

何万青

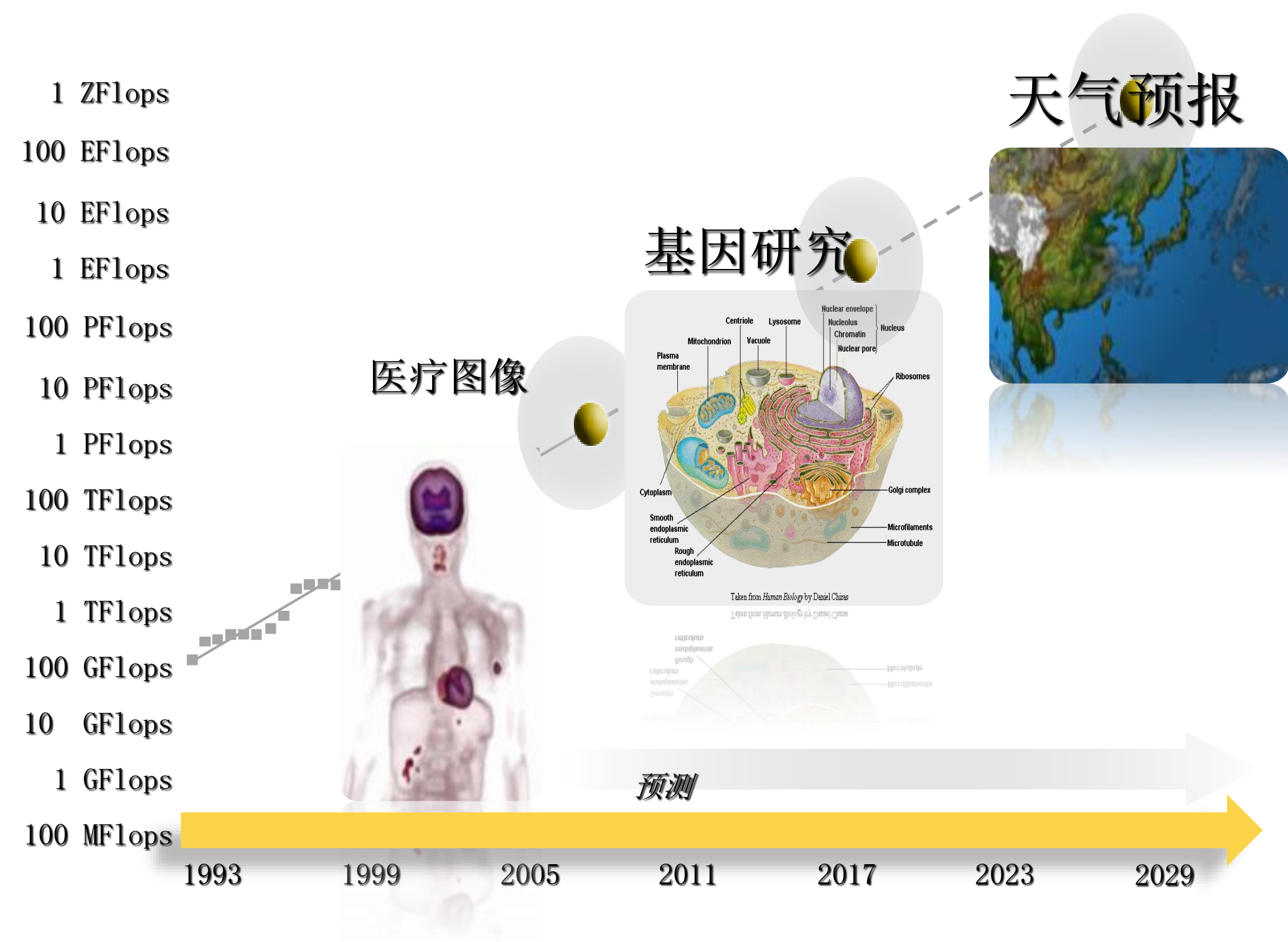
阿里云高性能计算总监

云上开发新时代，code up!

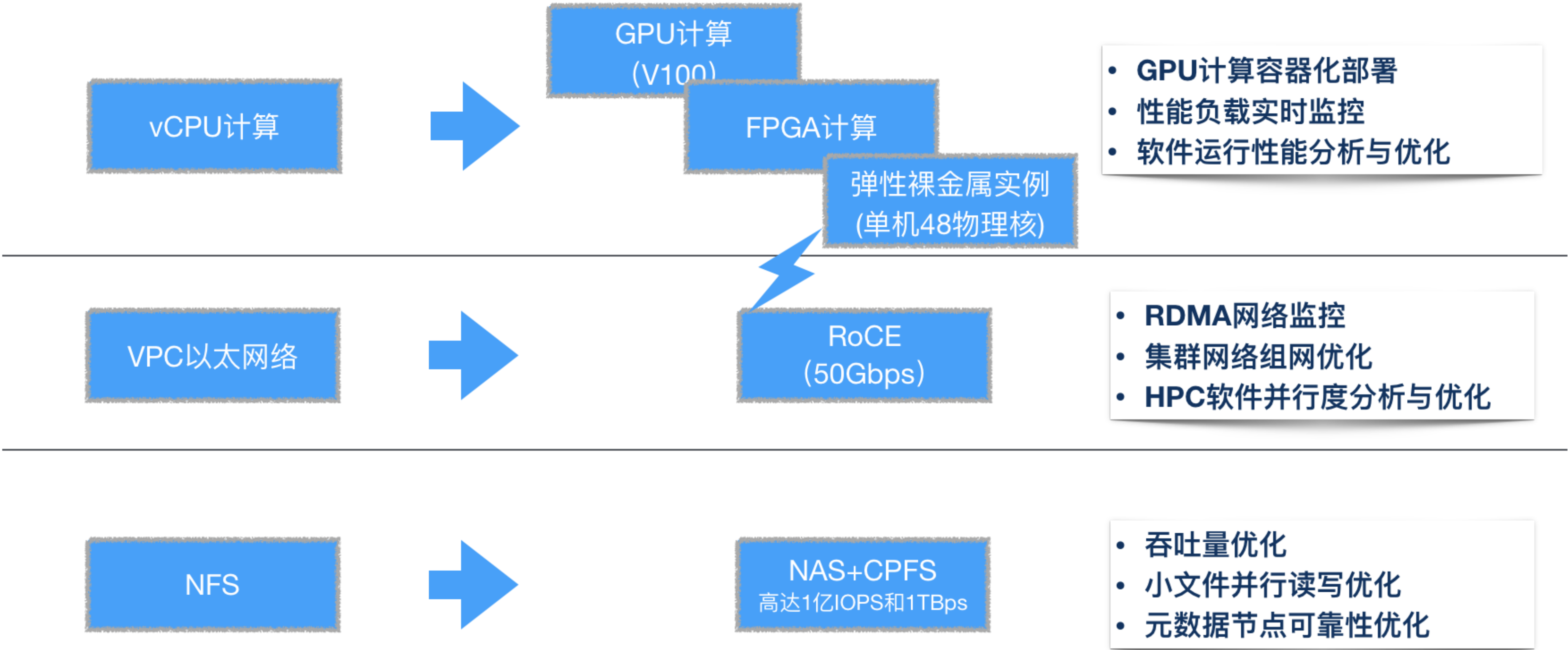
阿里云弹性高性能计算E-HPC：
一站式并行开发与优化

何万青

高性能计算(High Performance Computing)是计算机科学的一个分支。主要聚焦于解决大规模科学问题的计算和海量数据的处理，如科学研究、气象预报、计算模拟、军事研究、CFD/CAE、生物制药、基因测序、图像处理等等。构建高性能计算系统的主要目的就是提高运算速度，要达到每秒万亿次级的计算速度，对系统的处理器、内存带宽、运算方式、系统I/O、存储等方面的要求都十分高，这其中的每一个环节都将直接影响到系统的运算速度。



阿里云3招拆解超算上云障碍



面向“大计算”设计的弹性基础设施



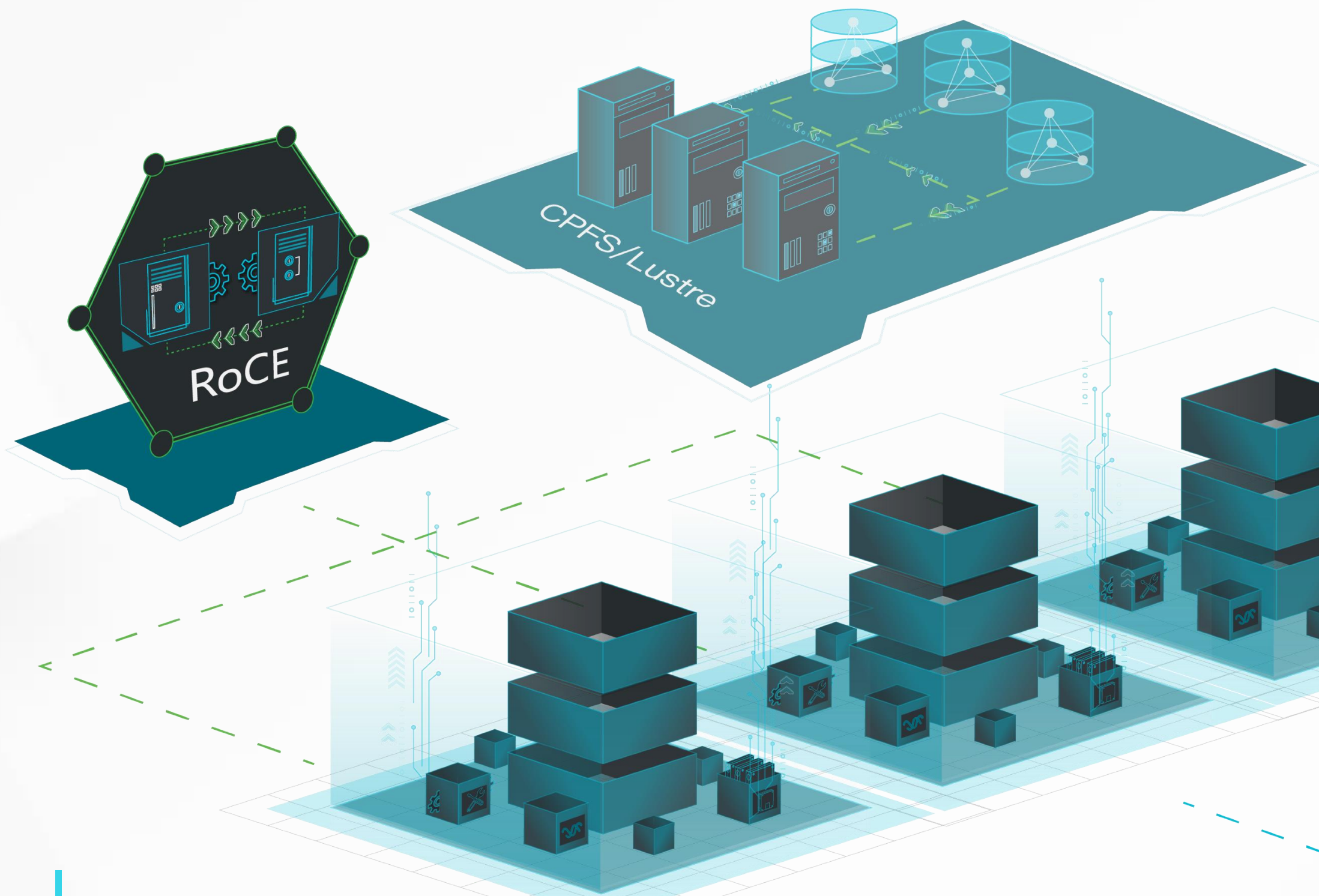
低延迟RDMA
50Gb 网络



基于Virtualization 2.0的
ECS弹性裸金属服务器



阿里云并行文件系统
NAS/CPFS (Lustre*)



ECS弹性裸金属服务器

- 无硬件性能/特性损失
- 硬件隔离/独占

超级计算集群SCC

- 弹性资源
- 快速服务部署，横向扩展和热迁移
- 运维自动化
- 统一服务接口

最高配置的单一实例



CPU
96 核



GPU
8xv100



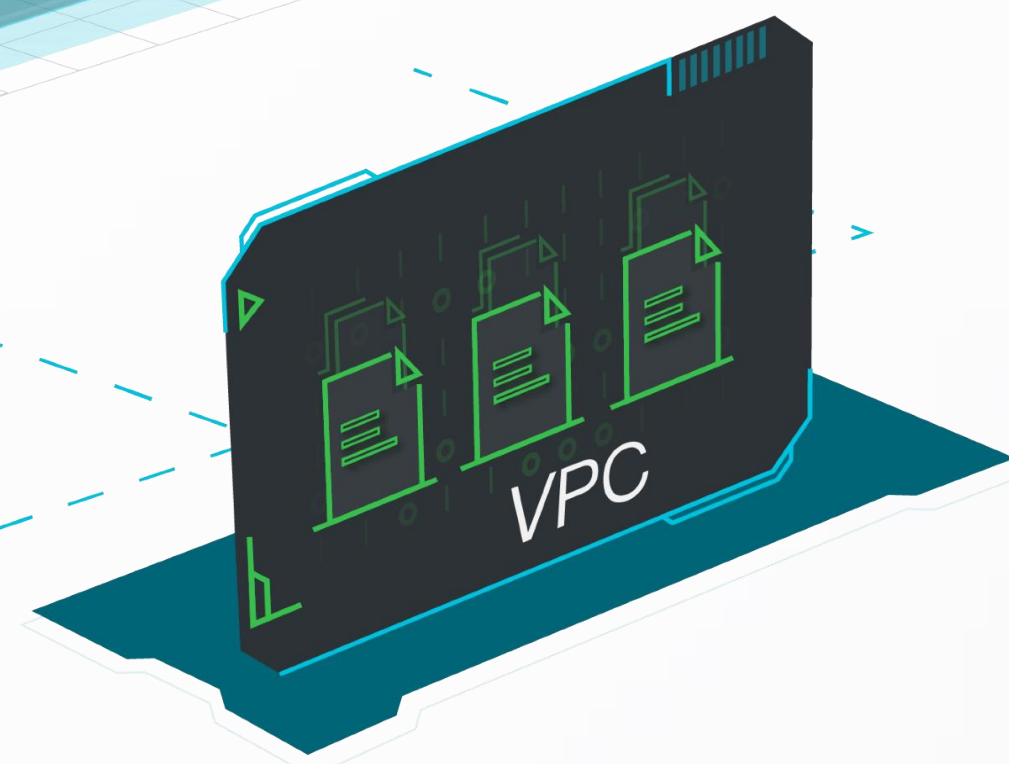
内存
384GB



网络带宽
50Gb

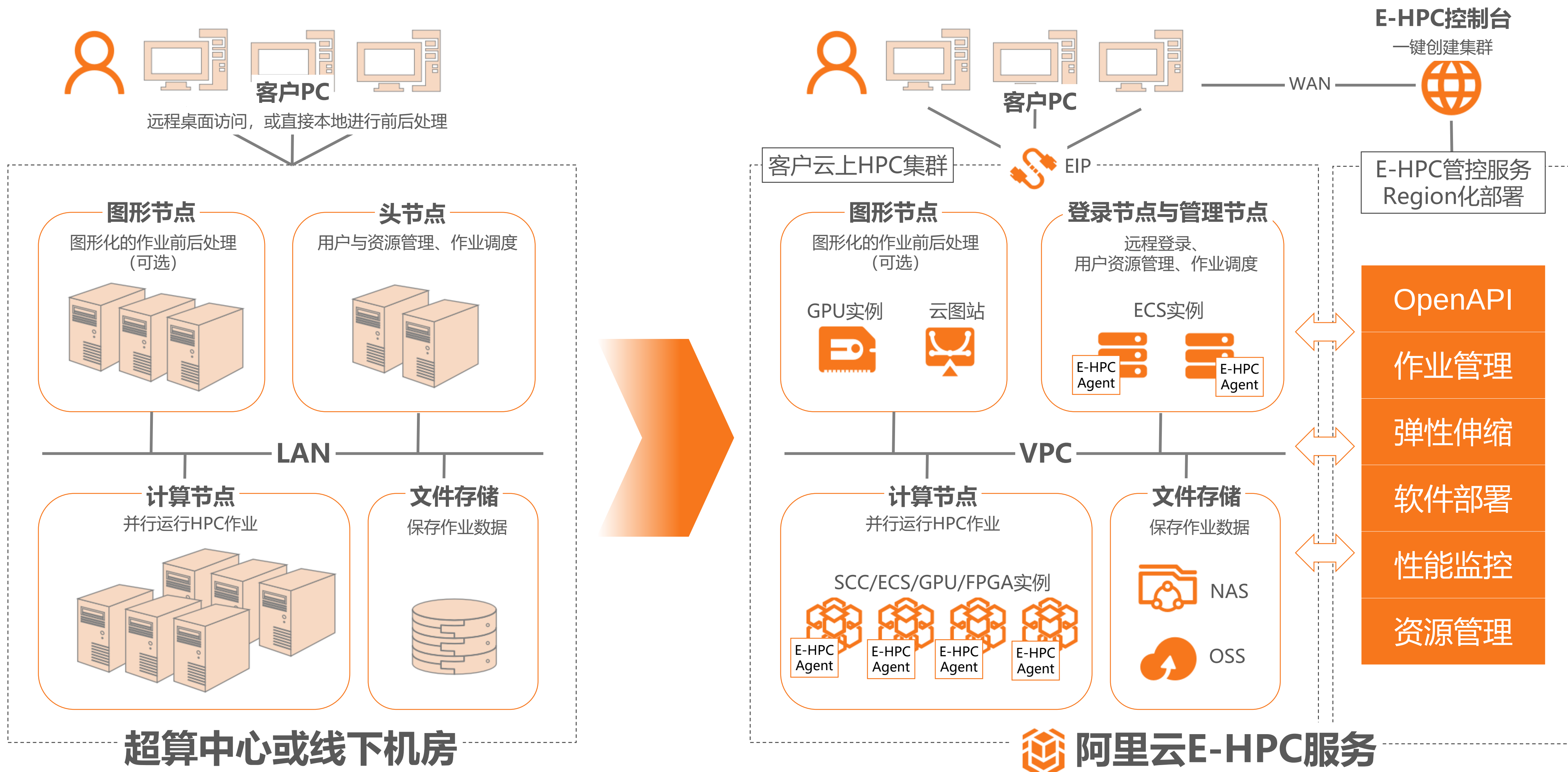


网络延迟
2.0μs



一张图描述E-HPC

从传统超算到云超算，将云服务整合为HPC集群



E-HPC环境搭建、部署与使用

01

部署集群

- 开通并创建NAS
- 硬件配置
- 软件配置

02

任务提交和数据导入

- 控制台提交作业
- 命令行提交作业
- 作业模版管理
- OSS数据导入

03

集谛性能分析工具

- 性能大盘
- 节点/进程性能
- 性能剖析

04

弹性伸缩和可视化

- 配置自动伸缩服务
- 自动伸缩实例
- 远程可视化

E-HPC让云超算强大易用

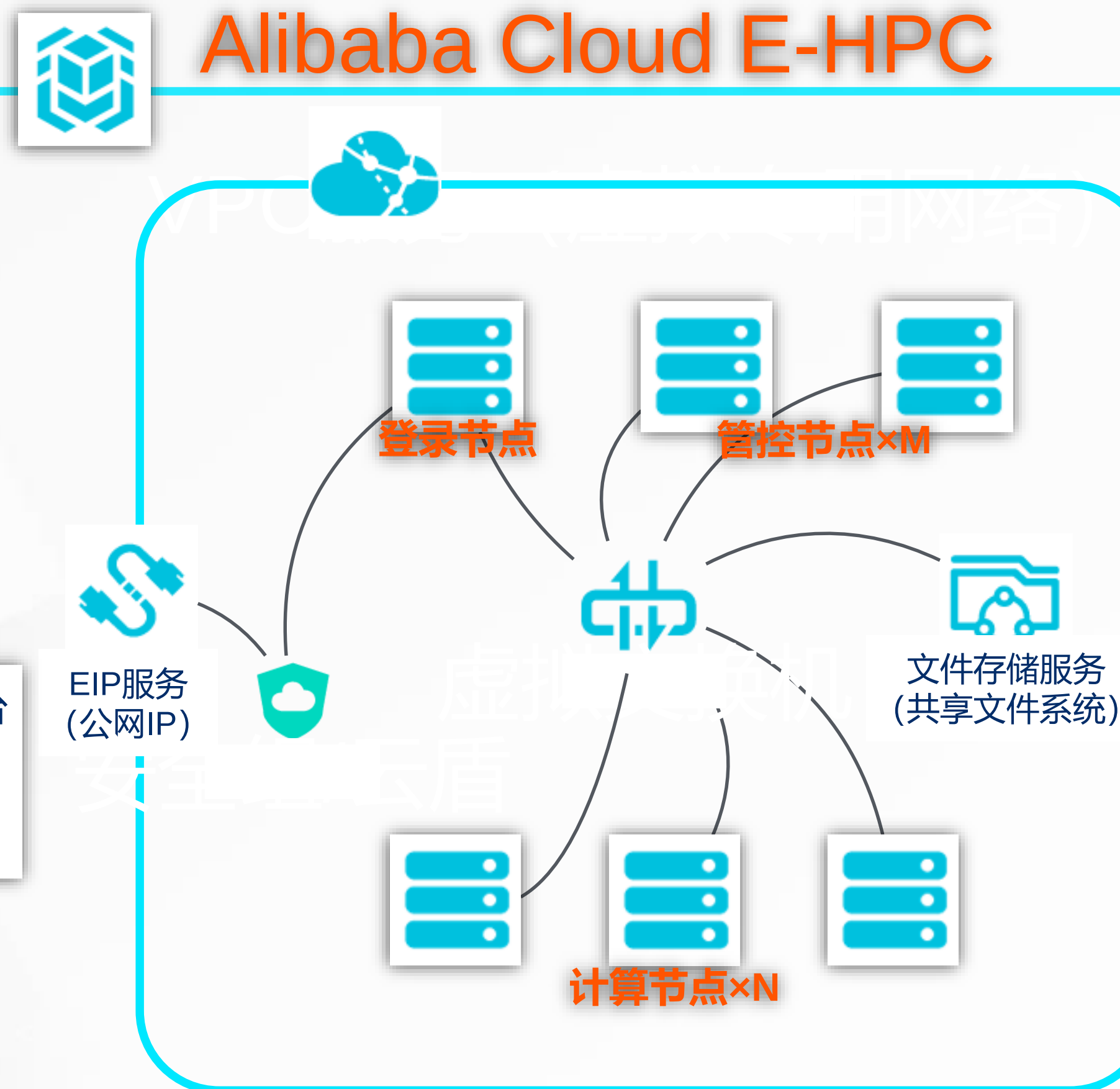
竞价实例：
低至1折，至少保持1小时可用时长

镜像市场与容器：
支持镜像市场镜像部署集群，支持HPC容器模式

按量付费：
精确到小时的付费粒度，按需使用随时释放



包年包月：
按周/月/年预付费，长期使用成本大减



作业调度器：
PBS/Slurm/GridEngine

应用软件：
GROMACS/LAMMPS/NAMD/GAUSSIAN.....

高性能运行时支撑：
MPICH/OpenMPI/BLAS/Boost/CUDA...

开发与调试工具：
GNU(gcc/g++/gfortran...)/Autotools(autoconf...)

筛选适合HPC的机型：
SCC/CPU计算型/GPU计算型/FPGA/网络增强型...

一键：
一键启动集群，一键修复异常集群

集群停机不收费

单集群支持扩容多种规格计算节点

多部署模式：
高可用，简易模式，OneBox模式

部署集群—基础配置

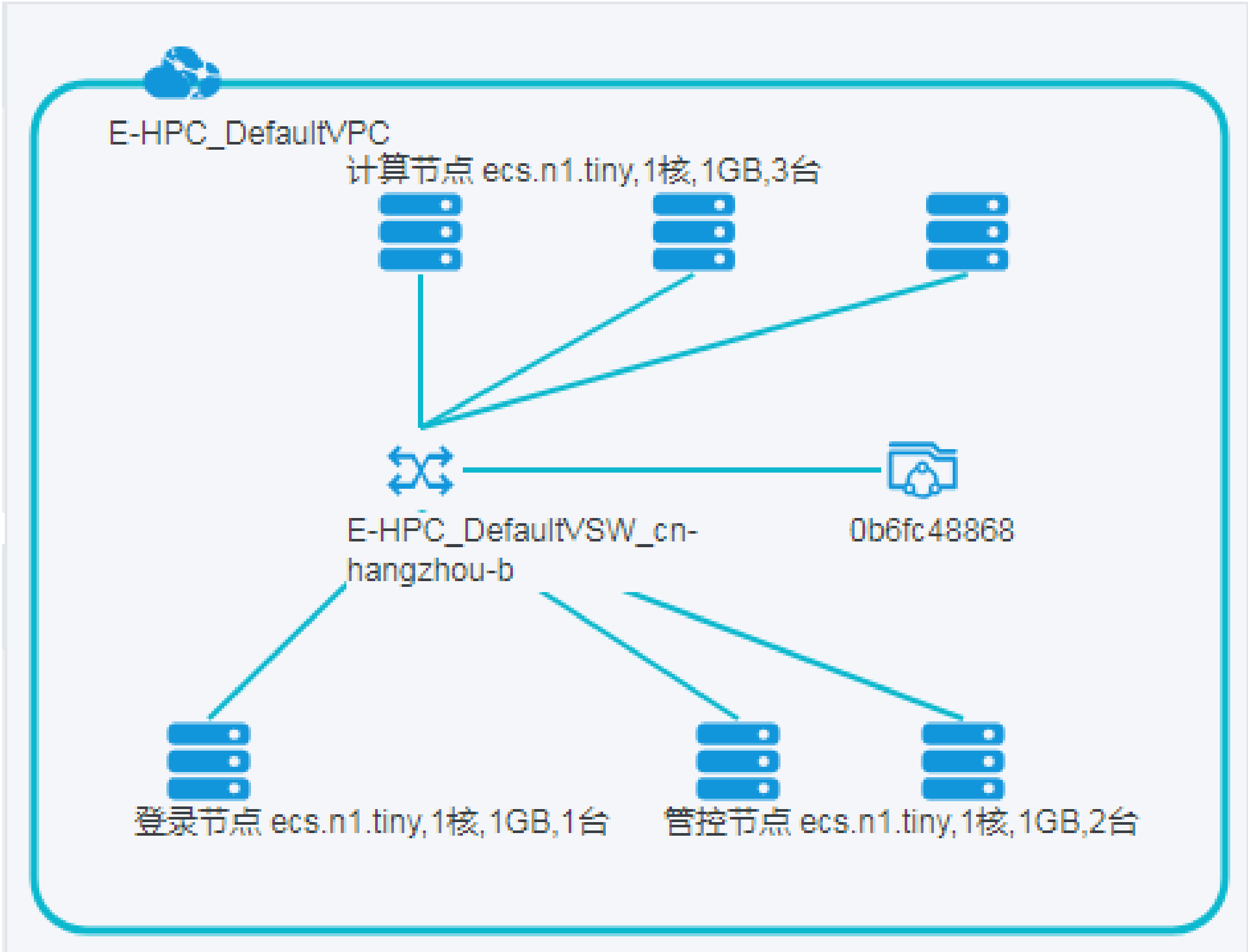
查看配置清单

查看配置拓扑图

配置清单

☒ 高级配置

集群名称	NewCluster
地域	华东1（杭州）
可用区	华东1（杭州）可用区 B
计算资源总核数	3
计算节点	ecs.n1.tiny,1核,1GB,3台
管控节点	ecs.n1.tiny,1核,1GB,2台
登录节点	ecs.n1.tiny,1核,1GB,1台
调度器	pbs
软件配置	OPENMPI_1.10.7,LAMMPS-OPENMPI-31Mar17
新建安全组	test_st
VSwitch	E-HPC_DefaultVSW_cn-hangzhou-b
共享存储	0b6fc48868-pty61.cn-hangzhou.nas.aliyuncs.com:/
操作系统	CentOS_7.2_64
产品版本	1.0.0
域账号服务	nis
系统盘大小	40 G



任务提交和查看

控制台

导入/导出作业文件

命令行

SSH 登陆提交作业

* 作业名	wrf.job	7/30
作业模板	请选择	
* 用户名 ②	alibaba@manget	
* 密码 ②		
* 作业执行命令	wrfun.pbs	
后处理命令 ②		
启动job array ②	☐	
作业优先级 ②	0	
调度器队列	请选择	
标准输出路径 ②	0/512	
错误输出路径 ②	0/512	
作业变量 ②	+ 添加	

编辑作业文件 使用OSS文件

新建文件 | 保存到本地 | 显示文件列表 隐藏文件列表

作业文件: /home/alibaba@manget/wrfun.pbs [重新编辑] [删除]

* 文件名: /home/alibaba@manget/wrfun.pbs

```
28 | cp -r $EHPC_WRF_MPICH .
29 | fi
30
31 module load mpich/3.2 wrf-mpich/3.8.1
32 export LD_LIBRARY_PATH=/opt/WRF_WPS-openmpi-3.8.1/util/lib:$LD_LIBRARY_PATH
33
34 export NODEFILE=/home/alibaba/Work/WRFV3/run/hostfile
35 export PBS_JOBDIR=$WRF_WORKDIR/WRFV3/run
36 export OMP_NUM_THREADS=1
37
38 # Copy the input data
39 cd $PBS_JOBDIR
40 if [ -r wrfbdy_d01 ]
41 then
42 | "INPUT DATA is already."
43 else
44 | cp $WRF_DATADIR/* .
45 fi
46
47 if [ -r hostfile ]
48 then
49 | "Hostfile is already."
50 else
51 | cp $NODEFILE .
52 fi
53
54 mpirun -np 8 -hostfile $NODEFILE ./wrf.exe
```

```
#!/bin/bash

#PBS -l nodes=2:ppn=4
#PBS -l walltime=24:00:00
#PBS -q high
#PBS -N wrf
#PBS -o wrf.out
#PBS -e wrf.err
#PBS -q workq

export EHPC_WRF_MPICH=/opt/WRF_WPS-mpich-3.8.1/WRFV3
export WRF_WORKDIR=$HOME/EHPC-WRF
export WRF_DATADIR=$HOME/data/

module load mpich/3.2 wrf-mpich/3.8.1
export LD_LIBRARY_PATH=/opt/WRF_WPS-openmpi-3.8.1/util/lib:$LD_LIBRARY_PATH

export NODEFILE=/home/jonathan/Work/WRFV3/run/hostfile
export PBS_JOBDIR=$WRF_WORKDIR/WRFV3/run
export OMP_NUM_THREADS=1

mpirun -np 4 -hostfile $NODEFILE ./wrf.exe

~
~
~
~
```


作业调度器



PBS



SLURM



SGE



OpenGrid



CUBE

PBS

```
1. $ cat > test.pbs
2. #!/bin/sh
3. #PBS -l ncpus=4,mem=1gb
4. #PBS -l walltime=00:10:00
5. #PBS -o test_pbs.log
6. #PBS -j oe
7. cd $HOME
8. test.py -i test.data
9. $ qsub test.pbs
```

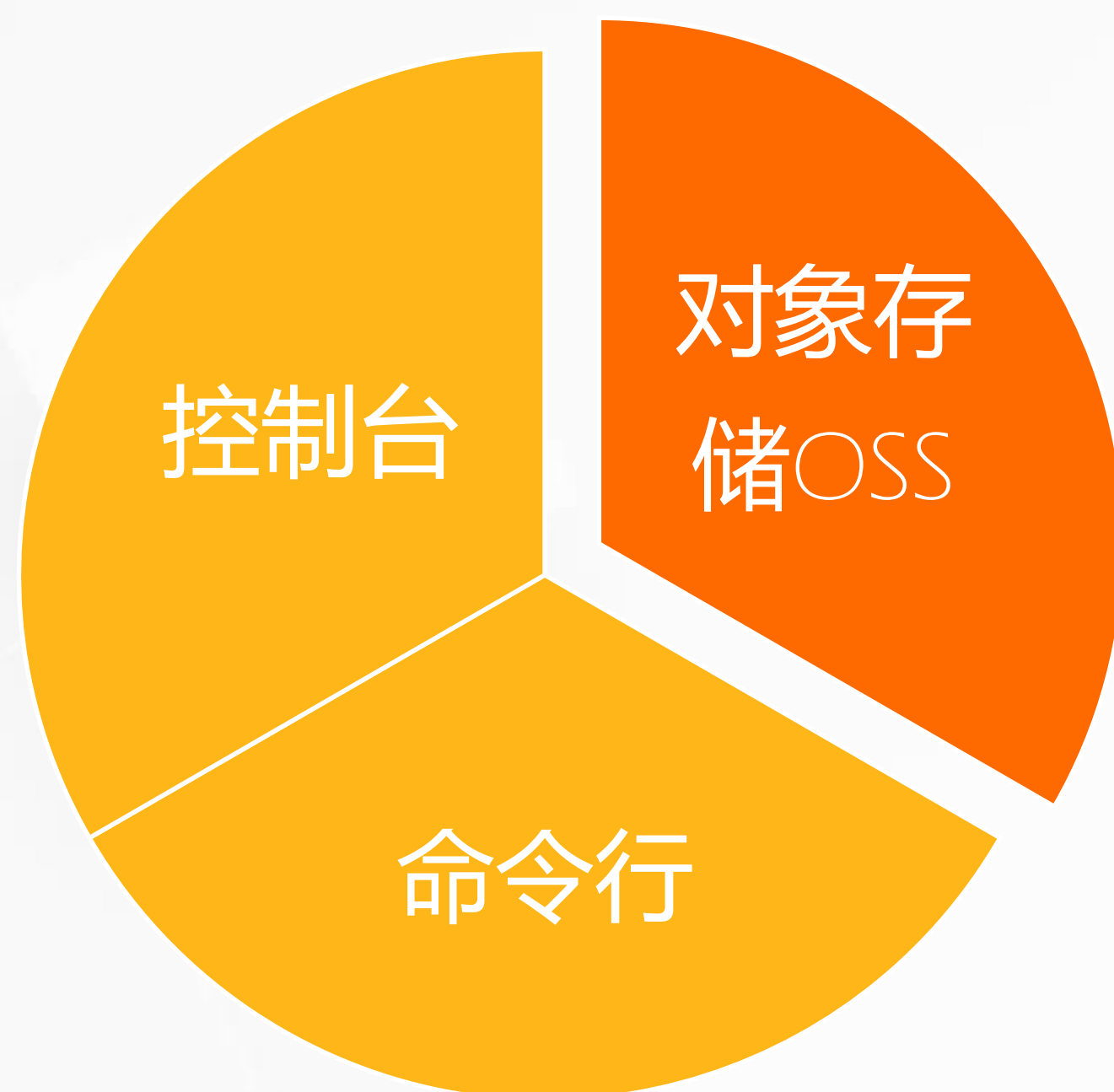
SGE

```
1. qsub -V -b y -cwd hostname
2. qsub -V jobscript.sh
3.
```

SLURM

```
1. $ cat > test.slurm
2. #!/bin/sh
3. #SBATCH --job-name=slurm-quickstart
4. #SBATCH --output=test_slurm.log
5. #SBATCH --nodes=1
6. #SBATCH --ntasks=1
7. #SBATCH --cpus-per-task=1
8. #SBATCH --time=00:10:00
9. #SBATCH --mem-per-cpu=1024
10. cd $HOME
11. test.py test.data
12. $ sbatch test.slurm
```

上传数据



通过OSS导入数据

- 通过HTTP访问VPC网络Endpoint导入数据

```
wget http://e-hpc-quickstart.vpc100-oss-cn-hangzhou.aliyuncs.com/test.data
```

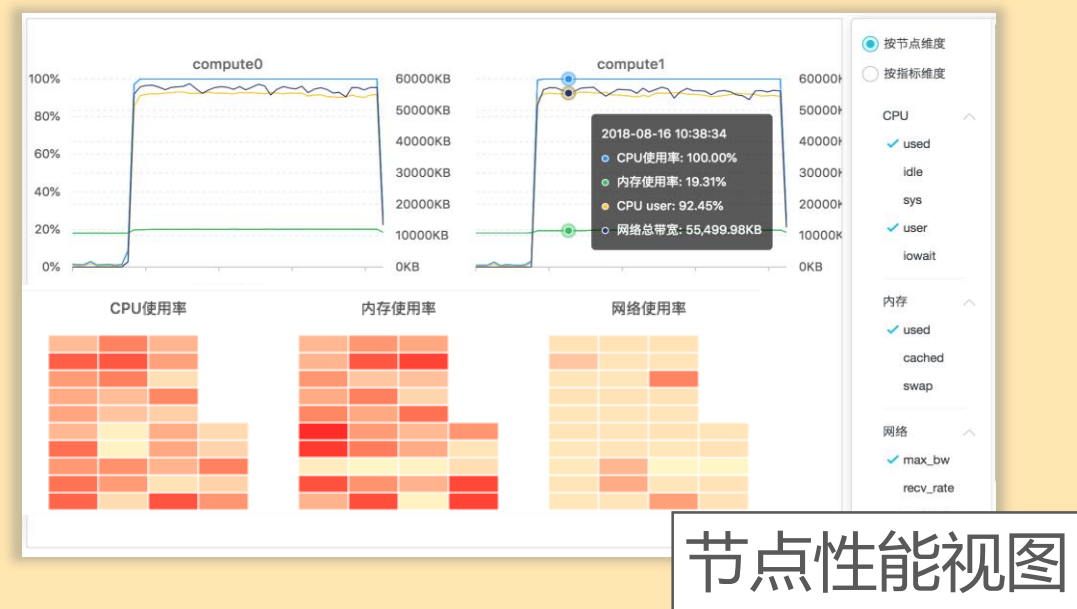
```
curl -O http://e-hpc-quickstart.vpc100-oss-cn-hangzhou.aliyuncs.com/test.data
```

私用Bucket

- 临时含签名的URL
- 通过HTTP访问VPC网络Endpoint导入数据

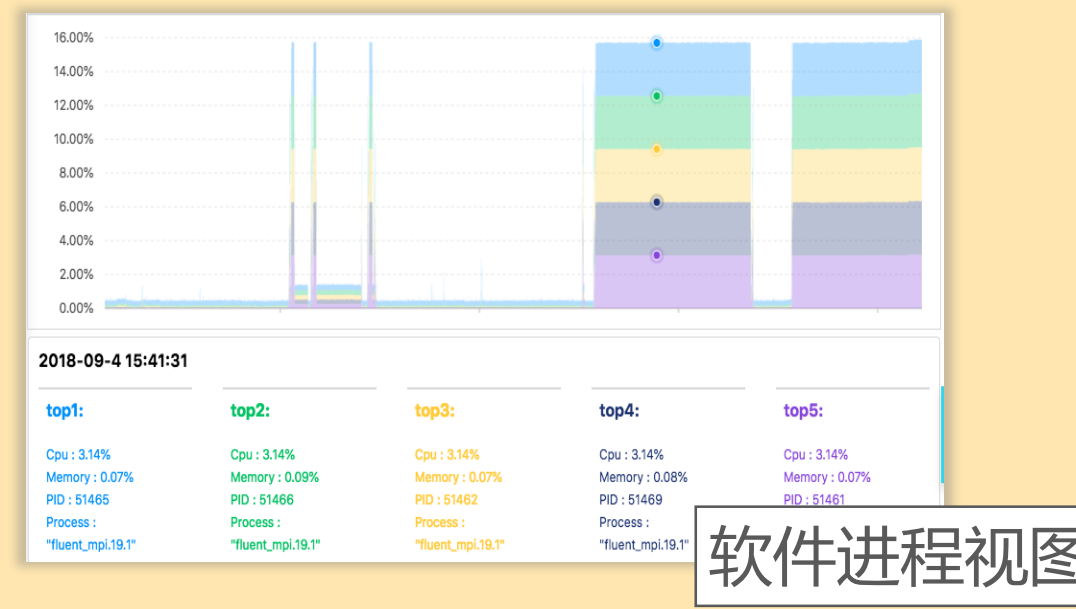
作业运行性能分析:CloudMetrics | 集谛

集群维度，定位实例热点



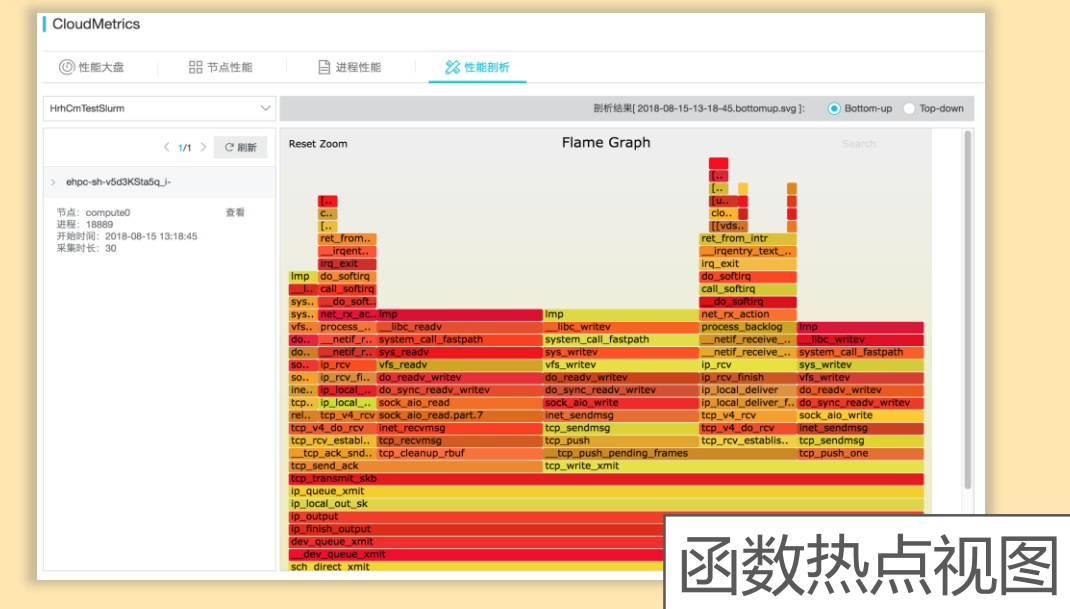
节点性能视图

实例维度，定位进程热点



软件进程视图

进程维度，定位函数热点



函数热点视图

作业进程与函数级

- 对指定进程进行性能剖析
- 显示各进程函数堆栈视图，分析函数热点

OS级

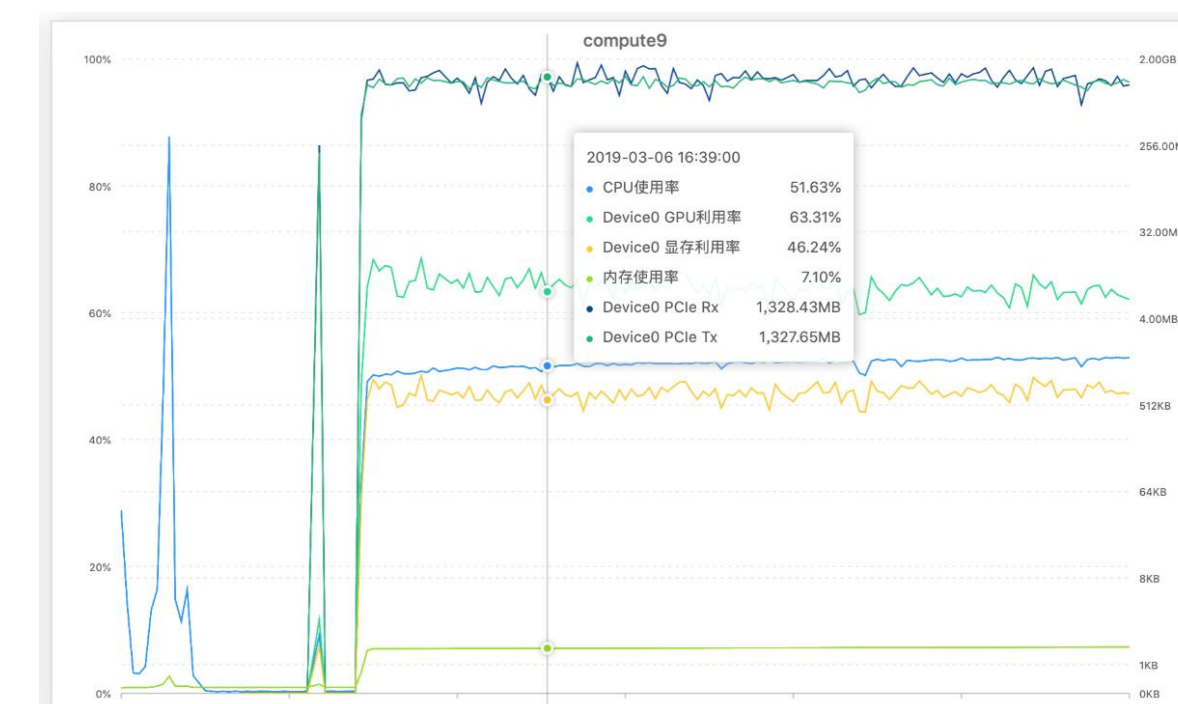
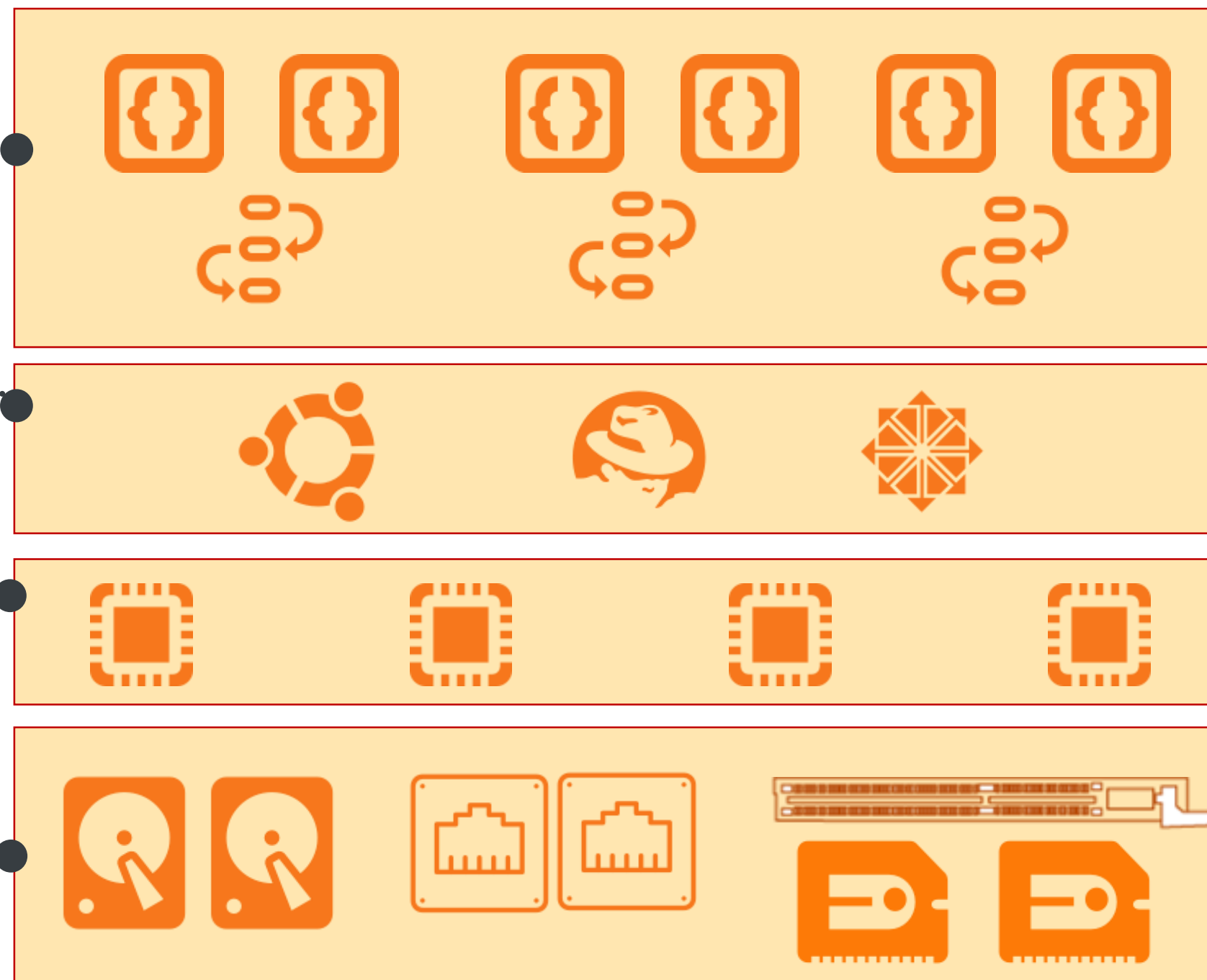
- 显示当前Top进程
- 显示Top进程CPU占用率和内存占用率

CPU级

- 显示CPU总体占用率、用户态和内核态占用率
- 显示CPU本地内存带宽和总内存带宽

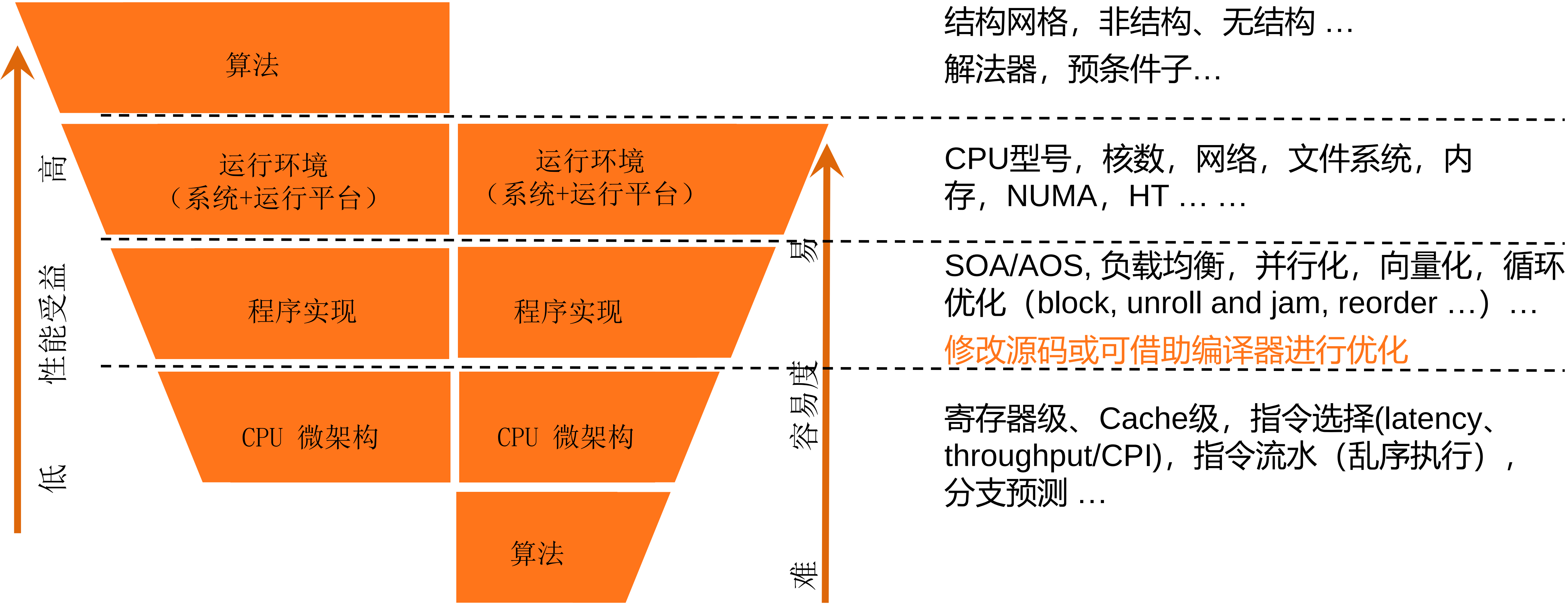
磁盘、网络与外设

- 显示磁盘吞吐量、磁盘利用率
- 显示网络总吞吐量、入方向和出方向吞吐量
- 显示GPU带宽、GPU利用率、显存利用率、功耗...



- 按作业运行检索，按作业周期进行查看
- 多指标同时时间轴显示，可对比分析

客户应用云上优化: 优化方法



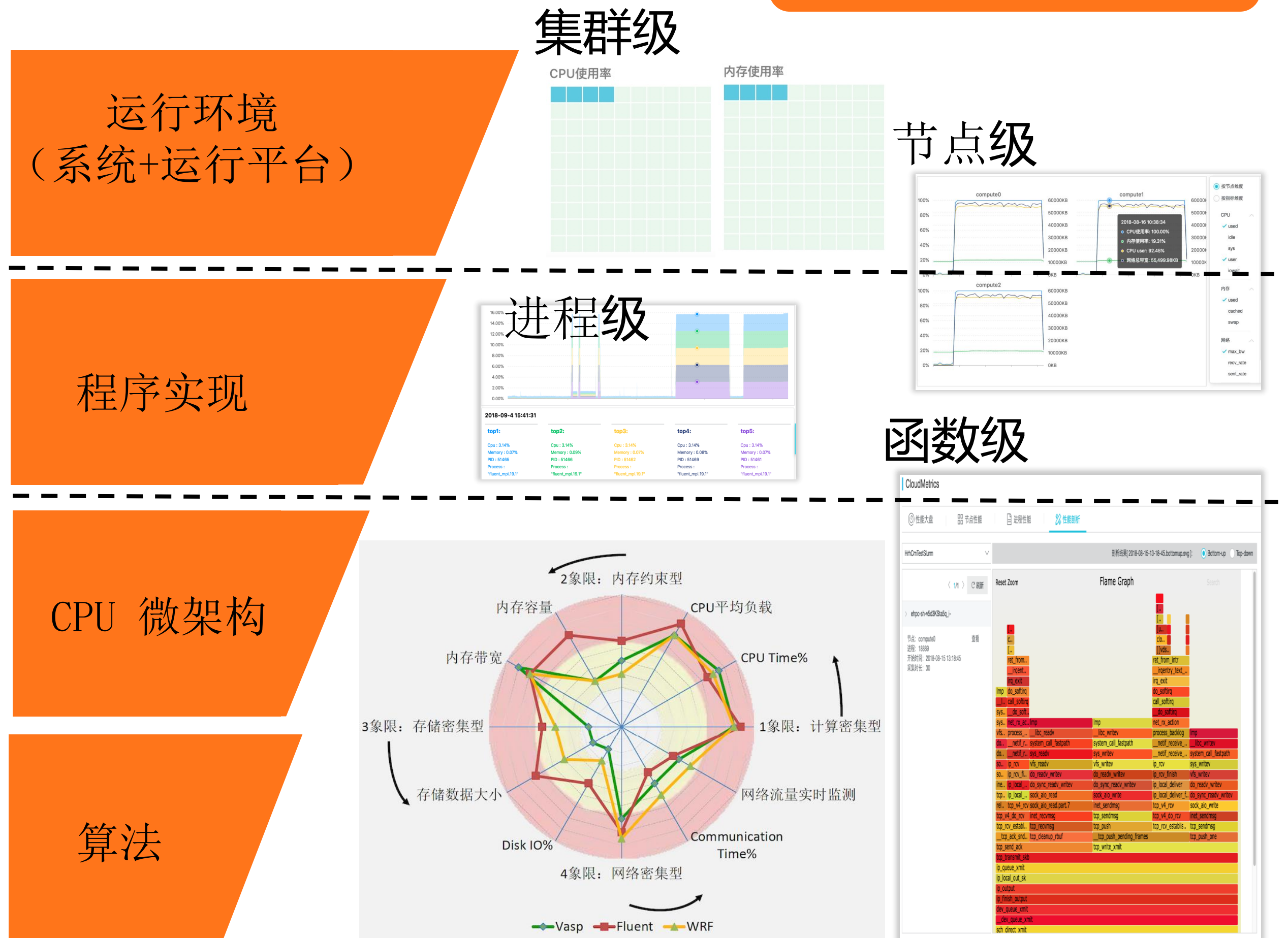
- 1. 运行环境优化, 收益高, 容易实现, 尤其是云上资源配置丰富、更换成本低的场景。
- 2. 性能优化的关键是定位性能问题, 确定优化点。

应用级性能监控工具——集谛

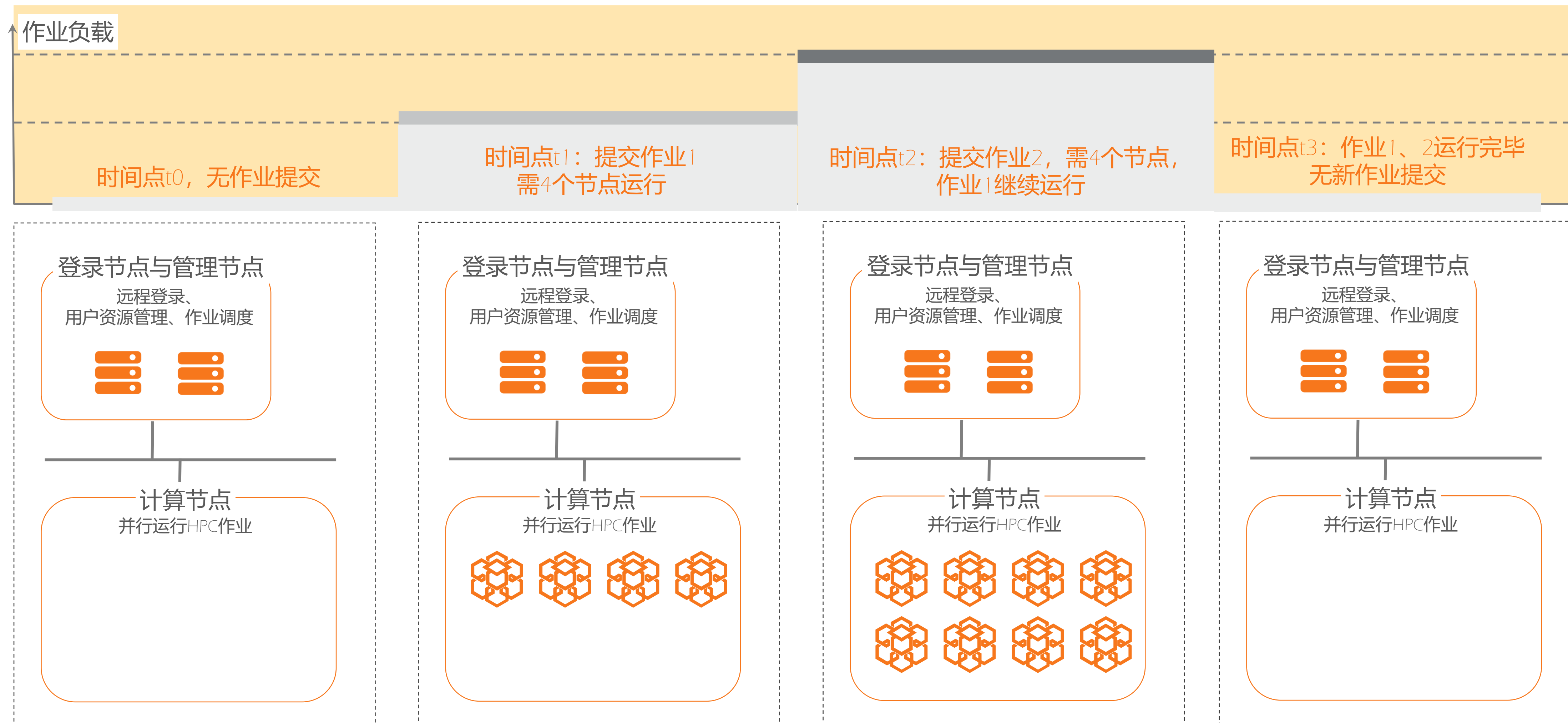
已上线，功能截图

• 层级递进的性能监控工具

- 集群级（性能大盘），了解集群所有节点负载；
- 节点级（系统资源饱和度），与作业调度系统绑定，查看节点相关度：
 - ✓ 按**节点维度**，查看指标间的变化关联；
 - ✓ 按**指标维度**，查看不同节点间的指标差异。
- 进程级（进程信息），查看节点占用OS时间片最高的进程信息；
- 函数级（热点函数），确定占用CPU资源最多的函数信息——**bottom-up**（热点函数）和**top-down**（逻辑调用关系）。

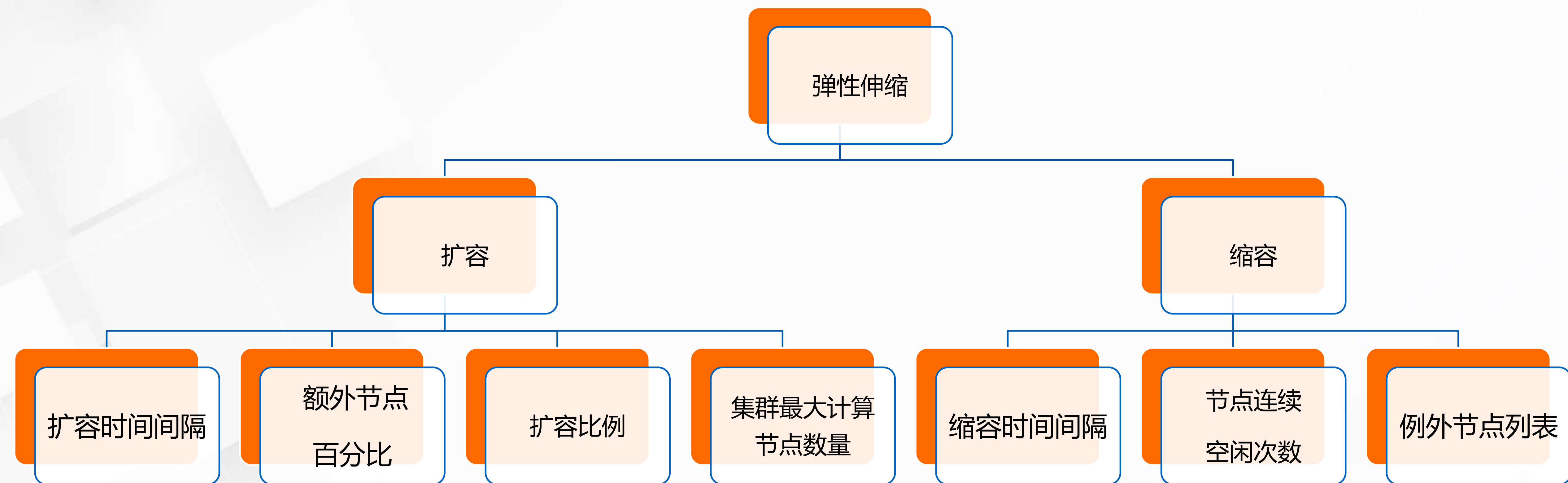


基于作业负载的弹性伸缩



- E-HPC 管控服务监控调度器作业资源需求, 动态创建/删除计算节点 (数据保留在共享存储), 按实际负载弹性计费。
- 可针对调度器队列制定差异化的伸缩策略, 支持抢占式实例, 支持跨可用区伸缩, 降低客户使用成本。

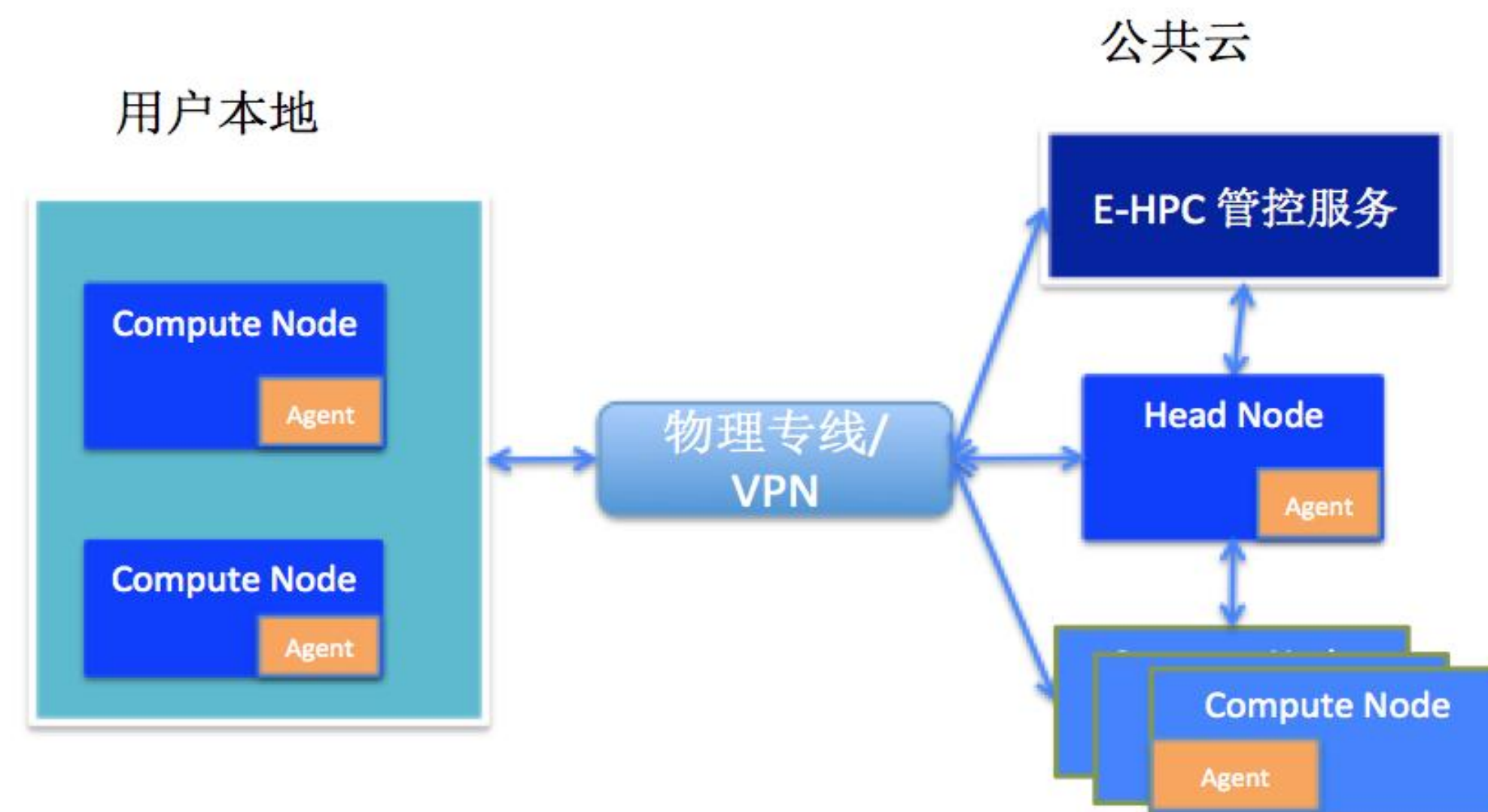
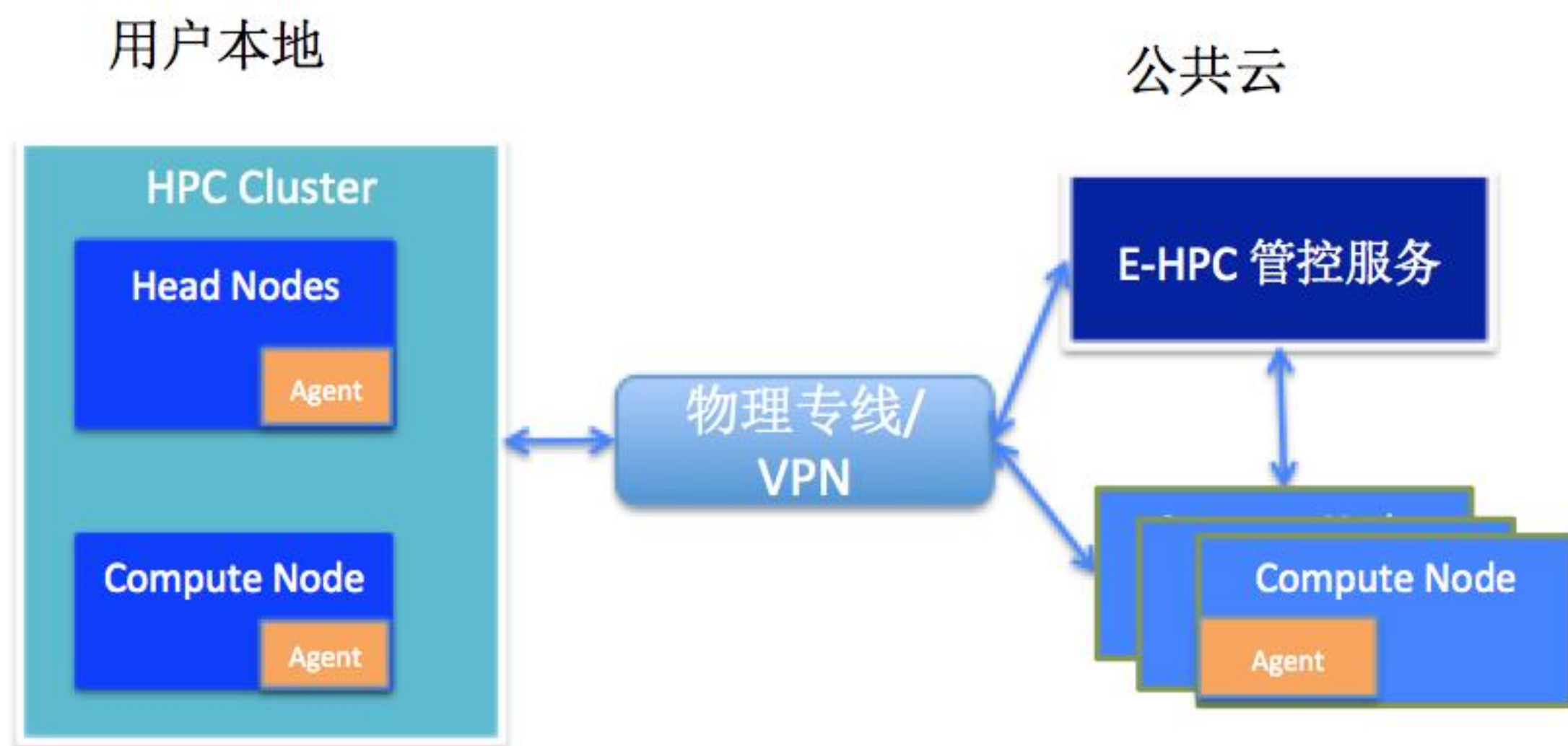
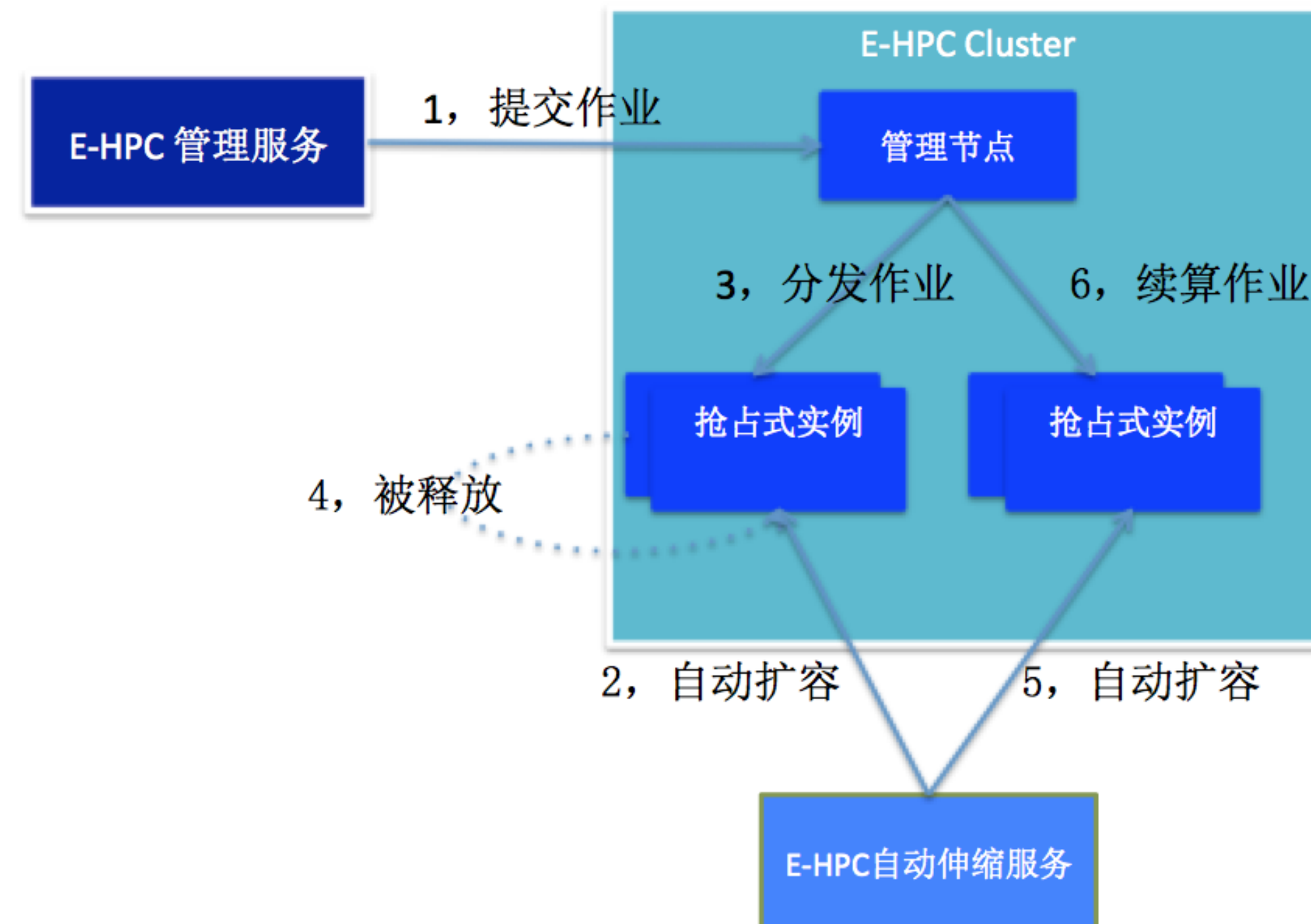
基于作业负载的弹性伸缩



E-HPC闲时计算方案与混合云

将成本作为重要因素的云上计算方案，利用以下技术将云上普惠的高性能计算服务提供给绝大部分HPC用户

- 抢占式实例：控制计算资源在期望的成本之下
- 自动伸缩：自动扩容抢占式实例，无需用户干预
- 断点续算：在计算资源被释放和重新获取之后，作业可以续算



- 在算例输入文件中设定步长
- 在作业提交脚本中设定checkpoint文件生成的时间间隔

```
#!/bin/sh
#PBS -l ncpus=4,mem=24gb
#PBS -j oe

export MODULEPATH=/opt/ehpcmodulefiles/
module load cuda-toolkit/9.0
module load gromacs-gpu/2018.1
module load openmpi/3.0.0

cd /home/testuser1/test/water-cut1.0_GMX50_bare/1536
gmx_mpi grompp -f pme.mdp -c conf.gro -p topol.top -o topol_pme.tpr
mpirun -np 1 gmx_mpi mdrun -ntomp 4 -noconfout -v -pin on
-nb gpu -s topol_pme.tpr -cpi state.cpt -cpt 2 -append
```

-cpi 选项用于指定checkpoint文件名，默认文件名是state.cpt
-cpt 选项指定checkpoint文件生成和更新间隔时间，用于下一次续算，单位是分钟
-append 选项指定续算产生的结果会继续写入先前的输出文件

```
vi water-cut1.0_GMX50_bare/1536/pme.mdp
设置nsteps=80000
```

断点续算应用算例：GROMACS

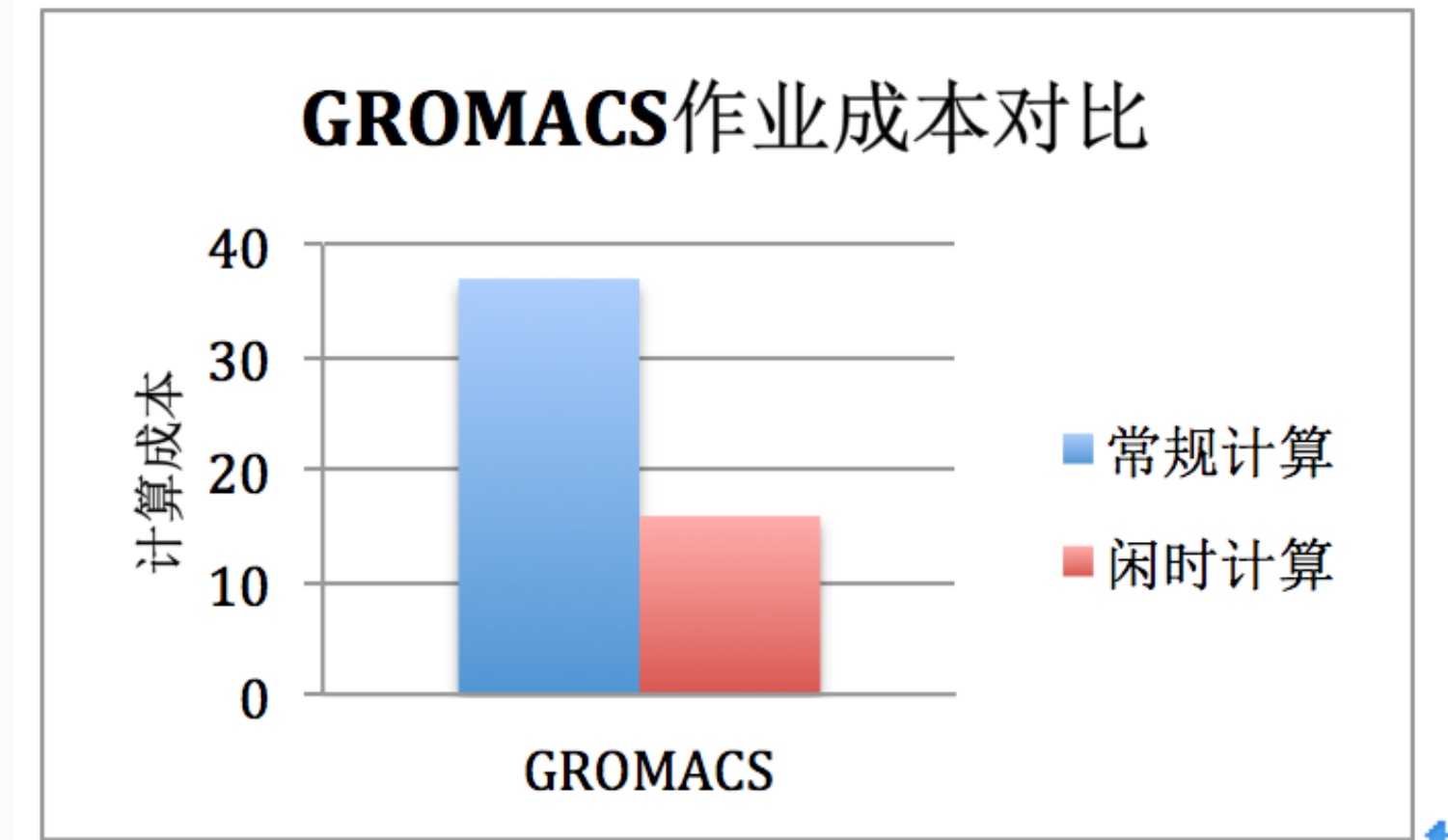


图 3|GROMACS 作业计算成本对比

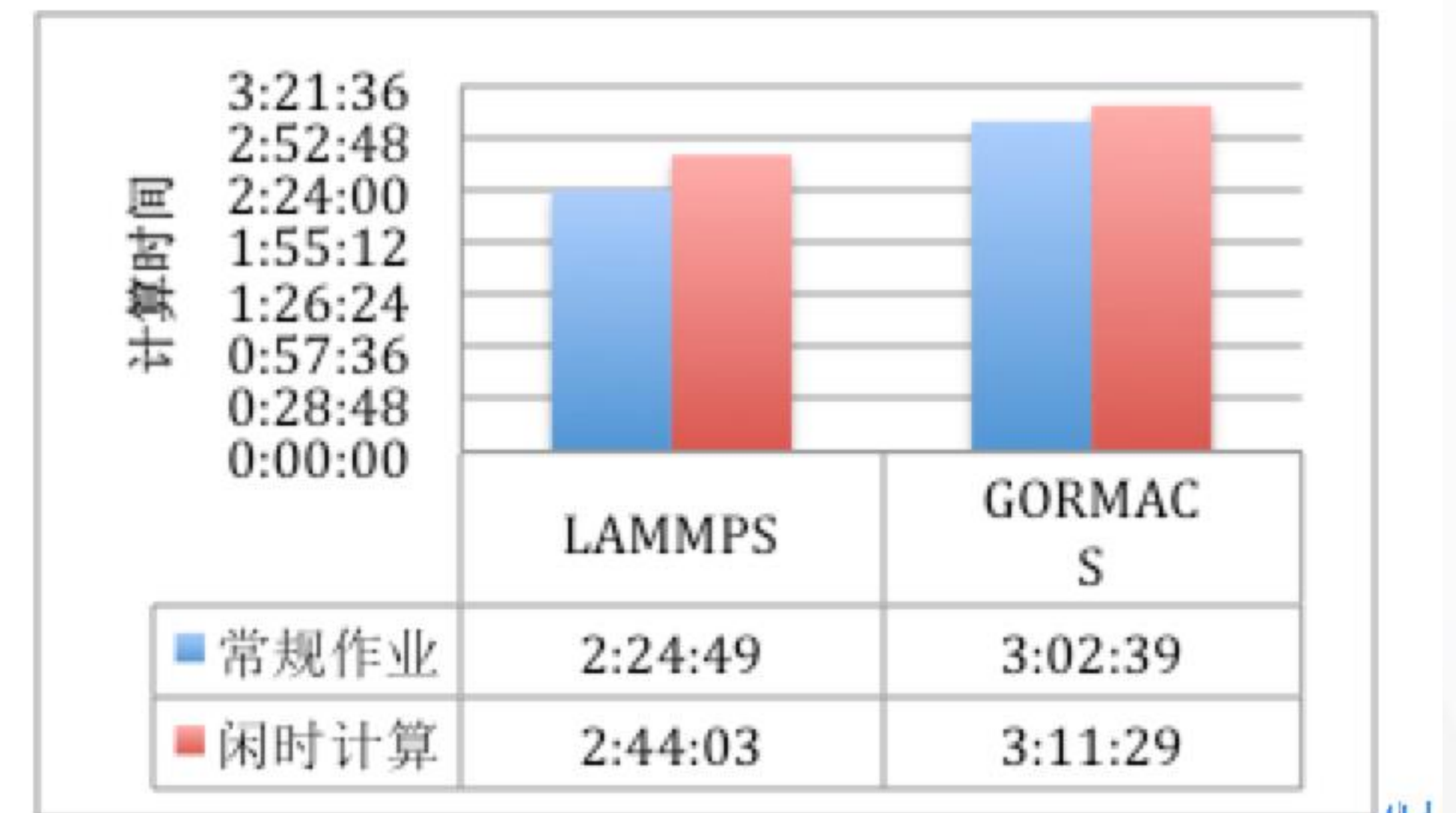


图 1 HPC 作业计算时间对比



创建集群

1.硬件配置

2.软件配置

3.基础配置

* 可用区: 请选择

* 付费类型: 包年包月 按量付费 竞价实例

部署方式: 标准 简易 One-box

推荐配置: 所有 裸金属服务器 GPU Skylake

计算节点 请选择 1

管控节点 请选择 2

登录节点 请选择 1台

若选择按量开通集群，8核及8核以上机型需要申请才能使用。[前往申请。](#)
[查看机型实例规格](#)

高级配置

显示拓扑

下一步

取消

配置清单

高级配置

集群名称	
地域	
可用区	
计算资源总核数	0
计算节点	
管控节点	
登录节点	
调度器	
软件配置	

总价: 0.00 /小时

登录节点将默认绑定公网IP，按流量付费，收费标准详见[《EIP计费说明》](#)

E-HPC集群使用NAS服务作为存储资源，NAS收费标准详见[《NAS计费说明》](#)

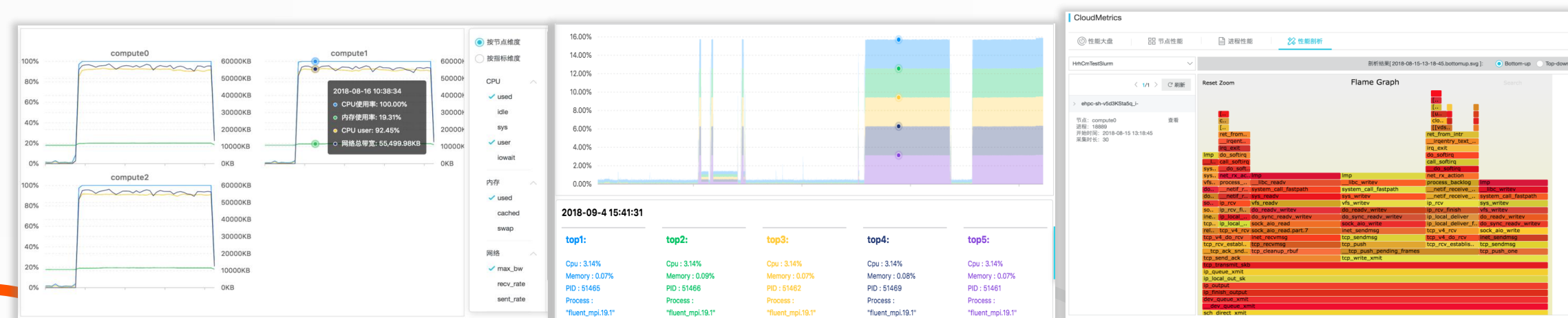
用户在一个地域的所有集群总核数小于200核(vCPU)不收费，高于200核(vCPU)的收费标准见[《E-HPC计费方式》](#)

咨询·建议

数据全流程可视化

作业运行——可视化性能分析

- 节点性能：按节点查看物理资源性能
- 进程性能：按进程查看软件运行性能
- 性能剖析：剖析进程，查看函数级性能数据



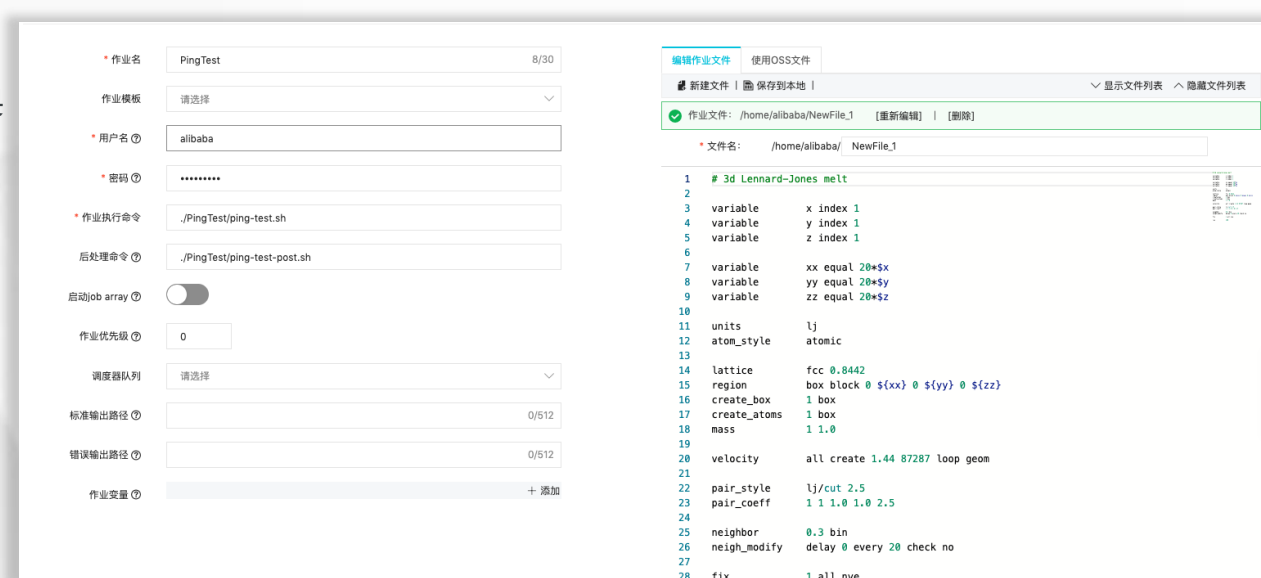
作业前处理

作业运行

作业后处理

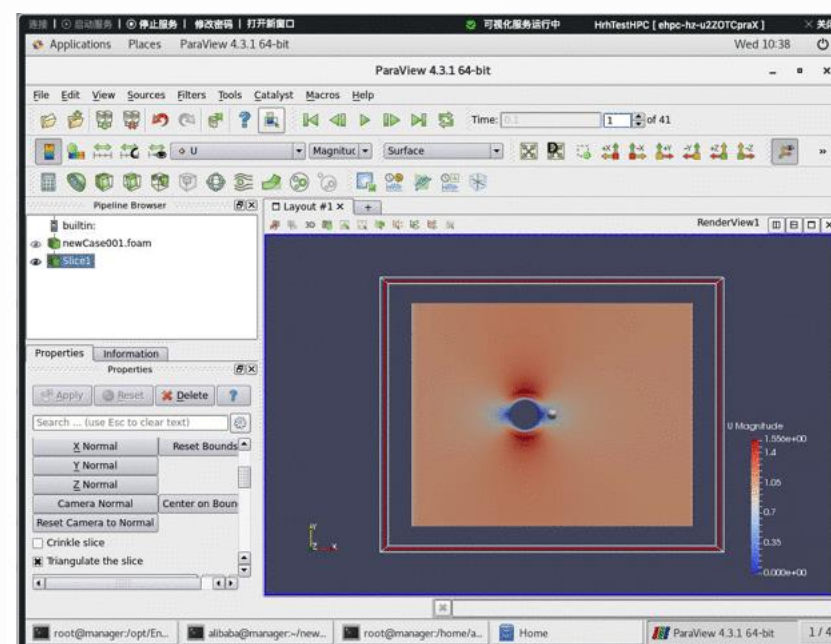
作业前处理——可视化编辑

- 调度器配置项可视化
- 控制台直接作业文件编辑器



作业后处理——计算结果可视化

- 控制台直接提供Web化的图形桌面
- 支持云图站等多种可视化服务

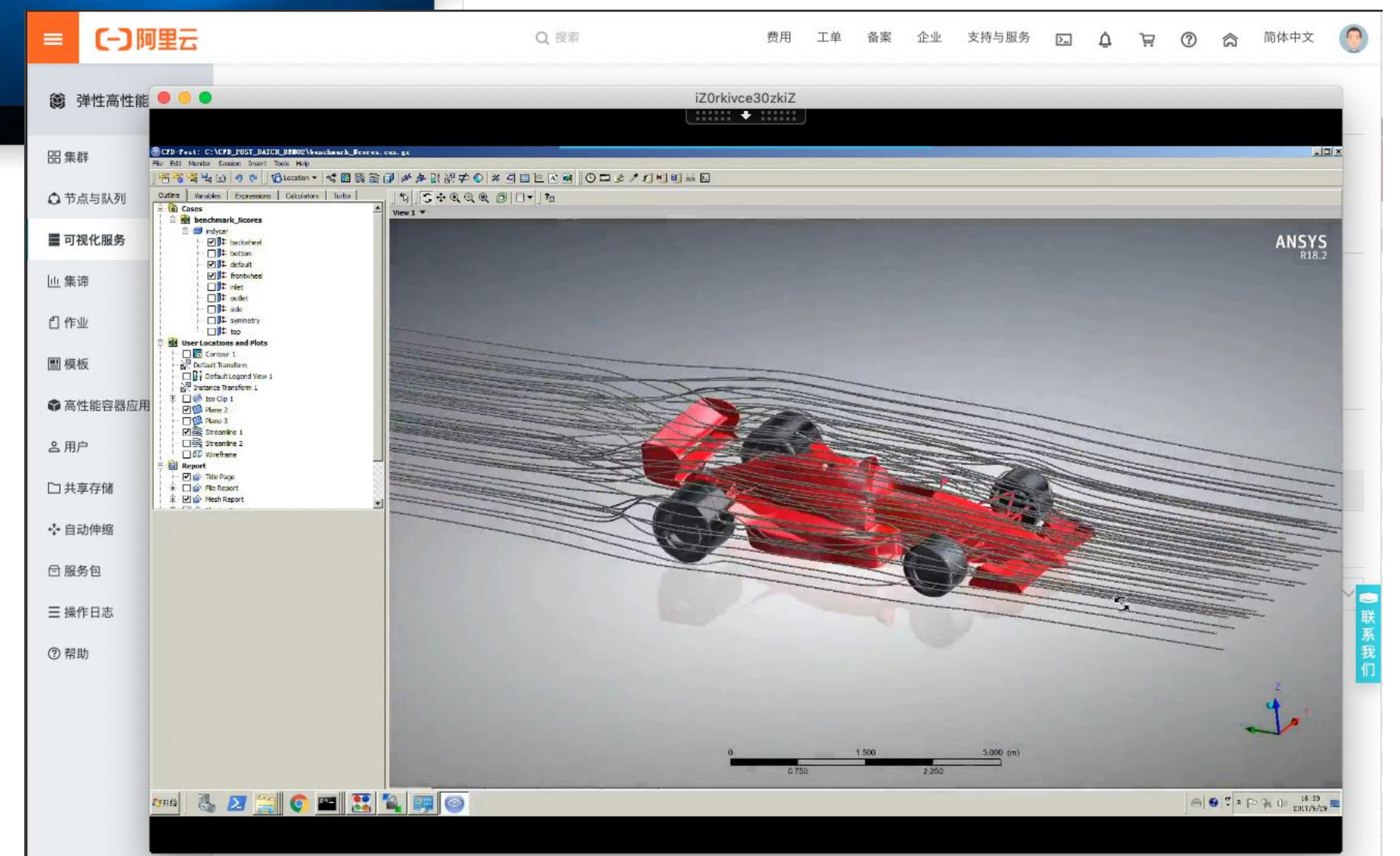
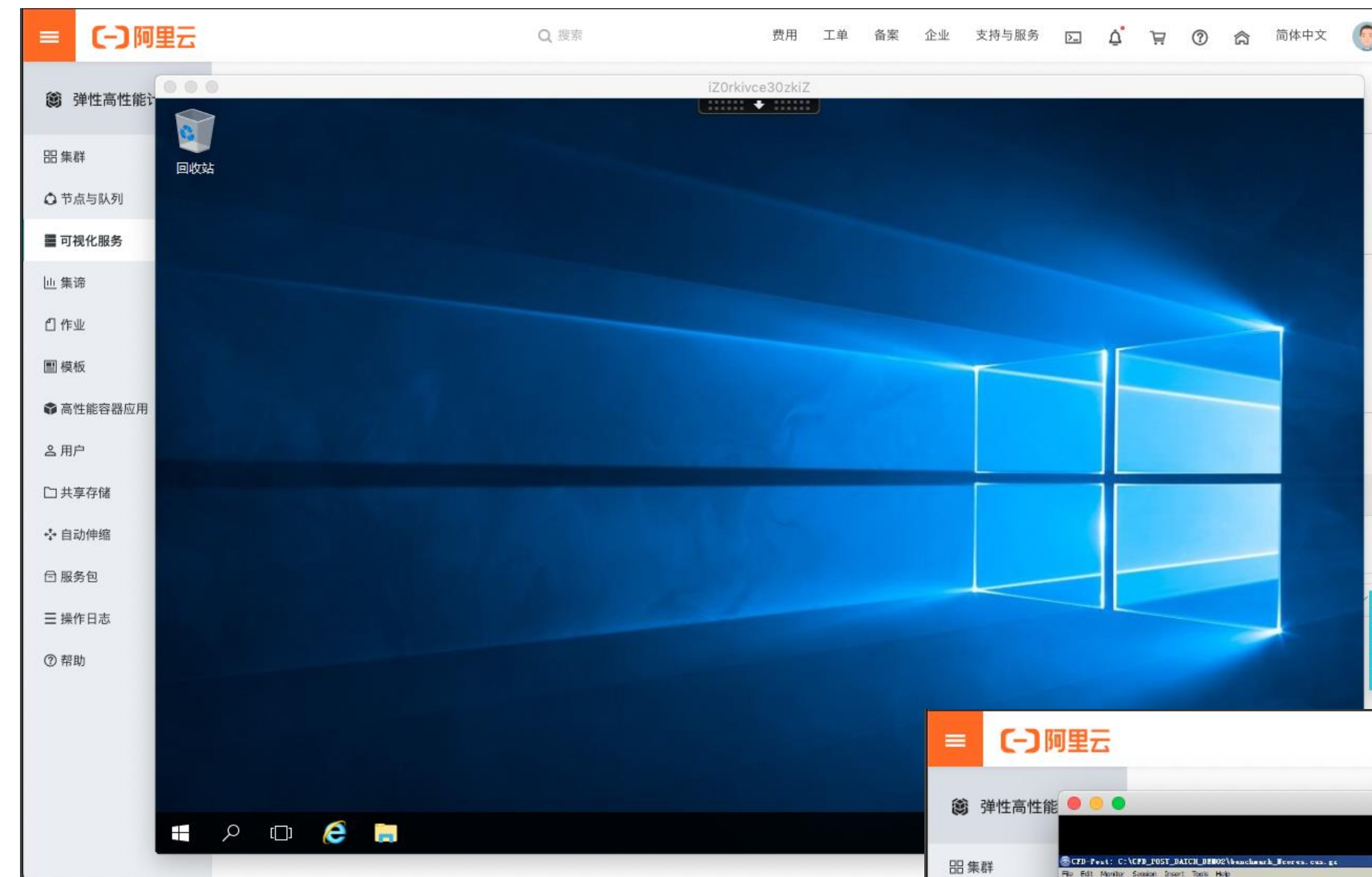


控制台远程可视化

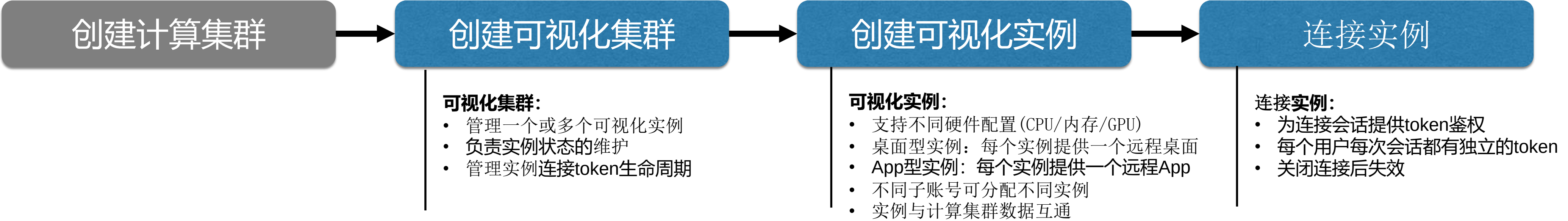
创建实例

连接实例

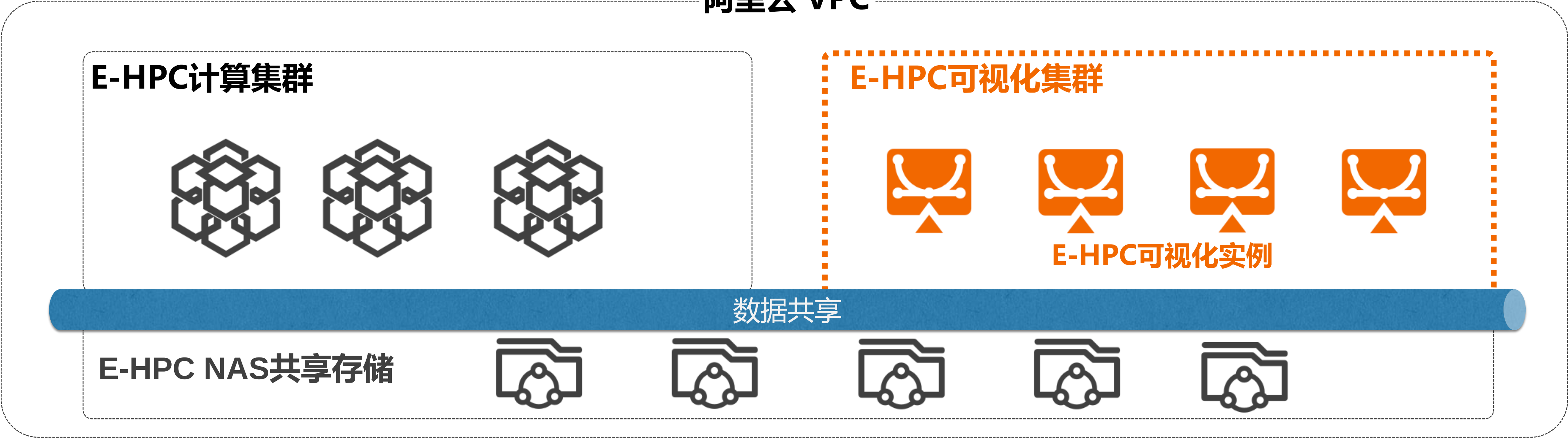
使用



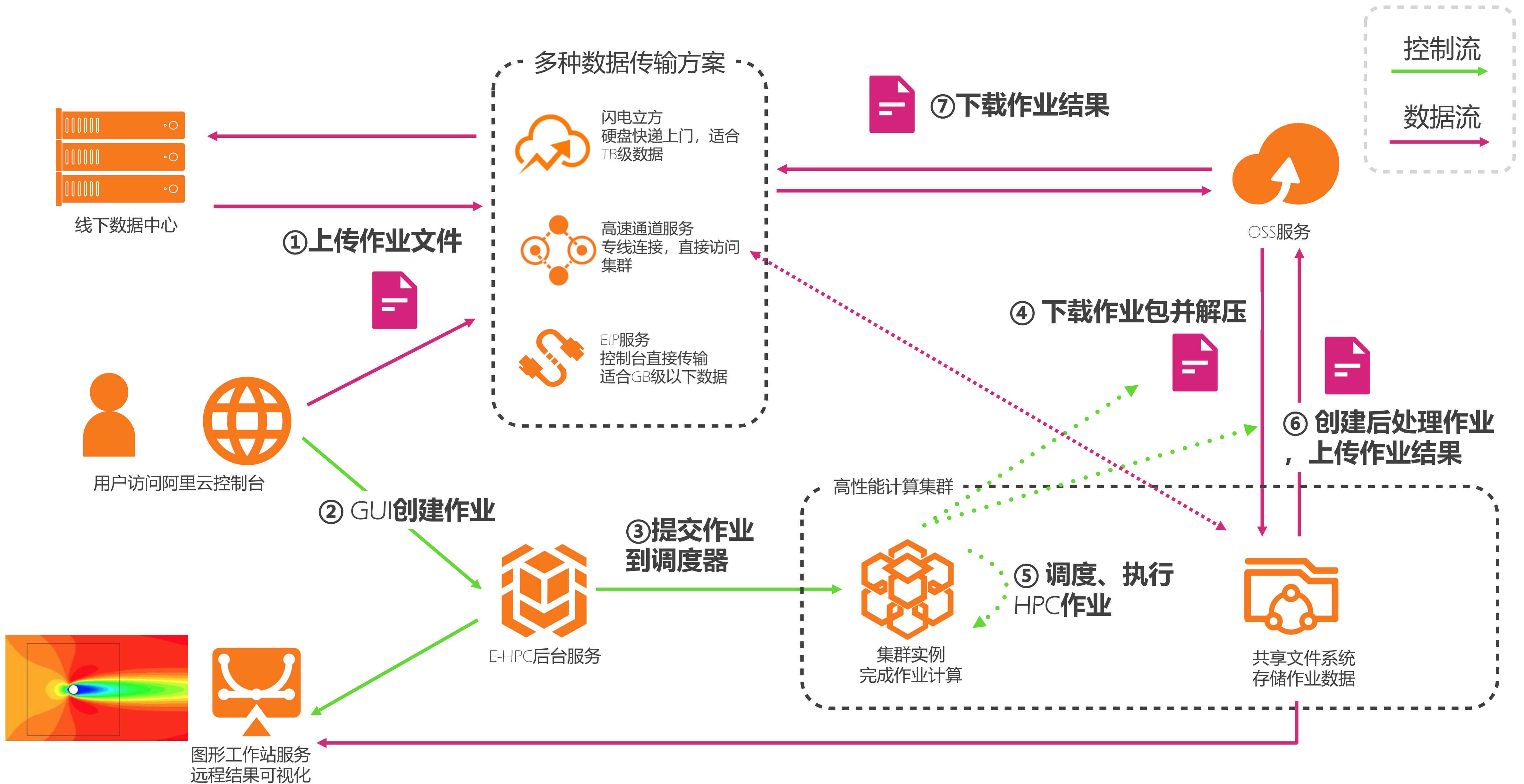
可视化设置流程



阿里云 VPC

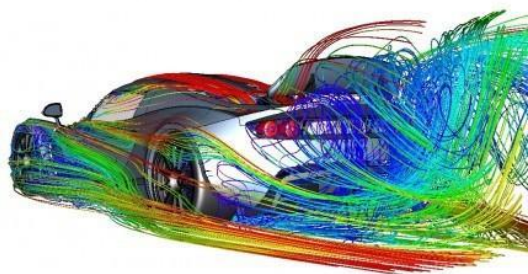


HPC workflows and data migration

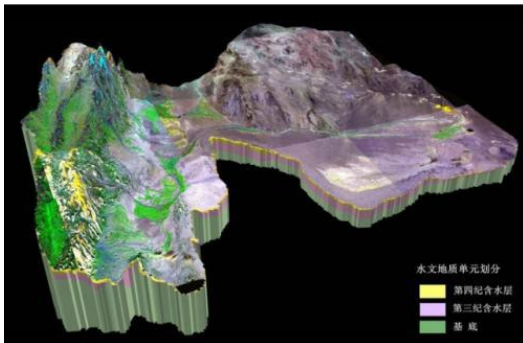


E-HPC 工业计算业务架构：被集成到SaaS

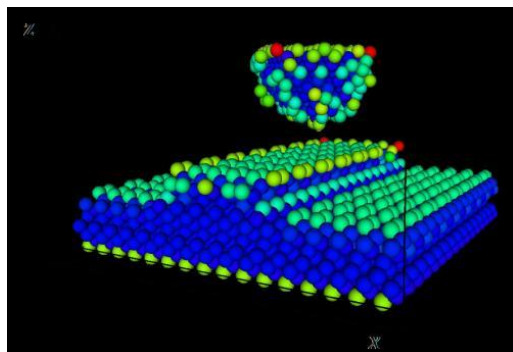
行业
客户



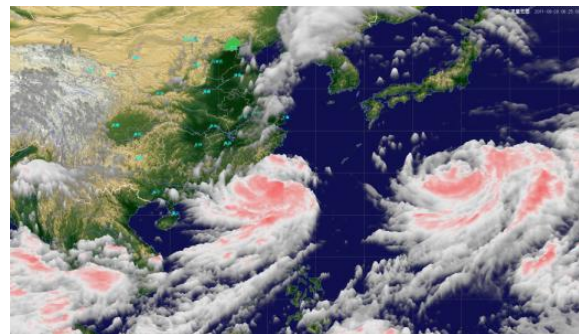
工业制造



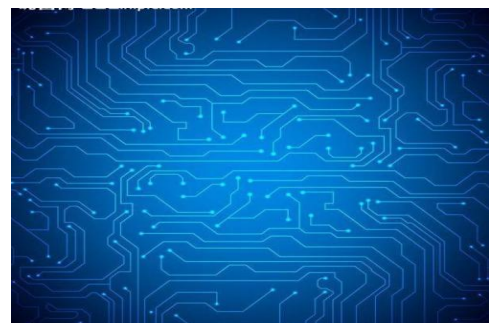
地质勘探



生物医药



气候气象



电子设计

E-HPC 通用解决方案支撑

混合云解决方案

可视化解决方案

SaaS
合作伙伴



ISV
合作伙伴



E-HPC OpenAPI

E-HPC ISV应用软件与运行环境管理

E-HPC 商业License管理

商业License生命周期管理

商业软件栈导入、部署与生命周期管理框架

商业License调度

E-HPC 作业管理与调度

E-HPC 资源生命周期管理



ECS



SCC



GPU



FPGA



VPC



Security



OSS



NAS/CPFS



闪电立方



阿里云开发者社区

扫码加入社群
与志同道合的码友一起
Code Up



云计算开发者技术群

谢谢！