



中国科学技术大学

University of Science and Technology of China

# 基于LNet的多网络互联实践

李会民

中国科大超级计算中心

<http://scc.ustc.edu.cn>

2019-08-24

# 中国科学技术大学简介



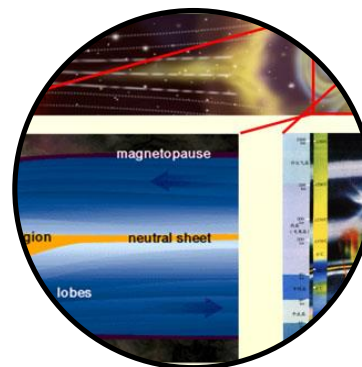
中国科学技术大学  
University of Science and Technology of China



隶属于中国科学院  
首批“985工程”C9高校之一  
首批“211工程”高校之一  
唯一参与国家知识创新工程的高校  
首批双一流建设高校



国家实验室/研究中心：2个  
其他国家级科研机构：10个  
中科院级科研机构：23个  
省部级重点科研机构：27个  
联合共建科研机构：25个



国家理科基础科学研究和教学人才培养基地：  
6个  
国家生命科学与技术人才培养基地：1个  
一级学科国家重点学科：8个  
二级学科国家重点学科：4个  
国家重点培育学科：2个  
安徽省A类重点学科：1个  
安徽省B类重点学科：19个



中国科学院和中国工程院院士：55人  
发展中国家科学院院士：17人  
国家万人计划领军人才：30人  
青年拔尖人才：13人  
国家杰出青年科学基金获得者：115人  
优秀青年科学基金获得者：108人  
教育部长江特聘教授：51人  
国家级教学名师：7人

以前沿科学和高新技术为主的综合性全国重点大学

# 中国科大超算中心概况



中国科学技术大学  
University of Science and Technology of China

2003年成立，作为校级科学计算平台，校六大公共支撑平台之一

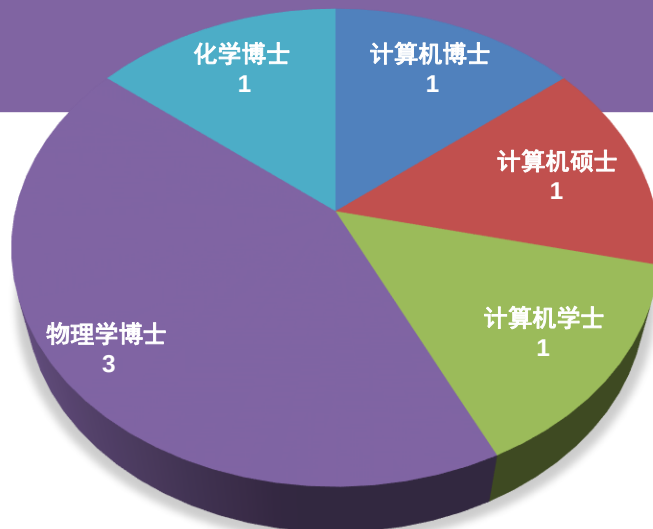
- 隶属于网络信息中心
- 为校双一流建设服务，面向校内所有的院系、科研机构、教师、学生提供高性能计算服务，以及设备托管、培训和应用优化等服务

建设以实际需求为主，不追求规模，注重应用效果，充分发挥资源价值

- 在用计算设备约2800万元，理论双精度计算能力合计为644万亿次/秒（国内高校偏上水平）
- 2018年建设经费约5000万元（2017年11月校六大公共实验中心评比第二，申请经费全额获批）

## 人员精干

- 现有人员7人
- 网络信息中心其他部门人员协助



# 中国科大超算中心概况



中国科学技术大学  
University of Science and Technology of China

中国科学院超级计算环境合肥分中心（中科大）

中国国家网格(CNGrid)合肥中心和合肥运行中心

中国教育与科研网格(ChinaGrid)中科大子节点

国家超级计算天津中心中科大分中心

安徽省高校科研协作高性能计算公共平台

超级计算创新联盟理事单位

# 为教学科研服务



- 常见使用问题
- 常见使用问题
- 培训讲座
  - 2016年秋季学期用户培训资料
  - 2016年6月Intel高性能计算培训
  - 2015年11月用户培训资料
  - 2015 MATLAB中文论坛现场交流会-合肥站
  - 2014年9月国家超算天津中心培训交流会

- 曙光TC4800百万亿次超级计算机系统
  - 曙光TC4800百万亿次超级计算机系统用户使用指南
  - 曙光TC4800百万亿次超级计算机系统用户使用指南-简略版
  - Intel 2017.0.093编译器、调试器、MKL及MPI资料
  - Intel 2016.0.109编译器、调试器、MKL及MPI资料
  - Intel 2015.1.133编译器、调试器、MKL及MPI资料

- ChinaGrid高性能计算集群
  - ChinaGrid高性能计算集群用户使用指南
  - ChinaGrid高性能计算集群MIC服务器使用指南
  - Intel 2013版编译器、调试器、MKL及MPI资料

- 刀片及胖节点超级计算机系统
  - 刀片及胖节点超级计算机系统用户使用指南
  - Intel 2011版编译器、MKL、调试器资料

- 托管课题组的超算集群设备
- 运维用户超算系统

用户超算系统托管

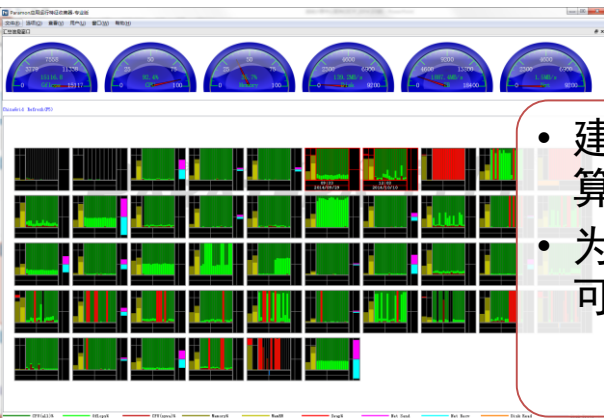
提供科学计算资源

用户培训技术支持

用户程序开发优化

- 建设和维护校级超算系统
- 为用户提供稳定、可靠的计算资源

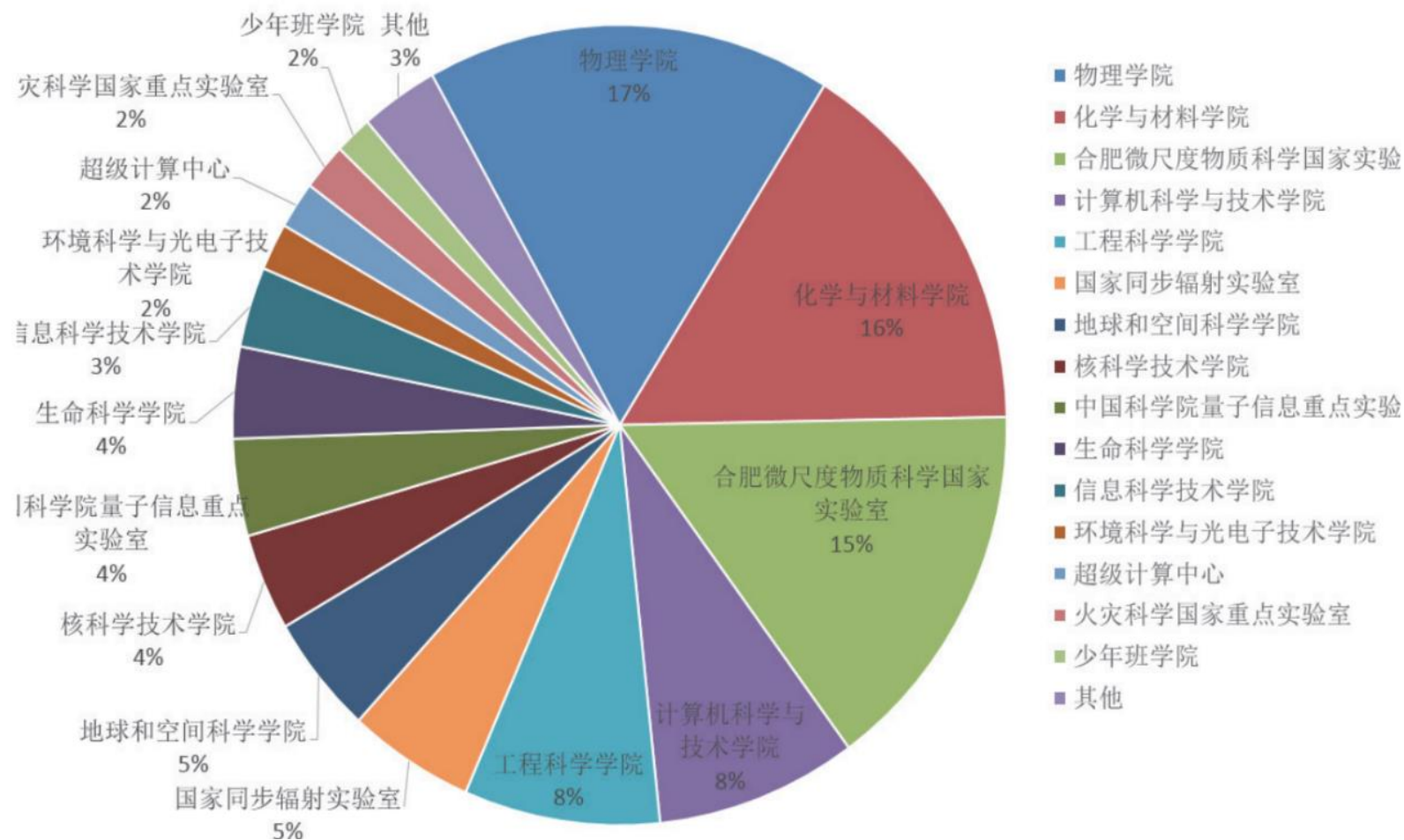
- 开发与优化用户的关键计算程序
- 开发校内高需求的计算模块与应用



```
Source Assembly CPU Time: Total Effective Time by Utilization Spin C Time T
684 int incw=1;
685 int incw=1;
686 //connect to cccoc* 31*ccoc* 31*ccococm1;
687 for(int i=0; i<incw; i++)
688 {
689     for(int j=0; j<incw; j++)
690     {
691         const int start1=incw*i+incw*j;
692         const int end1=incw*(i+1)+incw*(j+1);
693         const int start2=incw*i+incw*j;
694         const int end2=incw*(i+1)+incw*(j+1);
695         //for(int i=0; i<incw; i++)
696         for(int i=0; i<incw; i++)
697         {
698             const int start3=incw*i+incw*j;
699             const int end3=incw*(i+1)+incw*(j+1);
700             const int start4=incw*i+incw*j;
701             const int end4=incw*(i+1)+incw*(j+1);
702             int sum=0;
703             for(int k=0; k<incw; k++)
704             {
705                 // 1
706                 // 1
707                 int sum1=0;
708                 for(int l=0; l<incw; l++)
709                 {
710                     sum1+=1;
711                 }
712             }
713         }
714     }
715 }
```

# 超算用户

覆盖所有学院和主要实验室的教师和学生，  
目前有1606个系统账户（校外111个）



国家同步辐射实验室

合肥微尺度物质科学国家研究中心

量子信息与量子科技前沿协同创新中心

类脑智能技术及应用国家工程实验室

稳态强磁场科学中心

高能密度物理联合实验室

国家高性能计算中心(合肥)

核电子学国家重点实验室（联合）

热安全技术国家地方联合工程研究中心

大尺度火灾国际联合研究中心

火灾科学国家重点实验室

中国科学院合肥大科学中心

中国科学院量子信息与量子科技创新研究院

中国科学院星系宇宙学重点实验室（联合）

# 在用设备



中国科学技术大学  
University of Science and Technology of China

## 在用设备总投资约4000万元

- 机房：约1300万元（2013年建成，2016年扩建）
- 超算系统：约2756万元

理论双精度总计算能力644万亿次/秒，14512颗CPU核，理论可提供机时1亿CPU核小时/年

- CPU: 483万亿次/秒
- GPU: 140万亿次/秒
- MIC: 21万亿次/秒



示范项目奖

优秀教育数据中心奖

| 系统名                | 计算节点数 | 计算能力<br>(万亿次/秒)                       | 采购经费<br>(万元) | 建设时间<br>(年)        |
|--------------------|-------|---------------------------------------|--------------|--------------------|
| 曙光TC4600<br>超级计算系统 | 506   | 总519<br>(CPU:483, GPU:140,<br>MIC:21) | 2756         | 2014、2015、<br>2016 |



### 中国科学技术大学超级计算中心

数据中心设计等级：  
数据中心竣工时间：2013年12月  
数据中心用途：政府单位  
数据中心地址：安徽省合肥市



每年投入，累计超过1.1亿元，其中5000万元的新超算系统正在实施

1.5PB可用容量并行存储：DDN GS7990，GRIDScaler(GPFS)，实测IO：

- 聚合：1小时持续顺序读带宽20GB/s，1小时持续写带宽18GB/s
- 单流：1小时持续写带宽4GB/s，1小时持续顺序读带宽5GB/s



# 超算系统运行情况汇总



中国科学技术大学  
University of Science and Technology of China

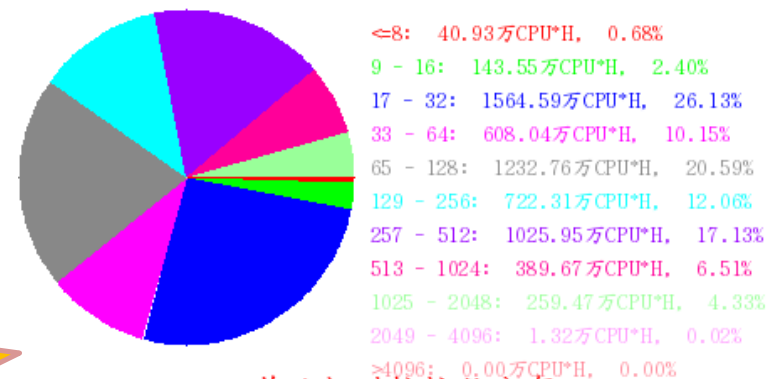
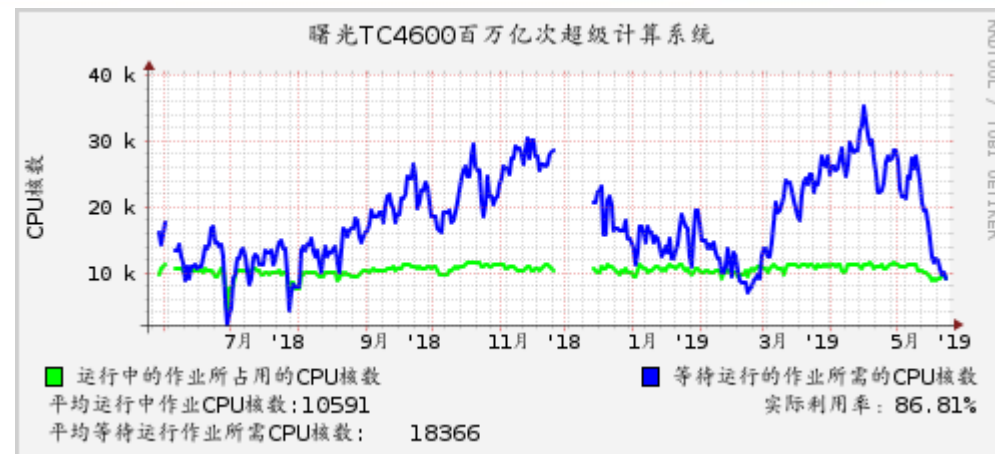
2018年8月22日-2019年8月23日

完成作业：32万个

实际使用计算机时：6421万CPU小时  
(校外615万)

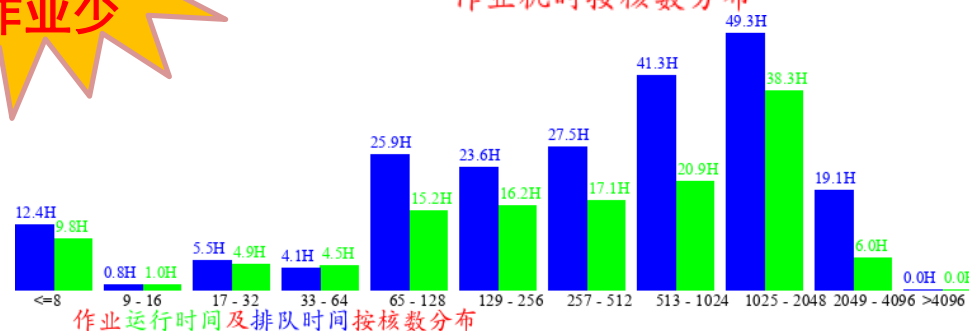
## 运行特点

- 实际利用率高：主要系统曙光TC4600百万亿次超级计算系统，在收费的情况下达**86.81%**
- 排队时间长：排队时间长，平均排队时间>运行时间，中等规模作业排队**2天以上**
- 作业规模有待提升：**超过千核CPU**的单个作业仍难于运行
- 目前用户需求旺盛：**现有资源远远无法满足需要**



千核  
作业少

作业机时按核数分布



# 用户成果产出与积分奖励



中国科学技术大学  
University of Science and Technology of China

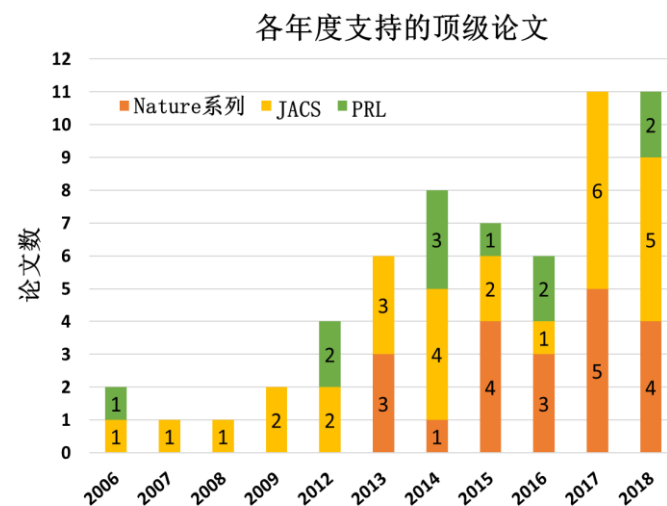
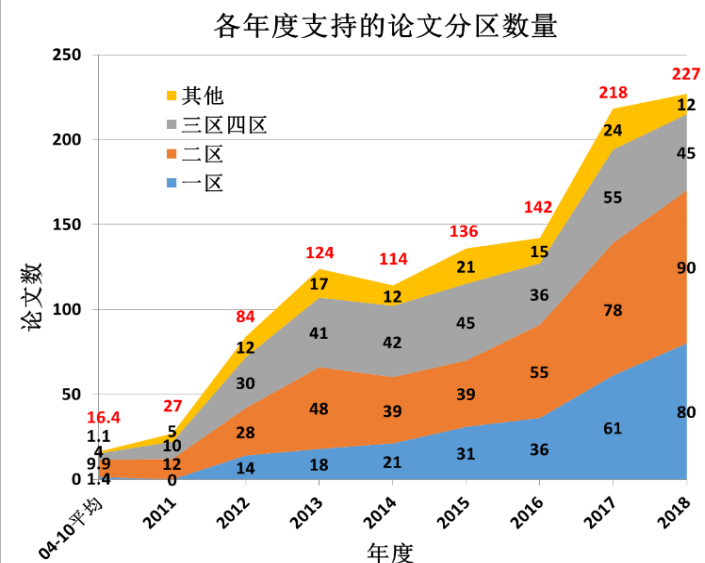
## 超算中心支持用户取得了丰硕的科研成果

据不完全统计，2018年支持了96项科研项目，总经费1.76亿元，产出227篇论文，一区、二区文章占比分别为35%、40%。

自2004年到2018年，支持发表致谢论文1171篇，其中Nature及其子刊、PRL与JACS三类顶级期刊59篇，2019年已支持发表2篇Science/Nature正刊论文、3篇子刊论文。

- 支持的高质量论文数量多，且分布广泛，支持的约20篇Nature/Science系列论文来自十多个课题组
- 发表论文的计算领域主要集中在原子分子材料计算领域，包括2018与2019年的全部顶级论文

2018年奖励1491积分（1积分奖励免费机时1万CPU核时）：化学与材料学院891积分（60%）；物理学院297积分（20%），合肥微尺度物质科学国家研究中心88分（6%），物理化学类占比86%。



论文目录: <http://scc.ustc.edu.cn/zclw/>

# 2016年底扩容：InfiniBand or OPA



中国科学技术大学  
University of Science and Technology of China

## 担心

OPA性能不好

无法与原有InfiniBand  
Lustre文件系统整合

老应用不支持



## 策略

允许OPA，至少让EDR降价

应用会发展，相信Intel实力

明确要求与原有Lustre系统整合，应标商承担风险

# 2016年底超算实际扩容方案



中国科学技术大学  
University of Science and Technology of China

## • CPU计算节点，194个：

| 数量  | CPU                       | 总CPU核数 | 内存    | 硬盘           | 计算网络                   |
|-----|---------------------------|--------|-------|--------------|------------------------|
| 150 | 2*E5-2680 v4, 共28核        | 4200   | 128GB | 1*240GB SSD  | 100Gbps Intel OPA      |
| 40  | 1*E3-1240 v5(3.5GHz), 共4核 | 160    | 32GB  | 1*500GB SATA | 100Gbps EDR InfiniBand |
| 4   | 8*E7-8860 v4, 共144核       | 576    | 1TB   | 2*240GB SSD  | 100Gbps Intel OPA      |

## • NVIDIA Tesla GPU计算节点，4个：

| GPU                | CPU                           | 内存    | 硬盘          | 计算网络              |
|--------------------|-------------------------------|-------|-------------|-------------------|
| 2*NVIDIA Tesla K80 | 2*Intel Xeon E5-2680 v4, 共28核 | 128GB | 2*240GB SSD | 100Gbps Intel OPA |

## • Intel Xeon融核计算节点，8个：

| CPU                                           | 内存   | 硬盘          | 计算网络              |
|-----------------------------------------------|------|-------------|-------------------|
| 1*Intel Xeon Phi 7210(64核, 1.3GHz, 2.66万亿次/秒) | 96GB | 1*160GB SSD | 100Gbps Intel OPA |

## • Lustre SSD闪存并行存储系统：

| 存储阵列    | 磁盘                | IO节点 | 网络         | FC         |
|---------|-------------------|------|------------|------------|
| 联想S3200 | 24*400GB EMLC SSD | 2    | 56Gpbs FDR | 2*16Gbps双口 |

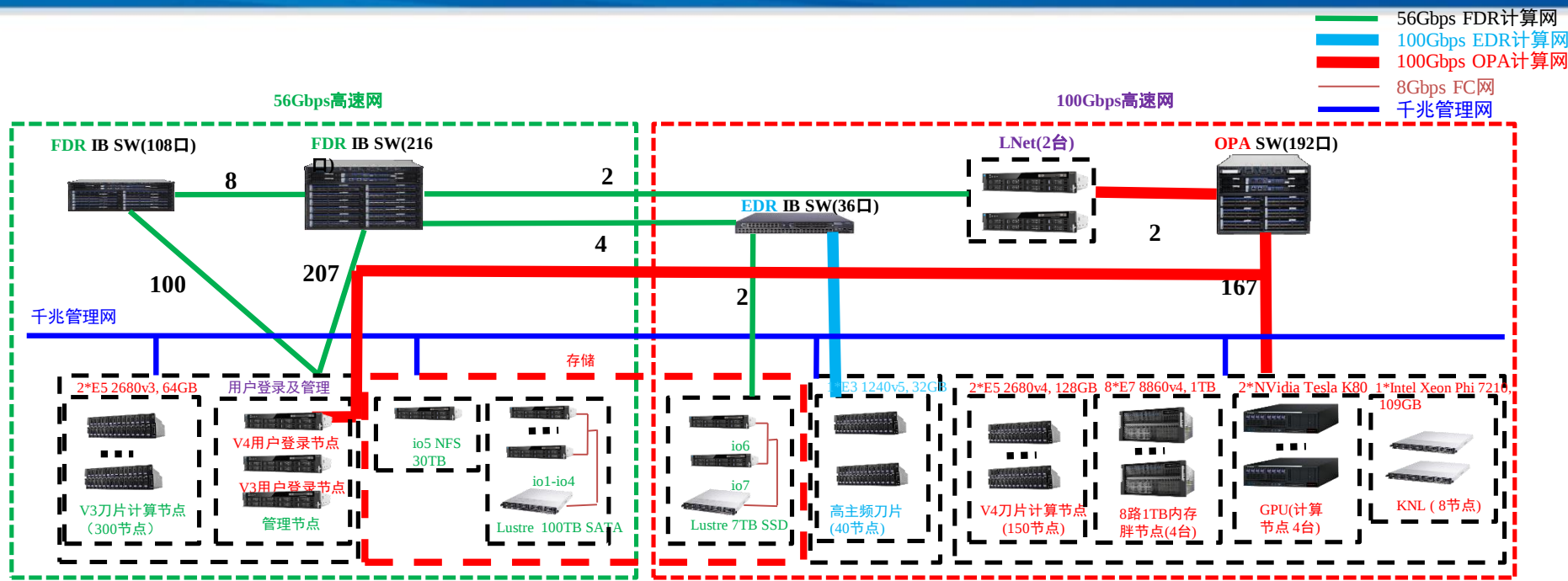
## • 其它：

- 100Gbps高速网络：192口Intel OPA交换机、36口Mellanox EDR InfiniBand交换机
- LNet路由节点、用户登录节点

# 扩容后系统拓扑图



中国科学技术大学  
University of Science and Technology of China



506个计算节点，共12200颗CPU核心，8颗Intel Xeon Phi融核(MIC) KNL和8颗NVIDIA K80卡，总双精度峰值计算能力为每秒519万亿次

- CPU: 482.82万亿次/秒
- GPU: 14.96万亿次/秒
- Intel Xeon Phi融核: 21.28万亿次/秒
- 100Gbps和56Gbps高速互联
- 1个管理节点
- 2个用户登录节点
- 2套Lustre存储+1套NFS存储:
  - 7(5+2)个IO节点
  - 2个Lustre LNet路由节点



# InfiniBand存储



中国科学技术大学  
University of Science and Technology of China

## Lustre SATA并行存储系统

- 曙光DS800-G10存储，34块4TB SATA硬盘，实际可用空间102TB
- Lustre并行IO节点4个：
  - 一块56Gbps FDR InfiniBand卡
  - 一块8Gbps FC卡

## Lustre SSD闪存并行存储系统

- 联想S3200存储，24块400GB SSD硬盘，实际可用空间7TB
- Lustre并行IO节点2个：
  - 一块56Gbps FDR InfiniBand卡
  - 两块16Gbps 双口FC卡

## 普通NFS存储

- 曙光I620-G20机架式服务器，实际可用空间30TB
  - 12块4TB 3.5寸7.2K NL-SAS硬盘
  - 1块2GB Cache RAID卡
  - 1块56Gb FDR InfiniBand双端口HCA卡

## Lustre 客户端-OPA

10.11.100.0/16 ib1/o2ib1



## LNet路由节点

10.10.100.11-12 ib0/o2ib0

10.11.100.11-12 ib1/o2ib1



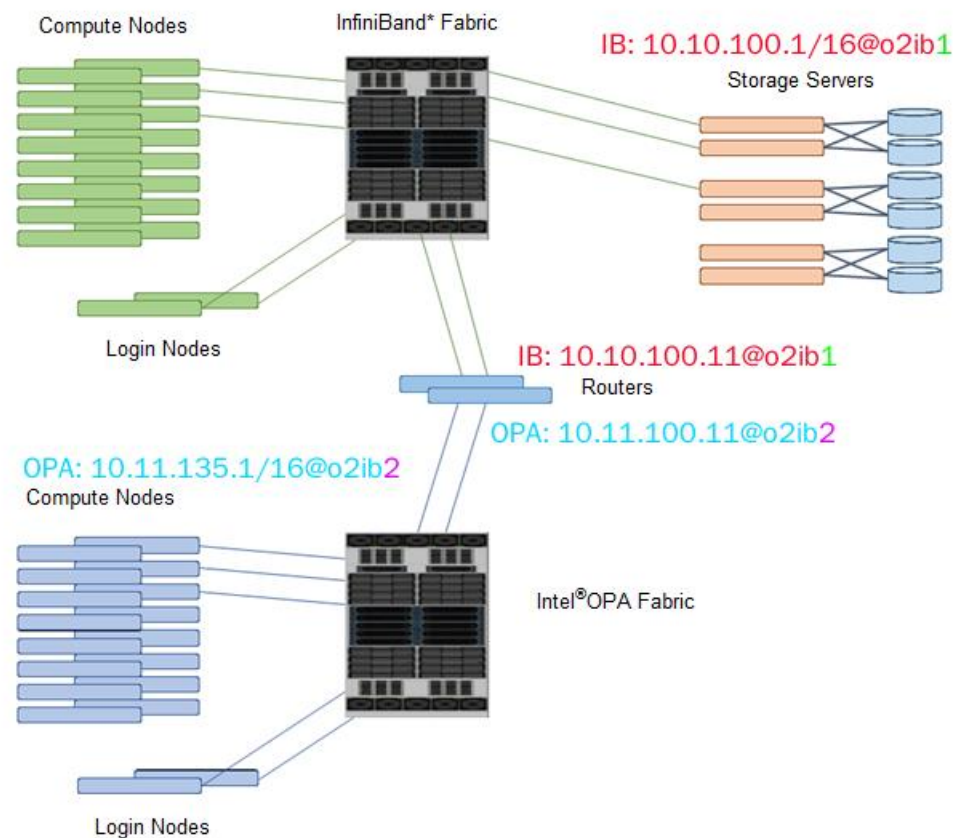
## Lustre服务端-InfiniBand

10.10.100.1-7 ib0/o2ib0

## 主要功能

Linux IP  
路由器

LNet  
路由器



## Intel Omni-Path Fabric Linux IP路由器

- 用于Intel OPA与InfiniBand高速网路之间的互联
- 通过IPoIB进行互联

## Intel Omni Path光纤采用IPoFabric

- 传统IB的采用IPoIB
- 从软件层面看，类似IPoIB，采用ib\_ipoib驱动，通过ib0和/或ib1端口发送接收IP通讯

# LNet路由器



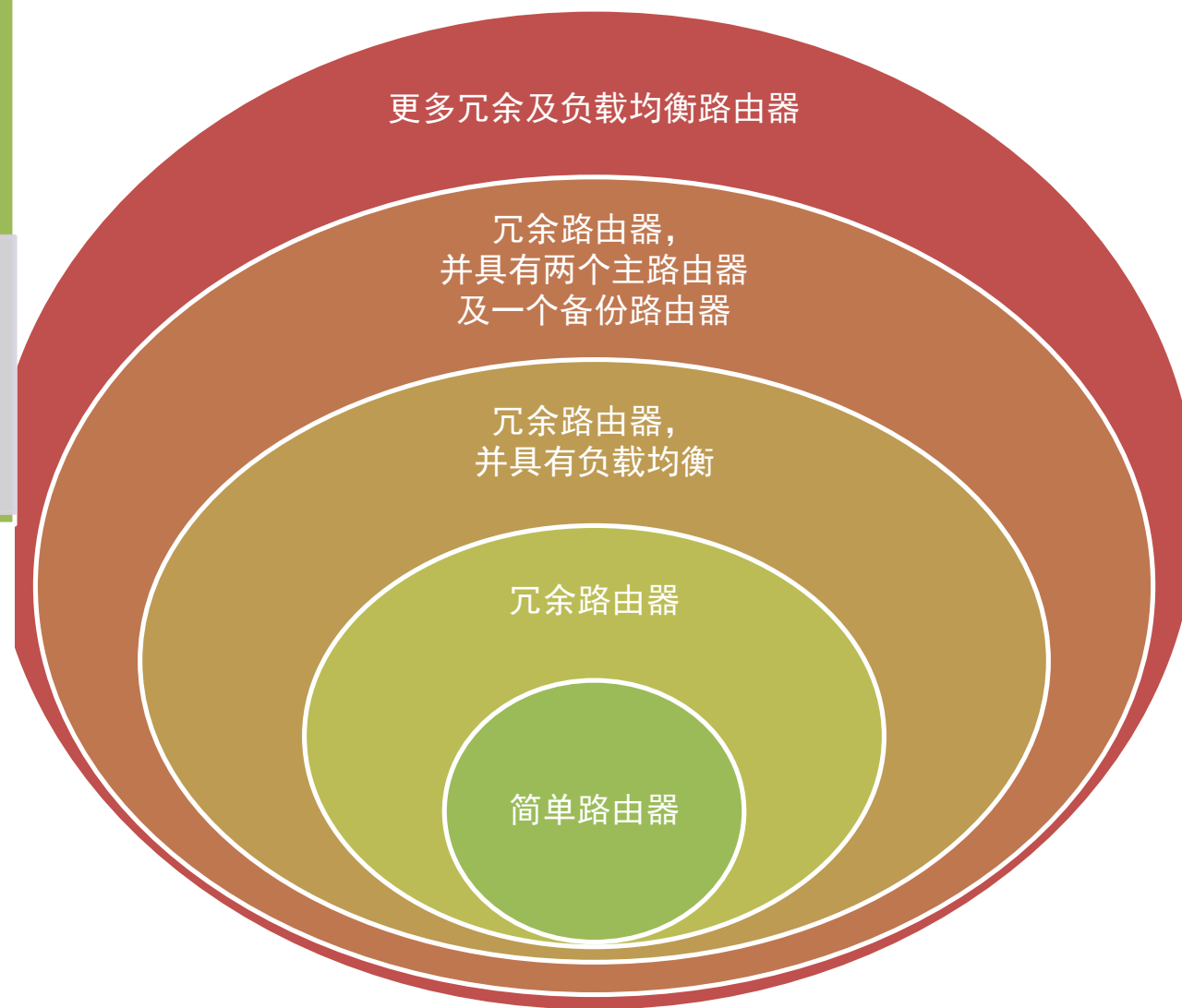
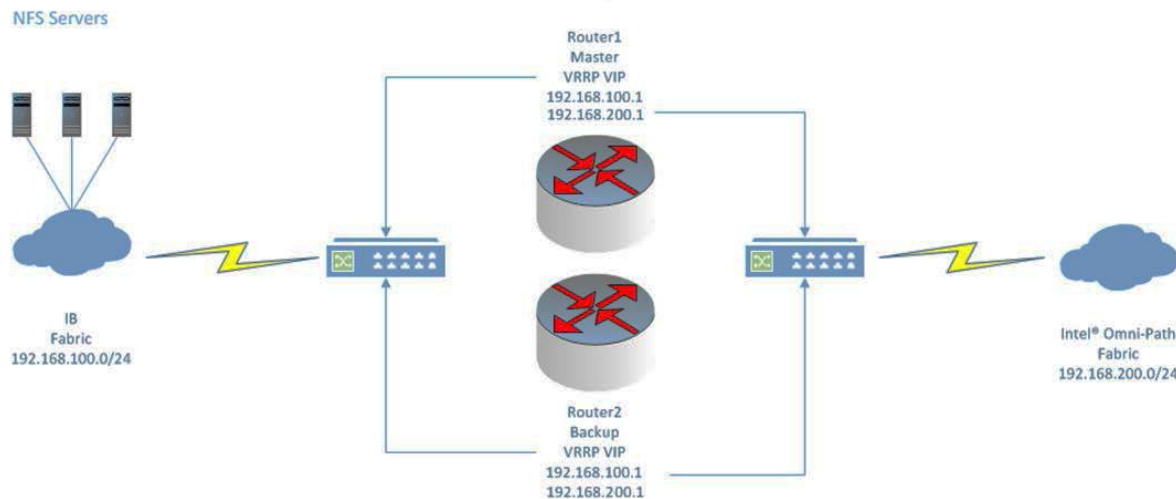
中国科学技术大学  
University of Science and Technology of China

支持冗余，可以设置LNet路由器池来做冗余、负载均衡及高可用性

LNet路由器被设计成能相互发现，并且作为一个池

可以监控彼此间健康  
和通讯状态

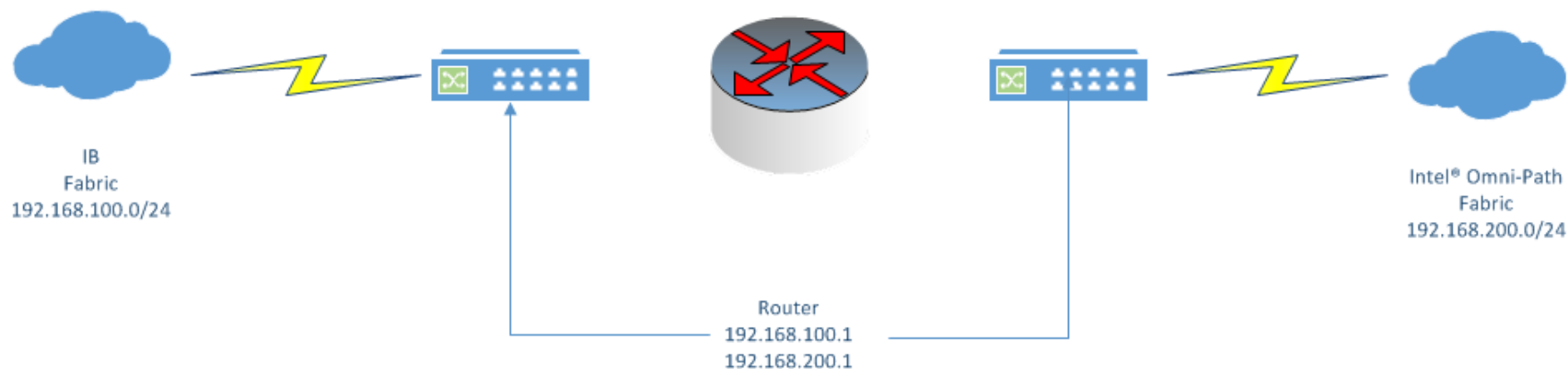
如某路由器失效，其它  
会自动接管



最简单的Intel Omni-Path与IB之间的路由器

不具有故障恢复、不具有负载均衡

不建议实际系统使用



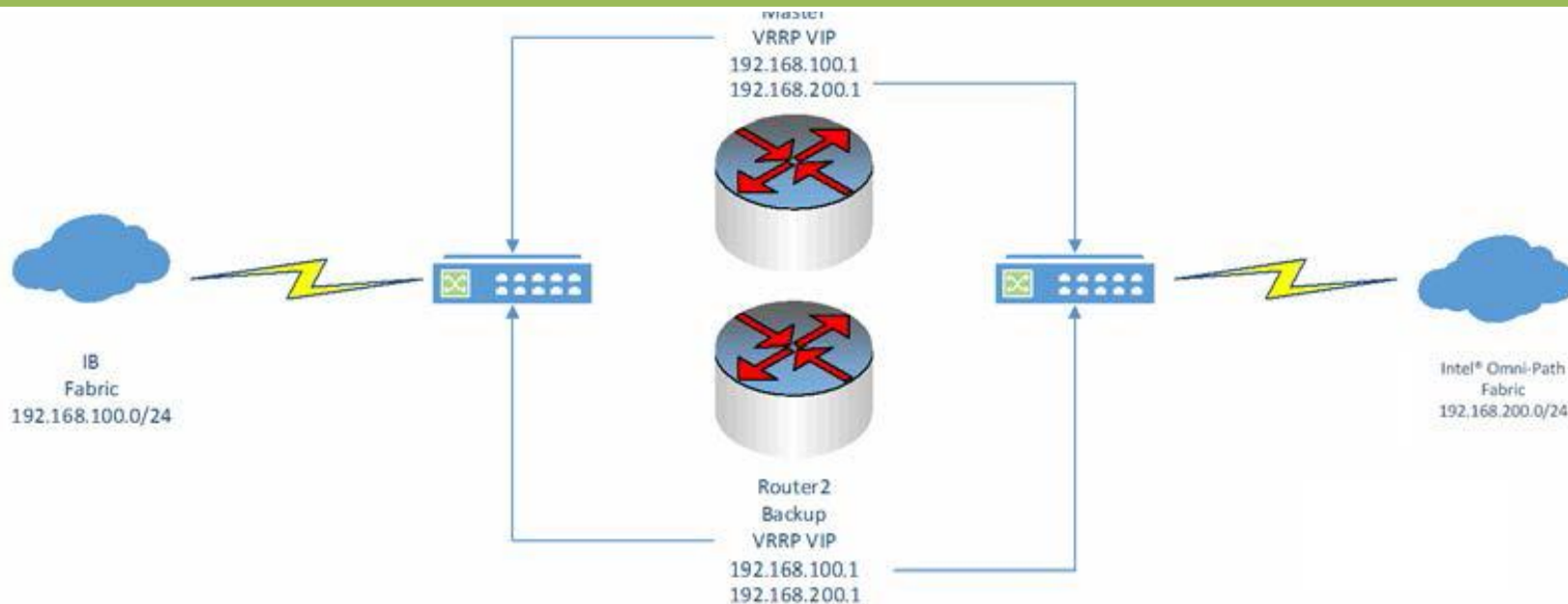
# 冗余路由器



中国科学技术大学  
University of Science and Technology of China

含有两个路由器

通过运行Virtual Router Redundancy Protocol (VRRP)提供故障冗余及恢复功能：一个路由器实效时，另一个自动接管



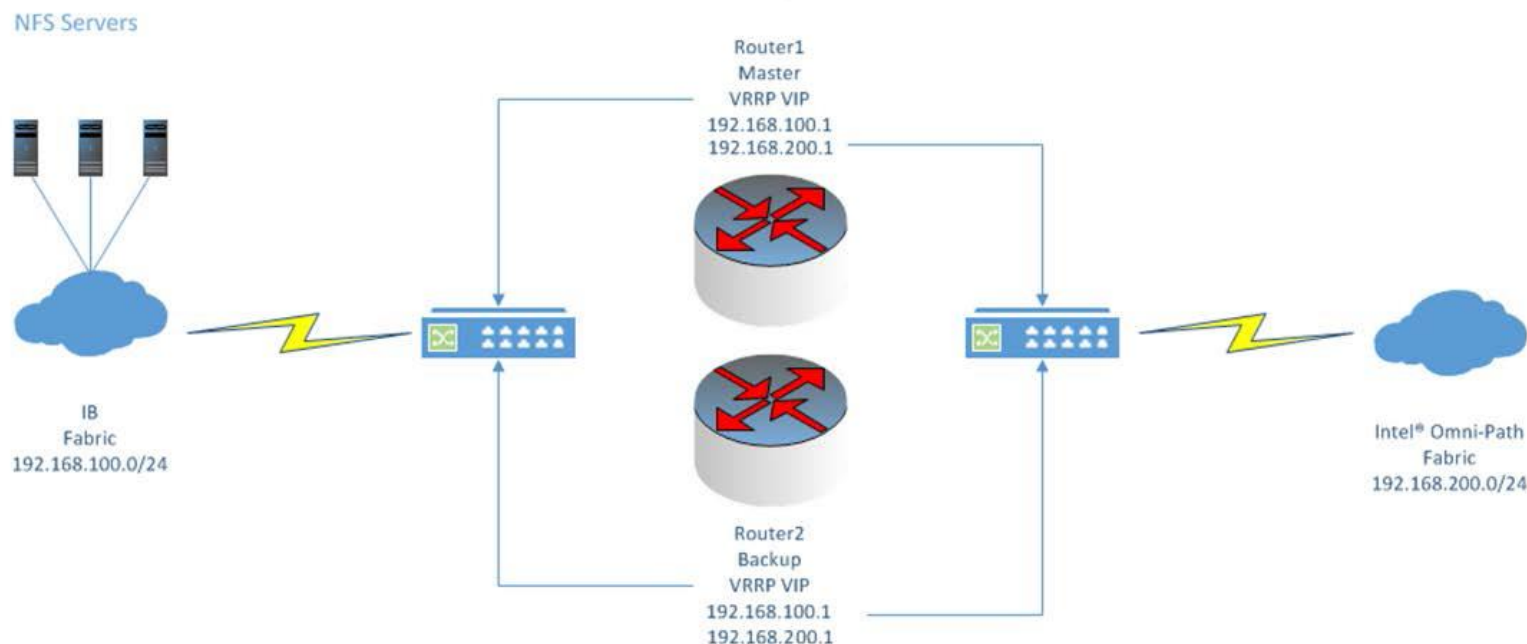
# 冗余路由器，并具有负载均衡



中国科学技术大学  
University of Science and Technology of China

两个路由器，通过运行VRRP和IP Virtual Server (IPVS) 提供故障恢复和负载均衡

利用Keepalived等为NFS等提供路由转发



# 冗余路由器，并具有两主一备

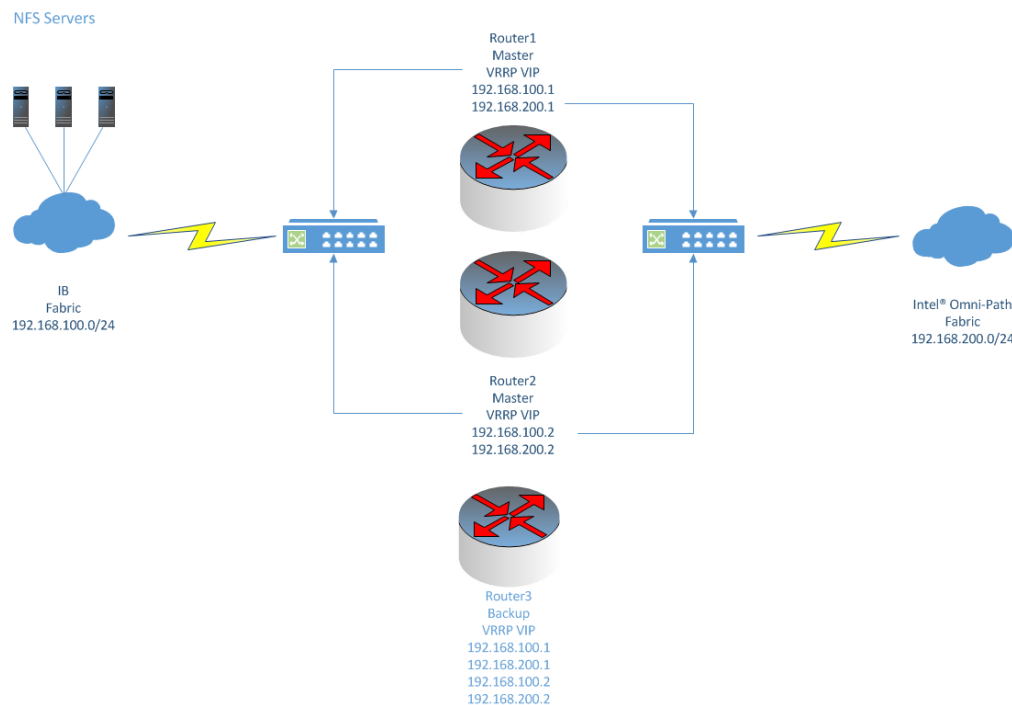


中国科学技术大学  
University of Science and Technology of China

含有三个路由器，两个主路由器，一个备份路由器

计算节点可以配置为使用不同路由降低路由负载

如果其中一个路由失效，备用路由自动顶替



# 路由节点设置路由转发



中国科学技术大学  
University of Science and Technology of China

临时设置，系统重启后实效，需重新设置

- # echo 1 > /proc/sys/net/ipv4/conf/ib0/forwarding
- # echo 1 > /proc/sys/net/ipv4/conf/ib1/forwarding

永久设置，系统重启后自动生效，修改/usr/lib/sysctl.d/00-system.conf

- net.ipv4.conf.ib0.forwarding=1
- net.ipv4.conf.ib1.forwarding=1

以上设置，只在IB/Intel® Omni-Path网卡启用IP转发

## 设置所需IP采用经过该路由

- `ip route add 192.168.200.0/24 via 192.168.100.1`

# LNet路由



中国科学技术大学  
University of Science and Technology of China

Lustre文件系统具有独特能力可以在不同网络拓扑上运行同样的全局命名空间

Lustre自带的LNet组件提供此抽象层

LNet是从Lustre中出来的独立项目，可以用于其他项目

LNet支持：Ethernet、InfiniBand、Intel® Omni-Path、legacy fabrics (ELAN和MyriNet)、其他特定计算光纤如Cray\* Gemini、Aries和Cascade

LNet是Linux Kernel空间的一部分，以支持完全RDMA通讯和zero copy通信

Lustre可以采用单个RPC初始化多OST读写，这将允许客户端利用RDMA访问数据，不管传输数据总量

LNet被设计成采用LNet路由提供最大灵活性连接不同网络拓扑

LNet的路由能力提供桥接两种不同网络的有效协议，如从Ethernet到InfiniBand或采用不同光纤拓扑，如Intel Omni-Path Architecture(OPA)和InfiniBand

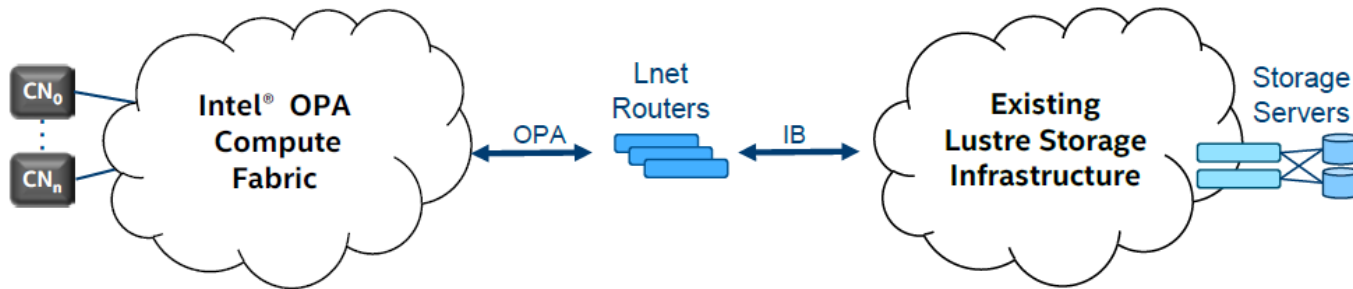
# LNet路由器节点



中国科学技术大学  
University of Science and Technology of China

LNet是一种Lustre客户端

在LNet路由器节点上只运行Lnet服务，无需运行Lustre计算节点客户端服务



## 两种语法

- routes=**dest\_inet** [hop] [priority] router\_NID@src\_inet; \  
  **dest\_inet** [hop] [priority] router\_NID@src\_inet
- **routes=dest\_inet:** [hop] [priority] router\_NID@src\_inet \  
  [hop] [priority] router\_NID@src\_inet

## 利用priority可以实现

- 一个LNet服务强于其他
- 一个为主，其余后备
- 分别对应不同节点
- 每个路由器将通讯移动到不同的物理位置

# LNet路由设置



配置文件: /etc/modprobe.d/lustre.conf

LNet路由器节点(IB: o2ib1(ib0), OPA: o2ib2(ib1))

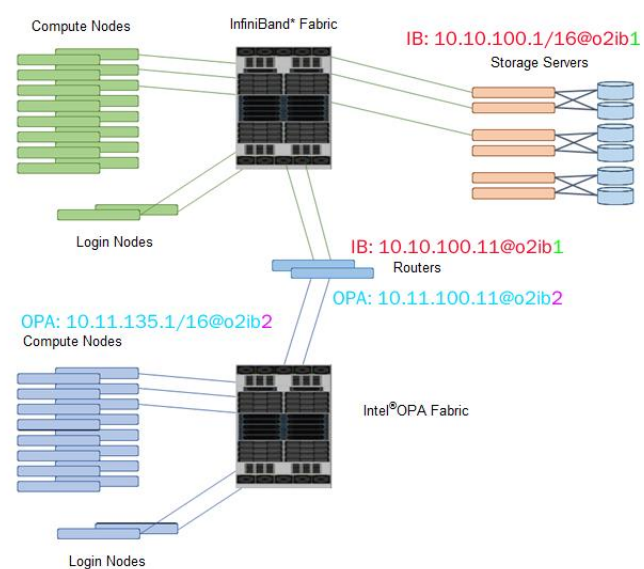
- #选项 Inet networks="本机IB网络类型(IB网卡),本机OPA网络类型(OPA网卡)" forwarding="启用"
- options Inet networks="o2ib1(ib0),o2ib2(ib1)" "forwarding=enabled"

IB服务端(o2ib0)

- #选项 Inet networks="本机IB网络类型(网卡)" routes="客户端OPA网络类型 LNet节点对应本地服务端IB的IP@网络类型"
- options Inet networks="o2ib1(ib0)" routes="o2ib2 10.10.100.11@o2ib1"

OPA客户端(o2ib2)

- #选项 Inet networks="本机客户端OPA网络类型(网卡)" routes="服务端IB网络类型 LNet节点对应客户端OPA的IP@网络类型"
- options Inet networks="o2ib2(ib0)" routes="o2ib1 10.11.100.11@o2ib2"



# 负载均衡/冗余设置: Keepalived



中国科学技术大学  
University of Science and Technology of China

## 功能

- 是集群管理中保证集群高可用的一个服务软件，其功能类似于heartbeat，用来防止单点故障。

## 工作原理

- 以VRRP协议为实现基础的，VRRP全称Virtual Router Redundancy Protocol，即虚拟路由冗余协议。

## 主要有三个模块，分别是core、check和vrrp

- core模块为keepalived的核心，负责主进程的启动、维护以及全局配置文件的加载和解析
- check负责健康检查，包括常见的各种检查方式
- vrrp模块是来实现VRRP协议的

## 配置文件

- 默认/etc/keepalived/keepalived.conf，主要包括几个配置区域，分别是global\_defs、static\_ipaddress、static\_routes、vrrp\_script、vrrp\_instance和virtual\_server

# LNet节点: /etc/keepalived/keepalived.conf

## IB NFS后端真实IP

- 10.10.100.5
- 10.10.100.254

## NFS服务虚拟IP (OPA计算节点实际挂载时采用的IP)

- 10.10.100.31

### LNet-1节点 IB主, OPA备

```
vrp_instance VI_1 {
    state MASTER
    interface ib0 #IB
    virtual_router_id 1
    priority 150 #优先级
    advert_int 5
    authentication {
        auth_type PASS
        auth_pass password
    }
    virtual_ipaddress {
        10.10.100.31
    }
}
vrp_instance VI_2 {
    interface ib1
    state BACKUP
    virtual_router_id 2
    priority 100 #优先级
    advert_int 5
    authentication {
        auth_type PASS
        auth_pass password
    }
    virtual_ipaddress {#虚拟IP, OPA计算节点挂载NFS存储时时使用
        10.11.100.31
    }
}
virtual_server 10.10.100.31 2049 {#虚拟前端: IP NFS端口号
    delay_loop 20
    lb_algo wrr #调度算法: 加权轮询
    lb_kind DR #直接路由(Direct Routing)
    persistence_timeout 50
    protocol TCP
    real_server 10.10.100.5 2049 {#NFS服务器1
        weight 100
        MISC_CHECK {
            misc_path /root/bin/misccheck.sh
            misc_timeout 100
        }
    }
    real_server 10.10.100.254 2049 {#NFS服务器1
        weight 50
        MISC_CHECK {
            misc_path /root/bin/misccheck.sh
            misc_timeout 100
        }
    }
}
```

### LNet-2节点 IB主, OPA备

```
vrp_instance VI_1 {
    state BACKUP
    interface ib0 #IB
    virtual_router_id 1
    priority 100 #优先级
    advert_int 5
    authentication {
        auth_type PASS
        auth_pass password
    }
    virtual_ipaddress {
        10.10.100.31
    }
}
vrp_instance VI_2 {
    interface ib1
    state MASTER
    virtual_router_id 2
    priority 150 #优先级
    advert_int 5
    authentication {
        auth_type PASS
        auth_pass password
    }
    virtual_ipaddress {#虚拟IP, OPA计算节点挂载NFS存储时时使用
        10.11.100.31
    }
}
virtual_server 10.10.100.31 2049 {##虚拟前端: IP NFS端口号
    delay_loop 20
    lb_algo wrr
    lb_kind DR #直接路由(Direct Routing)
    persistence_timeout 50
    protocol TCP
    real_server 10.10.100.5 2049 {#Backend
        weight 100
        MISC_CHECK {
            misc_path /root/bin/misccheck.sh
            misc_timeout 100
        }
    }
    real_server 10.10.100.254 2049 {
        weight 50
        MISC_CHECK {
            misc_path /root/bin/misccheck.sh
            misc_timeout 100
        }
    }
}
```

Inet1@tc4600: ~# ipvsadm

- IP Virtual Server version 1.2.1 (size=4096)  
Prot LocalAddress:Port Scheduler Flags  
-> RemoteAddress:Port Forward Weight ActiveConn InActConn  
TCP **Inet1**:nfs wrr persistent 50  
-> 10.10.100.5:nfs Route 100 0 0  
-> 10.10.100.254:nfs Route 50 0 0

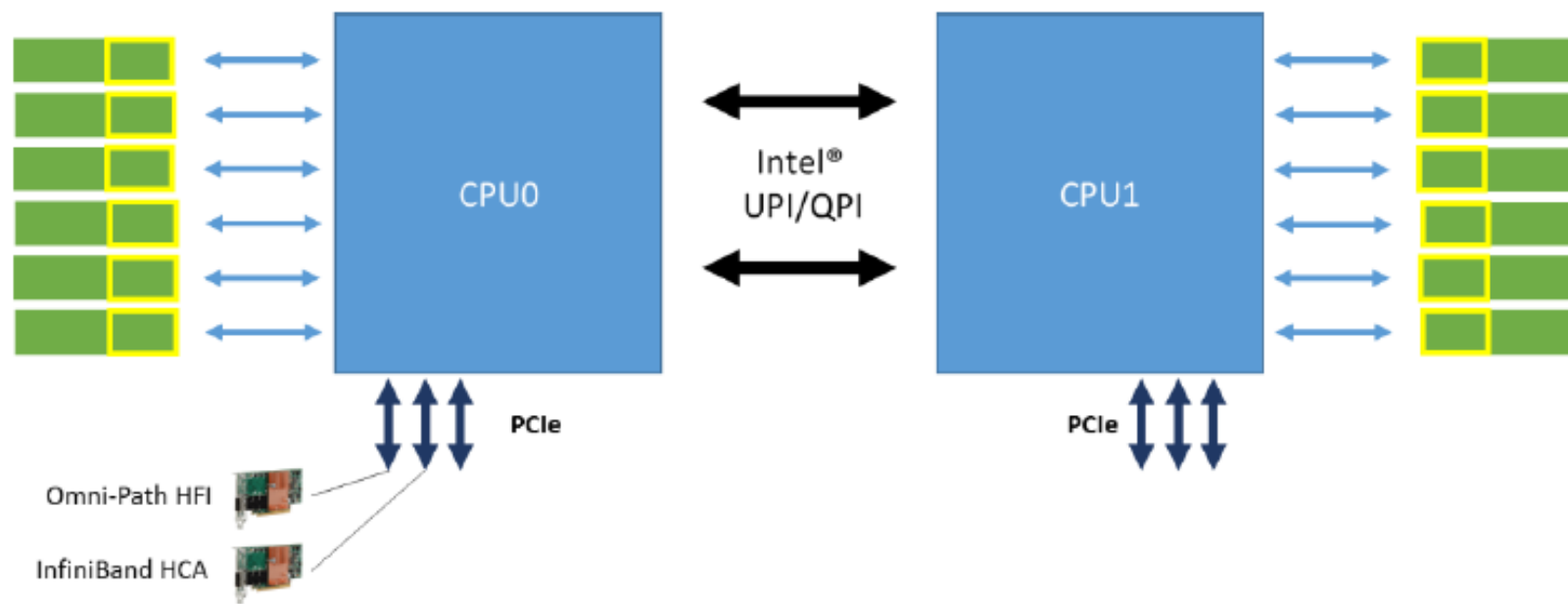
Inet2@tc4600: ~# ipvsadm

- IP Virtual Server version 1.2.1 (size=4096)  
Prot LocalAddress:Port Scheduler Flags  
-> RemoteAddress:Port Forward Weight ActiveConn InActConn  
TCP **10.10.100.31**:nfs wrr persistent 50  
-> 10.10.100.5:nfs Route 100 0 0  
-> 10.10.100.254:nfs Route 50 0 0

IB和OFI尽量插在同一颗CPU附近，以便缩短路径，提高效率

关闭CPU超线程

关闭CPU频率自动调节



## 稳定性和性能都挺好

- 某次一个LNet节点意外死机，Lustre和NFS（经IP）挂载正常，用户作业正常运行，说明LNet冗余功能正常



- <http://hmli.ustc.edu.cn/doc/linux/ib-lnet-opa/>



中国科学技术大学

University of Science and Technology of China

谢谢！

中国科大超级计算中心

<http://scc.ustc.edu.cn>