

华为TaiShan HPC解决方案 主打胶片

Intelligent Computing
HPC Solution Team



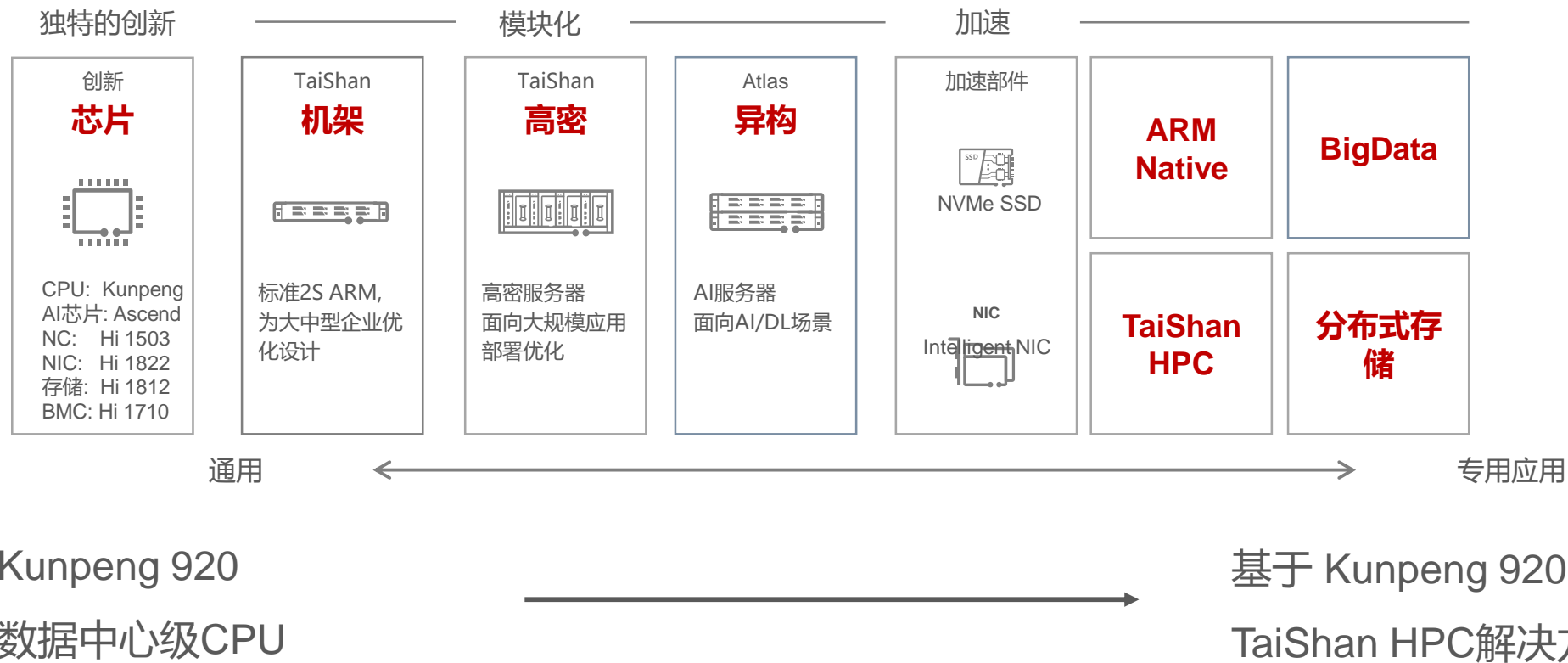
Security Level:



目录

1. 华为的持续创新
2. 华为TaiShan HPC解决方案
3. 华为TaiShan HPC方案亮点
4. 华为全面构建TaiShan HPC生态

持续创新，让计算变简单



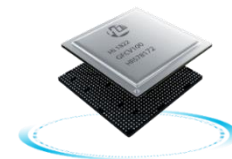
从“芯”开始，实现服务器芯片全面自研

昇腾AI芯片



华为Atlas 300 AI加速卡

智能网卡芯片



华为“IN”智能网卡

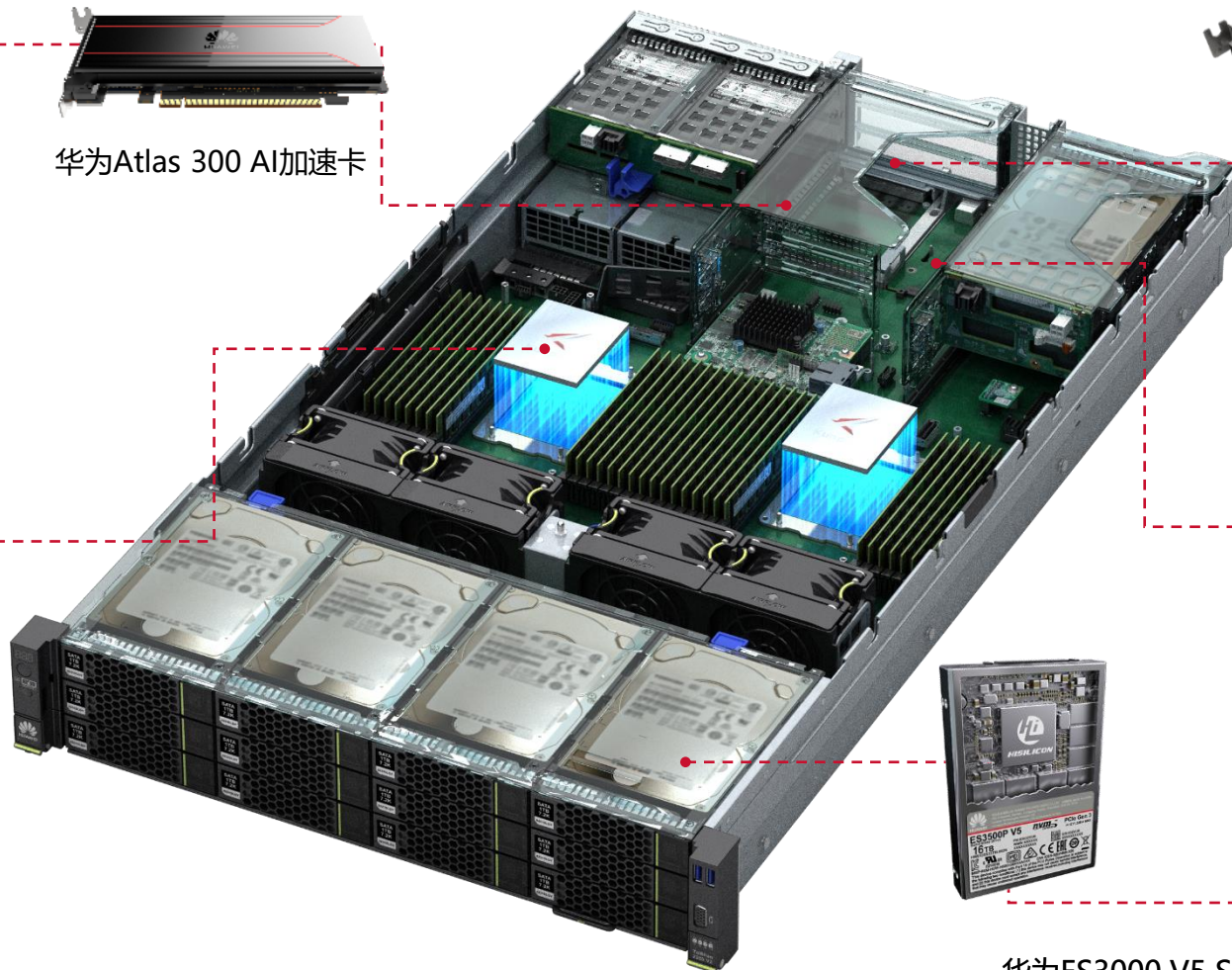
智能管理芯片



智能SSD控制器芯片

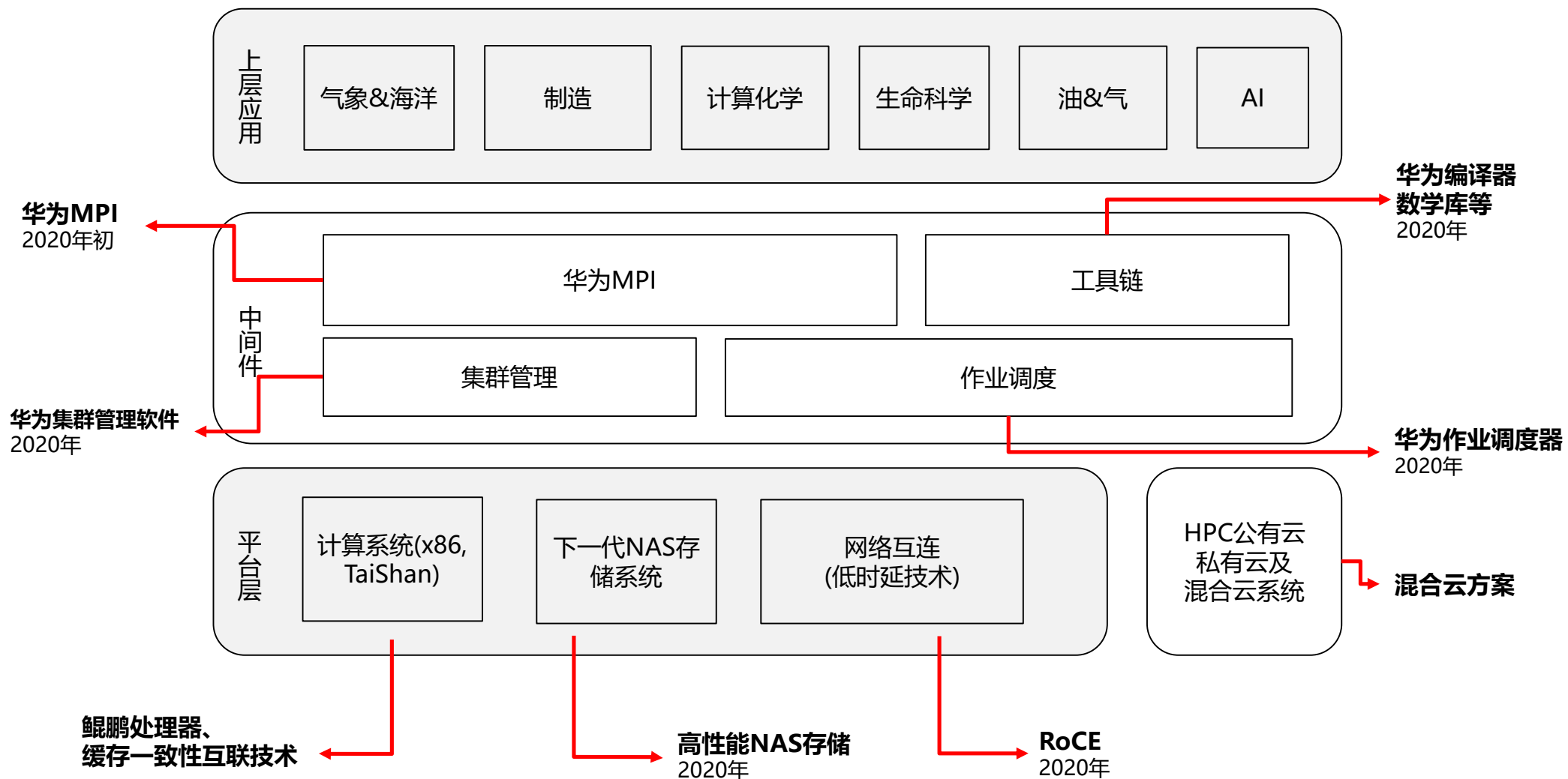


鲲鹏处理器



华为ES3000 V5 SSD

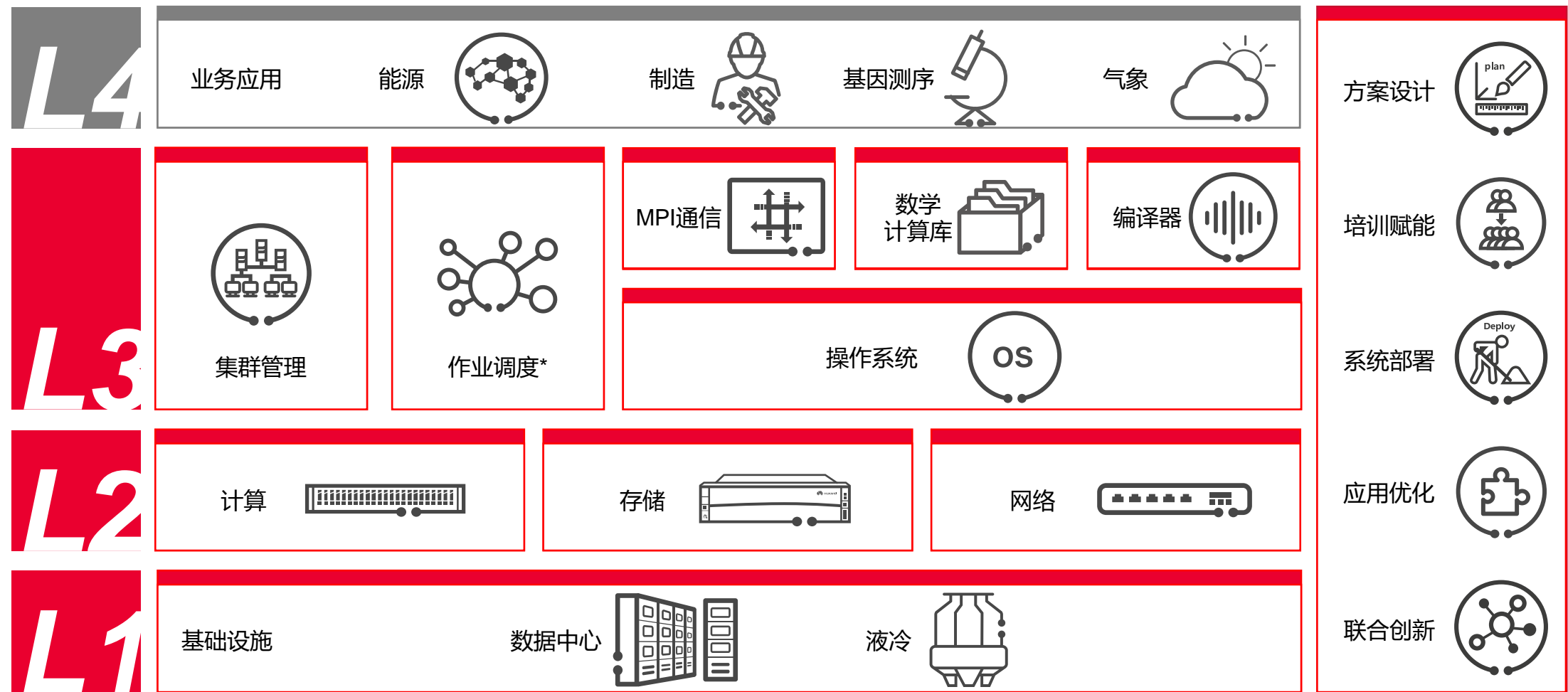
HPC领域长期战略技术投入



目录

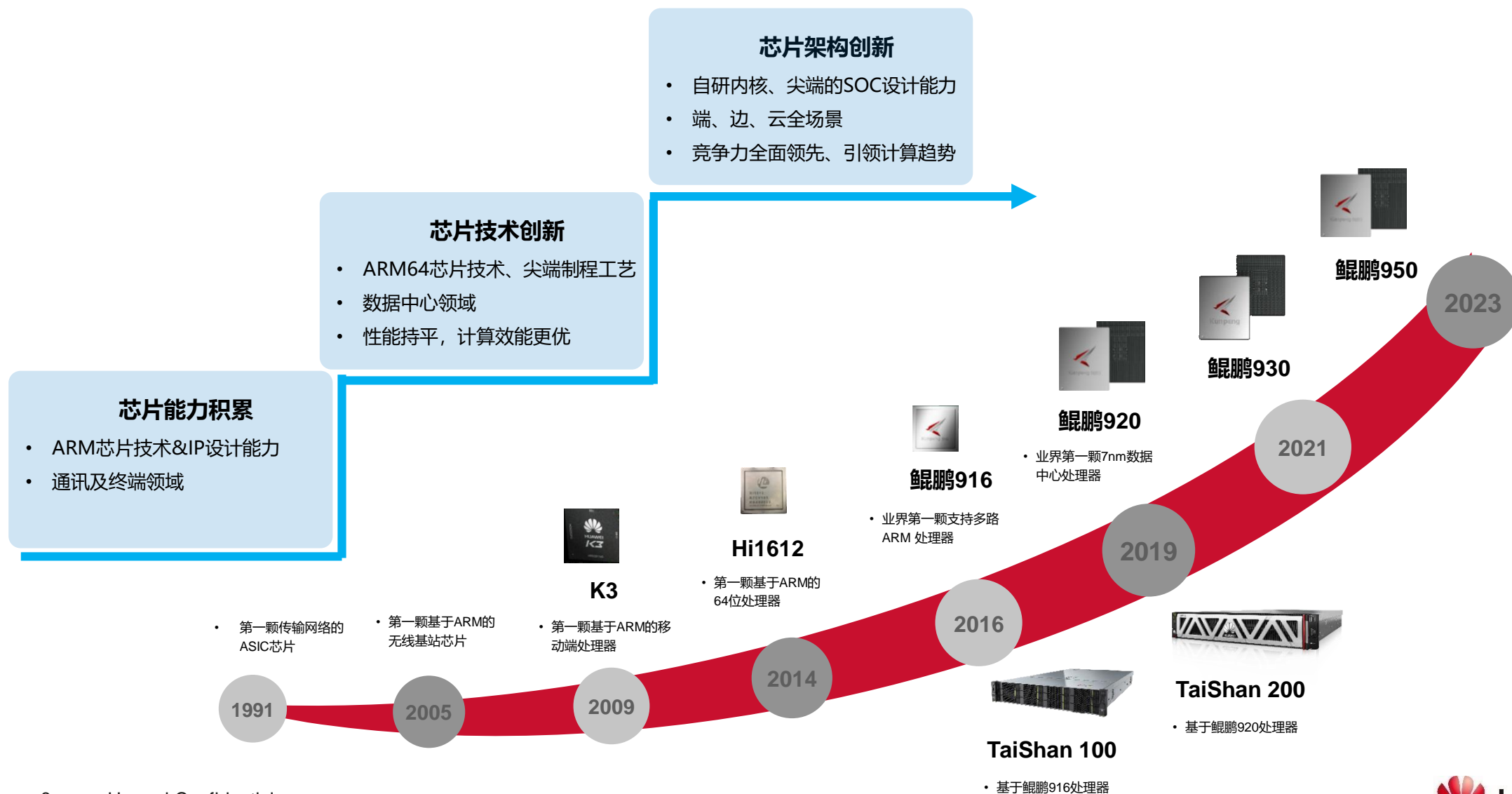
1. 华为的持续创新
2. 华为TaiShan HPC解决方案
3. 华为TaiShan HPC方案亮点
4. 华为全面构建TaiShan HPC生态

TaiShan HPC解决方案架构



*支持TaiShan/X86混合调度

创“芯”投入：鲲鹏处理器，华为长期、坚定的战略



鲲鹏920：数据中心高性能处理器

高性能

930+ **3x ↑**
SPECint®_rate_base2006 评估跑分

高吞吐

内存带宽: **2.4x ↑**
I/O 总带宽: **1.7x ↑**
网络带宽: **10x ↑**

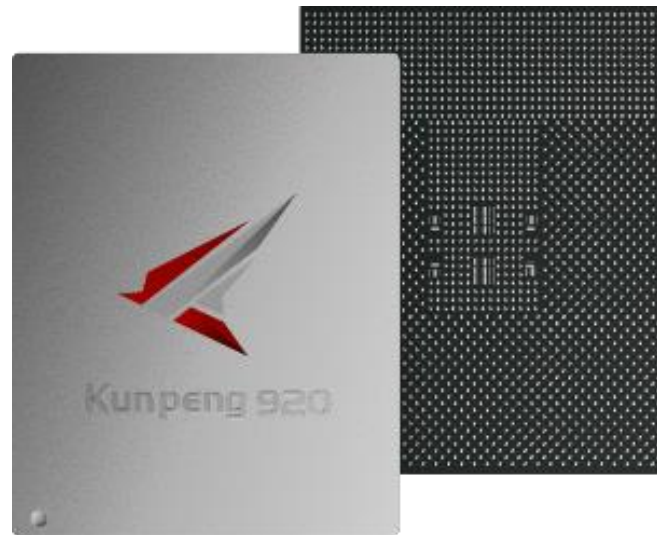
高集成

1 颗 = 4 颗芯片
(CPU, 南桥、网卡、SAS控制器)

高效能

35% ↑

*基于鲲鹏920-6426 vs 鲲鹏916处理器的华为实验室测试对比数据，
结果在不同环境中可能有偏差



工艺：7nm | 多核：64核 | 内存：8通道 | 接口：PCIe 4.0 & 100GE

华为TaiShan服务器，覆盖多样化应用场景

TaiShan 2280
均衡型服务器



适用多种应用

2U机架

2路

32*DDR4-2933 MHz

27*2.5英寸硬盘或16*2.5英寸NVMe SSD

CCIX, 8*PCIe 4.0

GE / 10GE / 25GE

风冷

TaiShan 5280
存储型服务器



单柜5.6PB容量

4U机架

2路

32*DDR4-2933 MHz

40*3.5英寸硬盘

CCIX, 8*PCIe 4.0

GE / 10GE / 25GE

风冷

TaiShan X6000
高密型服务器



HPC计算节点主打产品

单柜10240核

2U4节点

2路

16*DDR4-2933 MHz

6*2.5英寸硬盘或NVMe SSDs

CCIX, 2*PCIe 4.0

10GE / 25GE / 100GE RoCE

风冷或液冷

X6000高密型

极致高密的计算能力



主要亮点：

- 2U4机框支持4节点服务器
- 单节点支持2个鲲鹏916或920处理器
- 支持最多16个DDR4内存插槽
- 支持6个2.5英寸硬盘或NVMe SSDs*
- 支持100GE板载网络*
- 支持风冷或液冷散热*

形态	2U4节点/双路节点高密型
产品系列	TaiShan 200
型号名称	X6000
节点名称	XA320
处理器型号	2*鲲鹏920
内存插槽	16个DDR4-2933插槽
本地存储	最多6个2.5英寸SAS/SATA/SSD/NVMe SSD硬盘
RAID支持	支持RAID 0, 1, 5, 6, 10, 50, 60, 支持超级电容掉电保护 * TaiShan X6000 XA320 V2液冷服务器仅支持RAID 0,1
PCIe扩展	最多1个PCIe 4.0 x16和1个PCIe 4.0 x8插槽
板载网络	2*GE电口+1*100GE光口
电源	X6000超级机框： 2个热插拔3000W电源模块，支持1+1冗余
供电	支持100~240V AC, 240V DC
风扇	支持4个热插拔风扇模组，支持N+1冗余
操作系统	CentOS、EulerOS
工作环境温度	5°C ~ 35°C
散热	风冷以及液冷* *支持液冷会占用1个PCIe x16槽位
尺寸 (宽x深x高)	X6000超级机框： 436mm x 819mm x 86.1mm XA320 V2节点： 177.9mmx545.5mmx40.5mm

方案全景

计算

网络

存储

软件

小结

2280均衡型

均衡的计算、存储和网络能力
灵活的扩展性



主要规格：

- 2U机架支持2个鲲鹏916或920处理器
- 支持最多32个DDR4内存插槽*
- 支持16个3.5英寸或27个2.5英寸硬盘
- 支持NVMe SSD*
- 支持最多8个PCIe扩展槽位*
- 支持GE/10GE/25GE板载网络*

* 仅TaiShan 200支持

形态	2U双路机架均衡型
产品系列	TaiShan 200
型号名称	2280
处理器型号	2*鲲鹏920
内存插槽	32个DDR4-2933插槽
本地存储	最多16个3.5英寸或27个2.5英寸SAS/SATA/SSD硬盘，或16个2.5英寸NVMe SSD硬盘
RAID支持	支持RAID 0, 1, 5, 6, 10, 50, 60，支持超级电容掉电保护
PCIe扩展	最多8个PCIe 4.0 x8或3个PCIe 4.0 x16+2个PCIe x8插槽
板载网络	2个板载网络插卡，最多支持8*GE电口或者8*25GE/10GE光口或者4*GE电口+4*25GE/10GE光口
电源	2个热插拔1500W或2000W交流电源模块，支持1+1冗余
供电	支持100~240V AC，240V DC
风扇	支持4个热插拔风扇模组，支持N+1冗余
操作系统	CentOS、EulerOS
工作环境温度	5°C ~ 40°C
散热	风冷
尺寸 (宽x深x高)	447 mm*790 mm*86.1 mm

高速互连网络

方案全景

计算

网络

存储

软件

小结

长期的技术合作

背靠背技术支持

上百集群的成功案例



InfiniBand 技术演进

10Gbps SDR

20Gbps DDR

40Gbps QDR

56Gbps FDR

100Gbps EDR

200Gbps HDR

400Gbps NDR

XDR



ConnectX

网卡



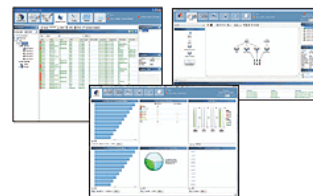
LinkX

高速线缆

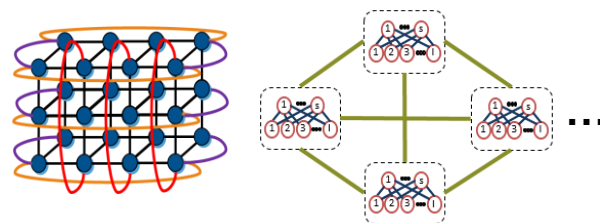


Mellanox Quantum

交换机



配套软件



支持多种网络拓扑

高性能存储

方案全景

计算

网络

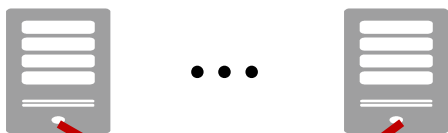
存储

软件

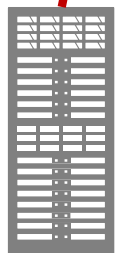
小结

方案1 NAS存储方案

TaiShan/X86计算节点集群



10GE

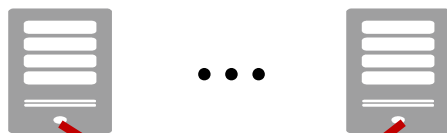


OceanStor NAS存储

适合轻量级集群

方案2 OceanStor9000 分布式存储方案

TaiShan/X86计算节点集群



IB/10GE



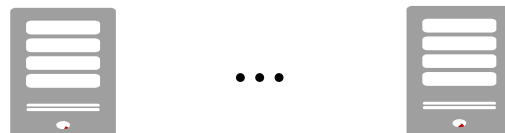
OceanStor DFS
分布式文件系统

OceanStor 9000
(存储节点集群)

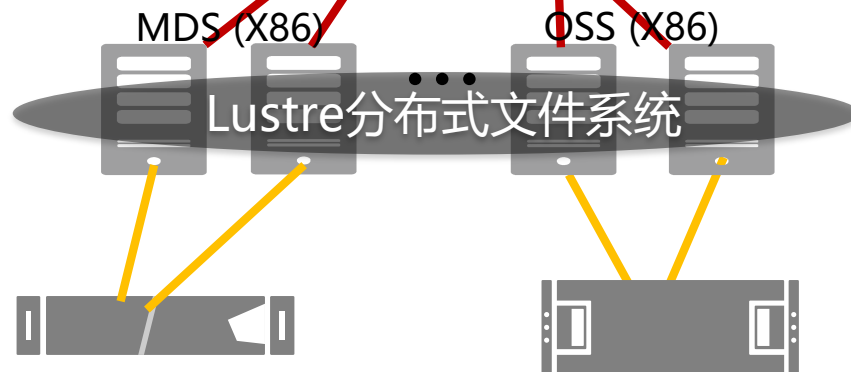
普通性能要求主选

方案3 Lustre并行存储参考架构

TaiShan/X86计算节点集群



IB/10GE



MDS (X86)

OSS (X86)

Lustre分布式文件系统

高性能要求主选

TaiShan已验证平台软件及应用



平台软件

软件类型	软件型号&版本
OS	CentOS 7.6 SLES 15.1
编译器	GNU 4.9 (OS内嵌), 5, 6, 7, 8, 9 LLVM, R
MPI	OpenMPI, Mpich3, HPC-X Mvapich
调度软件	Slurm, Open PBS Pro
集群管理软件	联科CHESS 并行科技 warewulf Nagios, ganglia



公司已验证应用列表:

<https://gitlab.com/arm-hpc/packages/wikis/categories/allPackage>



openHPC 软件集

<https://github.com/openhpc/ohpc/wiki/Component-List-v1.3.8>



制造

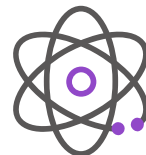
OpenFOAM
SU2
...



气象海洋

WRF
WRFDA
HDF5
NetCDF
WPS
NEMO
ROMS
FVCOM

MGLET
MITGCM
NorESM
CAMX
CMAQ
SMOKE
...



分子动力学

Lammps
Namd
Vasp
...



生命科学

GATK
GKL
Bwa
Bowtie
Bowtie2
Blast
blast+

Hisat2
Tophat
Stringtie
STAR
Flexbar
strelka
Cufflink
Control-Freec
Pysam
SAMtools
Text-Aligner
Novoalign
CNVnator
HipMer
assembler
Diamond
Spades
Jellyfish

Bzip2
Rapsearch
CANU
Minimap
SOAPdenovo
...

原则上所有有源码的HPC应用都可移植到TaiShan平台

• 小结

华为提供从基础硬件、软件生态到专业服务的完整解决方案

软件生态



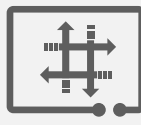
操作系统



集群管理



作业调度



MPI通信



编译器



数学计算库

基础硬件



多样化的计算资源



高性能存储



主流高速网络



专业服务



方案设计



系统部署



应用调优



培训赋能



联合创新

目录

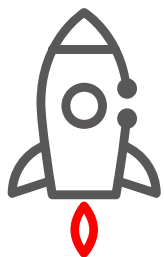
1. 华为的持续创新
2. 华为TaiShan HPC解决方案
3. 华为TaiShan HPC方案亮点
4. 华为全面构建TaiShan HPC生态

高性能

计算性能

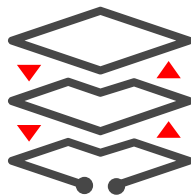
网络性能

能效



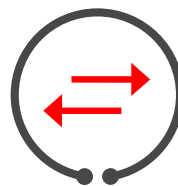
计算能力**强**

64核2.6GHz高性能处理器



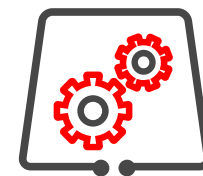
内存带宽**高**

8内存通道



IO吞吐**高**

PCIe Gen4
2倍于PCIe Gen3带宽



软硬件**协同优化**

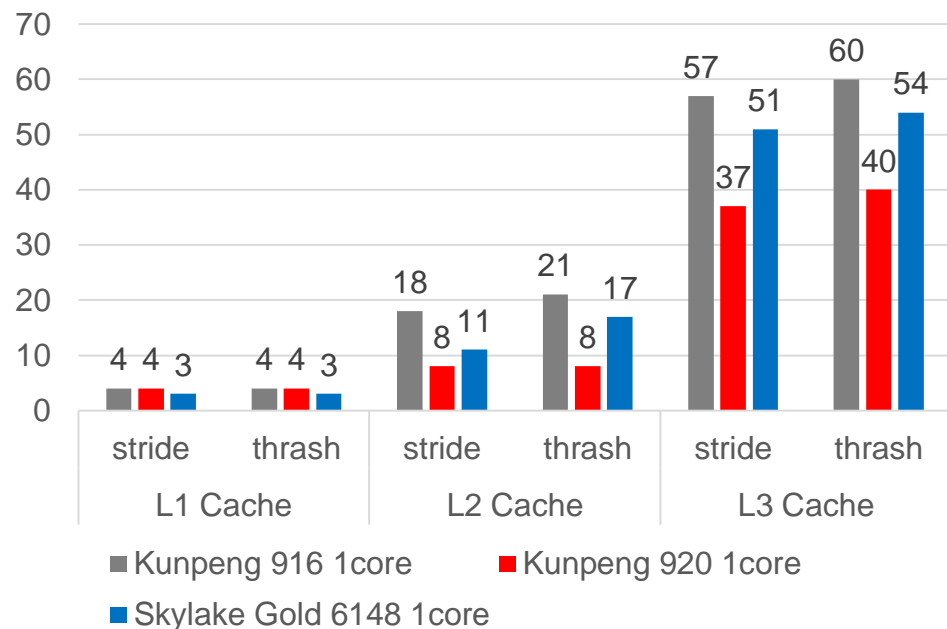
华为编译器、数学库
华为MPI

更高的内存带宽、更低的缓存时延

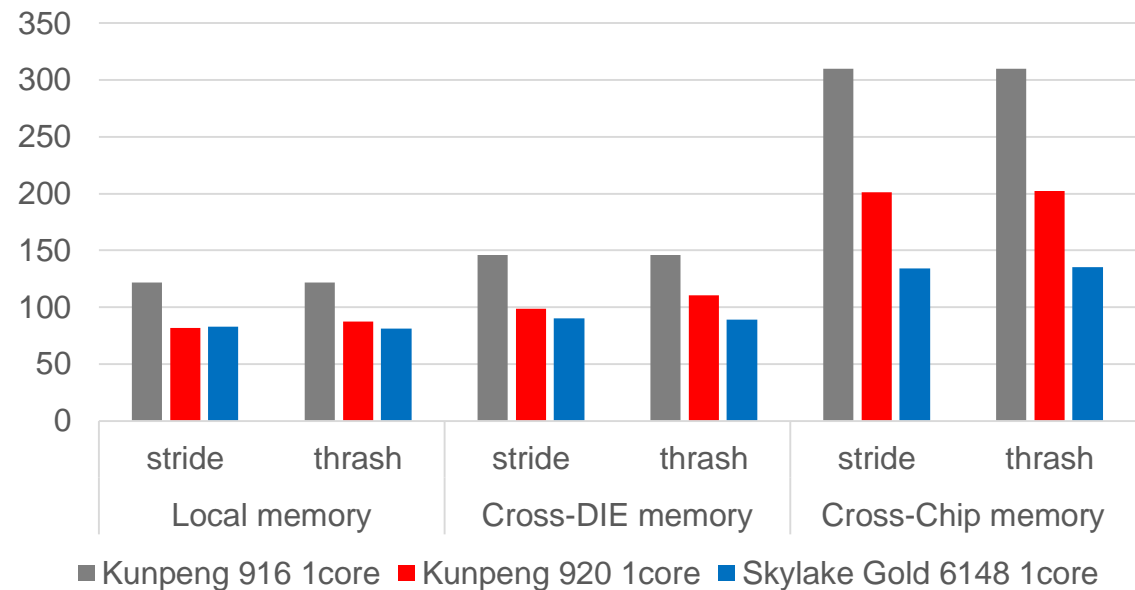
TaiShan加速访存密集型应用性能

	2P Kunpeng 920 (64 cores, 2.6 GHz, DDR4-2933)	2P Skylake 6148 (20 cores, 2.4 GHz, DDR4-2666)
STREAM	284 GB/S with 75.64% efficiency	197 GB/S with 76.95% efficiency

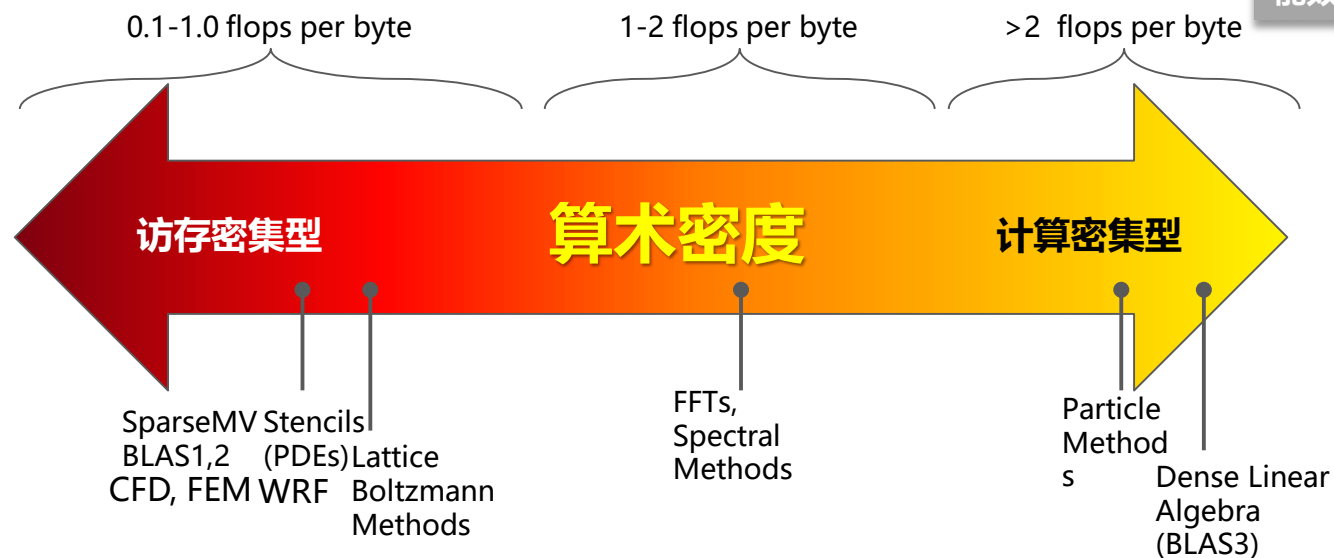
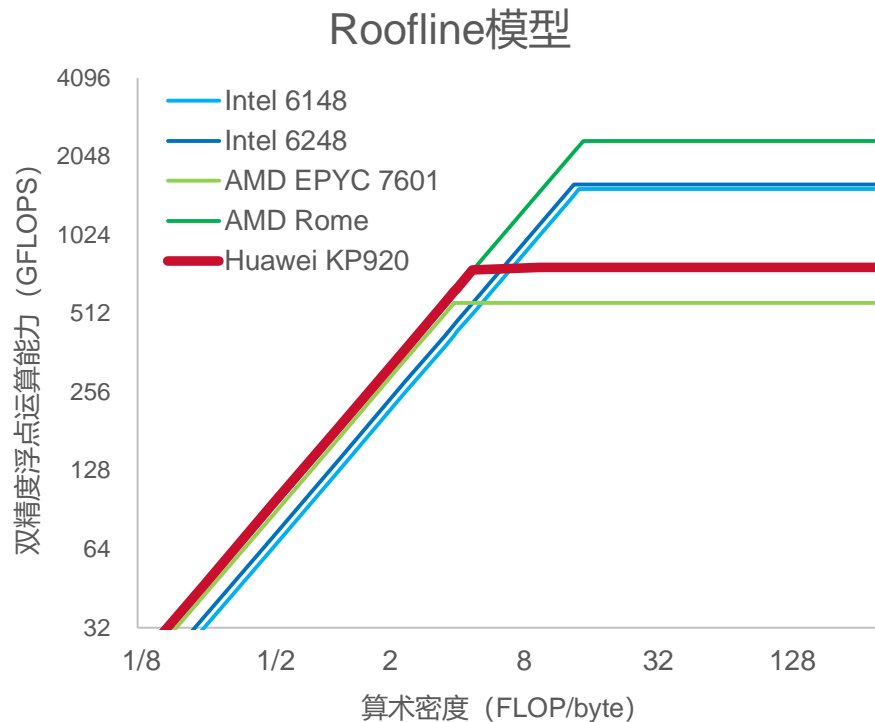
Imbench Latency (L1/L2/L3)



Imbench Latency(DDR)



关注应用实际性能，关注细分行业专业场景

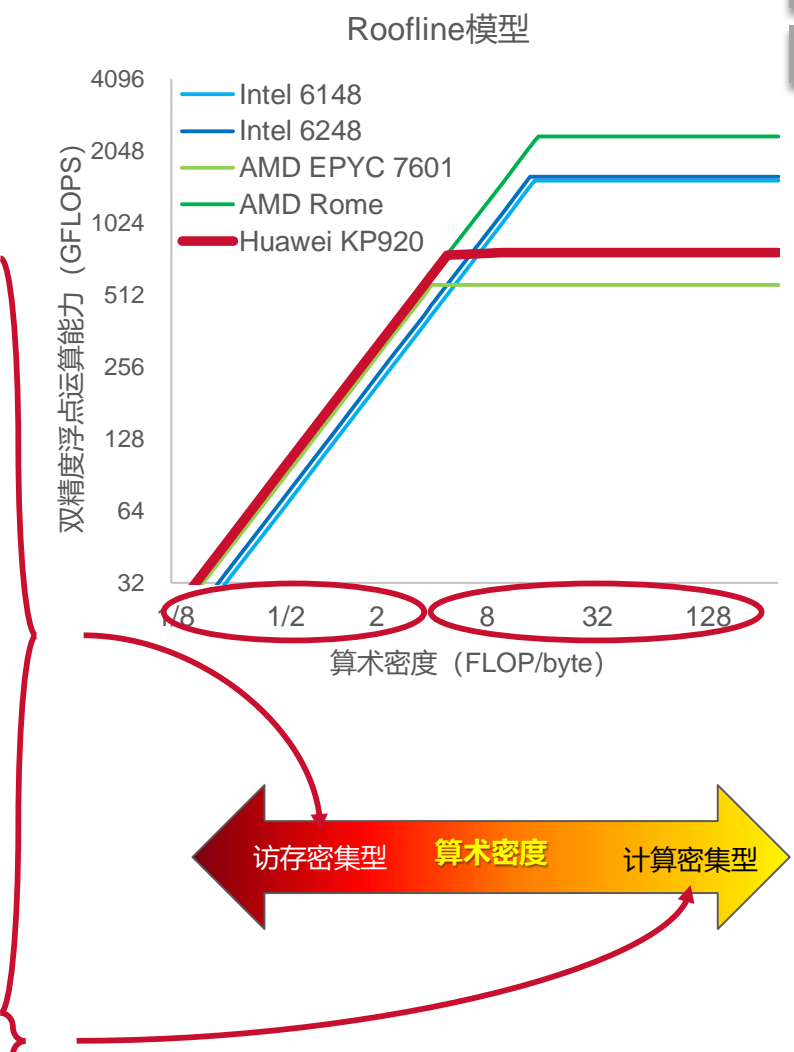


- **算术密度** (Operation intensity)
指单位时间内 (算术操作数) / (读内存字节数) 的比率
- **较低的**算术密度意味着**访存密集**
- **较高的**算术密度意味着**计算密集**

通过roofline模型分析，HPC领域大量的算法和应用属于访存密集型
TaiShan HPC瞄准**访存密集型**的应用, 比如CAE/CFD、气象、生命与油气

常见HPC应用为访存密集型应用

应用	场景	数值方法	算术密度
BCM	CFD	Navier-Stokes	0.14
OpenFoam		Finite Volumes – Finite Element	0.13
Turbine		DNS	0.56
MHD – FDM	Magneto Hydro Dynamics	Finite Difference Method	0.33
MHD - Spectral		Pseudo Spectral Method	0.45
QSFDm	Seismology	Spherical 2.5D FDM	0.46
SEISM3D		Finite Difference Method	0.47
Barotropic	Ocean Circulation Model	Shallow Water Model	0.51
BQCD	High-Energy-Physics	Hybrid Monte-Carlo	0.45
B-CALM	Electro-Magenetic Sim.	Finite Difference time-domain	0.3 (SP)
WRF	Weather Forecast model	Stencil code	0.5-1.5
HEPSPEC	SPEC2006 selection for HEP (CERN)	NAMD, DEALII, SOPLEX, POVRAY, OMNETPP, ASTAR, XALANCBMK	>=0.5
Gromacs	Molecular dynamics package	Bennett Acceptance Ratios	<1
KKRNano	Nanotechnology	Korringa-Kohn-Rostoker	4 (DP)



来源于华为内部测试数据



典型访存密集型应用 - OpenFOAM 案例

算例

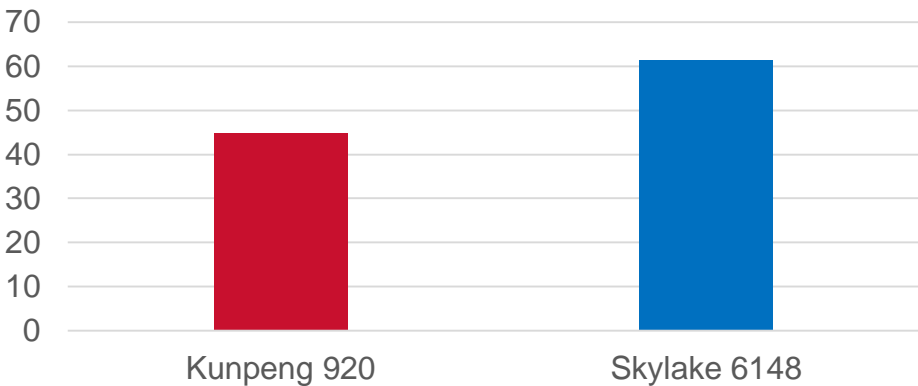
制造业客户算例，迭代次数: 200

性能

Kunpeng 920 比 Intel Skylake Gold 6148 **提升 32%**

OpenFOAM V1606+ 性能

运行时间越短越好



	2 路节点 Skylake Gold 6148 （2.4G, 20核）， 12*16GB 2666MHz内存	2 路节点 Kunpeng 920 （2.6G, 48核）， 16*16GB, 2666MHz内存
操作系统	RHEL 7.5 (Kernel 3.10)	EulerOS (Kernel 4.14)
编译器	Gcc 4.8.5	Gcc 7.3
MPI	Intel MPI 16.3.223	OpenMPI 3.0.1
运行指令	<code>./Example_runBenchmark.sh</code> <code>decomposePar</code> <code>mpirun -np 40 checkMesh -constant -parallel</code> <code>mpirun -np 40 pisoFoam -parallel</code>	<code>./Example_runBenchmark.sh</code> <code>decomposePar</code> <code>mpirun --allow-run-as-root -mca pml ucx -map-by core -np 96</code> <code>checkMesh -constant -parallel</code> <code>mpirun --allow-run-as-root -mca pml ucx -map-by core -np 96</code> <code>pisoFoam -parallel</code>

*华为实验室数据

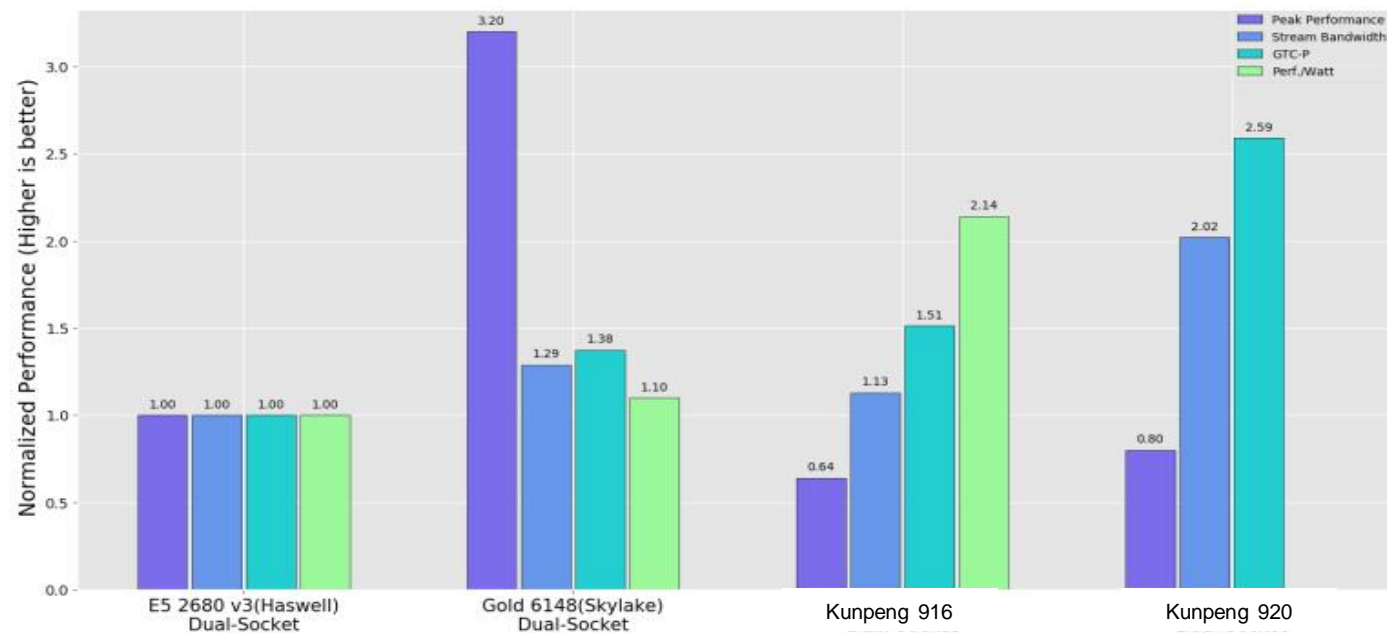


第三方测试数据 - 上海交通大学



GTC-P在鲲鹏920与6148上的性能对比

鲲鹏920的内存子系统设计使得其在GTC-P的运行中有较大优势。

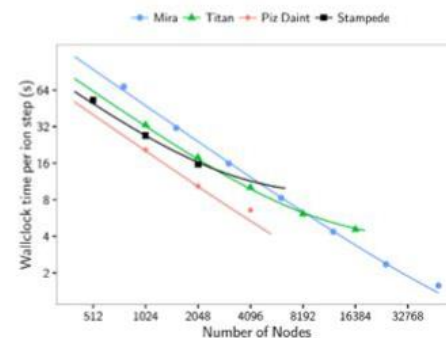


GTC-P: Gyrokinetic Toroidal Code - Princeton

- GTC-P is Particle-in-Cell code that delivers fusion simulations at extreme scales on the worldwide supercomputers including Tianhe-2, Titan, TaihuLight and etc., that feature CPU, GPU and many-core processors.



Supported by
NSF SAVI Project



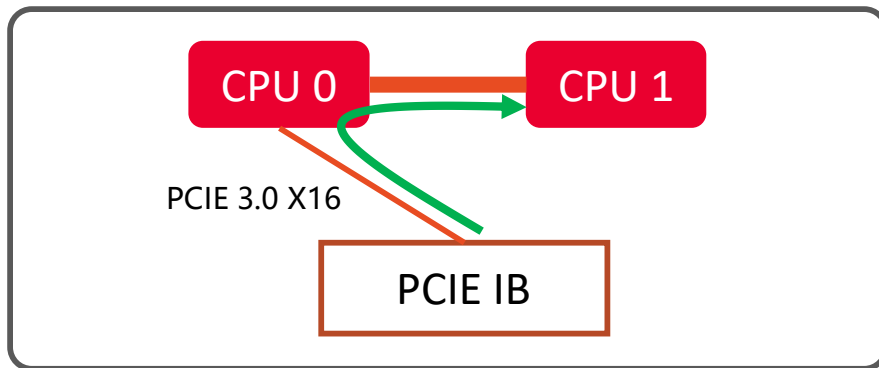
环境配置	TaiShan	Intel
处理器	2 * Kunpeng 920	2 * Skylake 6148
内存	16 * DDR4 2666 MHz	12 * DDR4 2666 MHz

X6000 面向HPC设计的高速网络接口

InfiniBand网络

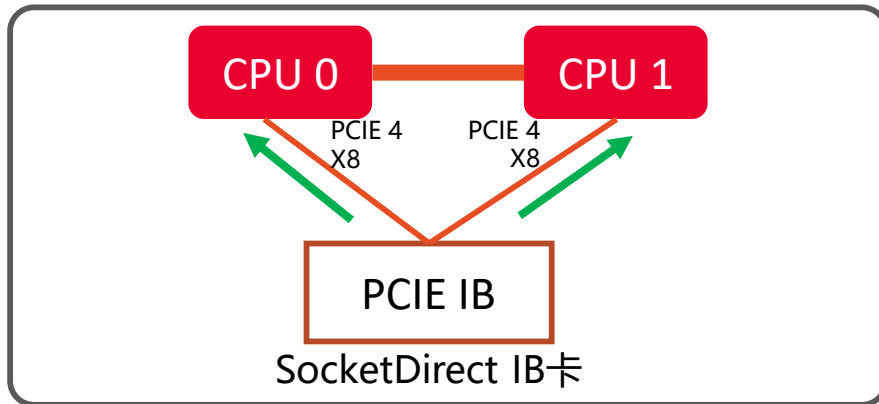
传统设计:

PCIe Gen3 网卡挂在单CPU上,
CPU1对外时延相比CPU0高XXX ns



创新设计:

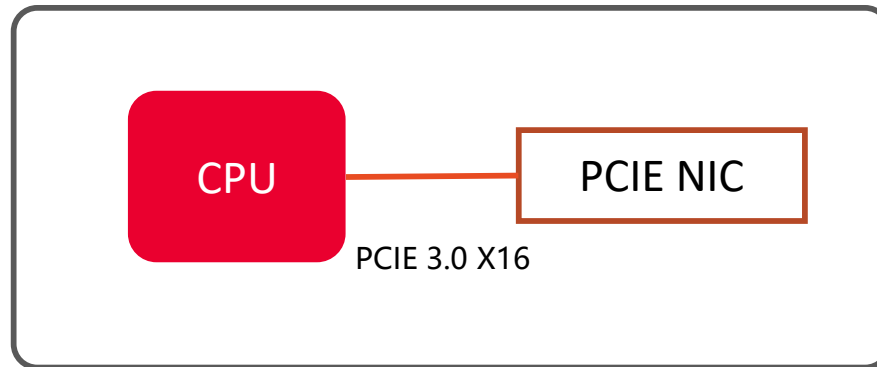
单插槽PCIe Gen4 x8 + x8分别连接到两个CPU
CPU1和CPU0对外时延相同



RoCEv2低时延计算网络

传统设计:

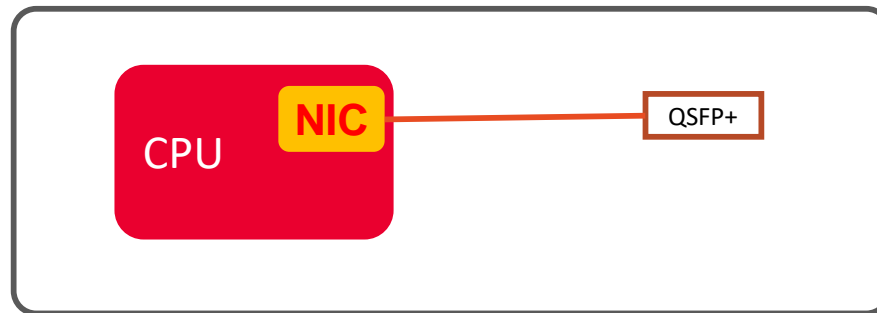
板载10G或者通过PCIe插槽插入RoCEv2网卡



创新设计:

外部信号直接到达CPU内部网卡

1. 减少PCIe信号处理, 降低链路时延
2. 免网卡, 降低网络投入,

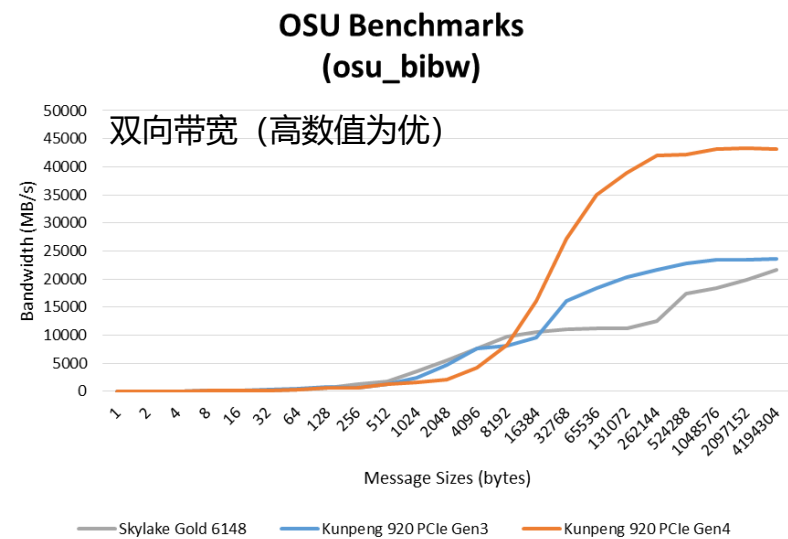
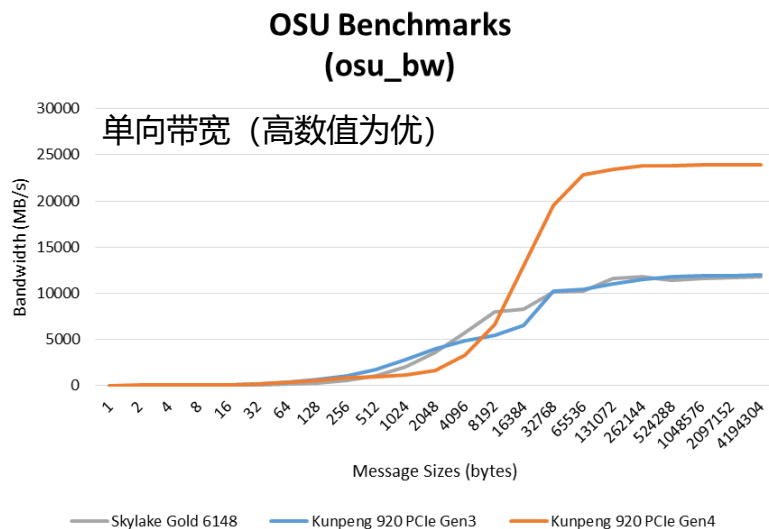
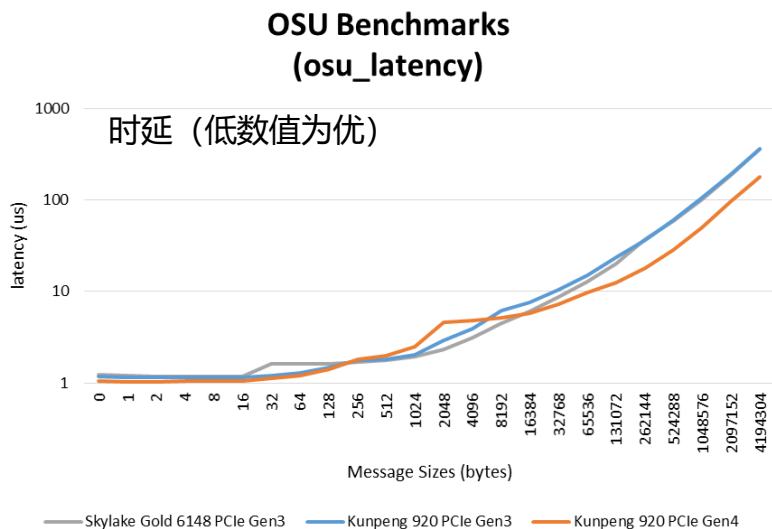


更高的链路带宽、更低的通信时延

Kunpeng920 支持**PCIe 4.0**

PCIe 4.0双口卡能带来**两倍带宽**和**更低时延**

华为与Mellanox公司联合对PCIe Gen4进行**深度性能优化**



高效能 – 板级液冷

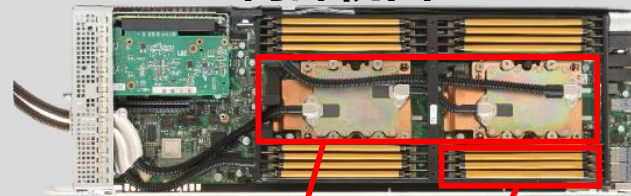
板级液冷：降功耗，省空间

- 板级液冷/全液冷，液冷占比高达95%，PUE降至1.05
- 支持超高密部署，节省机房空间80%；

XA320 V2

1. HPC场景CPU节能10%
2. 最高支持50°C进水
3. 高可靠、高抗压设计
4. 单节点插拔，易维护

内部视图



CPU金属冷板

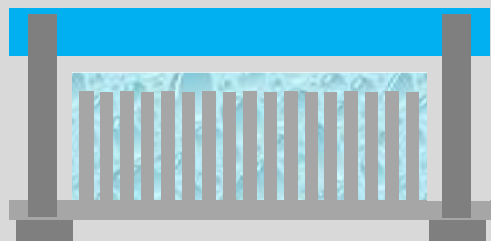
内存间冷却夹具

后置视图

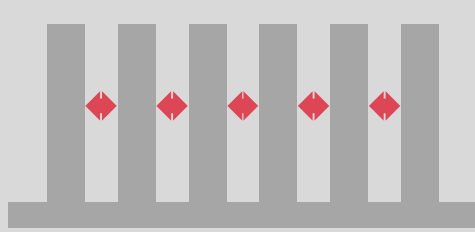


进出冷却水管

CPU冷却



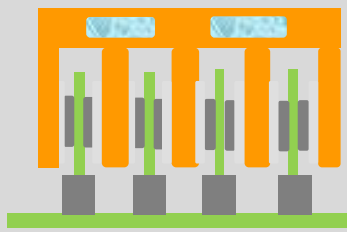
内部铲齿微通道设计



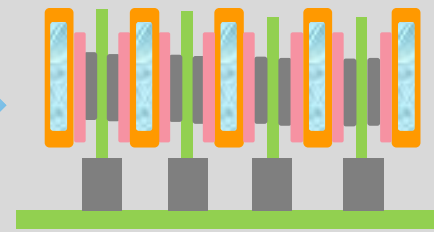
优化铲齿间距和通道流阻

优化铲齿设计，能效提升10%↑

内存冷却



传统内存冷板设计

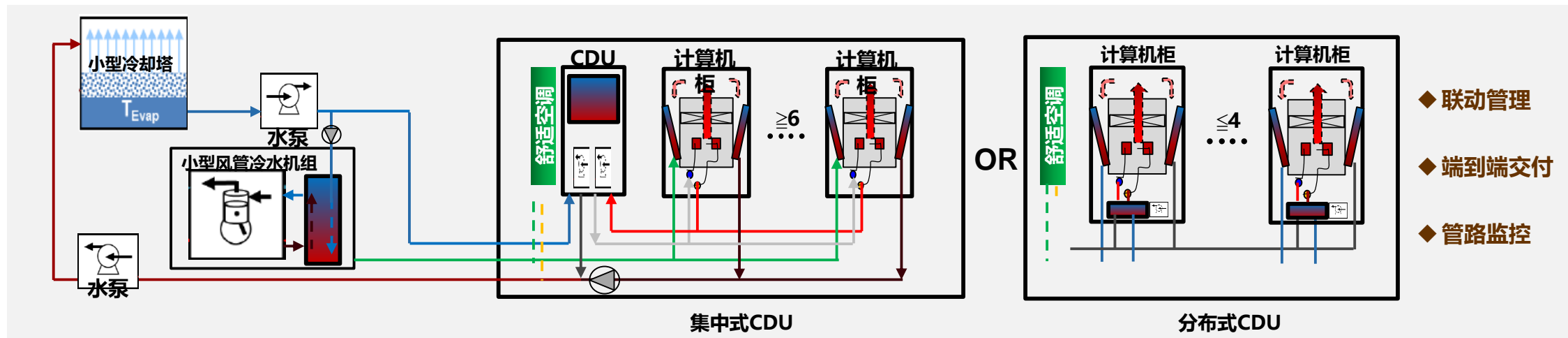


内存间走水设计

内存间进水，热传导路径缩短热阻减小65%
栅栏式夹具设计，减少风冷接触面积80%

高效能 – 全液冷解决方案

静音、节能、免维护，散热能力极致高效



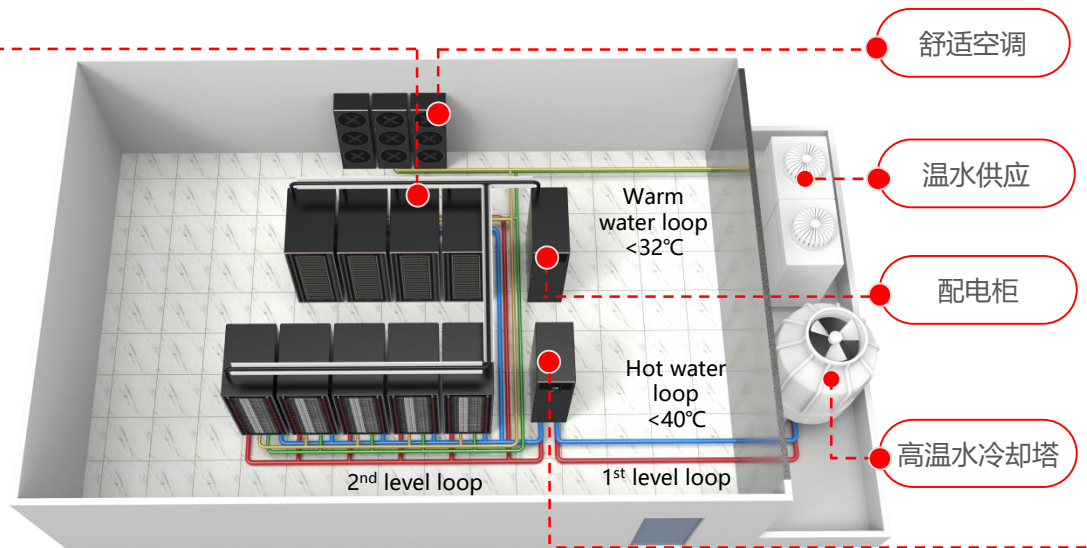
全封闭式液冷柜 + 分布式CCU



X6000

- ◆ 机柜尺寸: 800Wx1200Dx2200H
- ◆ 承载量: 20 * X6000
- ◆ 由分布式CCU监控管理: 漏液、温度、失效

CCU Monitoring device



CDU柜



- ◆ 机柜尺寸: 600Wx1000Dx2000H
- ◆ 重量: $<385\text{kg}$
- ◆ 供水量&散热能力: $<365\text{LPM}$ & 305kW
- ◆ 功耗: $<4.3\text{kW}$

高能效—小结

高能效液冷解决方案，PUE低至1.05

全液冷机柜100%热量由液体带走

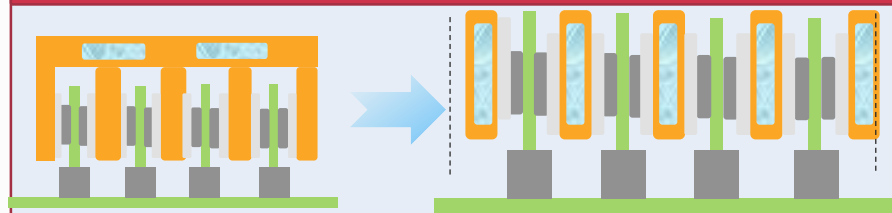
70%热量
板级水冷直接带走



30%热量
柜级风液换热后，由水带走



内存间走水冷却，内存液冷占比90%

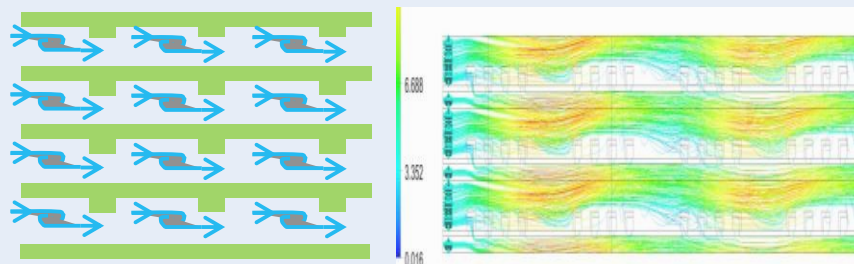


内存冷板设计

内存间走水设计

热传导路径缩短，热阻减小65%；单根内存条更换周期由3分钟减少到30秒

扰流强化换热，换热系数对比普通提升20%



PUE ≤ 1.05

目录

1. 华为的持续创新
2. 华为TaiShan HPC解决方案
3. 华为TaiShan HPC方案亮点
4. 华为全面构建TaiShan HPC生态

华为全面构建鲲鹏生态，使能TaiShan HPC解决方案

鲲鹏生态

- 水平生态：虚拟化，数据库，操作系统等
- 行业ISV合作：方案孵化，项目激励，联合推广等

- 产业组织
- 产业标准和政策
- 国家基础设施

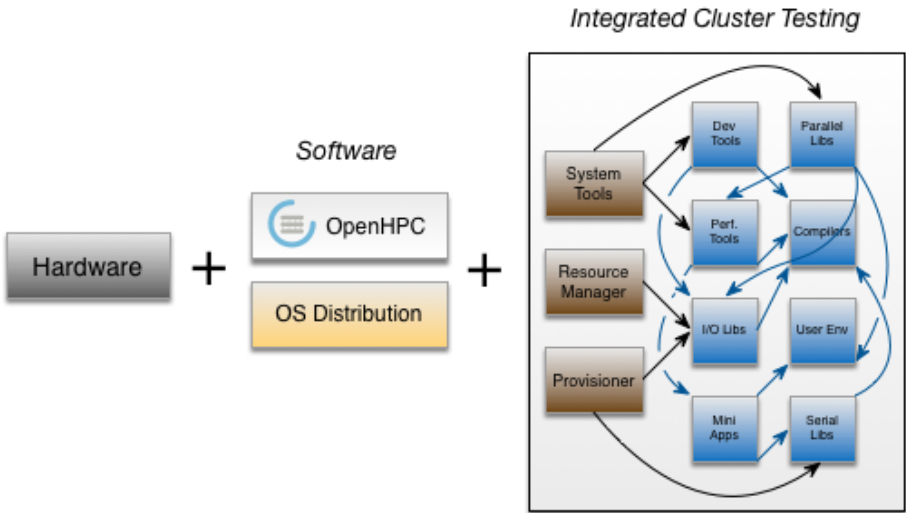


- 人才培养
- 开发者社区建设
- 开发者大赛
- 开源社区

开源分享，共筑TaiShan HPC开源生态

OpenHPC是完整的HPC开源软件堆栈
在华为TaiShan服务器上已经过全面测试

Functional Areas	Components	new in v1.3.8 release
Base OS	CentOS 7.6, SLES12 SP4	
Architecture	aarch64, x86_64	
Administrative Tools	Conman, Ganglia, Lmod, LosF, Nagios, NHC, pdsh, pdsh-mod-slurm, prun, EasyBuild, ClusterShell, mrsh, Genders, Shine, Spack, test-suite	
Provisioning	Warewulf, xCAT	
Resource Mgmt.	SLURM, Munge, PBS Professional, PMix	
Runtimes	Charliecloud, OpenMP, OCR, Singularity	
I/O Services	Lustre client, BeeGFS client*	
Numerical/Scientific	Boost, GSL, FFTW, Hypr, Metis, MFEM, Mumps, OpenBLAS, OpenCoarrays, PETSc, PLASMA,	

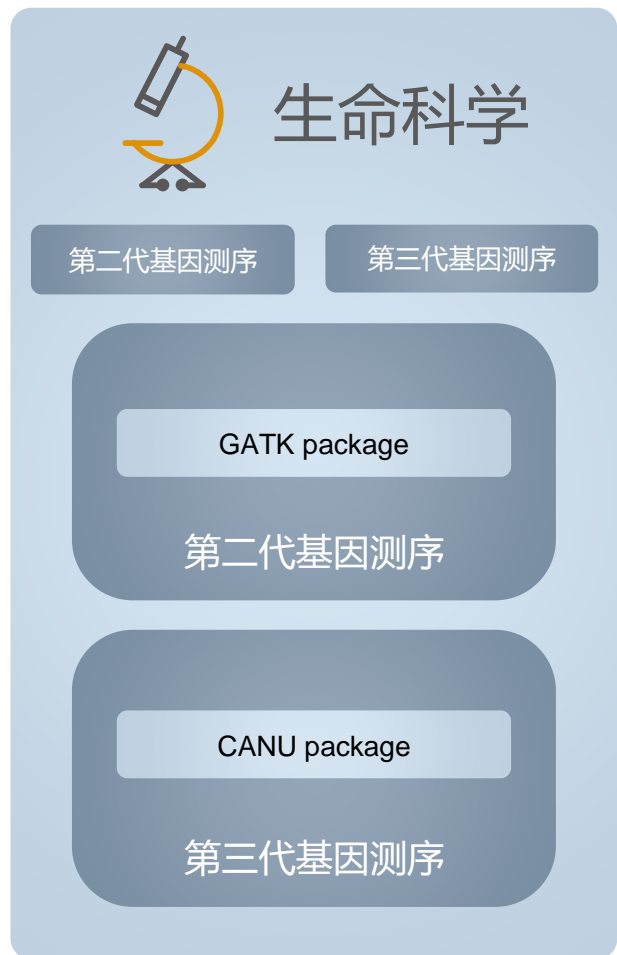


联合共建，携手共创TaiShan HPC应用生态



- 华为携手客户和合作伙伴投资于代码迁移，优化和基准测试的技术资源
- 华为联合投资OpenLab，用于代码迁移优化和基准测试的技术资源
- 华为与客户和合作伙伴共建开源/内部应用程序迁移的联合实验室

TaiShan HPC解决方案生态系统



华为在ARM生态中所作出的贡献



绿色计算产业联盟



Open Edge and HPC Initiative

促进基于ARM生态的开放及多样性

应对正在经历数字化的各个行业不断变化的需求



Thank you.

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

**Copyright©2018 Huawei 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. Huawei may change the information at any time without notice.

