传理论，LGA 最大的特点就是通过进化映射 (developmental mapping) 把基因型转化为表现型而实现局部搜索和遗传算法的结合。

29. VMD

VMD (http://www.ks.uiuc.edu/Research/vmd/) 是分子可视化程序，是一种观察和分析分子动力学模拟(molecular dynamics simulations) 的程序，由伊利诺斯大学研究人员研制并免费向广大用户开放。

VMD 用 OpenGL 提供高质量的 3D 分子图形，用于显示、动画演示和分析大的生物分子体系；能够控制大型的包括 1 百多万个原子的三维体系。此外，多重芯片(Multiple Alignment tool)也是 VMD 一个重要组成部分。

VMD 主要有如下特性：

① 多种分子显示方式和着色方法；

② 立体显示功能；

③ 包含了 Babel 程序，用于格式转化；

④ Tcl 语言编写的图形和字符用户界面，用户可以自己编写自程序用于分子的分析；

⑤ 用户可以用 C++ 编写模块；

⑥ 整合了 NAMD 程序，一起快速稳定地进行分子动力学并行计算，并显示计算结果；

⑦ 录制轨迹。

30. TINKER

TINKER (http://dasher.wustl.edu/tinker/) 是一些程序的集合，用于进行分子力学和分子动力学计算，以及一些用于生物聚合物计算的特殊功能。

TINKER 具有如下功能及特色：

① 可以支持的力场：AMBER-94/96、CHARMM27、MM2(1991)、MM3(2000)、OPLS-AA 和 OPLS-UA。

② 包含多种算法：新型距离 - 结构度量化方法；Elber 的反应路径方法；用于全局优化的几种势能曲面平滑和寻找 (PSS) 方法；高效的势能曲面扫描程序；高效简化的 Newton (TNCG) 定位优化程序；简单的自由能微扰工具；速率 Verlet 随机动力学；改善的球面能量截断方法；最新的长程静电反应场处理方法；模拟退火；分子动力学的 Andersen 随机碰撞恒温等。

③ 经过改善的晶体最小化算法。

④ 可以结合 RasMol、ChemDraw、Chem3D、gOpenMol、MOLDEN、GAMESS-US 和 ReView 等一起使用。

⑤ PDB 文件可以转化为 TINKER 文件，使得分子结构的输入有丰富的来源。

⑥ TINKER 可以串行和并行方式运行。

31. OpenMX

OpenMX (http://www.openmx-square.org/index.html) 是材料模拟程序包，用于实现基于密度泛函理论的大标度从头计算。在密度泛函计算中有三个部分是相当耗时的：求解哈密顿矩阵元素，求解泊松方程，以及对角化广义久期方程。而在 OpenMX 中，根据计算量和内存，几乎可以用 O(N) 标度完成每一步。OpenMX 是涵盖生物材料和复合材料的纳米尺度材料科学中具有强大效用的工具。

OpenMX 具有如下功能：

① 用 Fourier 变换求解双中心积分；

② 用固定实空间网格求解三中心积分；

③ 用 FFT (快速傅立叶变换) 求解泊松方程；

④ 交换 - 相关项使用局域密度近似 (LDA, LSDA) 和广义梯度近似 (GGA)；

⑤ O(N) 标度“分而治之”方法和广义的分而治之方法，求解本征值问题；

⑥ 计算使用赝原子轨道和赝势 (ADPACK 原子密度泛函程序计算产生)。

此外对于大标度的计算，OpenMX 用 MPI 方式进行并行计算，每个节点都是动态分配内存，实现了在分布式内存系统上的资源均衡。

32. VENUS

VENUS （http://homepage.mac.com/fujioizumi/visualization/VENUS.html) 程序主要用于进行晶体结构和电子 / 核密度 (如来自 WIEN2k，GAMESS，VASP (或其图形界面 MedeA)，ABINIT，Gaussian 和 SCAT 的电子结构计算) 的三维显示和操作。VENUS 还能显示 GAMESS，ABINIT，Gaussian 和 SCAT 计算的静电势和波函。

此外，VENUS 还包含两个用最大熵方法 (MEM) 的程序，用于三维显示电子密度，核密度和 Patterson 函数。

VENUS 由四个独立的程序组成：VEND 显示电子密度和核密度；VICS 显示晶体结构；PRIMA 进行迭代 MEM 分析；ALBA 进行此后的 Le Bail 分析。

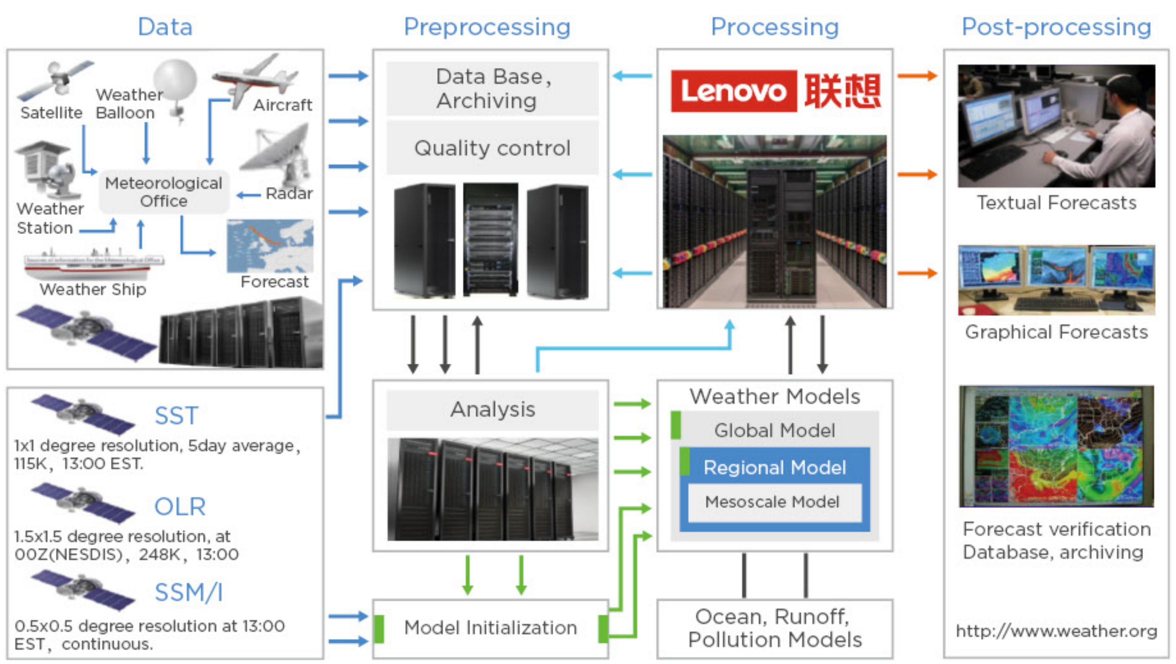
VICS 可以输入 / 输出多种格式的结构数据文件和三维网格数据 (电子密度，静电势等) 文件。文件格式转化工具包括：Alchemy、contrd、wien2venus.py、Cut3D 和 ELEN。

支持的图片输出格式有：BMP、EPS (像素数据)、EPS (矢量图数据)、JPEG、JPEG 2000、PPM、RAW、RGB (SGI)、TGA 和 TIFF。

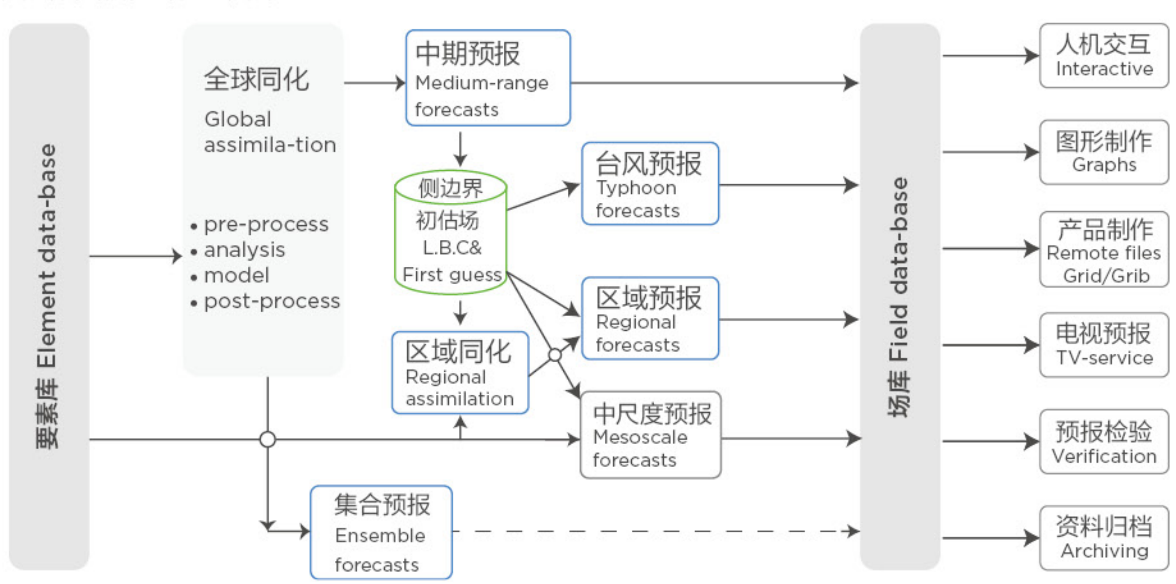
联想气象和海洋 HPC 解决方案

数值预报业务流程

Weather/Climate－工作流程



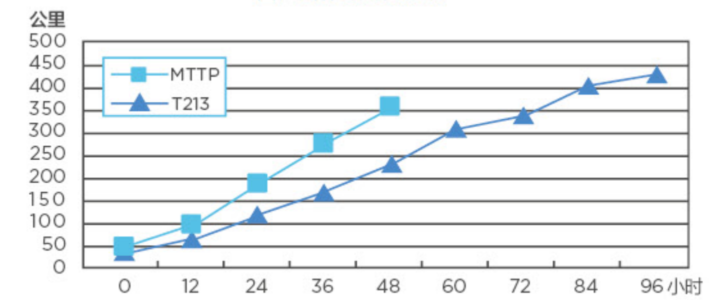
典型数值预报业务基本流程



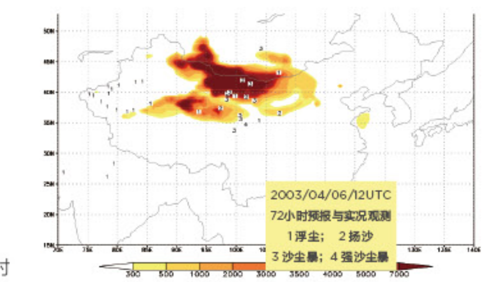
数值预报业务的热点问题

- 资料同化
- 精细化数值预报
- 集合预报
- 多模式预报
- 滚动预报
- 加密预报
- 灾情应对
- 智能数据分析

台风路径预报系统



沙尘暴数值预报系统



数值预报业务解决方案面临的挑战

数值预报对存储和计算资源的依赖	• 精细化预报、集合预报等等，需要更高效率的计算和处理
数值预报业务流程复杂	• 接收数据，数据同化，模式计算，图形可视化，决策分析 • 目前很多用户采用脚本方式组织本单位的业务流程
行业专家如何从复杂的IT系统中解放出来	• 服务器、存储、网络、软件越来越复杂 • 需要简洁、易用、智能的系统，用户更专注自身业务
现代高性能计算系统带来的挑战	• 供电、散热、系统日常维护 • 准关键业务和实时性特点，对系统可靠性要求较高 • 系统服务器节点数量增多，带来维护和难度几何倍数增长

典型模式计算量评估

主模式:

二重嵌套WRF模式，格点数分别为419×387和316×307（经向×纬向），模式垂直层51层，第一重区域积分步长45秒，预报时效36小时，每1小时输出一次预报结果；第二重区域积分步长15秒，预报时效36小时，每1小时输出一次预报结果。

前处理:

包含geogrid，ungrib，metgrid和real四部分，在一个节点内完成测试。

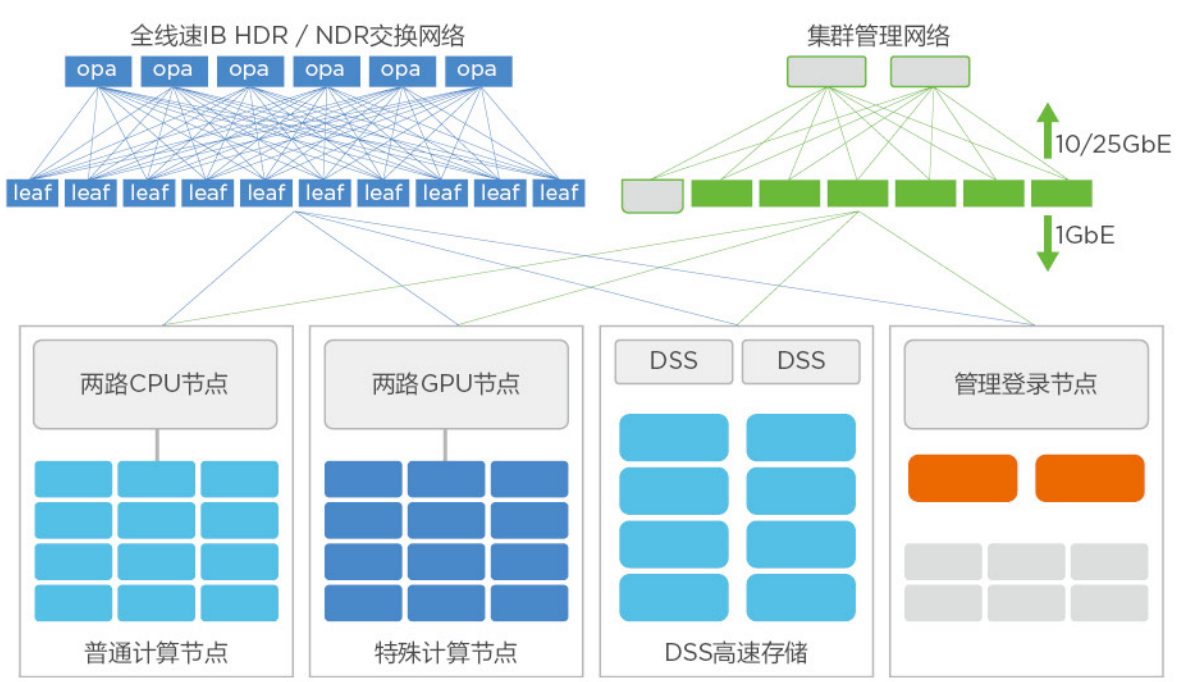
后处理:

包含RIP4、ARWpost_V3和UPPV2.1三部分，可以并发执行。其中RIP4处理d01和d02数据，ARWpost和UPP只处理d02数据（37个预报时次）。

作业数量	每作业CPU核数	运行时间（s）				
		序号	前处理	主模式	后处理	前后累计
1	896	1	120	1304	870	2294
2	448	1	119	1663	874	2656
	448	2	120	1663	872	
4	224	1	133	2369	890	3398
	224	2	129	2383	886	
	224	3	127	2363	885	
	224	4	128	2371	882	
8	112	1	155	3577	920	4691
	112	2	158	3614	919	
	112	3	158	3608	921	
	112	4	155	3592	927	
	112	5	158	3583	930	
	112	6	159	3605	921	
	112	7	154	3593	927	
	112	8	158	3563	920	

系统参考架构

典型数值预报HPC系统架构



数值预报系统的 HPC 方案设计建议

基于最新技术，我们对数值预报系统的设计建议

01 使用每处理器2-3个内存槽的计算节点，保证内存带宽

02 使用相互独立和自身冗余的计算网络和管理网络

03 使用IB直连存储、四路节点、GPU加速器等技术，简化系统架构

04 需要高速文件传输时，使用万兆网络直连节点，而不是网络交换机

05 关键功能使用商用HPC系统软件

06 使用温水水冷等更加节能和环保绿色的部件

联想石油石化 HPC 解决方案

石油行业 HPC 需求

在石油勘探中，像地震资料处理这类计算类数据密集型应用，传统上只能由超级计算机来满足。随着高性能计算技术的不断成熟，高性能计算 (High Performance Computing Cluster) 解决方案成为一个更理想的选择，原因在于它为获得所需的处理性能提供了一种更经济有效的方式。

高性能计算机在石油工业中的应用非常广泛，主要表现在科学计算、数据处理、油藏模拟、勘探开发数据库建立等方面。

作为高性能运算应用的主要领域—石油地震数据处理系统是一个要求计算机系统运算性能、运行稳定性极高的计算环境，其原因在于石油行业中的数据处理非常困难和庞杂。以最为直接的“找油”工作来说，在油田区找油打井，一般采用三维地震勘探的办法：先在地面放炮，通过仪器测量地面震动反射波的情况，把采集所获得的地震波在地层中传播的运动学和动力学信息进行去伪存真处理，转化为地质信息，经过计算分析，反推出地下的生油、储油、盖层的地质构造特点，再确定在哪里打井。这种企业计算环境随着石油勘探新技术的不断应用、信息技术的不断向生产中渗透，迫使企业不断采用性能优异、技术先进、价格便宜的高性能计算机处理系统来占领市场。

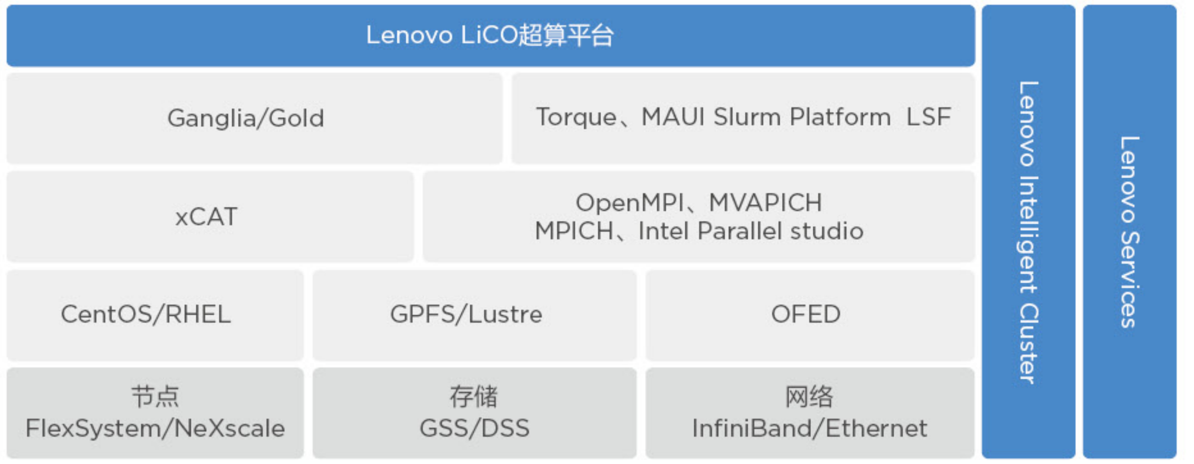
业界根据目前高性能计算环境性能和价格方面的要求，地震数据处理应用系统呈现出如下特点：

① 广泛使用计算机集群系统 (PC CLUSTER) 来提升处理中心的计算能力，满足各种运算量极大的特殊地震数据处理算法需求的同时，极大地降低企业的成本。这样企业的传统采用大规模计算机处理系统进行特殊运算的需求已经转变成采用性能高、价格便宜的计算机集群系统来实现。

② 随着采集技术的不断改进、地震数据体的数据量也在剧增，迫使企业的计算环境以服务器为中心转变成了以存储为中心。地震数据处理系统也是一个需求存储空间极大的应用，一般二维数据可达 1-2TB，三维数据可高达几百 TB 甚至几十 PB 级，近年来数据可以达到上百 PB。地震资料处理的作业的运行速度与计算机系统的 I/O，也就是使用的并行存储系统有着很大的关系。

③ 由于大量计算能力需求，各种新的高性能计算技术和方法能够在地震处理系统中率先使用推广，比如采用最新的 CPU、GPU 产品。GPU 的使用，大大加速了地震资料数据的快速产生，逆时偏等在传统的 CPU 集群上非常难以实现的目前已经在 GPU 集群得以实现并规模应用。

联想石油行业 HPC 解决方案



联想 HPC 方案组成

1. 节点系统

计算节点采用联想刀片服务器，提供强大的计算能力，以完成地震资料处理中巨大的计算任务。

2. 网络系统

管理网采用联想千兆以太网，用于 HPC 集群系统的管理和监控。

监控网采用联想千兆以太网，用于各节点和设备硬件管理和监控，如 IPMI 管理等。

计算网络采用线速万兆以太网或有阻塞的 InfiniBand 网络，承担并行计算时的数据通讯。

3. 存储系统

存储系统采用联想 DSS-D/DSS-G 系列存储。联想 DSS-D/DSS-G 系列存储是基于 GPFS 文件系统和 GPFS Native RAID 技术构建的一款存储产品。具有以下几个优点：

① 大容量 高 I/O 吞吐率：单集群高达 400GB/s 的 I/O 吞吐率 (真实实施案例)；高达 28PB 存储容量 (真实实施案例)；单存储单元 2.x PB 的可用磁盘容量 (具体视 RAID 方式而定)；单存储单元高于 20GB/s 的 I/O 吞吐率；

② 内置 IBM GPFS 技术：内置 IBM GPFS 分布式文件系统，集高性能和高可靠性于一身；加强的容错性能，可以实现 De-clustered RAID 6 的 8+2，8+3 和 2 路，3 路副本模式的容错方案，并可共存于一个阵列中；丰富而完善的异步磁盘与系统诊断功能；多路径设计，无单点失效；

③ 广泛的使用场景：20 年来被各行业及核心机构广泛验证的文件系统；为大数据计算提供整体存储解决方案；为云环境提供文件、块、对象存储统一存储解决方案；易用的 GUI，提供硬件到软件的监控管理平台。

4. 系统软件

系统采用联想 LiCO 高性能计算平台。LiCO 是联想研发的高性能计算系统平台软件。具有丰富的集群管理功能，包括：集群管理和监控功能；集群资源管理和作业调度功能；用户和账户管理功能；集群系统监控功能。

联想 CAE 仿真计算 HPC 解决方案

CAE 仿真 HPC 解决方案广泛适用于航空，航天，汽车，船舶，核，电等领域。

CAE 仿真应用分析

CAE 高性能运算的应用可以分为隐式有限元分析（IFEA）、显式有限元分析（EFEA）和计算流体动力学（CFD）三个子学科。采用隐式算法的软件主要有 ABAQUS/Standard、ANSYS、MSC.NASTRAN 等，适合求解静力、模态、屈曲等问题；采用显式算法的软件主要有 ABAQUS/Explicit、LS-DYNA、PAM-CRASH 等，适合求解接触、碰撞、冲击等问题。

几乎所有的 CAE 高性能计算都依赖于独立软件开发商（ISV）提供的商业软件，只有流体动力学算题中结构网格计算类型的部分软件是用户自己开发的。因此制造行业用户在购买硬件平台的同时通常会购买相应的科学计算软件产品。而在某种程度上，往往是应用软件的特性决定了硬件平台的选择。

从对计算资源的需求来说，隐式解法的基本特点是内存占用多、磁盘 I/O 大、进程通信量大，因此，隐式解法要求系统的内存容量大、访存带宽高、磁盘 I/O 速度快、通信延迟低；相对而言，显式解法对内存、磁盘 I/O 和通信延迟的要求要低一些。

从软件的扩展性上来说，隐式算法和显式算法有明显的区别。采用隐式算法的软件，扩展性相对较差，计算性能在 4 到 8 个物理 CPU 以上就很难获得进一步的提升。而采用显式算法的软件，扩展性就要好得多，在 64 到 128 CPU 以内都能获得较好的并行性能，伴随设计和仿真的需求，目前联想已经在国内用户部署了单户超过 1000 台服务器的案例。

下表给出了常用的 CAE 软件，并列出这些软件的特点，包括并行方式和可扩展性。

CAE 应用软件分类	应用软件	并行方式	扩展性
静态隐式有限元分析（IFEA Statics）	ABAQUS	pthreads	
	ANSYS	OpenMP,MPI	
	MSC.Nastran	pthreads,MPI	
动态隐式有限元分析（IFEA Dynamics）	ABAQUS	pthreads	低
	ANSYS	OpenMP,MPI	
	MSC.Nastran	pthreads,MPI	
显式有限元分析（EFEA）	LS-DYNA	OpenMP,MPI	
	PAM-CRASH	OpenMP,MPI	
	RADIOSS	OpenMP,MPI	高
计算流体动力学（CFD）	FLUENT	MPI	
	STAR-CC	MPI	
	PowerFLOW	OpenMP,MPI	

常用分析软件

从上表中我们可以了解到 CAE 应用软件具有以下特点：

① IFEA 类应用软件（如 ABAQUS、ANSYS 和 MSC Nastran）硬件平台支持的可扩展性不是很好。Nastran 对内存，I/O 性能要求高；

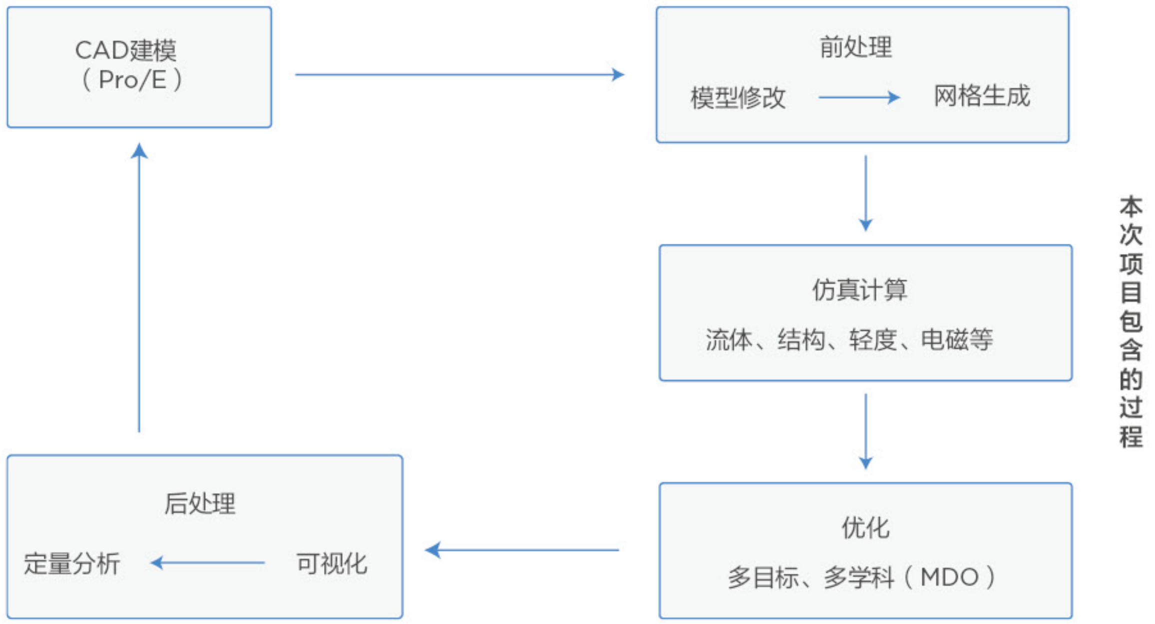
② IFEA 类应用软件通常使用共享内存方式（pthreads 或 OpenMP），进行并行处理，其中 ABAQUS 不支持消息传递方式（MPI）的并行；

③ EFEA 类应用软件（如 RADIOSS、LS-DYNA 和 PAM-CRASH）和计算流体动力学软件（如 FLUENT、STAR-CD 和 PowerFlow）的硬件平台支持的扩展性相对较好。RADIOSSS/LSDYNA 对 CPU，I/O 性能要求高；

④ EFEA 类应用软件和 CFD 软件以采用消息传递并行方式（MPI）为主。

CAE 仿真应用处理流程

在数字化设计过程中，主要包含几个过程：



前处理主要用于对工程或产品进行建模，建立合理的有限元分析模型（包括建立问题的几何模型、进行网格划分、建立用于计算分析的数值模型、确定模型的边界条件和初始条件等）。

仿真计算主要是对有限元模型进行单元特性分析、有限单元组装、有限元系统求解和有限元结果生成，该过程涉及对所建立的数值模型进行求解，经常需要求解大型的线性方程组，这个过程是 CAE 分析中计算量最大、对硬件性能要求最高的部分。而后处理部分则根据工程或产品模型与设计要求，对有限元分析结果进行用户所要求的加工、检查，并以图形方式提供给用户以便对所得的计算结果进行检查和处理，辅助用户判定计算结果与设计方案的合理性。

结合数字化设计流程，仿真高性能计算以计算服务器、机架式图形服务器为中心，将企业已有的设计、仿真（计算、前后处理）应用集中到高性能仿真中心上面。因此，涵盖数字化设计的整个过程。

联想 CAE 仿真高性能计算平台提供并通过浏览器，用一个统一的门户实现以下功能：

① CAD 设计和 CAE 前后处理等应用软件的使用和调度管理；

② 仿真类及优化类应用软件的使用和调度管理；

联想 CAE 仿真高性能计算平台涵盖设计、仿真、前后处理的统一 IT 综合应用服务平台，为设计师提供高效的计算、设计环境。

CAE 仿真 HPC 参考系统架构

① 联想 CAE 仿真高性能计算平台分为 5 个分系统：计算分系统、管理分系统、前后处理分系统、网络分系统、存储分系统；

② 计算分系统主要用于仿真计算和优化类应用软件的模型求解，设备包含 2 路计算服务器、4 路计算服务器和 8 路计算服务器也可以包含 GPU 加速节点；

③ 管理分系统主要是资源管理、作业调度、集群管理等管理平台，向设计师和管理员提供统一使用、管理平台，设备包含管理服务器、资源管理软件、集群管理软件，以及搭建并行环境所需的软件；

④ 前后处理分系统主要用于 CAE 前后处理及 CAD 模型设计类应用软件的交互式设计操作，设备包含机架式图形服务器；

⑤ 网络分系统主要用于搭建该项目所有基础网络，设备主要包含各种网络交换机；

⑥ 存储分系统主要用于向计算分系统和前后处理分系统提供数据存储服务，设备主要包含存储设备、存储节点及并行文件系统。

联想仿真一体机解决方案 LICP

随着“中国制造 2025”国家计划的提出，中国制造业正在由“中国制造”走向“中国质造”。如何运用数字化技术缩短研发周期、降低研究成本、提高产品质量，具有十分重要的意义。具体到航空、航天、兵器、船舶、汽车、手机、机械等多个行业，其科研、生产广泛使用的仿真计算与设计平台建设存在以下问题：

- 传统“孤岛式”IT 系统
- 大量使用桌面工作站，利用率低，投资大
- 各种信息系统独自建设，用户数据多点存放，不同系统间需要频繁切换
- 仿真计算、三维设计、研发流程脱节，使用不流畅
- 专业软件工具分散部署，低效使用，版本不统一
- 不利于协同研发的开展，管理维护也非常困难

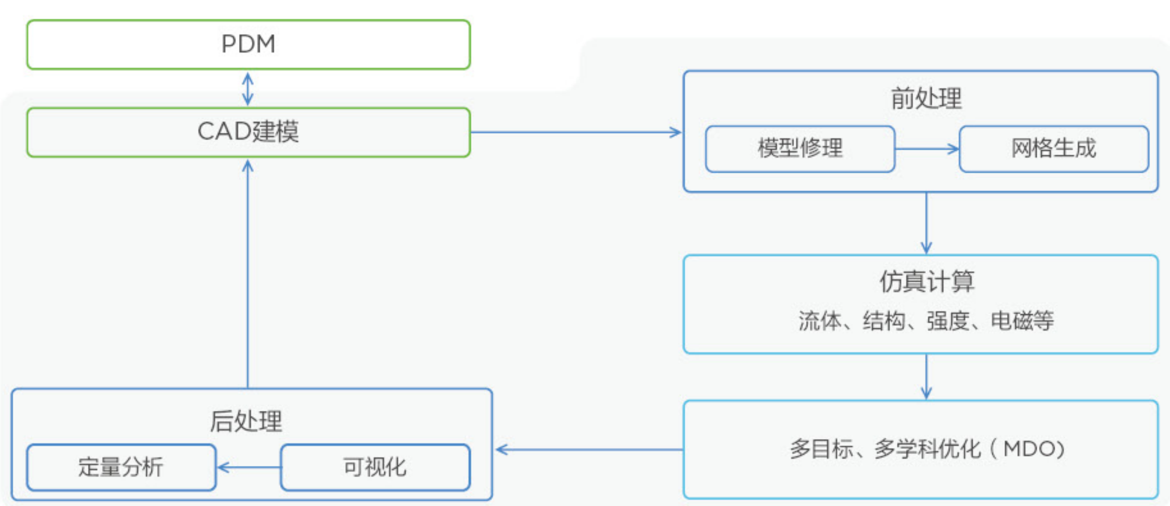
联想针对行业现状与痛点提出了“联想智算管理平台”（“Lenovo Intelligent Computing Platform”，简称 LICP）解决方案，为用户提供一站式业务平台，LICP 为实现高质量、高效率的仿真与设计工作提供以下服务：

- 简单方便的“一站式”应用平台，而非计算平台
- 直接从研发流程中调用资源进行仿真、设计、优化等工作
- 桌面轻量化，从工作站到瘦客户机
- IT 运维简化，桌面机零维护、零安装
- 研发工具软件标准化、云化、全面覆盖研发需求
- 复杂环境和网络条件下集中管理和使用
- 安全保密
- 图形服务器空闲时用于仿真计算，充分利用 GPU 卡的运算能力
- 仿真设计过程中的快速协同
- 更好的用户体验、尊重用户使用习惯
- 能适应向大数据和人工智能平台的扩展

联想总结多年制造业 HPC 系统经验，致力于解决用户使用图形工作站和高性能计算系统的痛点，提升用户协同研发能力和效率，提高投资回报。

方案内容

联想智算管理平台说明

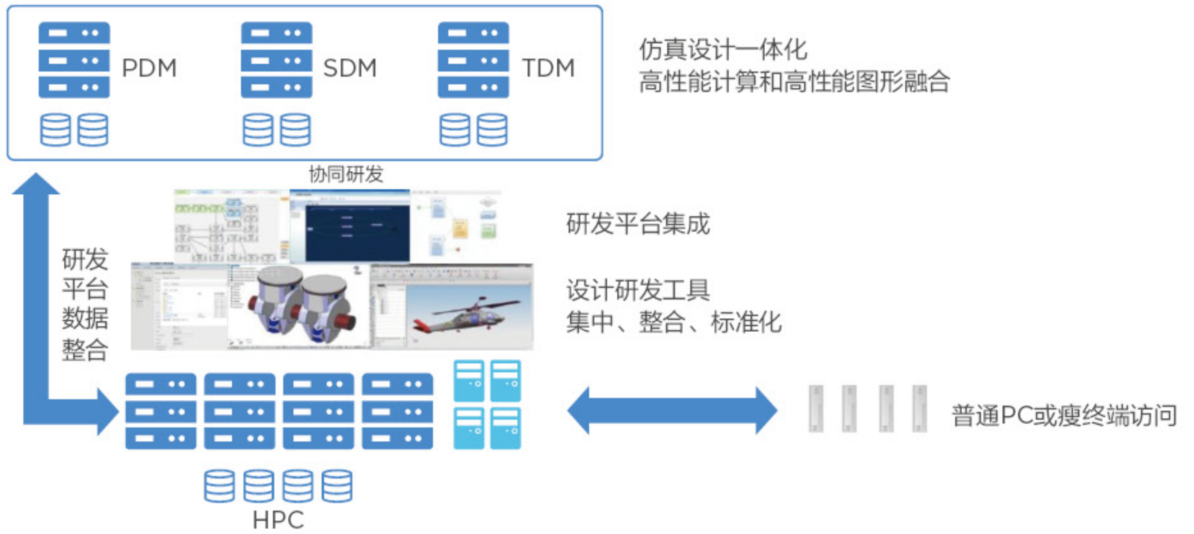


平台覆盖范围

仿真与设计的一体化主要体现在以下几个方面：

- HPC+HPG，高性能计算和高性能图形融合
- 仿真设计一体化
- 涵盖数字化设计全流程
- 数据无缝流转，单一数据源，避免频繁上传下载
- 应用集中，客户机零安装、零维护，开机即用

联想智算管理平台与研发体系融合：

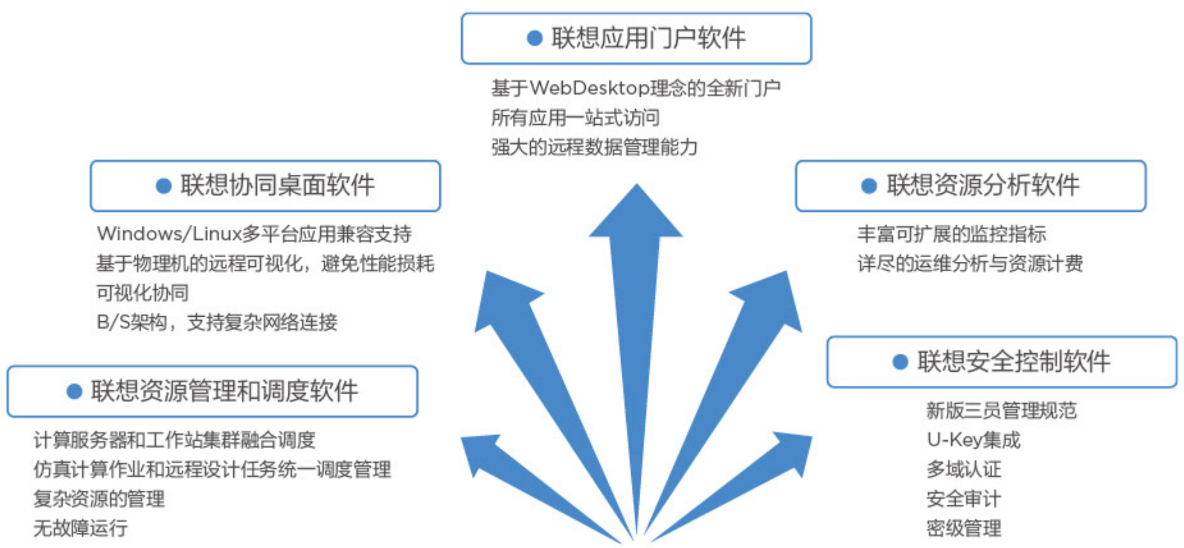


融合后的仿真设计一体化研发体系

联想智算管理平台功能架构：

LICP 功能架构

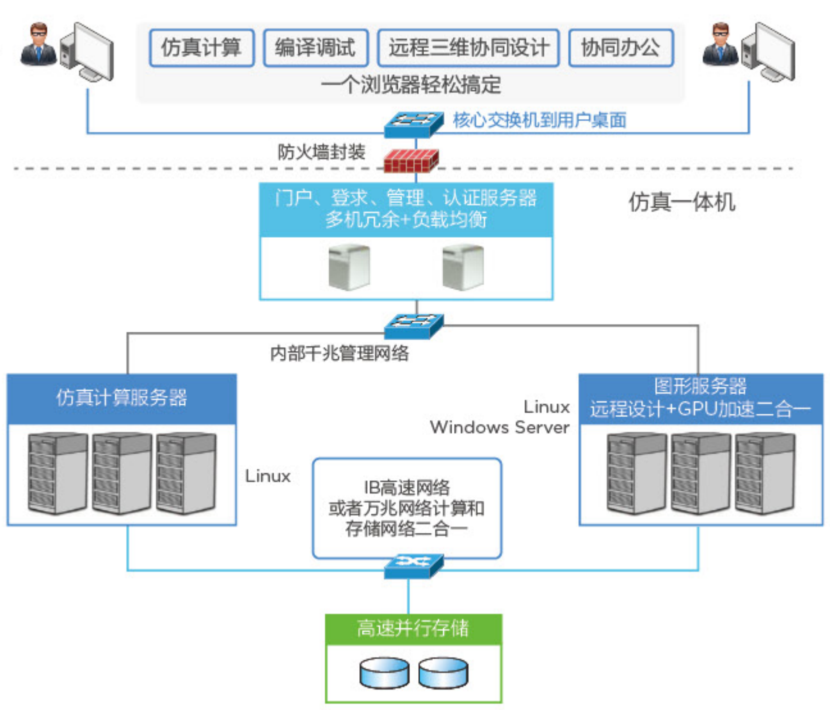
联想智算管理平台产品组合：



LICP 产品组合

联想智算管理平台典型部署架构：

设计仿真计算一体化平台



网络架构：

- 一个核心 — 以数据为核心
- 两种能力 — 提供远程协同设计能力和仿真计算能力
- 五个原则 — 安全性原则、易用性原则、先进性原则、经济性原则、可分析性原则

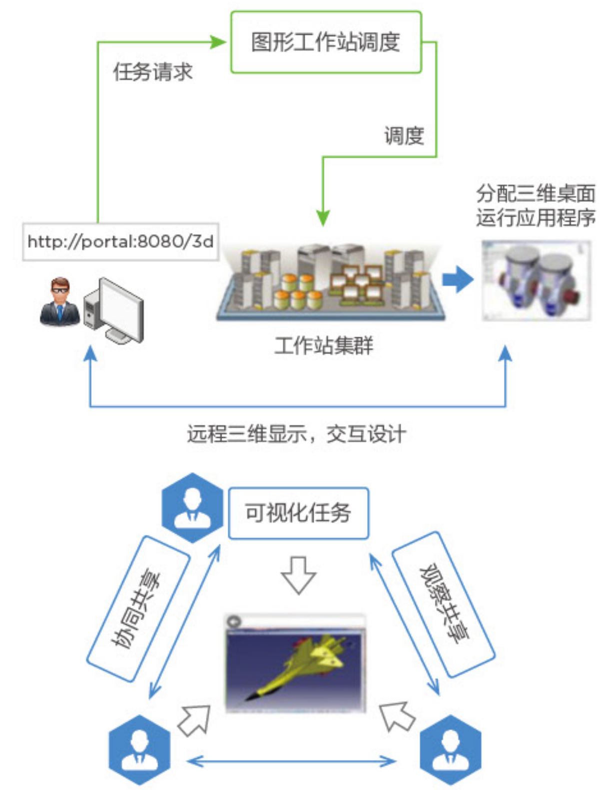
LICP 典型部署架构

体现以下特点：

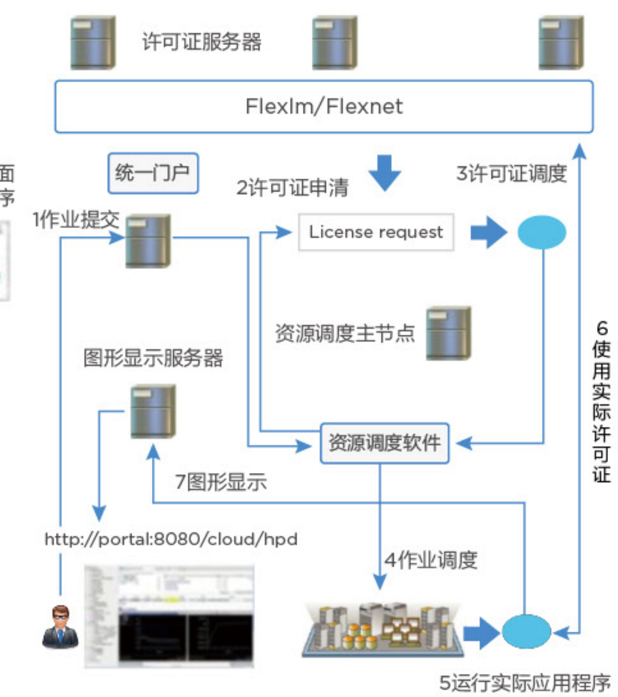
- 一个核心——以数据为核心
- 两种能力——提供远程协同设计能力和仿真计算能力
- 五个原则——安全性原则、易用性原则、先进性原则、经济性原则、可分析性原则

仿真设计一体化应用场景：

- 远程三维协同设计



- 仿真计算



仿真设计一体化应用场景

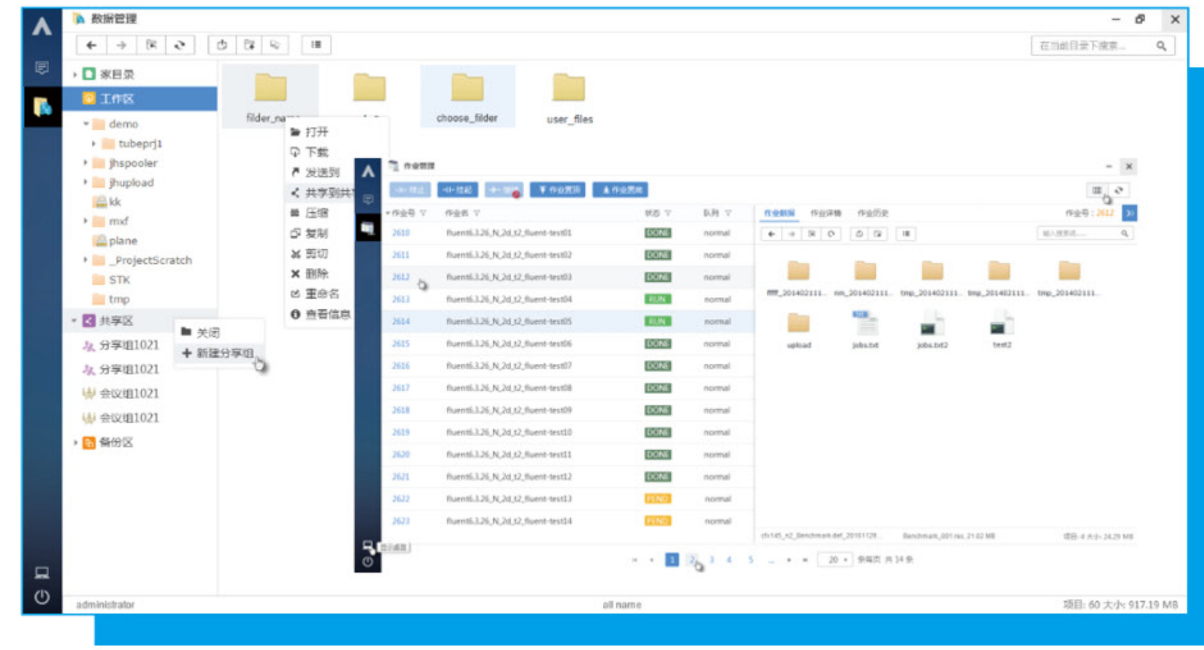
联想智算管理平台产品组成

联想智算管理平台的实现基于以下软件产品：

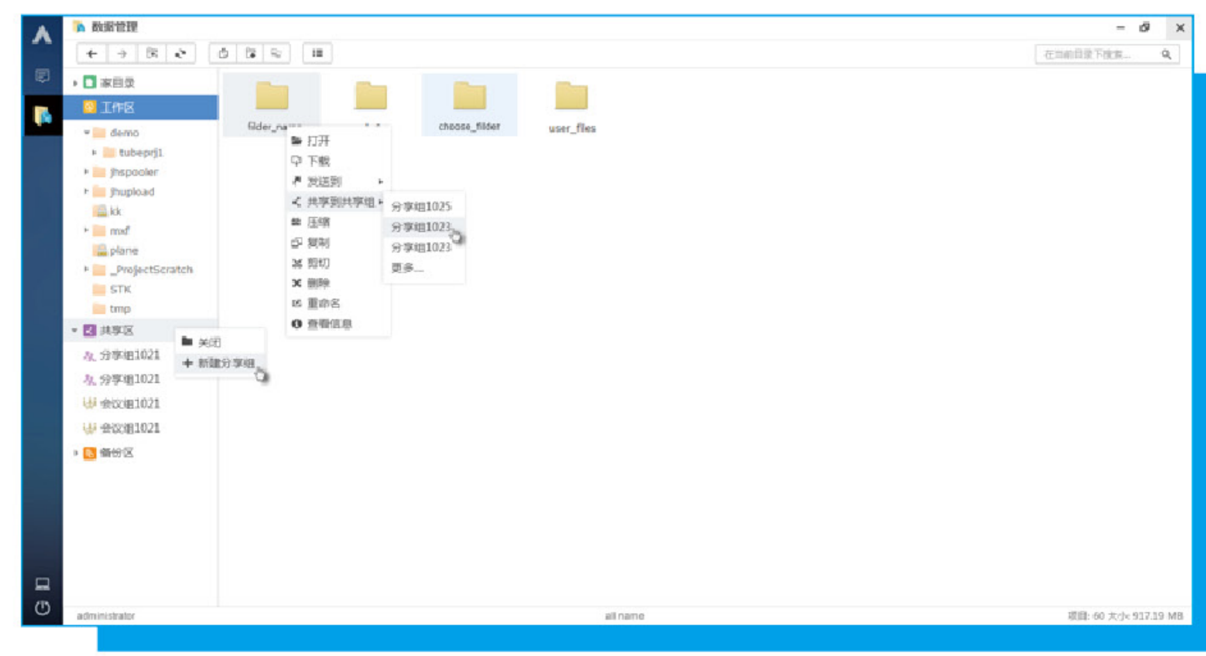
产品名称	型号规格及主要技术参数
联想智算管理平台应用门户软件	面向三维设计和仿真计算的一体化门户软件
联想智算管理平台协同桌面软件	远程可视化协同软件。支持远程三维 / 二维可视化协同设计
联想智算管理平台资源管理与调度软件	集群管理和任务调度软件
联想智算管理平台资源分析软件	集资源监控管理和智能化统计分析为一体的软件
联想智算管理平台安全控制软件	集群安全管理和控制软件

应用门户软件：采用“Web Desktop”和微客户端的全新设计，实现高效、绝佳的用户体验；支持浏览器、桌面客户端、瘦客户机等多种访问方式，易拓展；实现高性能计算和三维设计任务与数据统一管理；具备强大的数据上传下载功能；支持应用管理和应用仓库用户可自行挑选使用不同应用；支持 Web 服务集群部署模式，实现 HA 和负载均衡。

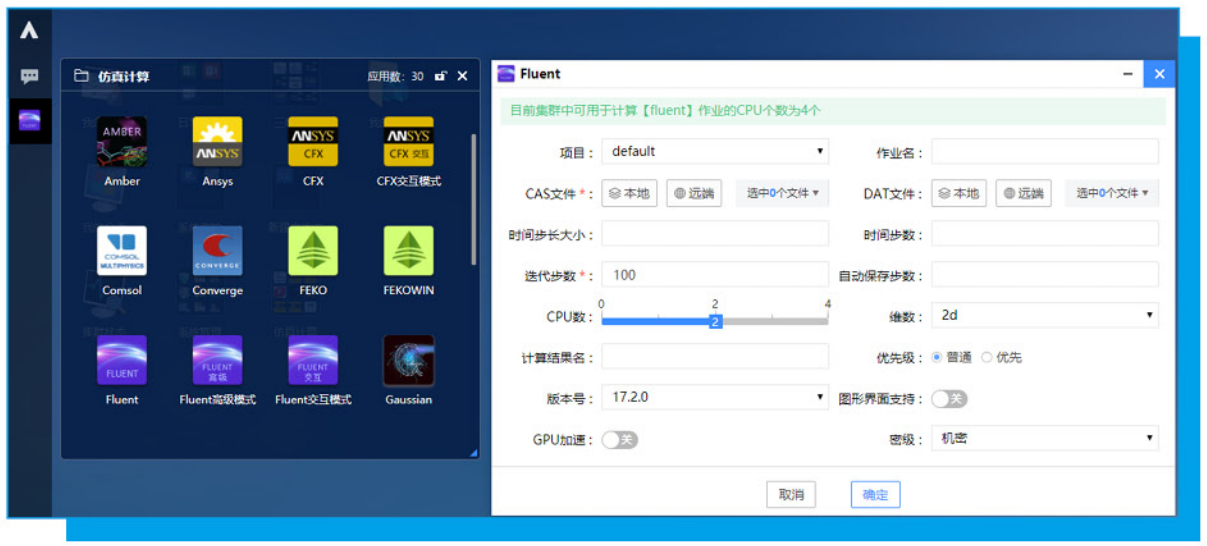
LICP 门户



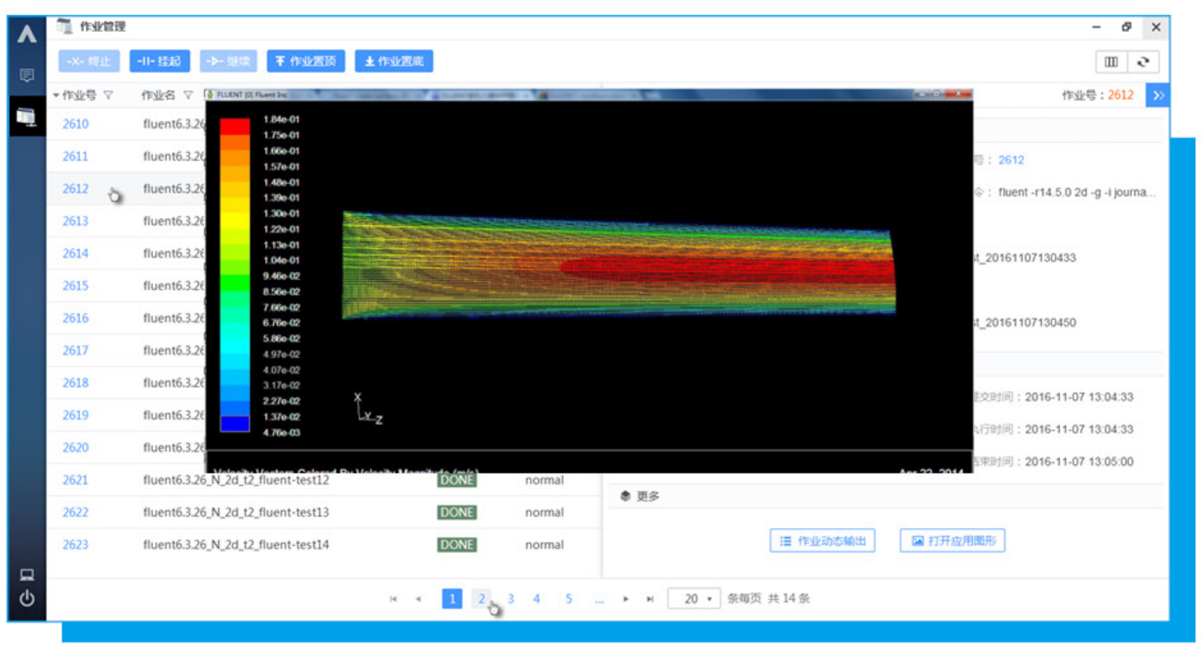
数据管理



云端数据协同管理

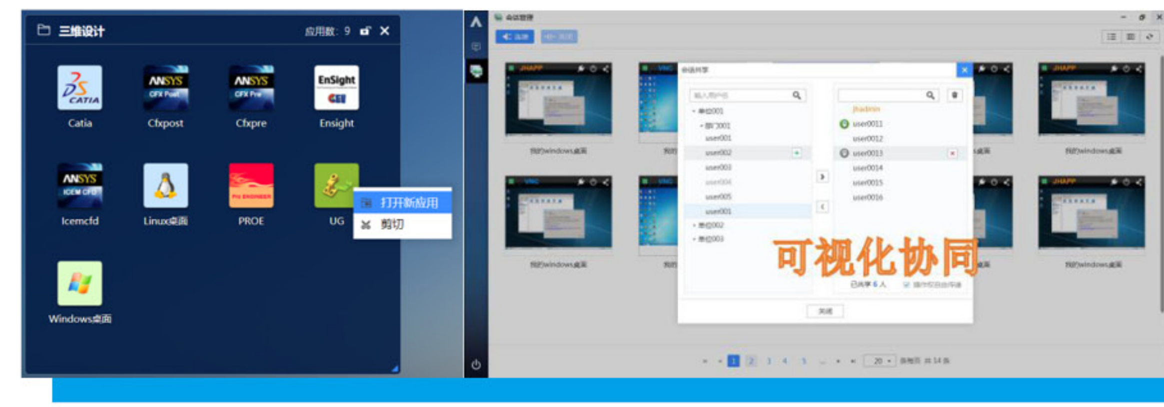


作业提交

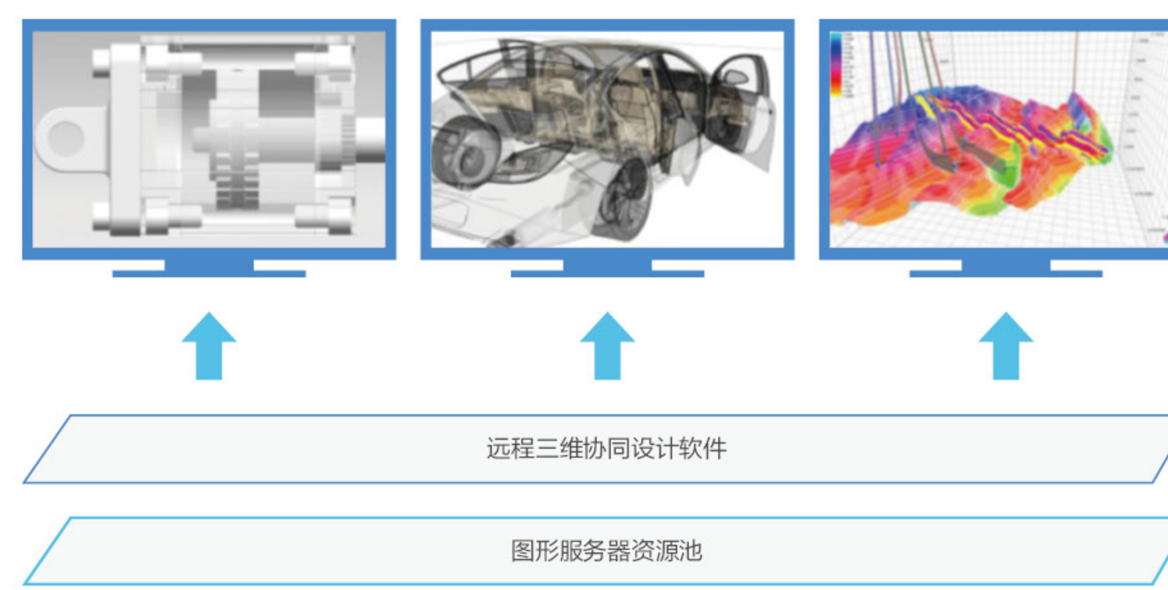


仿真计算过程可视化

协同桌面软件： 基于自研远程三维可视化技术，工程 / 工业设计人员和研究人员可从桌面终端远程使用数据中心高端图形工作站 / 服务器来进行高效协同设计。支持近乎本地的使用效果；先进的远程可视化技术，支持物理机和虚拟机，具备低延时、低带宽的特点，能应用于局域网和互联网环境；支持主流的 OpenGL 和 Direct3D 三维设计分析、前后处理、解释反演、平面设计和办公等软件；支持 Linux 和 Windows 版本的应用发布和桌面发布；支持多协同共享，提供观察和交货协同工作模式；适合小型工作室 / 会议室到大型企业设计 / 办公等多种场景。

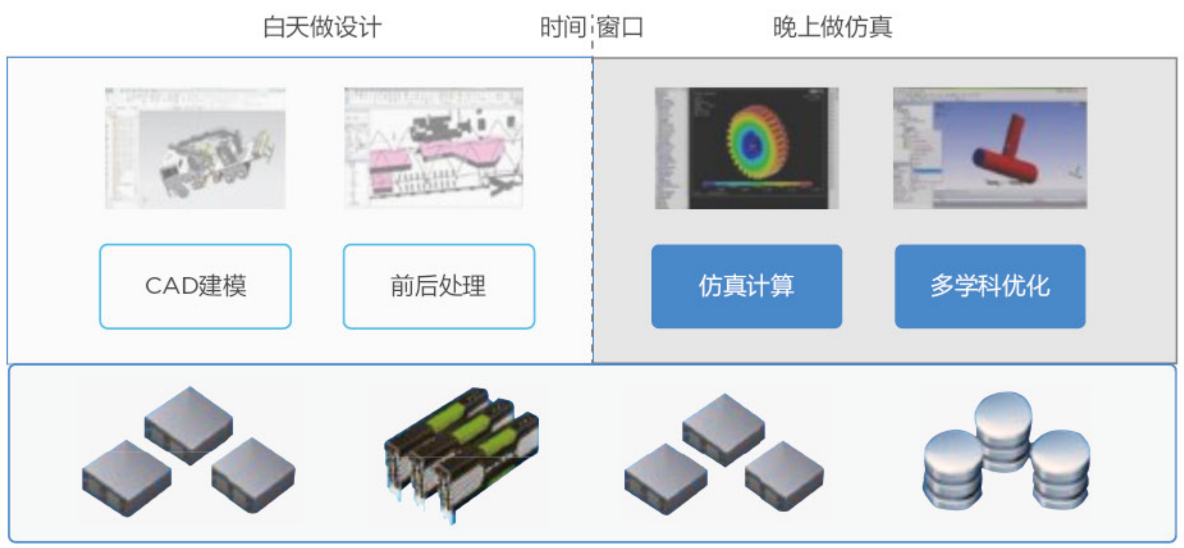


远程三维可视化与协同



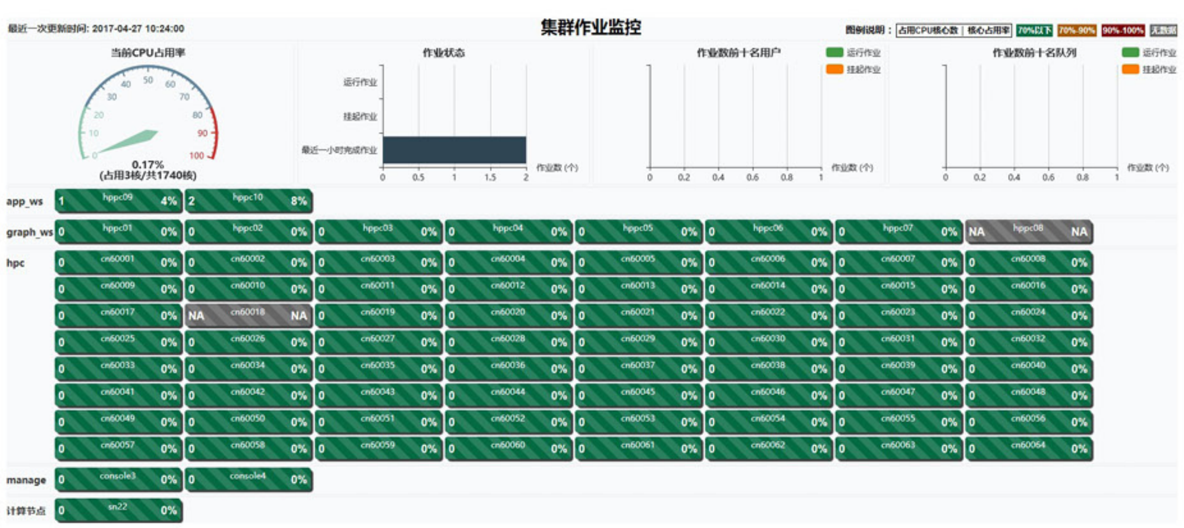
远程三维可视化与协同架构

资源管理与调度软件：强大的云计算基础架构管理中间件，集资源管理和分布式应用调度为一体，能实现计算资源、图形资源、许可证资源、网络资源的统一调度管理。

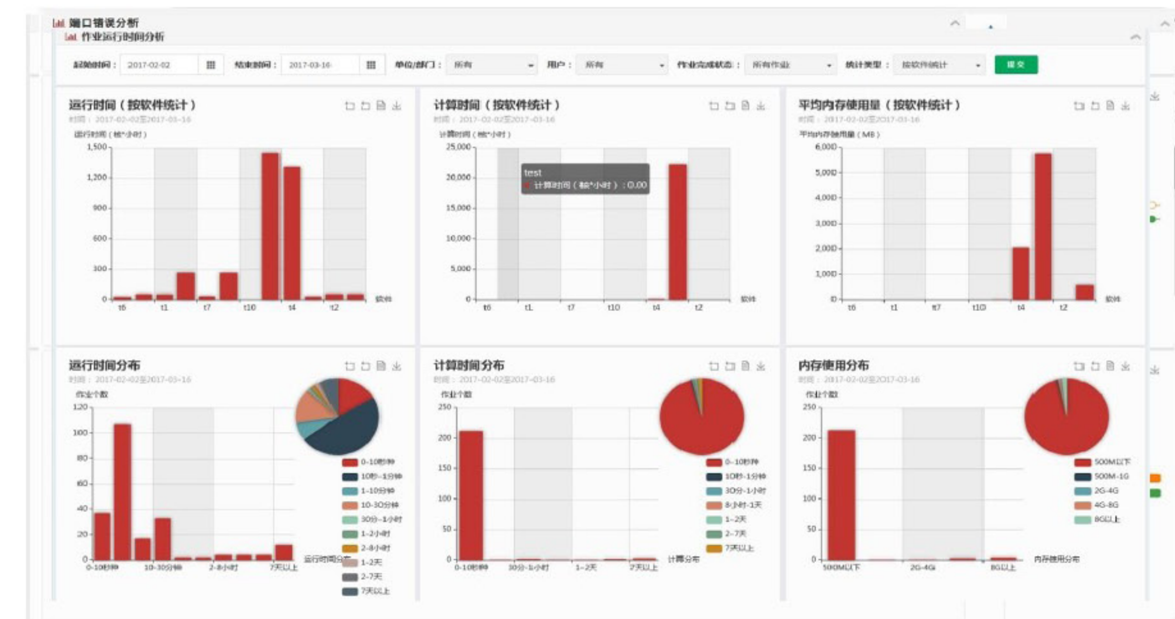


灵活的资源调度策略

资源监控与统计分析软件：集资源监控管理和智能化统一分析为一体，可监控、管理和分析网络上数十台到数千台不同操作系统的各类服务器、工作站和虚拟机、许可证以及相关网络和存储设备。

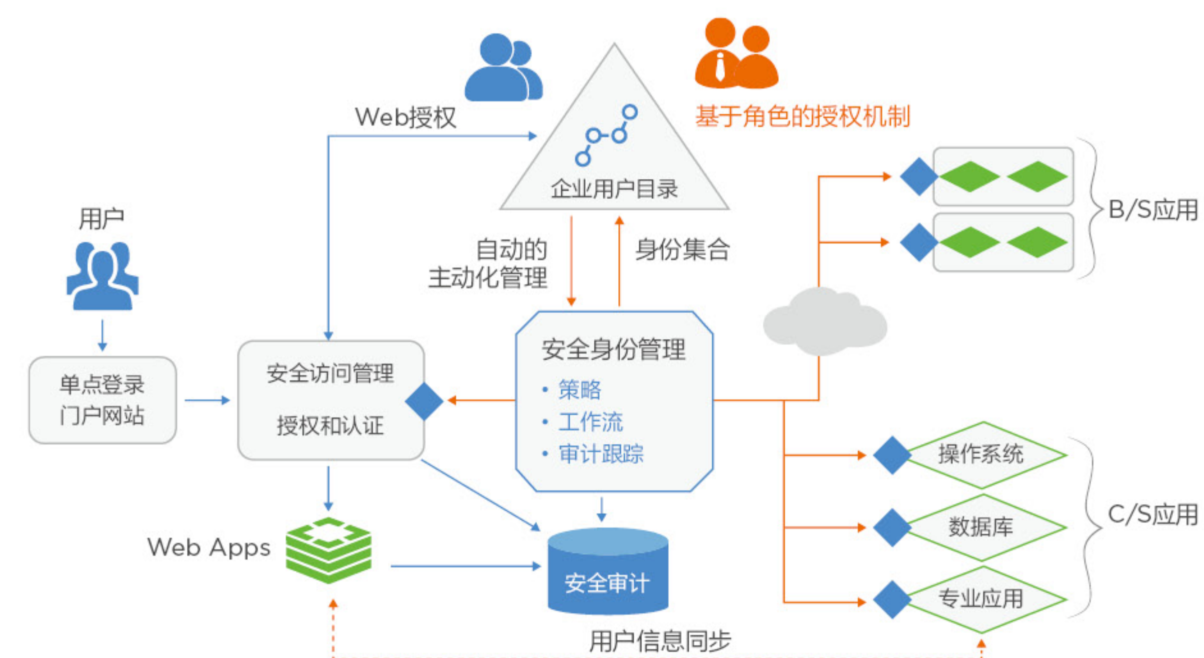


集群作业监控

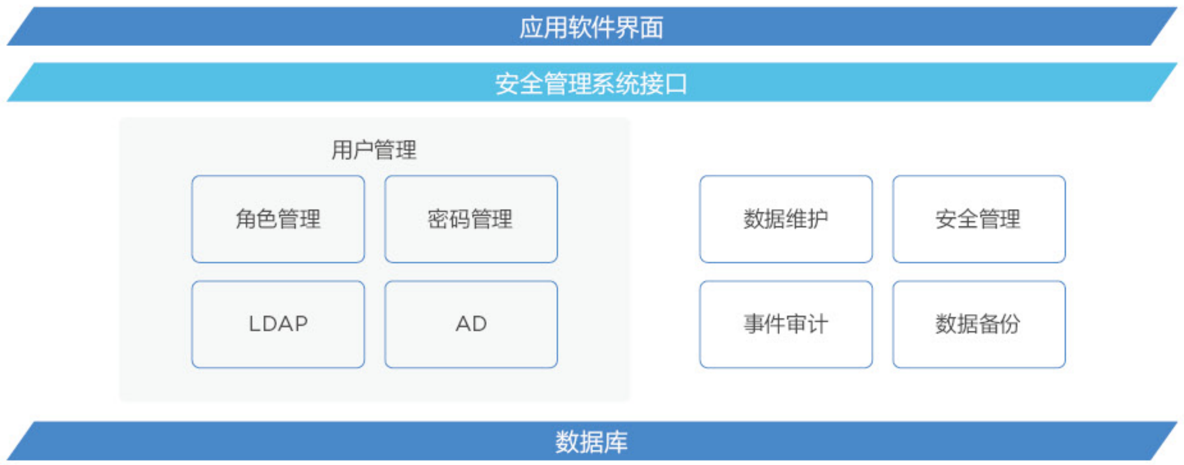


统计分析

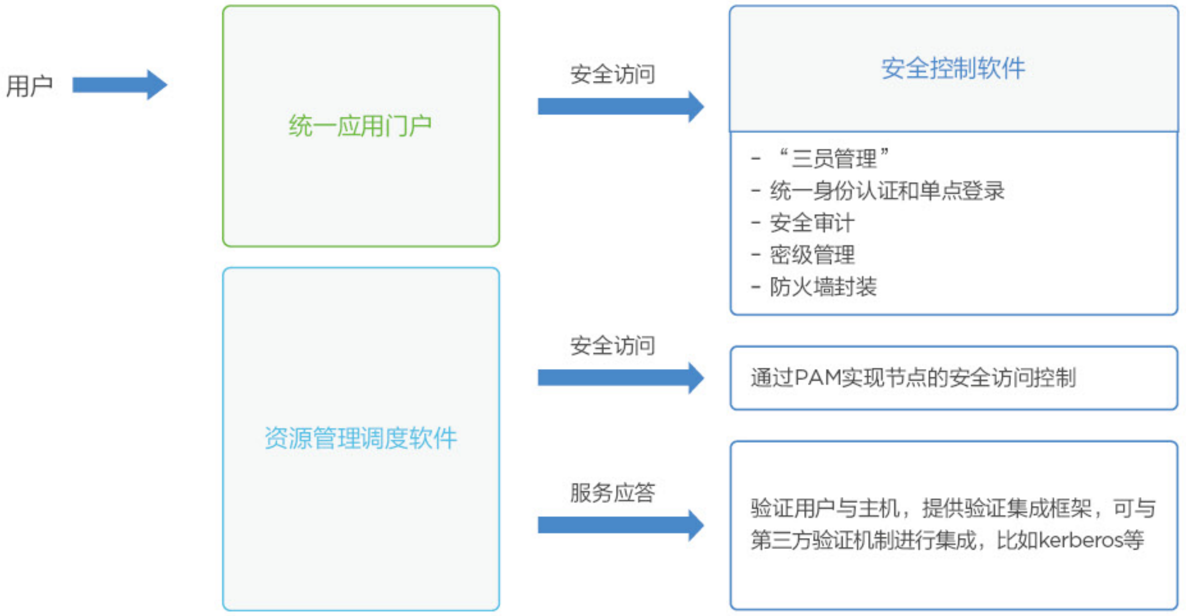
安全控制软件：对联想智算管理平台进行统一用户身份认证、密级管理和全面的安全审计、控制以及权限分配管理。可实现统一身份认证与密钥集成的单点登录；对防火墙友好；支持三员管理和型号项目管理以及密级管理。



统一身份认证系统



安全控制软件架构



安全访问机制

应用场景

- 航空、航天、兵器、船舶等军工研究所的信息化部门、仿真中心（室）
- 汽车、轮胎等制造业的设计中心、研发中心
- 高校及科研院所的信息中心、网络中心、数字化中心

凡涉及三维设计、高性能仿真计算及仿真设计一体化的业务部门均存在使用需求。

方案优势

- 覆盖客户数字化设计全流程
- 解决现有的软硬件投资模式，共享经济、共享资源、充分利用最新的信息化技术、协同与协作
- 自适应的仿真设计能力，减少无效工作，提高研发效率，释放脑力
- 提供统一的集群资源管理功能，集中管理集群中所有的计算资源、存储资源和图形工作站资源，帮助用户实现最优的资源利用率和作业吞吐量

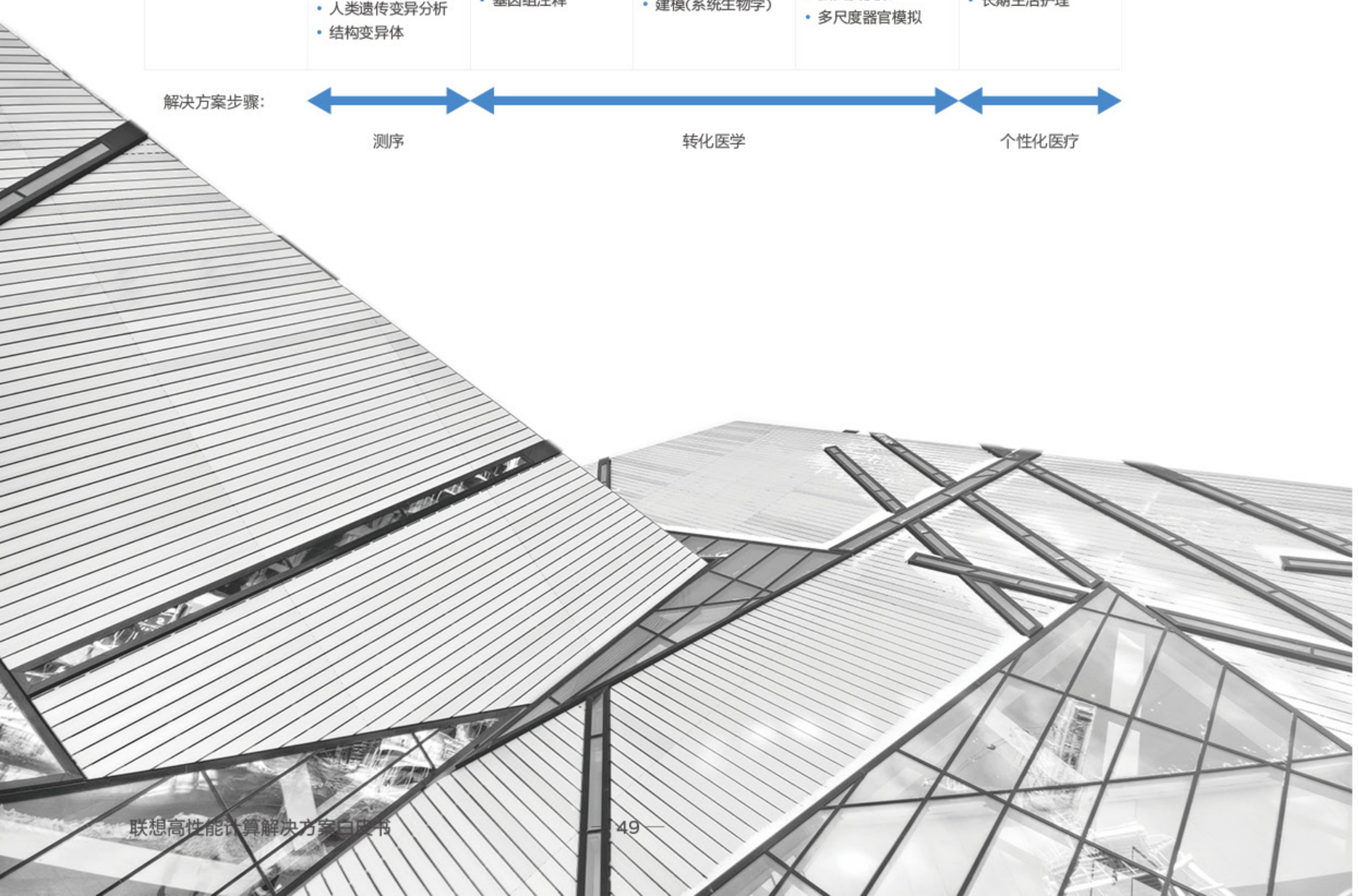
客户收益

- 提供数字化统一研发环境，帮助用户整合服务器、软件、存储资源池
- 支持 Linux/Windows 应用程序和全面协作型屏幕共享 / 控制的远程协议
- 通过 Open GL 和 Direct X 为 3D 图形提供最新支持
- 使用简单易用的应用使用门户，提供保密相关的功能要求
- 打通 HPC（高性能计算）+HPG（高性能图形）仿真设计流程
- 提供 IPT 数字化协同设计团队 IT 环境支撑
- 全面的服务器端监控功能

联想生命科学 HPC 解决方案

生命科学 HPC 应用分类

基因组医学发展路径



生命科学的主要软件

生物领域中的高性能计算最大的特点就是种类多。

① 生物分子的结构测定：Illumina 公司的 bcl2fastq，Genome Analys，HiSeq 2000，GAPipeline 等；ABI 公司的 Solid3，Solid4，Bioscope。电镜 EMAN，SPIDER；利用 Xray 方法测量用软件 CCP4、ARP/wARP，CNS 进行解析等；质谱仪 tandom(X!tandom) 等。

② 生物信息学类计算：最常用的有 BLAST，FASTA，HMMER，ClustalW，DNASTAR，PHYLIP，PAML，PAUP，T-Coffee，EMBOSS 等。

③ 分子动力学类计算：主要软件有 AMBER，Charmm，NAMD，Gromacs 等，这类应用非常适合大规模并行。

④ 计算机辅助药物设计类计算：应用软件有 DOCK，AutoDock，FlexX，Discovery Studio，ZDOCK，RDOCK，MORDOR 等。

主要生命科学学科的工作负载

学科	解决方案	数据/应用程序特点	主要应用程序
生物信息学 -序列分析	搜索、对齐生物序列(DNA和蛋白质)并对其进行模式匹配	结构化数据整数占主导，频率依赖性，大量缓存和内存BW并非关键，一些算法适合进行SIMD加速	NCBI BLAST、wuBLAST、ClustaW、HMMER、FASTA、Smith-Waterman
生物信息学 -NGS	对齐并合并DNA短片段，以重建原始序列	• 一些需要大量内存 • 许多为IO密集型任务 • 字符串分析和匹配算法	Phrap/phred、CAP3/PCAP、Velvet、ABySS、SOAPdenovo、Newbler、MAQ、BOWTIE、BFAST、SOAP、BioScope、SAM工具、GATK、Bowtie、Bwa等二代测序的比对工具，Fastqc等数据预处理工具
生物化学 -药物研发	筛选大型数据库，查找具有所需生物活性的潜在药物	• 大多为浮点数据 • 计算密集型极高 • 高度并行任务	Dock、Autodock、GLIDE、FTDock、Ligandfit、Flexx
计算化学 -分子模拟和量子力学	使用分子动力学和量子力学技术对生物分子进行建模	• 需要处理大量浮点数据 • 延迟至关重要 • 频率依赖性 • 可缩短到100秒以下	CHARMM/CHARMm、GROMACS、Desmond、AMBER、NAMD、Gaussian、GAMESS、Jaguar、NWCHEM
蛋白质组学	解读质谱分析数据，将频谱与蛋白质数据库进行匹配	• 整数比较 • FFT频谱分析具有浮点数据 • 对通信要求不高	Mascot、Sequest、ProteinProspector、X!Tandem、OMSSA
结构生物学	用物理学方法，配合生物化学和分子生物学方法研究生物大分子结构与功能的新学科	• 数据量大 • 计算密集型	Relio为典型代表(GPU优化)

生命科学应用特点及解决方案

1. 仪器后处理类的应用：

① 计算量和数据量巨大，可靠性要求高；

② 大容量存储 + 并行文件系统。

2. 基因序列比对，组装；分子对接，药物筛选：

① 计算任务相对独立，进程间通信量很小，计算网线速千兆或非线速万兆即可；

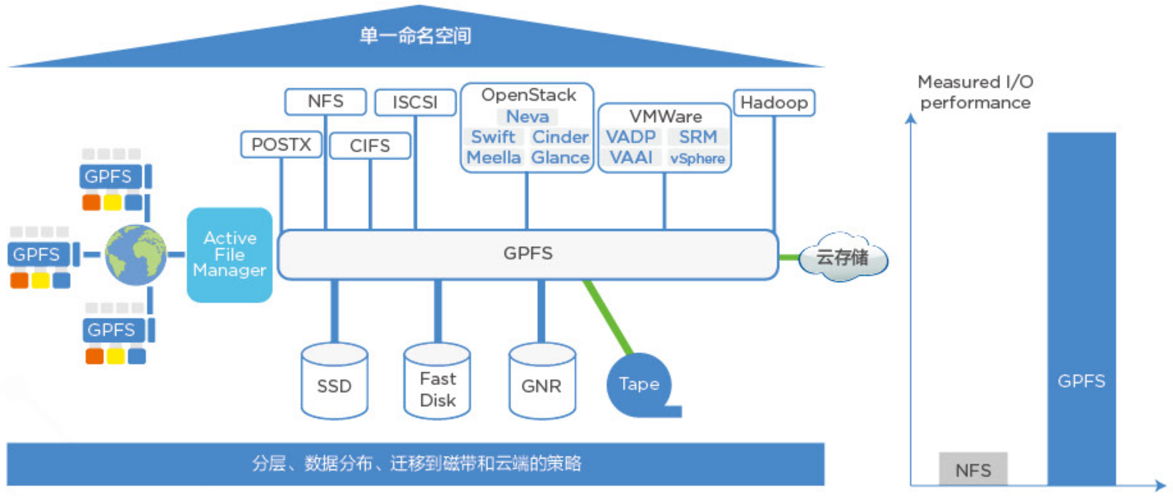
② 较大内存要求，较高的数据 I/O 吞吐量；

③ 集群系统：普通两路服务器节点 SRXX 系列服务器，高密度刀片服务器 SD630V2，或 SMP 胖节点 SR8XX/SR9X，或者水冷服务器 SD650V2；

④ 大容量存储 + 并行文件系统。

联想并行存储解决方案

使用DSS-G的联想并行存储解决方案，可用于在站点间快速访问、分层和迁移生命科学数据



- 测量数据表明，可显著提高性能，进而缩短生命科学分析时间
- 可以在联想合作伙伴的帮助下，通过联想HPC团队实施基于IBM GPFS的归档解决方案

联想并行存储解决方案

DSS高性能并行存储联想解决方案

- 冗余双路2U机架式服务器
- 84插槽JBOD存储设备
- 首选高速IB/OPA网络技术IBM GPFS 具有管理GUI并支持CIFS/SAMBA
- 支持NVMe
- 支持SSD和HDD混合
- 超过800GB/s的持续读写带宽
- 优异无与伦比的单流性能

特点

- 并行访问
- 冗余组件
- 可扩展性能和容量(PB)
- 存储分层
- 进行加密以提高安全性(仅限GPFS)
- 开发人员API (GPFS)
- 元数据优化
- 分钟级数据重构技术

- 以智能群集的形式交付预先集成的GPFS存储服务器
- 高容量和接近线性的扩展能力，面向大型环境
- 分摊RAID重建时间，降低网络受到的影响

3. 分子动力学：

① 集群系统： 高密度刀片服务器集群，4GB/Core，两个本地硬盘；

② 线速无阻塞 InfiniBand。

生命科学 HPC 参考架构

联想生命科学解决方案

联想加速生命科学工具——GOAST 基因组优化和可扩展工具

Genomics

Optimization

And

Scalability

Tool

双管齐下的解决方案：

1、GOAST架构

2、GOAST可扩展工具

Mileidy Giraldo, Ph.D.

Global Lead, Genomics Research and Development
Lenovo HPC &AI

Previous affiliations：Bioinformatics Scientist (15+ years) NCBI, NIH.
National Cancer Institute (NCI), NIH
National Institute of Allergy and infectious Diseases (NAID),NIH

Expertise：

- Large-Scale Sequence Analysis
- Next-generation sequencing (NGS)
- High Performance Computing for Life Sciences
- Large-scale Biological Databases
- Biomarker discovery
- Cancer Genomics
- Infectious Diseases

Current Focus：

Dr.Giraldo leads Lenovo's Genomics R&D Group focusing on evaluating the latest HPC technologies to bring innovation and reduce time to scientific insight. With a vertical team of researchers, engineers, and expanding partnerships, Dr. Giraldo's group has built robust capabilities around Population-level Next Generation Sequencing and Precision Medicine.

GOAST架构

在GATK4上的性能表现

序列读取　基因组学分析　Variants

	WGS	WES
Typical genomics user	150 hr.	4 hr.
Intel's Original Architecture (NVMe + FPGA)	10.8 hr.	25 min.
Lenovo's GOAST Architecture (Open source software, No NVMe, No FPGA, no specialty hardware)	**5.5 hr.**	**4 min.**
GOAST Speed up	**27X**	**40X**

First solution provider to obtain

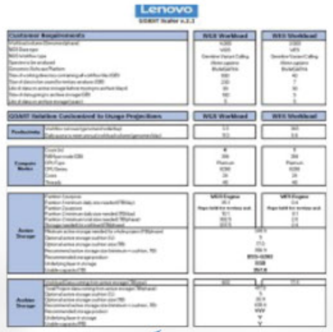
intel select solution

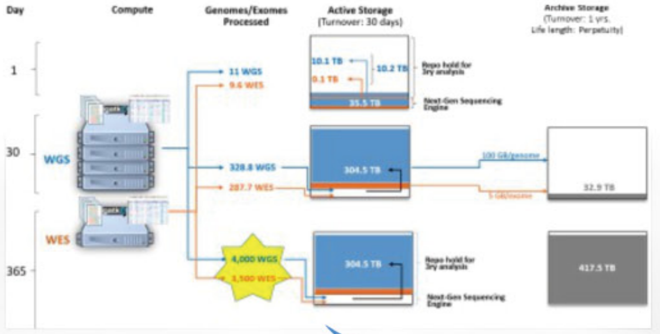
获得Intel认证的
用于基因组分析

基因组数据中心可以处理
4.5全基因组或者343全外显子
4.5 Whole genomes or 343 Exomes
在GOAST架构上一个节点一天

GOAST可扩展工具

我们如何为生命科学确定HPC的规模，并针对不同的容量、工作负载组合、预算和需求扩展GOAST体系结构。





- 了解客户工作负载的特点
- 计算所需的HPC：从计算到存储再到存档

- 显示其数据将如何超时增长的快照的示意图
- 为跨集群的数据流和数据管理提供建议

03

联想高性能计算产品及特点介绍

联想高性能计算硬件和环境配套

高性能计算服务器产品

1. Lenovo ThinkSystem

Lenovo® ThinkSystem™ 是对 Lenovo ThinkServer® 系列的一个有益补充，它极具经济效益，以创新的方式实现更大可用密度。经过优化的 ThinkSystem 可处理大量需要灵活性的工作负载，通过快速采购、部署和灵活的选件加快业务速度。这款简单但功能强大的系统可处理各种应用程序，从技术计算、GRID 部署、分析工作负载到大型云和虚拟化基础架构。

服务器产品

管理节点	SR630 V2 1U 2S机架	SR650 V2 2U 2S机架 (New)		
胖节点	SR850 V2 2U 4S 机架	SR860 V2 4U 4S 机架	SR950 V2 4U 4-8S 机架	
高密节点	SD530 2U 4节点 高密刀片	SD630 V2 2U 4节点 高密刀片 (New)	SD650 6U 12节点 高密	SD650 V2 6U 12节点 高密 (New)
异构节点	SR670 2U 4卡 异构	HG680 4U 8卡 异构	SG670 4U 8卡 异构	SR670 V2 4/8卡 异构 (New)

2. 水冷服务器

液冷刀片：联想直接水冷刀片系统是业界新的高密度计算解决方案。以其灵活、开发和简化数据中心的特点为高性能计算、网格计算、仿真和分析以及大规模云计算等应用提供了优秀的基础架构。

● 直接水冷服务器

ThinkSystem SD650 V2
支持高密度的双路Intel Xeon
高TDP与最新的Gen4 PCIe

ThinkSystem SD650-N V2
混合架构耦合高TDP Intel
Xeon与NVIDIA NVLink
SxM4 GPU

- Lenovo Neptune旗舰产品
- 同样的6U12节点机箱，支持SD650 V2和SD650-N V2
- 在混合架构中提供百亿亿次的性能，提供低成本的方法
- 行业领先的及经过验证的基础设施设计的液体冷却系统

BSC　CINECA　lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften　HARVARD UNIVERSITY

直接水冷刀片系统面对目前数据中心的局限性进行密度和性能优化而设计，可以在一个传统 42U 机柜内轻松部署 72 个节点，以及相应的网络设备。

直接水冷刀片系统设计运行于“温水”环境，节点冷却进水温度在 18-50 摄氏度。这意味着用户可以大幅降低 IT 系统的制冷能源消耗，最大化提高能效，降低 TCO 总拥有成本。

液冷型机箱：由于采用液冷设计，使得刀片机箱可以抛弃传统刀片机箱中转为刀片散热而提供的大量散热风扇，这一方面降低了风扇消耗的能耗，另一方面大幅降低了系统运行的噪音，测试表明在 0.5 米距离上，刀片机箱的运行噪音仅为 56 分贝。

联想下一代海神水冷服务器

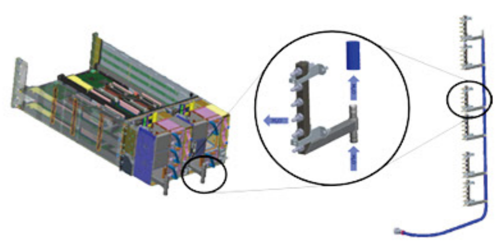
单柜性能：~3 PFlops

ThinkSystem SD650-N V2/ SD650V2

高密异构：SD650-N	50°温水 / 冷板设计与优化 / 功耗部件全覆盖设计	高密同构：SD650 V2
6U6N Intel NG *2 up to 270W 16 DDR4 DIMMs PCIe Gen4 4x NVLINK A100	50°温水 TDP 增加48% 冷板设计与优化 并行水路设计 全铜，硬连结 功耗部件全覆盖设计： CPU/GPU, MEM, SSD, NIC, PSU	6U12N Intel NG *2 up to 270W 16 DDR4 DIMMs PCIe Gen4 Mellanox HDR200 SharedIO

Think 机箱后视图

机箱后部已取消散热风扇，只有六个电源模块，管理模块和供水、回水接口。



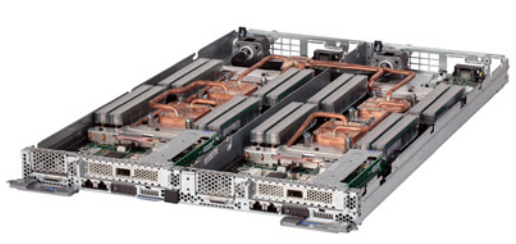
机箱液冷组件连接

节点插入机箱后，直接通过快速接头与液冷组件连接。

机柜视图

一个机柜中六个机箱的液冷组件直接连接，构成整个机柜的供水与回水通道，正两个管路再分别连接到 CDU（冷水分配单元）的供水与回水管路，构成完整的二次侧水循环。

液冷节点：

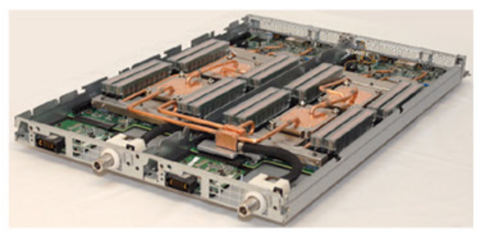


液冷节点内部视图

液冷节点采用通用的 Intel CPU，具有 16 条内存插槽，支持 TruDDR 内存；可以安装 2 个 SSD 硬盘，集成双千兆网卡。1 台液冷机箱内部可以安装 12 个液冷节点，每两个节点为一组。

液冷是“冷板式”设计，即发热部件不直接接触液体，而是先把热量传输给装有液体的铜制冷板，再通过液体循环带出设备。

节点的主要发热部件：CPU、内存和 I/O 板卡，本地硬盘，机箱电源都有导热件通过，节点中大部分的热量都通过冷却液体带走。



液冷节点后视图

每对内存中间插入一个金属导热件，以高效吸收热量，并传递到水冷管件。每对内存条上有一个塑料盖，用以确保没有空气流入来吸收热量。

液冷节点总体效果：通过在机箱和节点上的多重液冷设计手段，ThinkSystem 液冷系统可以将 85%-90% 的发热量通过液冷系统带走，而剩余的 10%-15% 的热量只需要机房环境空调即可轻松解决。

整机柜交付：联想 ThinkSystem 采用整机柜交付方式，一个机柜提供 6 个水冷机箱，可安装 72 个水冷节点，管理网和监控网的千兆连接在机柜内实现，整个机柜对外连接最多可达 16 个 10Gb 端口。

整机柜交付水冷单柜

高性能计算 GPU 产品

协处理器是提升应用性能的重要手段，联想高性能计算系统的 GPU 节点通常选用 NVIDIA GPU。协处理器的性能提升效果是严重依赖于应用本身代码的改造和数学模型 / 算法的特点的，因此不能保证协处理器会在任意应用和算法上都带来与标称性能相当的应用性能提升幅度。

高性能计算存储产品

1. 联想高性能存储系统

联想HPC存储解决方案概览

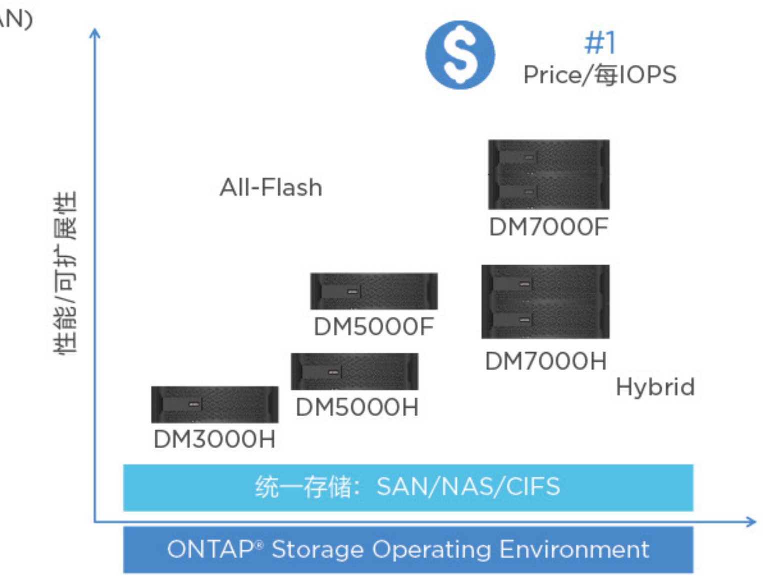
联想DM系列统一存储 (iSCSI/NAS/SAN)

极致性能

- ONTAP 数据管理软件
- 延迟敏感型应用
- 云分层/关键数据远程容灾
- 支持集群，12对HA pair
- 650k 随机IO(8k)，单套最大4.8M

最佳用途

- 人工智能
- 大数据分析
- EDA工程设计



2. Lenovo (DSS-D) 存储系统

Lenovo DSS-D 整合了联想凌拓的高性能存储和 GPFS 文件系统，以及灵活扩展的存储系统，可以为客户带来高性能、高带宽和高容量的存储系统解决方案。DSS 是基于联想 DE 系列存储和 GPFS 技术的软硬件一体化解决方案。

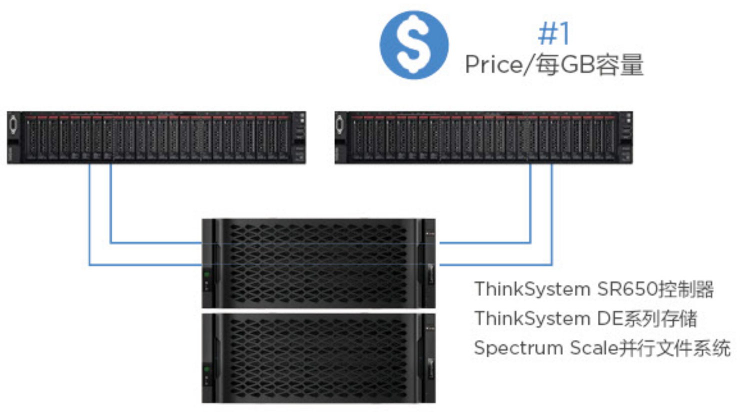
联想DSS-D系列高性能存储

高性价比

- 高存储密度，4U/60硬盘
- 卓越的性能
- 单个构建模块6.72PB容量
- 单个构建模块>20GB/s
- 灵活的构建模块，支撑横向扩展

最佳用途

- 高性能计算
- 数据密集型场景
- 高性能NAS
- 适用于中小型HPC/AI场景



3. Lenovo DSS-G

联想的 DSS（Distributed Storage Solution）。DSS 基于内置先进 RAID 技术的 GPFS 软件，和以可靠性和性能著称的联想 ThinkSystem 服务器，不需要专用磁盘阵列控制器，简化了存储系统结构，提高了可靠性。同时 DSS 不像普通的分布式存储系统需要使用多份数据拷贝方式来保证数据安全性，保证了磁盘利用率与基于磁盘阵列的解决方案相同。

联想DSS-G分布式存储

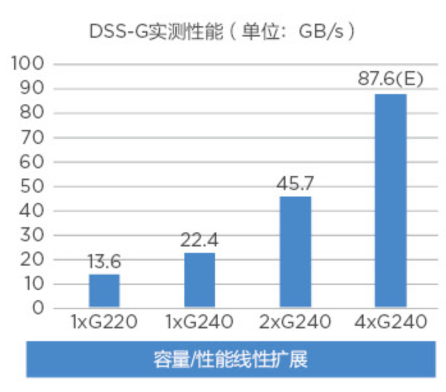
DSS-G是适用于HPC/AI/大数据/云的可扩展存储解决方案

极致性能

- 卓越的性能
- 快速重建，重建时不影响应用性能
- 实测性能超过 2500GB/s
- 灵活的构建模块，支撑横向扩展
- 线性容量/性能扩展
- 多种数据访问方式：客户端/块/文件/对象

最佳用途

- 大容量分布式存储
- 极致存储性能需求
- 适用于中大型HPC/AI场景



DSS-G 的另一个优势是秒级的 rebuild 时间，这使得磁盘故障和修复过程导致的系统性能下降，预报作业时间延长等现象不再存在。

借助于 GPFS 方便灵活的分级存储功能，选择不同配置的 DSS 组合在一起，就可以获得一个具有优秀分级存储功能的存储集群。

磁盘盘组中的磁盘和 2 个存储节点上的多个 SAS 接口构成多个环路，磁盘故障不会中断 SAS 环路，不会导致其余磁盘不可访问。磁盘盘组本身无控制器电路，只提供磁盘安装和环路连接，以及冗余的磁盘供电和风扇，也不构成单点故障点。

由于我们提供了热备空间和分布 RAID，我们的 rebuild 方式更加先进，从被动变成主动。每台 DSS 的一半盘中同时损坏三块磁盘后开始进行 critical rebuild，现有配置下需时 20-50 秒，通常为 30 秒。此过程不须更换磁盘，整个存储系统基本不会因为磁盘故障有性能损失，整个过程全自动化。

内置 Decluster 技术，硬盘故障时，所有磁盘参与重建，将重构过程中的读写操作分散到各磁盘中，实现并行 I/O，从而提高重建速度及减小对存储性能的影响。块数据支持 8+2P，8+3P 的 Reed-Solomon 算法，元数据支持 3 副本，4 副本复制。

在数据安全性方面：

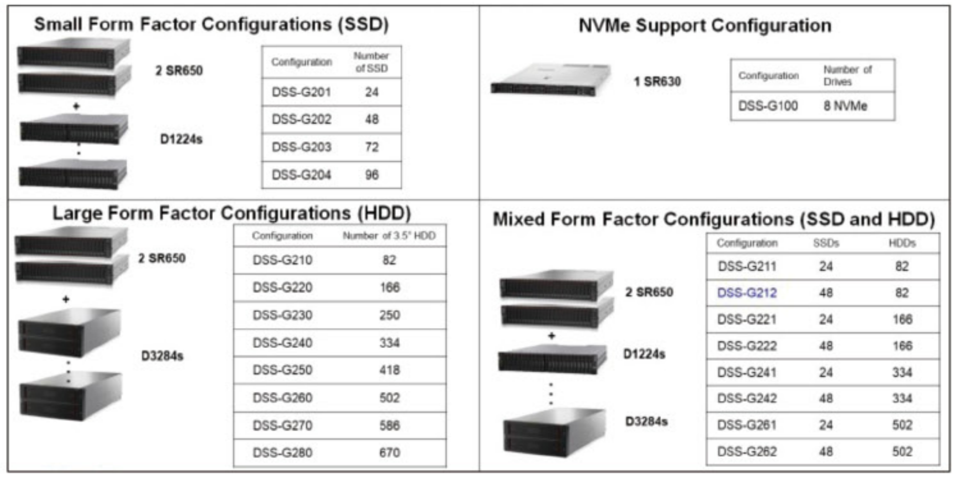
软件 RAID 模式支持同时坏掉 3 块硬盘而不中断业务，允许一个扩展柜损坏而数据不丢失、应用不中断；

自动检测硬盘响应时间，当出现 timeout，I/O error 等情况能自动判断是硬盘本身还是路径原因导致，与此同时，系统重建由于错误而丢失的数据信息；

自动检测和维修硬盘的扇区错误，端对端的检错功能，完全避免磁盘“无声数据失效”。

可以灵活配置和任意扩充的DSS-G产品系列

- 21 base configurations available in 1410 rack or 7X74 rack independent configuration
- Customer choice (Spectrum Scale, JBOD, HDD / SDD, NVMe, connectivity)



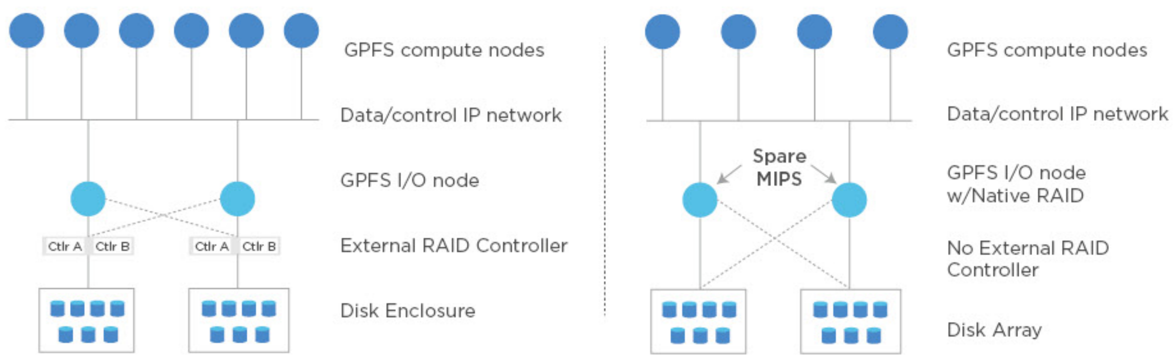
Configuration	Number of SSD
DSS-G201	24
DSS-G202	48
DSS-G203	72
DSS-G204	96

Configuration	Number of Drives
DSS-G100	8 NVMe

Configuration	Number of 3.5" HDD
DSS-G210	82
DSS-G220	166
DSS-G230	250
DSS-G240	334
DSS-G250	418
DSS-G260	502
DSS-G270	586
DSS-G280	670

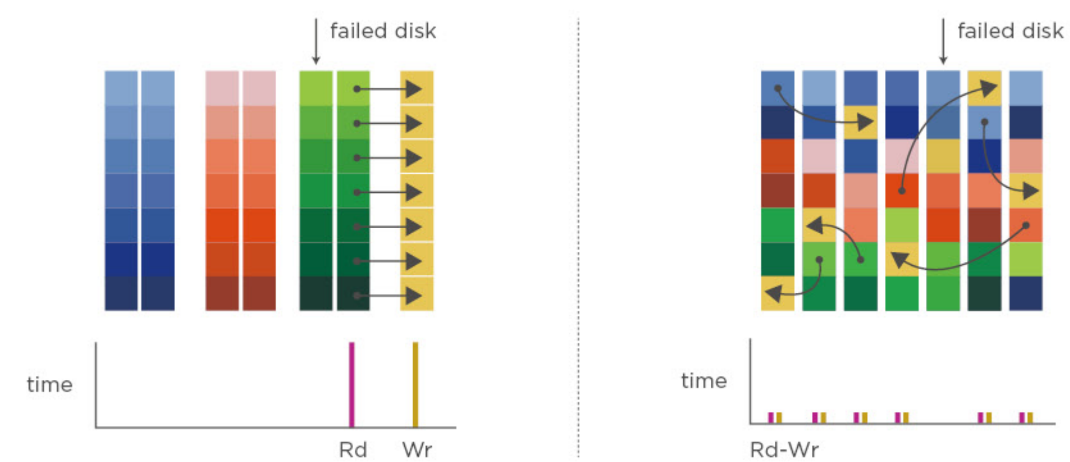
Configuration	SSDs	HDDs
DSS-G211	24	82
DSS-G212	48	82
DSS-G221	24	166
DSS-G222	48	166
DSS-G241	24	334
DSS-G242	48	334
DSS-G261	24	502
DSS-G262	48	502

基于 GPFS GNR (General Native RAID) 技术的存储系统：



传统GPFS　　GPFS GNR

GPFS GNR 较传统的 GPFS 而言，省去了磁盘阵列控制器。而是采用 JBOD 的方式直接连接磁盘。这种连接方式不仅节约了成本，还在 I/O 性能和可靠性方面有更好的表现。尤其是在磁盘出故障需要重构时，GPFS GNR 的重构时间较传统磁盘阵列有明显优势。



传统RAID　　GPFS GNR

较传统 RAID 而言，GPFS 采取 d-cluster 方式组织磁盘。这种方式能大大缩短磁盘损坏后系统重构的时间。上图分别描述了传统 RAID1 和 GPFS GNR RAID1 配置。黄色磁盘为预留的 spare 盘。在传统 RAID1 系统重构过程中，所有数据将从同一磁盘拷贝到 spare 磁盘。所需时间较长。而 GPFS GNR 中，每一对数据拷贝都分散分布在不同磁盘对中。当某个磁盘失效时，可以从各磁盘中读取相应的数据块，复制到预留的数据块中即可完成重构。其优势在于能将重构过程中的读写操作分散到各磁盘中，实现并行 I/O，从而缩短重构时间。

GPFS GNR 运行在 x86 平台下，采取的是 8+3p 模式。即 8 个数据盘加上 3 块校验盘。每份数据保存有 4 分拷贝，采用 d-cluster 的方式分布在 47 块磁盘上。当一组磁盘失效时，GPFS GNR 能很快重构，几乎不影响用户数据访问。即使有 3 组磁盘同时失效（非常少见），GPFS GNR 也能很快将其恢复在 2 组磁盘失效的模式，然后进行重构。GPFS GNR 在 x86 平台下，较传统的 RAID 而言在系统重构方面有很大优势。

研究发现，现行存储系统由于数据传输路径较长。从主机 CPU，到主机内存，光纤卡，存储控制器，最终到磁盘。在数据的传输和存储过程中，有一些细微的数据错误（如比特位翻转）很难通过常用的工具检测到。GPFS GNR 能提供企业级的点到点数据校验检测。存放在磁盘上的所有数据块都有相应的校验位。GPFS GNR 能对数据在传输和读写过程进行全面的检测，保证数据有效性和可靠性。

GPFS GNR 提供统一管理接口。磁盘重构，负载均衡，点对点数据校验检测等都有后台自动进行。很少需要用户手动干预。GPFS GNR 还能动态配置系统重构对整个磁盘带宽占用的比例，能保证关键应用的数据带宽。

GPFS GNR 提供全面的磁盘管理功能。能检测并管理磁盘连接，介质错误，磁盘失效等问题。针对介质错误或者点对点校验失效等错误，GPFS GNR 能在后台自动修复。根据磁盘工作状态，记录其“健康”状况，能在磁盘失效前，进行预警。

GNR 功能概览：

Spectrum Scale - 高性能，高度可扩展的分布式文件系统

DSS-D和DSS-G集成了最新的Spectrum Scale分布式文件系统

消除性能瓶颈

确保数据可用性、完整性和安全性

支持QoS(Quality of Service)

支持本地写缓存(HAWC)/读缓存(LROC)

策略驱动在线压缩

分布式的元数据，支持闪存加速

支持最新的低延迟、高带宽NVMe

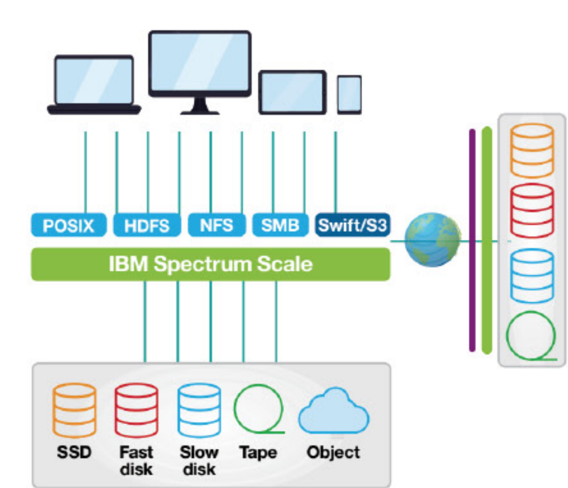
多协议支持

- POSIX, GPFS, NFS v4.0, SMB v3.0
- Big data and analytics: Hadoop MapReduce
- Cloud: OpenStack Cinder (block), OpenStack Swift (object), S3 (object)

支持云和本地混合

- Amazon S3, OpenStack Swift 和 与Amazon S3兼容的云存储

高性能分布式文件系统，用于大规模管理数据，具有执行归档和分析的独特能力。



4. Lenovo Intelligent Cluster 的介绍：

A. 要点

a. 复杂解决方案变得简易

b. 行业出色的技术，优化的解决方案设计

c. 享受 Lenovo 端到端支持的端到端解决方案

部署适用于技术计算、高性能计算 （HPC）、存储和云环境的解决方案可能会给 IT 带来极其沉重的负担。Intelligent Cluster 利用 Lenovo 数十年的丰富经验，凭借预先集成、交付、得到全面支持的解决方案来降低部署复杂性，这些解决方案将行业出色的组件与优化的解决方案设计紧密结合。借助 Lenovo Intelligent Cluster，客户可以集中精力最大限度提高业务价值，而不是耗费宝贵的资源来设计、优化、安装和支持满足业务要求所需的基础架构。

B. 行业出色的技术，优化的解决方案设计

Intelligent Cluster 解决方案采用的 Lenovo® System x® 服务器、存储设备、软件和第三方组件，可以在集成交付的解决方案中支持广泛的技术选项。 Lenovo 针对可靠性、互操作性以及最高性能彻底测试和优化了每个解决方案，因此客户可以迅速部署系统并使投入工作，从而实现他们的业务目标。

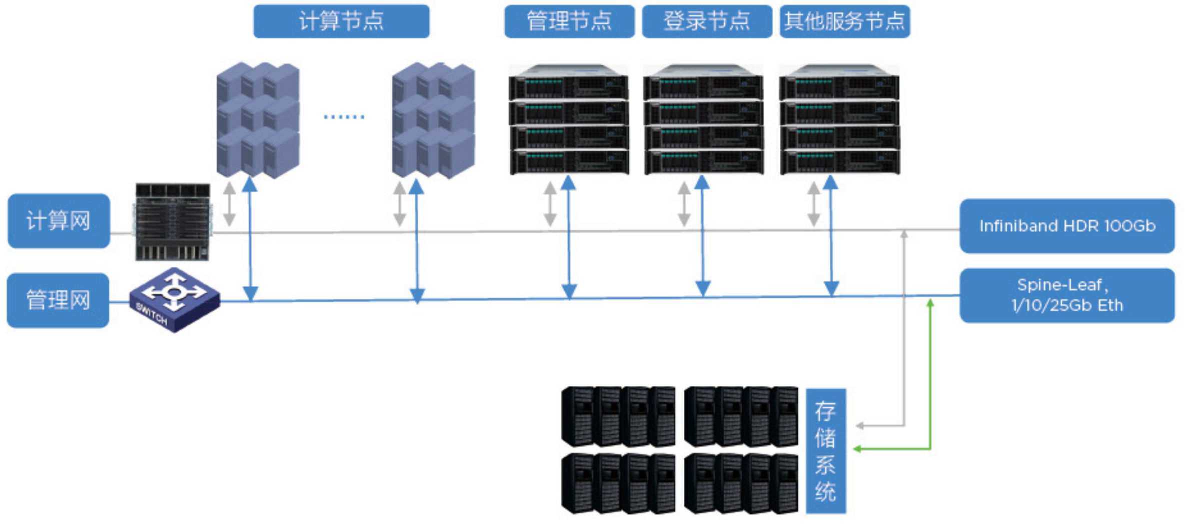
C. 享受 Lenovo 端到端支持的端到端解决方案

Intelligent Cluster 解决方案由 Lenovo 构建、测试、交付和安装，并作为单一解决方案进行支持，而不是作为数以百计单独组件进行处理。Lenovo 提供了包括 Lenovo 和第三方组件二者、单一联系点、解决方案层面的支持，可以在系统的整个生命周期内实现最高的系统可用性，因此客户可以花较少的时间维持系统，花更多的时间实现成果。

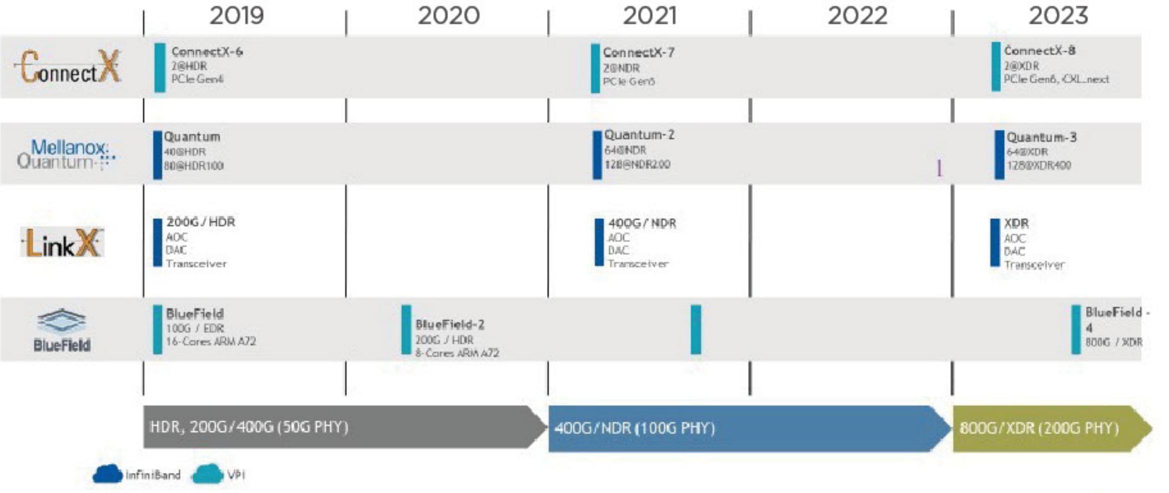
高性能计算网络产品

高性能计算网络部分包含多个不同协议产品，他们分别是：（1）高速网络 100Gb/200Gb InfiniBand 网络：包括 36 端口；216 端口和 648 端口交换机系列；也包括 40 端口 HDR200 和 800 端口 HDR200 交换机以及 NDR400 128 端口和 2048 端口交换机系列；（2）高速网络 Intel 100Gb OPA：包括 48 端口和 768 端口的交换机；（3）16/32Gb FC 网络；（4）12Gb SAS 网络；（5）万兆网络；（6）千兆网络。

HPC网络概况



INFINIBAND技术线路图



高性能计算机房环境配套产品（配套的水冷和模块化数据中心）

1. 联想模块化数据中心

1. 产品概述

目前数据中心基础设施形态千差万别，但一体化、标准化及模块化已经成为数据中心建设的主流趋势。在此背景下，联想推出了模块化数据中心（MDC），它将数据中心所必需的电气、制冷、机柜、监控、消防、布线、IT 设备及操作平台等软硬件集成在一个封闭的模块化空间内，在高度集成了计算能力的同时，还大大降低了对空间和能耗的需求，在具备高可靠性的同时提供极其灵活的可扩展能力。

联想针对不同客户类型，提供了完整的模块化数据中心解决方案，包括了模块化数据中心 Lenovo Smart Aisle、小型数据中心 Lenovo Smart Row，以及微型数据中心 Lenovo Smart Cabinet 解决方案等。

智慧数据中心

Lenovo Smart Cabinet	Lenovo Smart Row	Lenovo Smart Aisle
		
机柜内的整合的智能型基础架构，适用于小型网络间，远程站点。 容量：3.5KW，单机柜	整排机柜的整合的智能型基础架构，适用于小型数据中心，灾后恢复。 容量：12.5KW，1-6个机柜	针对中大型数据中心基础设施的标准化、模块化解决方案，可快速部署、在线扩容、在线改造。 容量：25-200KW，7-35个机柜

2.Lenovo Smart Cabinet 介绍

Lenovo Smart Cabinet 是联想数据中心解决方案产品，应用于微小型数据中心、办公区域等室内环境。为客户提供完整的 UPS 不间断电源、机架配电、热量管理、机柜支撑和监控管理；600mm 宽 1200mm 深 2000mm 高，42U 高机柜（IT 可用空间 29U），重量约 300Kg，满足 3.5kW 设备散热及供电需求；采用机架式 UPS，满足后备 30 分钟；全封闭机柜，带来机房的进一步节能。

具有如下特点：

- 防尘降噪，高效节能
- 智能监控
- 高度集成、节省空间
- 良好的人机界面
- 快速交付，品牌无忧

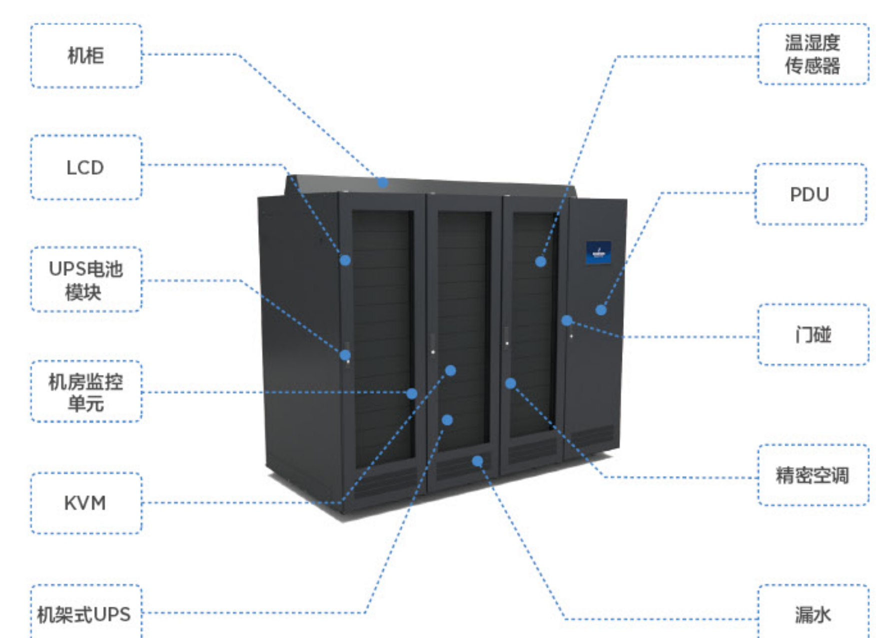
3.Lenovo Smart Row 介绍

Lenovo Smart Row 系列解决方案，是在已获得巨大市场成功的 Lenovo Smart Row 系列的基础上进行重大升级的全新解决方案，其应用仍然定位于 10~100 平米小型数据中心。符合用户对移动管理的需求、对机房环境的依赖度更低、适应更宽广的 IT 热密度范围。

为客户提供完整的 UPS 不间断电源、机架配电、热量管理、机柜支撑和监控管理，满足客户机房的基础架构需求。采用变频 12.5kW 的制冷空调，机架式 UPS，满足后备 30 分钟；采用全封闭机柜，带来机房的进一步节能；整体化方案具有高可靠、高智能和快速部署等优势；可为用户提供多种搭配方案。

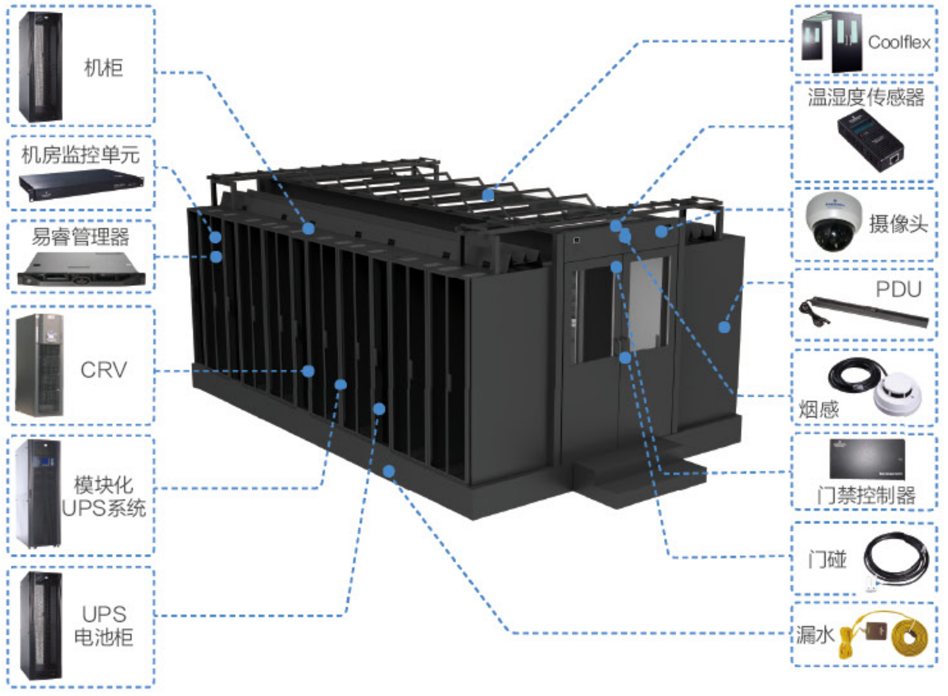
具有如下特点：

- 灵活设计与扩容
- 低建设成本
- 低运营成本
- 交付周期短
- 良好的人机界面
- 绿色节能



4.Lenovo Smart Aisle 介绍

Lenovo Smart Aisle 按照不同机柜数量、机柜密度，集成行间制冷，冷通道封闭，监控，机柜及配电并保证消防的一体化设计；供电上可灵活的采用集中 UPS 供电、分散模块化 UPS 供电等多种方式，实现 UPS 单路、双总线供电模式，并根据后期业务需要可逐渐扩容。



Lenovo Smart Aisle 模块化整体方案设计包括四大系统：供配电系统、空气调节系统、服务器机柜及附件系统和动力环境监控系统。具有如下特点：

A. 降低系统复杂度，提高系统可靠性 — 与传统数据中心相比，模块化设计可预先验证，可复制性强，可靠性不随系统数量增加而降低；

B. 部署速度快 — 模块化数据中心可以非常迅速地部署，通常只需要几周时间；相比之下，建造传统的数据中心通常得需要 1 到 2 年，甚至几年时间；

C. 降低运维成本（OPEX）— 相同的模块可以采用相同维护规程，降低维护人员要求，共享备品备件；采用行间空调及封闭冷通道技术，可以大大降低 PUE 值，带来数据中心整体运营成本降低；

D. 易于部署 — 模块化数据中心可采用集装箱或预制单位构建方式部署在指定的任何位置，甚至可以移动；易于扩展 — 如果需要，可以添加更多的模块。

5. 模块化数据中心整体方案特点

经济性

降低机房TCO。采用分期建设的思路，减少初期建设成本CAPEX占用；可提升单机柜容量，带来占地40%，方案美观紧凑；运营成本OPEX降低30%，更高效节能。

智能性

管理简单方便。通过统一监控平台实现机房智能监控、智能管理、智能供电和智能制冷，保障了对IT环境的智能控制。

高可用性

可用性超过99.999%。通过工程产品化，实现整体设计及交付，避免各系统之间耦合问题，大大提高数据中心整体可用性。

协同性

高度一体化。各系统风格、协议统一，一站式服务，实现快速部署和在线扩容，保障客户关键业务永远在线。

售后服务

联想提供整体机柜三年售后服务，且全国均有售后网点。

2. 直接水冷解决方案

1. 设计规范、标准或依据

《Think M5 WCT solution implementation general guide V3》

《Thermal Guidelines for Liquid Cooled Data Processing Environments》

2. 主要设计参数和设计工况

设计指标及工况如下表：

NO.	项目	参数	备注
1	单机柜额定热负荷	33.5KW	节点按照最高功耗 CPU 算（165W） 规定其中 85% 的热负荷需要由水冷设备承担制冷
2	机柜数量	2	
3	单机柜最小热负荷	10KW	以 2*105W CPU、8*16GB MEM、1*IB、2*2TB 节点的 25% 负载估算
4	机柜进水温度	45℃	
5	单机柜额定水流量	≥36lpm	36x compute trays (dual node), 6x switches：Minimum flow rate = 36 liters per minute
6	单机柜水阻力	2.4psi	rack manifold and 6x DWC chassis
7	机柜工作水压范围	≤50psi	Pressure should be below 50psi
8	室外温度（额定点）	40℃	
9	电源形式	380V 50Hz 3N	最大为 4 路输入
10	室外温度范围	-14℃ ~ 40℃	在该范围内满足全负荷制冷要求，冬季防冻
11	冷水机组配置	不标配	实现完全自然冷
12	水质要求	参看文件	
13	机柜进出水接口形式	参看文件	
14	水过滤器要求	50μm	300 目

设计参数

3. 通用设计要求

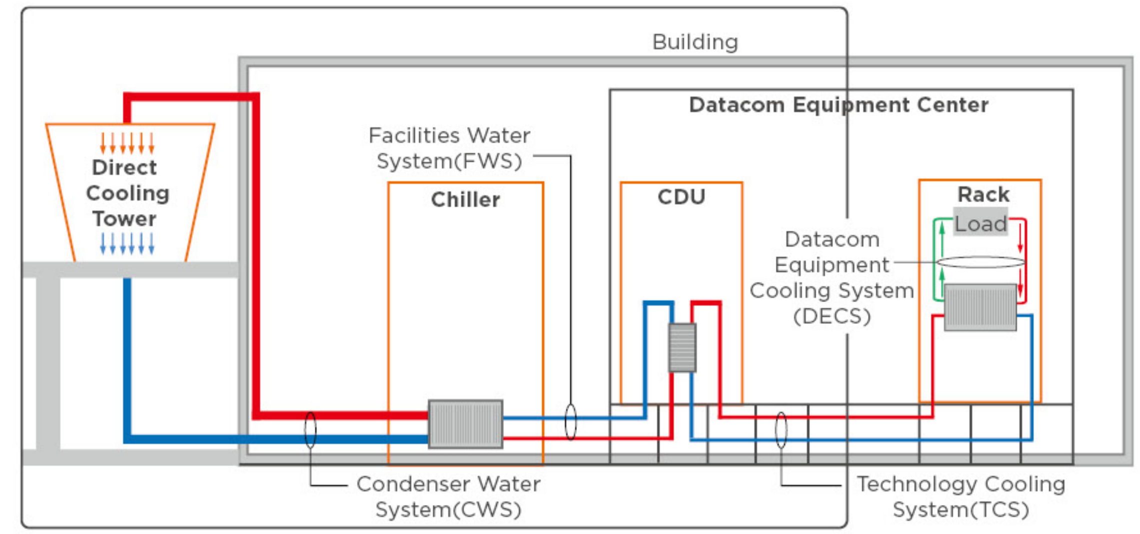
制冷系统具备高可靠性，满足液冷服务器长时间不间断运行。

制冷系统能够有效预防水温过低，防止在液冷服务器内出现凝露。

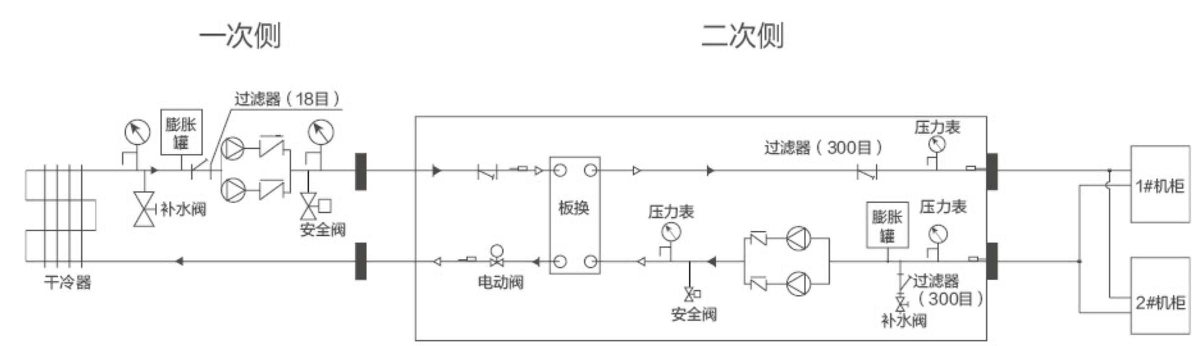
制冷系统具有必要的智能监控功能，能够对系统出现的异常情况作出告警和处理。

4. 制冷系统设计方案

液冷服务器配套制冷系统如下图所示，本文的设计范围包括左边框体内的所有设备或功能，具体即中间换热单元、室外干冷器等设备或功能。



液冷服务器配套制冷系统

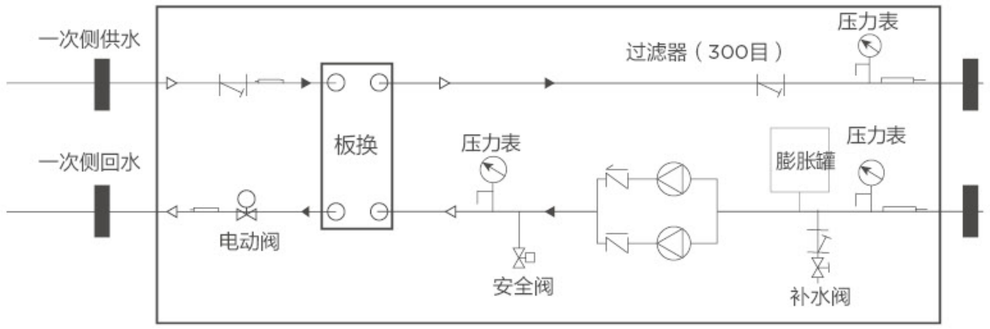


液冷服务器配套制冷系统设计方案

图所示为设计方案，中间换热单元提供一定流量、一定温度的冷却水进入液冷服务器的换热芯体，带走 CPU 发出的热量，被加热的冷却水流回中间换热单元的板式换热器，将热量释放到室外侧循环水，该部分热量再通过室外干冷器或者冷水机组释放到室外环境中，完成对液冷服务器的热量管理。

5. 中间换热单元方案

中间换热单元在二次侧通过调节送入液冷服务器换热芯体的水温和水流量从而向末端机柜提供冷量，起到冷量分配的作用，此外为了便于实施最佳的室外侧方案及保持液冷服务器换热芯体内的水质，中间换热单元也起到系统隔离的作用。



中间换热单元方案

6. 中间换热单元系统设计

中间换热单元主要包括二次侧板式换热器、循环水泵、过滤器、膨胀罐、阀件、管路及电气和控制系统，如所示。

7. 二次侧板式换热器

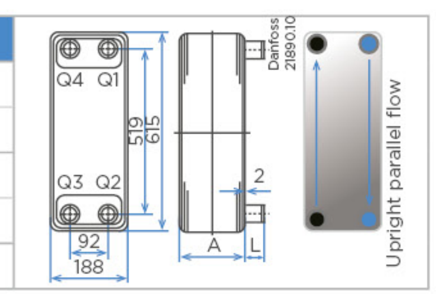
板式换热器选用钎焊式，采用国际知名品牌，不低于 SWEP、Danfoss。

为保证纯水水质，板式换热器采用 316L 不锈钢材质、换热器与外管路采用螺纹连接。

板式换热器表面敷有保温棉，消除壁板结露现象。

为了最大限度的使用自然冷源，中间换热单元按照 45℃出水、55℃进水进行选型计算，2 台机柜的散热量为 57kW（假设单机柜 33.5kW 的发热量 85% 全部被冷却水带走）。

Item	二次侧	一次侧
进水温度 ℃	55	43
出水温度 ℃	45	48
水流量 L/min	83	175
总压降 kPa	2.5	10.7
换热量 kW	68	



中间换热单元板换选型（一次侧工质为 30% 乙二醇水溶液）

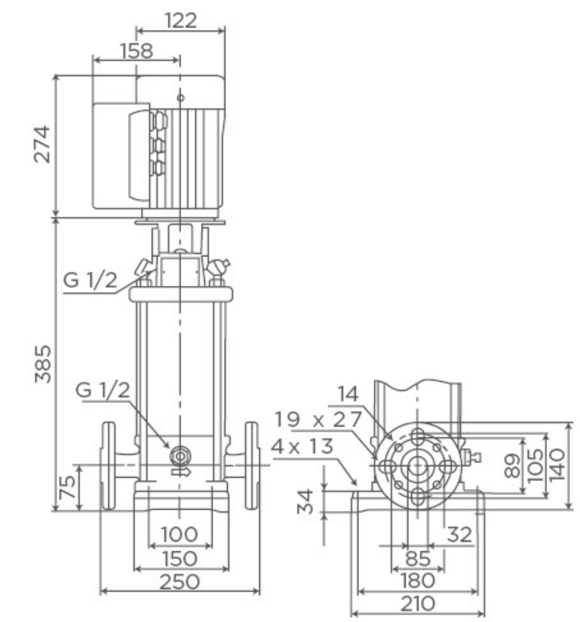
8. 二次侧循环水泵

采用一主一备双泵设计。

水泵采用高效、节能、紧凑的离心式清水泵。

水泵选用国际知名品牌，水泵电机可调速。

设计流量为 5.1m3/h，克服软管（假设 30m）、过滤器、机柜等器件的阻力，设计扬程为 33m。



二次侧循环水泵

9. 电动调节水阀

电动调节阀采用国际知名品牌：Siemens、Belimo、Johnson Controls 等品牌。

电动调节阀采用等百分比球阀形式。

执行机构应采用 24V AC/DC 供电，0.5 ~ 10VDC 或 4 ~ 20mA 信号控制，具有位置信号（0 ~ 10VDC 或 4 ~ 20mA）反馈。调节精度高，稳定可靠。

10. 二次侧膨胀罐

膨胀罐选用隔膜式膨胀罐，作为定压膨胀装置，提供系统定压、缓冲水泵启停产生的水锤、容纳水温变化导致的管路内容积变化。

膨胀罐选用国际知名品牌，不低于 Aquasystem 和 Varem。

11. 过滤器

过滤器选用 Y 型过滤器，不低于埃美珂、OR、天津格莱式等品牌。

12. 橡胶软管

橡胶软管材质应为三元乙丙橡胶（EPDM）。

橡胶软管满足系统可能出现的最高压力。

橡胶软管采用国际知名品牌：Paker、Hansa、Manuli 等。

13. 中间换热单元控制系统

中间换热单元采用艾默生机房空调专用控制器。

中间换热单元控制系统具有信号采集、逻辑计算、自检告警等功能，以保证系统的正常运行。

中间换热单元具有人机操作界面、数据通讯传输等交互途径，提供快捷的控制方式。

14. 信号采集

一次侧进出水温度采集；

二次侧进出水温度采集；

二次侧进出水压力采集；

漏水检测采集；

远程温湿度采集；

电动调节水阀开度采集。

15. 二次侧出水温度控制

通过对一次侧电动调节水阀的控制，调节一次侧的水流量，使二次侧的出水温度达到设定的要求。

16. 二次侧水流量控制

通过二次侧供回水压差、温差及进水温度对二次侧循环水泵的转速进行控制，向末端服务器提供充足的水流量。

17. 防凝露控制

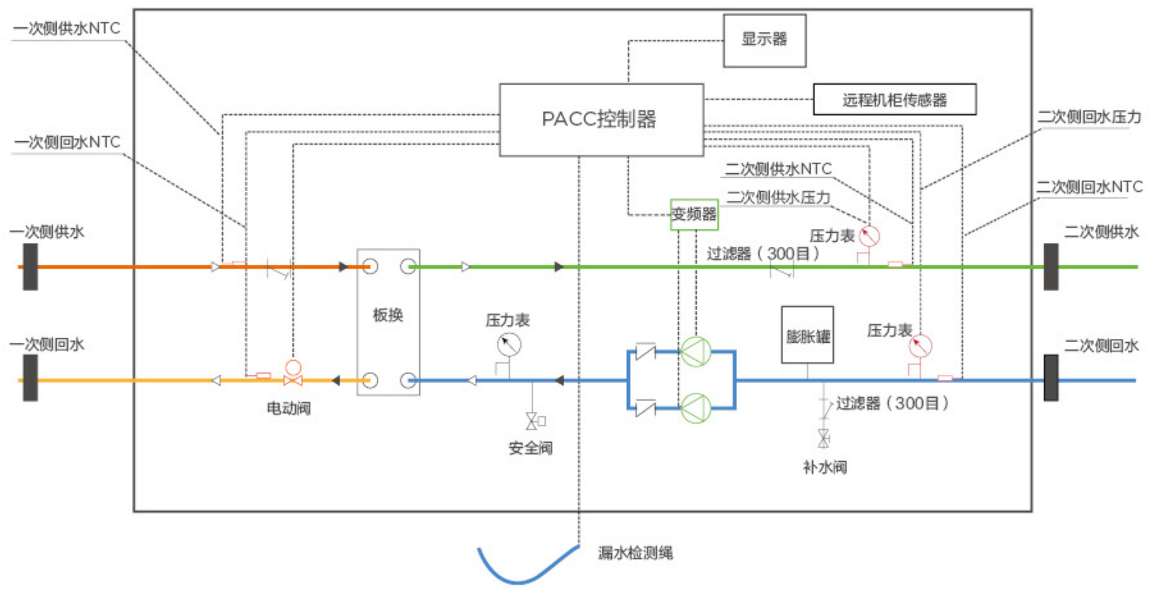
通过采集服务器周围环境的温湿度，实时监测出水温度使其高于露点温度。

18. 告警

可设置高低水温告警、漏水告警、泵故障告警等多种属性告警。

19. 水泵主备切换

当检测到运行水泵故障告警时，自动切换至备用水泵运行。



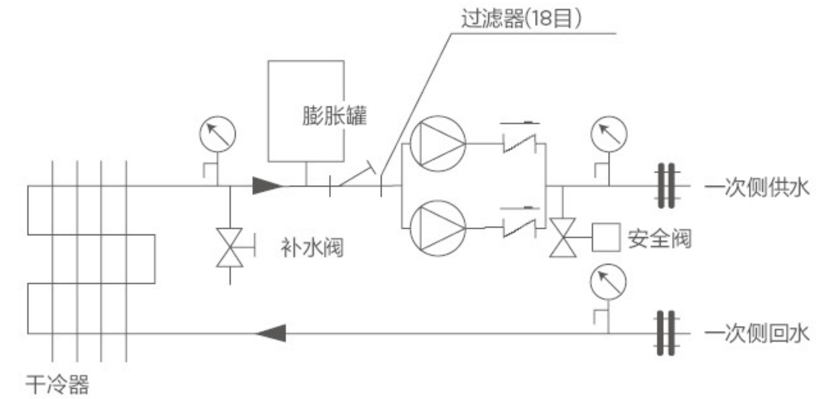
中间换热单元控制系统

20. 室外散热方案

室外散热系统将中间换热单元传递出来的热量释放到室外环境中。

21. 室外散热系统设计

室外散热系统主要包括干冷器、一次侧膨胀罐、水泵、过滤器、管路、电动调节水阀及电气和控制系统。如图所示。



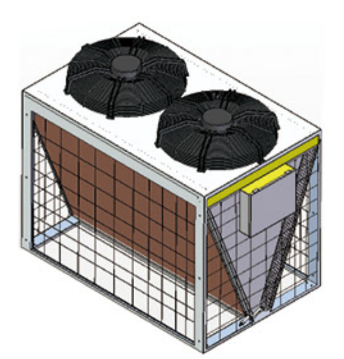
室外散热系统

22. 干冷器设计

按照室外温度 35℃条件下能够提供满负荷散热量进行设计计算。

40℃室外温度干冷器设计（30% 乙二醇水溶液）

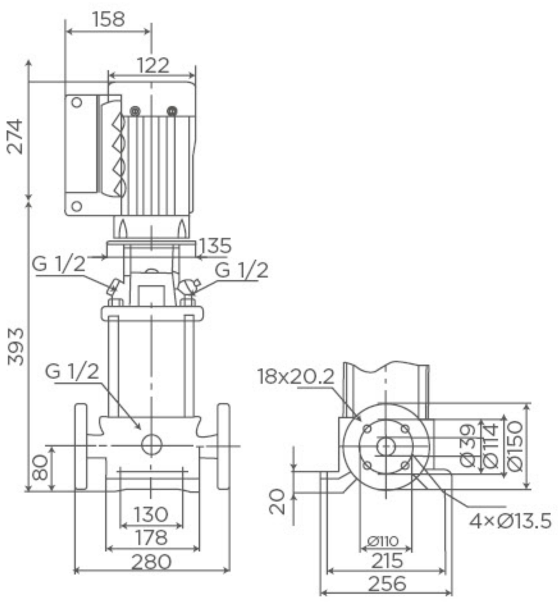
Item	空气侧	水侧
进水（风）温度 ℃	40	48
出水（风）温度 ℃	43.6	43
流量	54,000m³/h	203L/min
换热量 kW	65.3	



干冷器

23. 一次侧循环水泵

采用一主一备双泵设计。设计流量为 12.2m³/h，克服管路、过滤器、干冷器等器件的阻力，设计扬程为 22m。其他要求同二次侧水泵。



一次侧循环水泵

24. 室外散热控制系统

室外散热系统的运行逻辑要尽可能的使用自然冷，以二次侧的出水温度为控制目标，一次侧循环水泵、干冷器风机、水阀按照由高到低的优先级进行调节，尽可能的降低能耗。

25. 水质处理

二次侧循环水需要线监控水质、PH、硬度、腐蚀率等参数，一旦超出水质要求，就定期更换或者采用一部分水旁滤并进行补水的方法使总体水质满足要求。

对二次侧系统进行补水时，先经过一定的水处理，经过软化、去离子、反渗透（RO）等技术处理，由专业的水处理设备实现。

26. 干冷器和闭式冷却塔的对比

干冷器和闭式冷却塔都是典型的室外散热设备，最本质的不同在于干冷器直接与室外空气进行热交换，而闭式冷却塔是在干冷器的基础上增加了一套喷淋设备，利用喷淋水的蒸发起到增强换热的作用。鉴于该特点：

A. 闭式冷却塔需要专用的喷淋水供水系统，包括水泵、管路等，以使其能利用源源不断的喷淋水来制冷，同时该系统还需要水处理设备，以保证喷淋水不会在换热器上产生结垢、腐蚀等问题。因此闭式冷却塔的优点在于强化了换热，在室外温度较高时也能带来相当的制冷量，尤其是在炎热干燥的地区效果更明显，但相应的也带来了明显的缺点，冷却塔需要持续供应喷淋水，需要有水源，并且也产生了用水的费用，如果水质达不到要求，则需要定期的除垢换水，日常的维护工作多，控制较复杂，另外在冬季低温时，还要将喷淋系统放水防冻。

B. 干冷器没有喷淋系统，因此不存在由此带来的工程施工、额外设备投入等问题，不需要提供水源以及不会产生持续的用水费用，干冷器日常的维护也较简单。但由于干冷器没有喷淋水强化换热，冷源仅来自于与空气的强制对流换热，在室外温度较高时，需要更大的换热器才能达到更大的制冷量。

对于液冷服务器的散热需求，从应用上来看，当机柜数量较少即散热需求较小时，适合使用干冷器，投入小，施工维护简单；当机柜数量较多即总体散热需求较大时，需根据应用场地的环境温度、水源有无、占地面积等综合考虑选择冷却塔或是干冷器。

模块	设备名称	参数
中间换热单元 系统部分	板式换热器	68kW（一次侧 55/45℃，二次侧 43/48℃）
	变频水泵	5m³/h，330kPa　2-9m³/h 调节
	膨胀罐及安全阀	2L，泄压阀
	补水阀	1 1/2
	止回阀	DN40
	水过滤器	300 目（50μm），DN40
	水过滤器	18 目，DN65
	球阀及执行器	DN50 螺纹，Kvs=40m³/h
	橡胶软管	EPDM，1”，15m
	快速接头	1”
中间换热单元 控制部分	控制器 & 显示器	PACC 控制器
	NTC	一次侧供回水，二次侧供回水
	远程温度传感器	1# 机柜，2# 机柜
	压力传感器	二次侧供回水
	漏水检测绳	
	配电器件	主空开、变频泵空开 / 接触器、汇流排、成套线缆
室外干冷器	风机	27000m³/h，50Pa
	翅片管换热器	
室外水力模块	水泵	12.2m³/h，220kPa
	单向阀	DN65
	补水阀	1 1/2
	膨胀罐及安全阀	8L
	水过滤器	18 目，DN65
	压力表	
室外控制系统	控制板	PACC 控制器
	温度传感器	室外温度
	配电器件	主空开、泵空开 / 接触器、风机空开 / 接触器、冷水机组空开 / 接触器、汇流排、成套线缆

关键设备清单

联想高性能计算软件

联想智能高性能计算平台 LiCO

联想LiCO (Lenovo Intelligent Computing Orchestration) 智能高性能计算平台是联想企业科技集团基于用户实际需求和联想大量行业应用经验所开发的一套超算中心管理软件，同时 LiCO 是联想高性能计算产品 x9000 的核心。LiCO 是联想 HPC 的一站式解决方案，适用于各种规模的高性能集群，使用 LiCO 可以快速安装部署好一个 HPC 集群，同时 LiCO 针对管理员和普通用户提供易用的管理平台。

软件平台由节点操作系统、集群软件系统、编程环境、应用软件、存储系统等 6 个子系统组成。


中华人民共和国国家版权局
计算机软件著作权登记证书

联想智能高性能计算平台 LiCO-HPC 包含但不仅限于以下功能:

部署功能

集群管理软件平台系统应具有批量部署操作系统的功能，以方便同时部署多个服务器的操作系统。

集群管理软件平台系统应具有批量更改系统配置的功能，比如根据规则设置主机名、IP 地址等。

集群管理软件平台系统应具有备份操作系统或者操作系统配置的功能，以便需要重新部署环境时，能够迅速恢复原有环境。

集群管理软件平台应为图形化的使用方式，以方便系统管理员使用。

监控功能

集群管理软件平台应具有对硬件系统的监控功能。

对于服务器，应监控到以下指标:

① 系统整体运行状态

② CPU 数量、内存数量、网络状态、硬盘状态

③ 系统运行温度，包含但不限于 CPU、内存、硬盘

④ 系统状态异常情况

控制功能

集群管理软件平台应具有远程控制功能，控制功能主要包含:

① 远程开关机 (上电、下电) 操作

② 动态调整 CPU 的运行频率

③ 动态调整风扇的运行速度

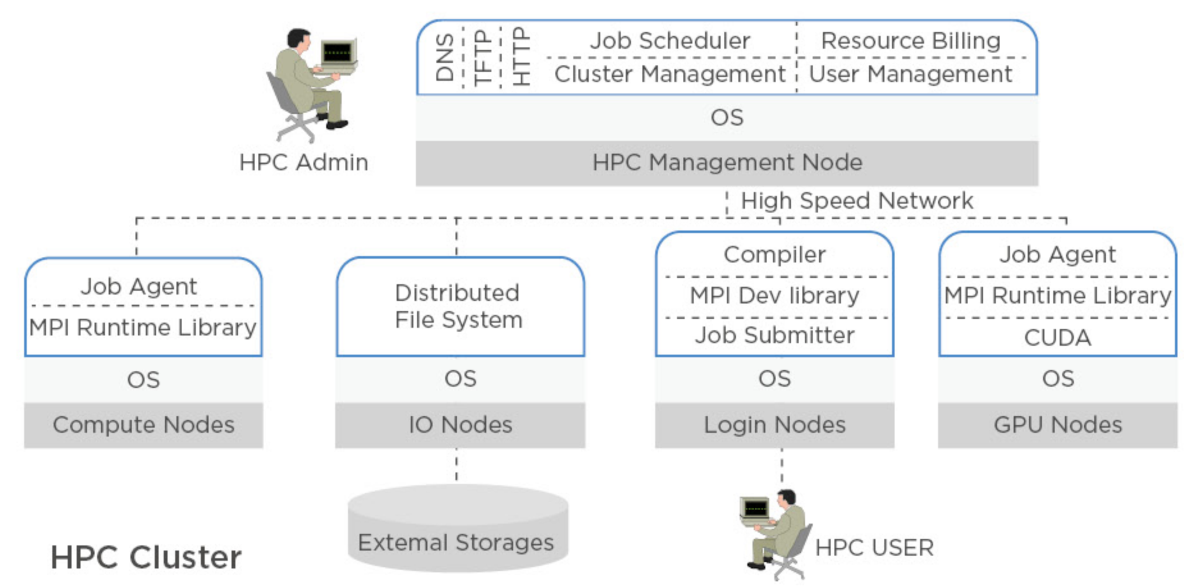
④ 动态调整能耗，以使系统运行在兼顾性能和节能的状态下

分布式文件系统管理功能

分布式文件系统的创建，磁盘空间管理，存储节点 HA 配置，文件的操作。

二次开发能力

所有 HPC 软件接口直接向用户开放，用户可以非常方便的要求定制或进行二次开发。LiCO 解决方案架构示意如下:



LiCO 解决方案示意图

针对智能计算平台的联想 LiCO-AI

随着互联网的普及和 IT 业的高速发展，高性能计算（HPC）已经不再是少数大公司或大型科研机构的专属要求，而是被越来越多的包括政府，教育科研，石油石化，制造，军工和生命科学类的客户所需要和接受， 针对 HPC 方案的复杂性，用户需要一个易用的 HPC 集群管理和作业管理平台。

同时，最近几年人工智能（AI）快速发展，大多数人工智能模型的训练需要使用 GPU，而如 NVIDIA Tesla A100 等 GPU 通常比较昂贵，不同的部门和人员有不同的模型训练需求，为每个部门和人员购买独占的 GPU 是对资源的浪费，所以一个集中共享式的 AI 模型训练平台也成为了越来越多客户的需求。

联想智能高性能计算平台 AI 增强版本（Lenovo intelligent Computing Orchestration 以下简称 LiCO）是联想企业科技集团（ISG）开发的，针对高性能计算（HPC）和人工智能（AI）的一站式解决方案，在一套集群中通过统一的资源调度，可以同时支持 HPC 作业和 AI 作业的运行。

LiCO-AI 集成了集群需要的集群调度软件、监控软件、计算库、分布式文件系统等，使用 LiCO-AI 可以快速的部署好一个 HPC 和 AI 集群。

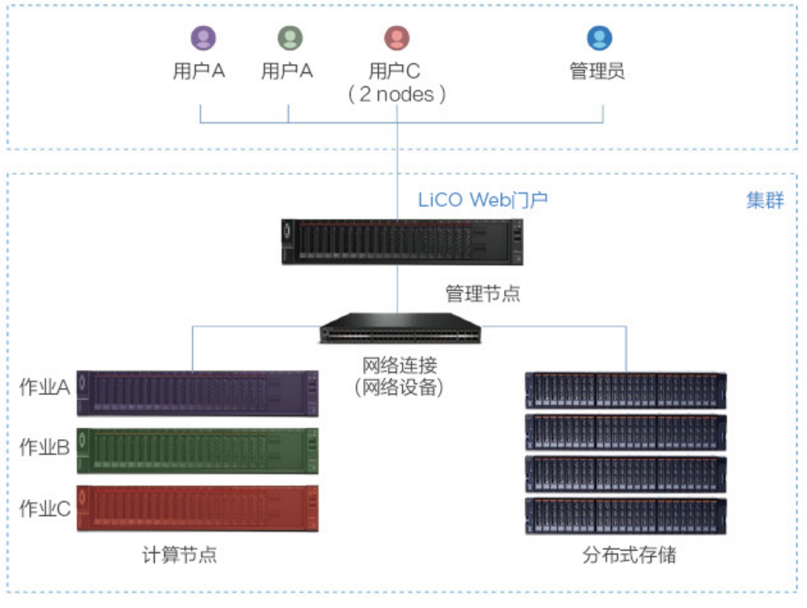
LiCO-AI 提供了统一的 web 访问接口，集群管理员可以使用 LiCO-AI 方便的管理集群，HPC 用户可以使用 LiCO 方便的提交和管理 HPC 作业，AI 用户可以使用 LiCO-AI 进行 AI 模型的训练。（以下介绍 LICO-AI 简称 LiCO）

在如下的一个物理集群上：

LiCO 可以作为一个 HPC 的平台；

LiCO 可以作为一个 AI 模型训练的平台；

LiCO 可 以 作 为 一 个 HPC+AI 模型训练的平台。



LiCO 平台提供了统一的 Web 门户：

- 集群管理员可以使用 LiCO 管理集群；
- HPC 用户可以使用 LiCO 提交和管理 HPC 作业；
- AI 用户可以使用 LiCO 进行 AI 模型的训练。
- LiCO 平台提供了开放 API。

节点操作系统

联想高性能集群系统中的服务器节点支持的操作系统有：Red Hat Enterprise Linux、CentOS、SUSE 、Windows Server。

集群软件系统

整个集群的部署、监控、管理和计算资源的分配和计算任务的调度。集群软件系统需要包括集群部署、集群监控、集群管理、作业调度等几个主要的功能模块。这些功能组件合力对集群系统的软硬件资源整合管理，使得用户无需关心某一资源单元，而只要集中于生产应用即可。

1. IBM Platform 解决方案

原 Platform 公司拥有 60% 世界 500 强的用户，用户行业遍布航空航天，电子，制造，教育科研，生物，海洋，能源等。

IBM（国际商业机器）公司于 2011 年 10 月 11 日宣布收购 Platform 公司。位于加拿大多伦多、中国北京和西安的研发团队，也全部并入 IBM 国际商业机器（中国）投资有限公司。于 2012 年 7 月 1 日，IBM 完成了与 Platform 公司的整合，完成了 IBM Platform 这个产品品牌的简历，这次强强联合，使 IBM 拥有了在高性能计算领域从硬件到软件的整体解决方案。2014 年 10 月，联想完成了对 IBM System X 的收购，从此联想转售 Platform 软件产品。

IBM Platform 产品系列包括多款适用于 HPC、大数据、云计算等各大领域软件产品系列，如 Platform Cluster Manager，用于部署、监控和管理 HPC 集群；Platform LSF，强大的负载管理平台等。

其中 Platform LSF 是 HPC 领域非常权威和应用广泛的一款产品。Platform LSF 产品家族实现了软硬件资源共享调度，将所有软硬件资源有机地组合在一起，根据事先定义的调度策略，统一管理，结合 Platform LSF 产品家族成员，建立先进的高性能计算平台，提高软硬件资源的利用率。

什么是IBM Spectrum Computing LSF Suite?

IBM Spectrum LSF套件提供了高度可扩展、高可用的体系结构，
支持传统的高性能计算和高吞吐量工作负载，以及大数据、感知、GPU机器学习和容器化工作负载。

Suite for Workgroups

Up to 128 servers，25K jobs per server/user，subscription/perpetual

Suite for HPC

Up to 1024 servers.250K jobs per server/user，subscription/per

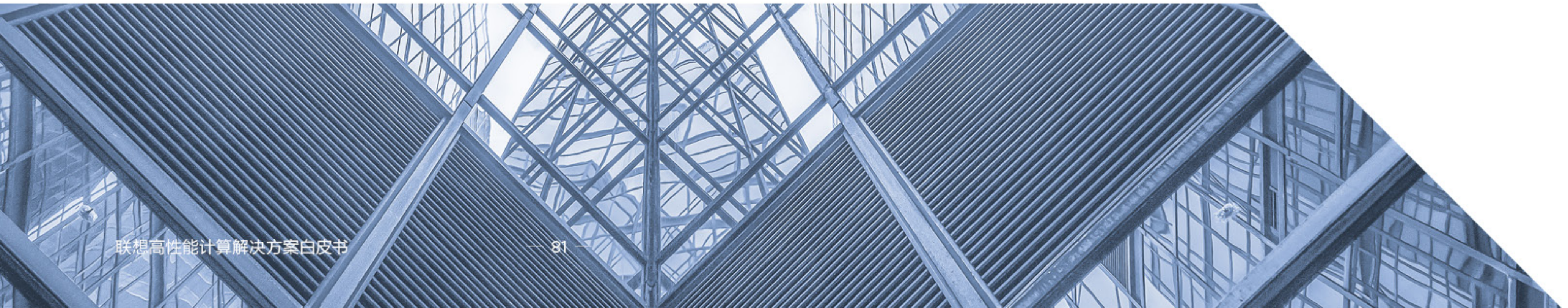
Suite for Enterprise

Unlimited servers and jobs per server/user，subscription/perpetual

IBM Spectrum LSF

	IBM Spectrum LSF Community Edition	IBM Spectrum LSF Suite for Workgroups	IBM Spectrum LSF Suite for HPC	IBM Spectrum LSF Suite for Enterprise
Workload Scheduling	✔	✔	✔	✔
Container Support (Docker, Shifter, Singularity)		✔	✔	✔
Web Portal	✔	✔	✔	✔
Remote 3D Visualisation Support		✔	✔	✔
Monitoring & Reporting		✔	✔	✔
Web Services/Restful API		✔	✔	✔
Mobile Support	✔	✔	✔	✔
Systems Monitoring	✔	✔	✔	✔
Bare Metal Provisioning	✔	✔	✔	✔
Receive workload from other clusters			✔	✔
Cloud Scaling Policies (Amazon EC2, Google Computing Engine, IBM BlueMix, Microsoft Azure, OpenStack)			✔	✔
Job Dependencies	✔	✔	✔	✔
Graphical computational workflows			✔	✔
Transfer of files before and after job execution	✔	✔	✔	✔
Intelligent pre/post staging of data			✔	✔
Deduplication of data transfers			✔	✔
Scheduling based on the availability of ISV software licenses	✔	✔	✔	✔
ISV License Allocation and Prioritisation Policies				✔

Platform 功能列表



Platform LSF 是一个管理和加速高性能计算的系统软件、尤其对用户的关键任务和作业提供运行保障。借助 Platform LSF，将能够智能地调度并行和串行作业，轻松解决大型、复杂的计算问题，同时最大限度地利用可用的计算资源。Platform LSF 全面地利用了高性能计算机的高性能网络互连资源，让用户得到利益的最大化。Platform LSF 是一个经过实践证明的可靠的解决方案，在 TOP500 高性能计算项目中，目前，在排名前 10 的超级计算机都安装和运行了 Platform LSF。

Platform LSF 主要特点如下：

基于拓扑结构的调度算法 — 具有一致性的、更优的性能

Platform LSF 与系统厂商和第三方具有高性能互连结构的应用进行了高度集成，以期达到更佳性能的负载调度。基于拓扑的调度方法可根据每个处理器间的高性能互连的特性来确定哪些处理器的连接距离最短。这保证了每次运行一个应用时都能确保最佳的运行性能和一致性。

IT 资源虚拟化 - 自我管理和可靠的服务

用 Platform LSF，用户可共享虚拟的 IT 资源池，无需为每个批量作业指定特定的 IT 资源。Platform LSF 的自适应功能能够动态地分配资源，以适应不断变化的 IT 基础设施，减少了管理成本和相关成本。

采用强大、可靠、基于标准的高级技术

Platform LSF 是创建在经生产证明的、开放的、具有网格能力的虚拟执行机（VEM）结构上，拥有 11 年的分布计算经验，可提供强大、可靠的负载管理解决方案。Platform LSF 最大限度地利用了全球分布的异构计算资源。Platform LSF 的机遇商业的高级技术提供了无可比拟的可扩展性和性能，为企业组织提供了他们所需的速度和灵活性，使他们能够满足预期的功能和项目要求。

智能的、由策略驱动的 HPC 负载处理

通过 Platform LSF 管理和加速高性能计算（HPC）中重要的、计算密集和数据密集的负载。以智能方式对并行和串行负载进行排程，以便克服主要困难，并最大限度地利用计算资源。Platform LSF 还提供了丰富的用户工具，如 PAC，WEB Portal，集成应用后可以直接从网页提交作业；Reports，提供报表功能，供用户以报表形式查看资源使用情况。

2. GPFS 并行文件系统介绍

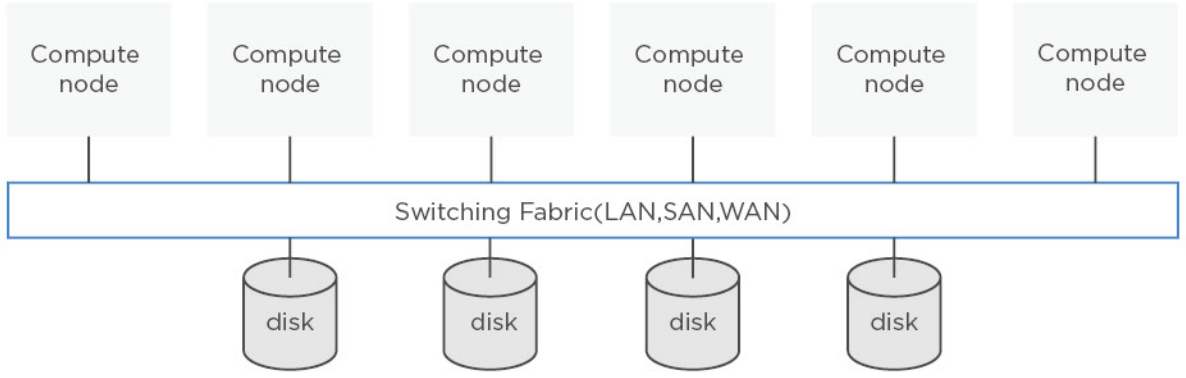
GPFS（General Parallel File System）是并行分布式通用并行集群文件系统，GPFS 从 1993 开始研发，1995 年投入商用（最早投入使用在多媒体处理行业，所以我们现在看到 GPFS 的很多目录和命令为什么都有 MM[multimedia] 两个字母，就是根据这个背景延续下来的），其中经过了多次版本的改进和修订，到 2013，GPFS 的版本为 3.5。到 2015 年，GPFS 已经到 4.2 版本目前 GPFS 的版本已经全面升级到了 V5.x。

GPFS 是一个共享的磁盘文件系统，集群内的所有节点 (NODE) 可以并行的访问所有共享磁盘，并通过分布式的令牌管理机制和条带化技术来管理和优化各个节点的访问。

GPFS的应用范围非常广泛，从多节点的文件共享服务到大型的高性能计算机群，都可以看到GPFS的优秀表现。GPFS 也应用在众多的企业应用系统中，比如高可用的解决方案：HACMP，ORACLE RAC，SAP 的共享文件系统等，GPFS 在这些应用里面都表现出了出色的性能和高可用性。

1. GPFS 是可扩展的、并行化的、高可用的、高性能的文件系统

- 群集：数千个节点、快速可靠的通信、通用管理域；
- 共享磁盘：用户可在分布式锁定服务的协调之下从任何阶段接入磁盘上面所有的数据和元数据；
- 并行：数据和元数据在所有节点及所有磁盘之间并行流动；文件分布在所有磁盘上。



2. GPFS 是高可扩展的：（实际超过个 20000 个节点）

- 对称的、可扩展的软件架构
- 允许轻松地逐渐扩展系统（节点和磁盘空间）
- 分布式元数据管理

3. GPFS 是高性能的文件系统

- 大规模数据块（可调）支持与（跨越所有节点和磁盘）的广泛分布。对于高带宽环境来说，GPFS 可通过单次操作读 / 写大块数据，从而最大限度地降低 I/O 运营开销；
- 从多个节点并行接入文件；
- 全面的令牌增强（扇区范围）和令牌管理；
- 高效的深度预提：预读和迟写。GPFS 能够从缓冲池中智能地预提数据，从而并行向足够多的磁盘发送 I/O 请求，以便充分利用基本的存储硬件基础架构的带宽；
- 识别接入模式（自适应机制）。GPFS 可识别多个 I/O 模式，包括顺序、逆序及其他各种形式的分布式接入模式；
- 高效的多线程后台程序；
- 面向 MPI-IO 及其他应用的数据迁移模式。

4. GPFS 是高可用的和容灾的

- 支持多个数据保护机制，包括日志记录、复制、镜像和影子等（都是标准的文件系统技术）；
- GPFS 可配置用于消除单点故障。您可将这个文件系统配置成在磁盘或服务器发生故障时自动保持可用性；
- 心跳机制可在发生多个磁盘、节点和连接故障时恢复正常运行。各层均已实施了恢复软件机制。

5. GPFS 提供明显的性能和扩展优势

- 具有公认可扩展性的单一文件系统；
- 专门开发用于高性能超级计算群集应用；
- 非常成熟和稳定 – 最基本的基础架构能够支持包含 20000 多个节点的群集，提供卓越的性能及稳定性；
- 当与 SSD 一起使用时，可提供巨大的性能优势；
- 当主数据丢失时可继续存活，自动使用复制好的数据；
- 群集文件系统，群集可在节点丢失时继续存活；
- 支持高可用性解决方案。

3. Lustre 并行文件系统介绍

Lustre 由三个部分组成：客户端（Client）、对象存储服务器（Object Storage Target，OST）和元数据服务器（MetaData Server，MDS）。

客户端通过标准的 POSIX 接口向用户提供对文件系统的访问。对于客户端而言，Lustre 是一个透明的文件系统。它无需知道具体数据所在的位置，可以透明地访问整个文件系统中的数据。

OST 负责实际数据的存储，处理所有客户端和物理存储之间的交互。这种存储是基于对象（Object -based）的，OST 将所有的对象数据放到物理存储设备上，并完成对每个对象的管理。OST 和实际的物理存储设备之间通过设备驱动程序来实现交互。通过驱动程序的作用，Lustre 可以继承新的物理存储技术以及文件系统，实现对物理存储设备的扩展。

MDS 负责元数据的管理。MDS 构造、管理文件分布的视图，允许客户端访问对象。通过 MDS 的文件和目录访问管理，Lustre 可以控制客户端对文件系统中文件的创建、删除、修改以及对目录的创建、删除、修改等访问控制。通过 MDS，客户端得到数据所在的 OST，并与其建立连接，此后的读写操作就在客户端同 OST 之间进行，除非有对命名空间的修改，将不再同 MDS 有关系，这样就降低了 MDS 的负载。在多个客户端的情况下，由于有多个 OST 存在，上述的工作模式就把对文件系统的访问转换为并行操作，从而可以较好地提高性能。

对象存储结构的优点：

性能：对象存储体系结构为计算节点提供可以直接、并行访问存储设备的方法。通过把文件数据 stripe 到多个 OST 上实现并行访问。

传统存储结构的元数据服务器有两个主要功能。（1）负责为计算节点提供存放数据的逻辑视图（vfs 层），提供文件列表以及目录的组织结构。（2）组织数据在物理存储媒介上的存放方式（inode 层）。

对象存储结构把存放数据的逻辑视图和物理试图分开，把 inode 层的工作分给具有处理能力的智能 OST 存储设备。实现了元数据的分布存放，避免了传统元数据服务器的性能瓶颈问题。

可扩展性：将负载分布到多个智能的 OST，并用网络和软件将它们有机结合起来，消除了可扩展问题。一个对象存储系统有内存、处理器、磁盘系统等，允许它们增加其存储处理能力而与系统其它部分无关。如果对象存储系统没有足够的存储处理能力，可以增加 OST，确保线性增加性能。

易管理性：智能化的分布对象存储结构可以简化存储管理任务，可以简化数据优化分布的任务。例如，新增存储容量可以自动合并到存储系统中，因为 OST 可以接受来自计算结点发出的对象请求。系统管理员不需要创建 LUN，不需要重新调整分区，不需要更新文件服务器等。RAID 块可自动扩展到新的对象，充分利用新增的 OST。

安全性：对象存储结构在每个级别都提供安全功能，主要包括存储设备的身份认证，计算结点的身份认证，计算结点命令的身份认证，所有命令的完整性检查，基于 IPSec 的私有数据和命令等。

高性能文件系统结构：文件系统由三部分构成，包括客户端 Client、元数据服务器 MDS 和存储服务器 OSS。元数据节点和 OSS 节点接入 InfiniBand 网络。

基于 Lustre 构建文件系统，对特定应用采用相应的策略优化，并侧重于易管理性和可恢复性。完成的高性能文件系统具有如下主要特色：

- 基于对象存储，使存储更具智能化；
- 对单个目录下的元数据操作进行优化；
- 极佳的可扩展性。基于对象存储的体系结构使系统可以支持几万个客户节点，存储容量支持 Petabytes 级别；
- 可靠性。由于 Lustre 已经部署到多个不同规模大小的实际机群系统中，可满足商业和安全应用对正常运行时间的要求；
- 开源及开放的标准。Lustre 作为一个开源软件开发和维护，满足开放的网络协议和 Posix 文件系统语义，对工业标准平台和异构网络环境进行广泛支持；
- 支持多种主流的网络连接，包括 TCP/IP，Myrinet，InfiniBand，Qrandrics 等。

4. 应用软件

联想高性能计算系统可以集成和支持多个领域的多种商业和开源的科学计算应用软件，包括：商业应用软件如 MOLPRO、GAUSSIAN、Q-CHEM、MATLAB、VASP、TURBOMOLE、WIEN2K、SPARTAN、Materials Studio、Linda、LS-DYNA、FLUENT、CRYSTAL、ADF 等。

开源应用软件如 GROMACS、GrADS、ABINIT、GMT、CASTEP、CPMD、DL_POLY、DOCK、GRAPES、MM5、NAMD、AutoDock、VMD、TINKER、OpenMX、VENUS、WRF 等。

编程环境

为用户提供应用程序运行环境、用户编程环境以及集群系统软件；
集群编程环境主要是为用户提供软件开发环境，同时为集群软件系统提供运行环境。除了有很开源的集群编程环境项目以外，各硬件厂商会根据自己的硬件架构设计会提供整套优化的编程开发工具。主流集群编程环境如下所示：

- Intel OneAPI
- NVIDIA CUDA Toolkit
- PGI Compilers & Tools
- GNU
- MPICH
- OpenMPI
- OpenMP

运行环境

应用软件：各领域用户生产软件

软件系统层次图

04

联想高性能计算集群实施服务

此高性能计算集群实施和培训服务会将已经完成上架的各种硬件设备进行系统软件的安装和配置，从而真正成为一套可用于实际工作的集群系统。联想企业级业务集团专业服务团队的高性能计算实施工程师通过快速配置集群以满足客户的需求，并且针对集群日常管理所需的技能对客户进行培训。集群可由各种联想自有产品或联想 OEM 的产品组成，包括 ThinkSystem 机架式服务器、高密度刀片式服务器、GPU 服务器以及存储产品和网络产品等。联想企业级业务集团专业服务团队拥有专业的高性能计算领域的知识和技能，快速帮助客户将各种硬件组合成集群系统。

在项目的前期阶段就可以引入专业服务团队的 HPC 实施工程师，和客户进行充分的沟通和交流，了解客户的需求和想法，同时可以基于我们的实践经验，为客户提供最佳实践的意见和建议。而后针对客户的需求和想法，我们会撰写工作说明书，以明确实施服务的主要工作内容和范围。在和客户确认工作说明书后，我们就可以评估整个项目实施服务所需的工作量，从而提供报价给到联想的销售人员。

在高性能计算集群的硬件和服务下单后，我们会得到通知。实施工程师即可安排和客户沟通，制定具体实施计划以确保实施过程顺利。在实施阶段，根据集群规模大小，一个或者多个实施工程师会到达客户数据中心，完成所需的实施服务。在项目尾声，实施工程师会将正常工作的集群系统交付给客户，并且提供相应的安装实施文档以及交付培训。

实施速度——我们的实施工程师会凭借他们丰富的经验以最快的速度完成集群的实施工作。他们会使用经过多次验证的最佳实施经验以及软件组件版本，避免反复的试验和错误，快速让集群系统可以进入工作状态。

实施质量——我们的实施工程师不断的积累最佳实践方法，改进安装配置技术。他们还与联想的研发团队以及各地客户紧密配合，以帮助完善产品和实施服务的质量。

减少风险——我们的实施工程师会帮助您避免典型错误导致的性能问题、兼容性问题或保修问题。

选择联想的专业实施服务将帮助客户充分利用新系统的能力，而不会忽略一些重要的特性。客户不需要花费大量的时间去学习集群的知识，而是快速的从实施工程师的丰富经验中得到所需的知识和信息。

高性能计算集群实施服务通常包含以下内容：

每一个高性能计算集群实施服务都有特殊的地方，但是多数通常都包含以下内容：

- 一个准备和计划会议
- 管理节点的安装和配置。包含 RAID 配置、操作系统安装、BIOS/UEFI 设置、微码更新、群集管理软件安装（LiCO/xCAT）
- 以太网配置和验证
- 节点（计算节点、存储节点、登录节点、其它用途节点等）RAID 配置
- 向节点（计算节点、存储节点、登录节点、其它用途节点等）分发操作系统，并配置 BIOS/UEFI、微码更新、驱动和软件包更新

根据硬件配置不同的可选服务包含：

- 高速网络（InfiniBand 或高速以太网）的配置和验证，以及相匹配的软件包安装，例如 OFED
- 存储系统配置和验证
  - 在管理节点上安装存储管理软件
  - 根据存储需求配置多路径软件
  - 存储控制器或 SAN 交换机 Zone 配置
  - 创建 RAID、LUNs 和文件系统
- 特殊用途节点配置，如 GPU 节点

根据软件配置不同的可选服务包含：

- 资源管理 / 调度软件的实施，如 Torque/Maui, Torque/Moab, PBS-Pro, LSF, Platform HPC, SLURM 等
- 集群监控或报警软件实施，如 Ganglia, Nagios
- GPFS 并行文件系统实施
- 特殊用途工具、软件、库的实施，如 Intel OneAPI, Intel MKL, 特殊的 MPI bindings.

额外可选的附加服务包含：

- 协助解决方案的架构设计，审阅建议的解决方案
- 硬件测试：加电开机测试，包括 STREAM, PALLAS, IOzone, HPL 在内的基准测试
- 协助性能测试和验收测试
- 对于集群中各种组件的深入技术培训
- 协助客户进行应用软件和许可证管理器的安装
- 协助项目管理：实施进度安排、硬件和软件的订购和交付、预算管理

05

联想高性能计算成功案例

Leibniz-Rechenzentrum

der Bayerischen Akademmie der wissenshaften

比风冷系统节省 35% 的运行成本合计每年降低 2000 万欧元

一期（2009 年）:

- 每秒 3.2 千万亿次
- 9216 个节点

二期（2012 年）:

- 每秒 3.6 千万亿次
- 3072 个节点

三期（2017 年）:

- 每秒 26.7 千万亿次
- 6500 个节点

北京大学

- 机型：ThinkSystem SD650 支持 45℃温水
- 节点数量：196
- 节点配置：Intel Xeon E5-2687A V4,16GB*8,OPA 100G
- 性能：260Tflops
- 项目规模：
  - 水冷机柜 3 台，单柜 30KW
  - 风冷柜 6 台，单机柜功耗 2.5 ~ 8kW
  - CDU 305KW
- 冷源：
  - 闭式冷却塔，供回水温度 38/43℃
  - 列间空调按 2+1 冗余设计
- 配电：小母线

国内首套温水水冷服务器超算系统

- 稳定可靠运行超过 4 年
- MDC+ 温水水冷服务器
- PUE 值 1.08
- Linpack 92.59%

北京市气象局

- 机型：ThinkSystem SD650
- 节点数量：800
- 节点配置：Intel Xeon Gold 6240，32GB*12，HDR 200G
- 性能：2.396Pflops
- 机柜规划
  - 水冷机柜 16 台，单机柜功耗 28kW
  - 风冷机柜 8 台，单机柜功耗 6 ~ 10kW
  - CDU 2 备 1
- 冷源：客户提供
- 配电：小母线

- 采用水冷 MDC+ 温水水冷技术
- 水冷服务器及空调系统共用一套冷源，降低 30% 的机房投资成本
- Linpack %：70.2%