inspur 浪潮 | Asia Supercomputer Community

ASC16规则解析
&
集群系统构建+HPL解读

王渭巍

http://www.asc-events.org
techsupport@asc-events.org

总则

- **目标**
  - 五人大学生参赛队伍+一名指导老师
  - 构建实时功耗3000W以内的小型集群，成功运行并优化若干应用，将一段代码并行化，取得最好成绩
  - Enjoy

总则

优化规则

1. 不得进行针对特定参数或输入数据的优化；
Optimization methods that are only applicable to specific parameters or input data are strictly prohibited.
2. 如果改动算法，则新算法和原算法必须在数学上严格等价；
If there are any modifications on the algorithm, the new algorithm must be mathematically equivalent to the original one.
3. 以上两条如有违反，该题将被判为零分。
If any rules given above are violated, a score of zero will be given for the corresponding task.
注：参赛团队可就特定的方案是否违规，预先书面提交组委会进行裁决。未经裁决的方案违规与否，以评审委员会裁决为准。
Note: when in doubt, a team may submit a query to the contest committee before the competition on whether a specific optimization method violates the rules, and a decision will be made by the contest committee before the competition. Otherwise the team will have no chance to provide further explanations when an optimization method is ruled out by the evaluation committee during the competition.

评分简则

初赛	决赛应用	决赛功耗
• 初赛通知包含评分方法 • 决赛名单由ASC评审委员会评选得出	• 运行结果通过正确性验证 • 第一名满分，其他队伍按照比例得分	• 实时功耗不得超过3000W • 功耗超过3000W则当前运行结果无效

荣誉与奖励

总冠军

赞助参加ISC15

最高性能大奖

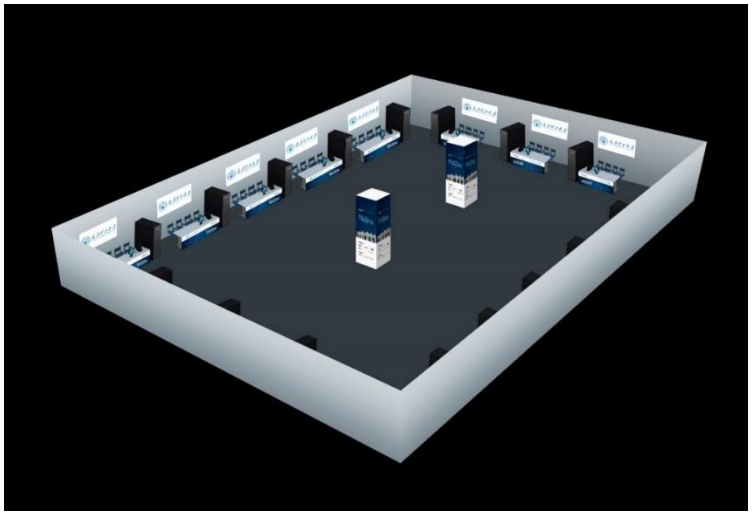
E-prize大奖

比赛阶段

- 初赛阶段
- 参赛队伍需要按照初赛通知的要求提交proposal，初赛的部分题目是决赛题目的一部分
- 初赛对理论知识的考察较强，各参赛队提交的proposal的技术题目需要有实验结果
- Tips: 指导老师和参赛队员通力合作，并取得学校（学院）的支持

比赛阶段

- 决赛阶段
- 参赛队伍在决赛现场根据决赛要求构建计算集群，在限定时间内运行决赛应用，根据运行结果以评分规则得分。并将参赛总结等呈现给评审委员会，此阶段指导老师不得在现场指导

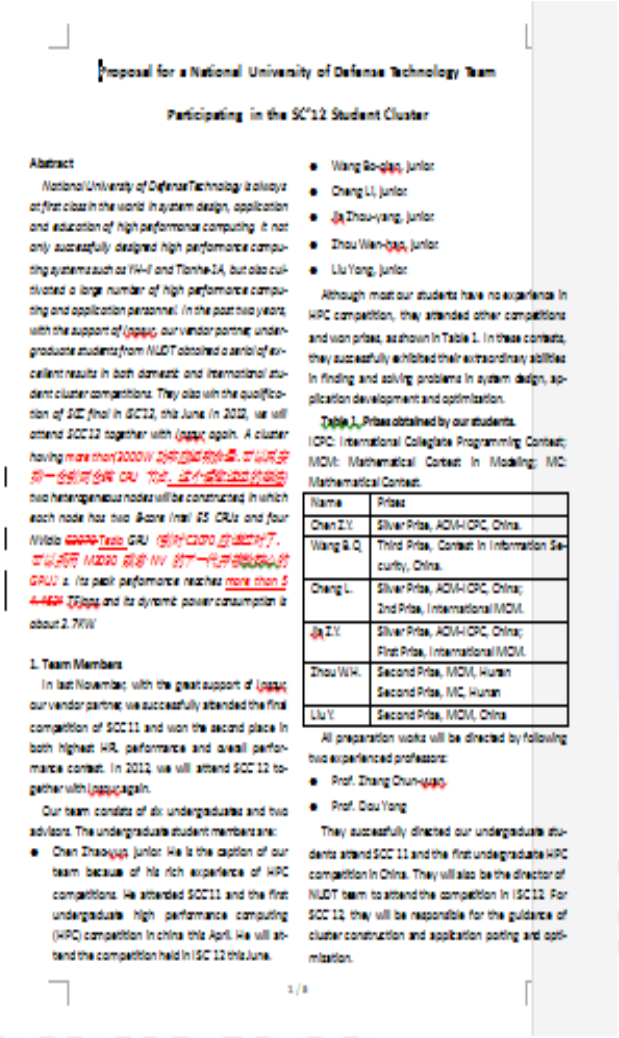


ASC14决赛现场

ASC16 初赛proposal

- Proposal内容
  1) 学校的超算背景介绍 5分

     超算中心，课程设置，兴趣小组，论文项目，etc
  2) 参赛组织及队伍组成 5分
  3) 技术要求部分 90分
     a) 集群构建——3000W内最高性能为原则 15分
     b) HPCG Test 15分
     c) MASNUM_WAM Test 20分
     d) DNN coding 40分



Proposal for a National University of Defense Technology Team

Participating in the SC'12 Student Cluster

Abstract

National University of Defense Technology is always at first class in the world in system design, application and education of high performance computing. It not only successfully designed high performance computing systems such as YH-I and Tianhe-1A, but also cultivated a large number of high performance computing and application personnel. In the past two years, with the support of [illegible], our vendor partner, undergraduate students from NUDT obtained a series of excellent results in both domestic and international student cluster competitions. They also win the qualification of SC final in SC12, this June. In 2012, we will attend SC'12 together with [illegible] again. A cluster having more than 3000W [illegible] two heterogeneous nodes will be constructed, in which each node has two 8-core Intel E5 CPUs and four Nvida [illegible] GPU. [illegible] GPU. Its peak performance reaches more than 5 4.464 TFlops and its dynamic power consumption is about 2.7KW.

1. Team Members

In last November, with the great support of [illegible], our vendor partner, we successfully attended the final competition of SCC11 and won the second place in both highest HPL performance and overall performance contest. In 2012, we will attend SCC'12 together with [illegible] again.

Our team consists of six undergraduates and two advisors. The undergraduate student members are:

- Chen Zhaoqun, junior. He is the captain of our team because of his rich experience of HPC competitions. He attended SCC11 and the first undergraduate high performance computing (HPC) competition in china this April. He will attend the competition held in ISC'12 this June.
- Wang Bo-qian, junior.
- Cheng Li, junior.
- Jia Zhou-yang, junior.
- Zhou Wen-tao, junior.
- Liu Yong, junior.

Although most our students have no experience in HPC competition, they attended other competitions and won prizes, as shown in Table 1. In these contests, they successfully exhibited their extraordinary abilities in finding and solving problems in system design, application development and optimization.

Table 1. Prizes obtained by our students.

ICPC: International Collegiate Programming Contest; MCM: Mathematical Contest in Modeling; MC: Mathematical Contest.

Name	Prizes
Chen Z.Y.	Silver Prize, ACM-ICPC, China.
Wang B.Q.	Third Prize, Contest in Information Security, China.
Cheng L.	Silver Prize, ACM-ICPC, China; 2nd Prize, International MCM.
Jia Z.Y.	Silver Prize, ACM-ICPC, China; First Prize, International MCM.
Zhou W.H.	Second Prize, MCM, Hunan Second Prize, MC, Hunan
Liu Y.	Second Prize, MCM, China

All preparation works will be directed by following two experienced professors:

- Prof. Zhang Chun-yuan.
- Prof. Dou Yong.

They successfully directed our undergraduate students attend SCC'11 and the first undergraduate HPC competition in China. They will also be the director of NUDT team to attend the competition in ISC'12. For SCC'12, they will be responsible for the guidance of cluster construction and application porting and optimization.

1 / 3

设备

竞赛设备

Inspur NF5280M4

Xeon Phi 31S1P

验证平台

远程验证平台

并行环境

```
: Column=055808 Fraction=0.695 Kernel=875900.71 1
: Column=063744 Fraction=0.795 Kernel=821069.06 1
: Column=071680 Fraction=0.895 Kernel=697580.79 1
: Column=079616 Fraction=0.995 Kernel=345927.73 1
================================================================================
   N    NB     P     Q               Time
--------------------------------------------------------------------------------
)000   256     1     1             388.65
:t time Sat Sep 27 18:53:03 2014

 time   Sat Sep 27 18:59:32 2014

--------------------------------------------------------------------------------
'(||A||_oo*||x||_oo+||b||_oo)*N)=        0.00305(
================================================================================

:ests with the following results:
:ests completed and passed residual checks,
:ests completed and failed residual checks,
:ests skipped because of illegal input values.
--------------------------------------------------------------------------------
```

ASC16远程登录平台

- 远程登录平台配备

Item	Name	Configuration
Server	Inspur NF5280M4	CPU: Intel Xeon E5-2680v3 x 2，2.5Ghz，12 cores Memory: 16G x8，DDR4，2133Mhz Hard disk: 300G SAS x 1 Power consumption estimation: E5-2680v3 TDP 120W,memory nominal 7.5W, hard disk nominal 10W
Accelerator card	XEON PHI-31S1P	Intel XEON PHI-31S1P （57 cores, 1.1GHz, 1003GFlops, 8GB GDDR5 Memory） Power consumption estimation: 270W
HCA card	FDR	Infiniband Mellanox ConnectX®-3 HCA card, single port QSFP, FDR IB Power consumption estimation:9W
Switch	GbE switch	10/100/1000Mb/s，24 ports Ethernet switch Power consumption estimation:30W
	FDR-IB switch	SwitchX™ FDR InfiniBand switch, 36 QSFP port Power consumption estimation:130W
Cable	Gigabit CAT6 cables	CAT6 copper cable, blue, 3m
	Infiniband cable	Infiniband FDR optical fiber cable, QSFP port, cooperating with the Infiniband switch for use

ASC16远程登录平台已经就绪

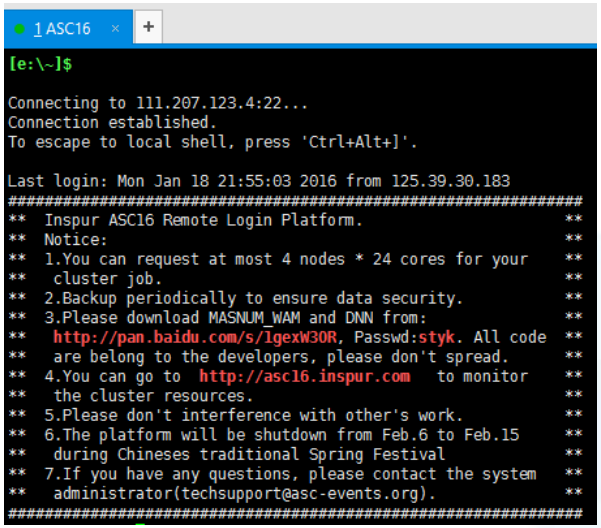
- 远程登录平台如何使用

ssh登录

- 由asc16.inspur.com域名统一登录
- 使用由techsupport@asc-events.org发送的用户名和密码登录

浪潮ClusterEngine网页监控

- 由asc16.inspur.com域名访问
- 使用由techsupport@asc-events.org发送的用户名和密码登录
- 可在该网页看到集群的使用情况

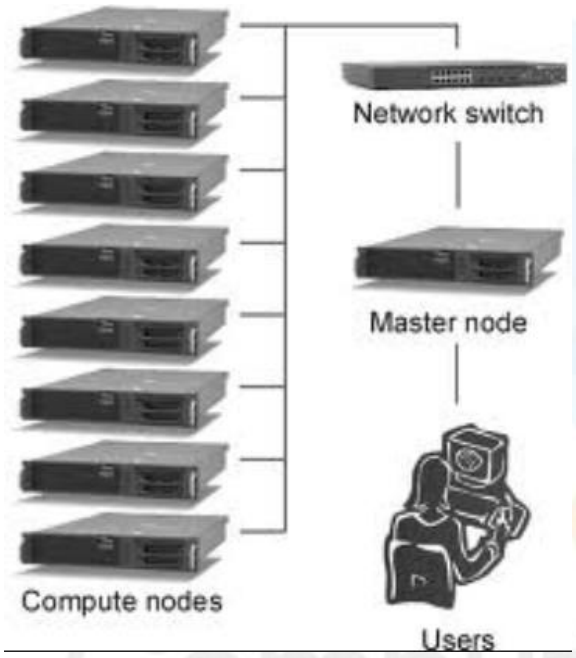


inspur 浪潮 | Asia Supercomputer Community

集群构建目标

HPC集群构成：

- 节点：
  - 计算节点
  - 登录和管理节点
- 网络：
  - 交换网络
- 存储
  - 本地存储
  - 网络存储



如何理解节点配置信息

计算节点的配置信息

Item	Model	SPEC.
Servers	INSPUR NF5280M4	CPU: Intel Xeon E5-2680v3 x2, 2.5Ghz, 12cores Memory: 16G x8, DDR4,2133Mhz Disk: 300G SAS x 1
Co-processor	XEON PHI-31S1P	Intel XEON PHI-31S1P (57 cores, 1.1GHz, 1003GFlops, 8GB GDDR5 Memory)

CPU Frequency

Memory Frequency

Disk capacity

Floating-point operations per second

如何计算节点性能峰值

- GFlops=(CPU freq) x (FP op per sec) x (core number) x (CPU number)
- **AVX 2.0 每秒钟16次双精度浮点运算**

Haswell新计算指令集

- 英特尔®高级矢量扩展指令集2 (Intel® AVX2)
  - 包括
    - 256-bit整数矢量
    - FMA:融合乘加
    - 全宽度元素置换
    - 聚合
  - 优势
    - 高性能计算
    - 音频和视频处理
    - 游戏处理
- 新的整数指令
  - 索引和散列指令
  - 加密/解密
  - 字节序转换 – MOVBE

	指令集	每周期单精度FLOPs	每周期双精度FLOPs
Nehalem	SSE (128-bits)	8	4
Sandy Bridge	AVX (256-bits)	16	8
Haswell	AVX2 & FMA	32	16

类别	指令
比特封装/解析	BZHI, SHLX, SHRX, SARX, BEXTR
变量比特长度流解码	LZCNT, TZCNT, BLSR, BLSMSK, BLSI, ANDN
比特聚合/分散	PDEP, PEXT
随机精度算法 & 散列	MULX, RORX

协处理器

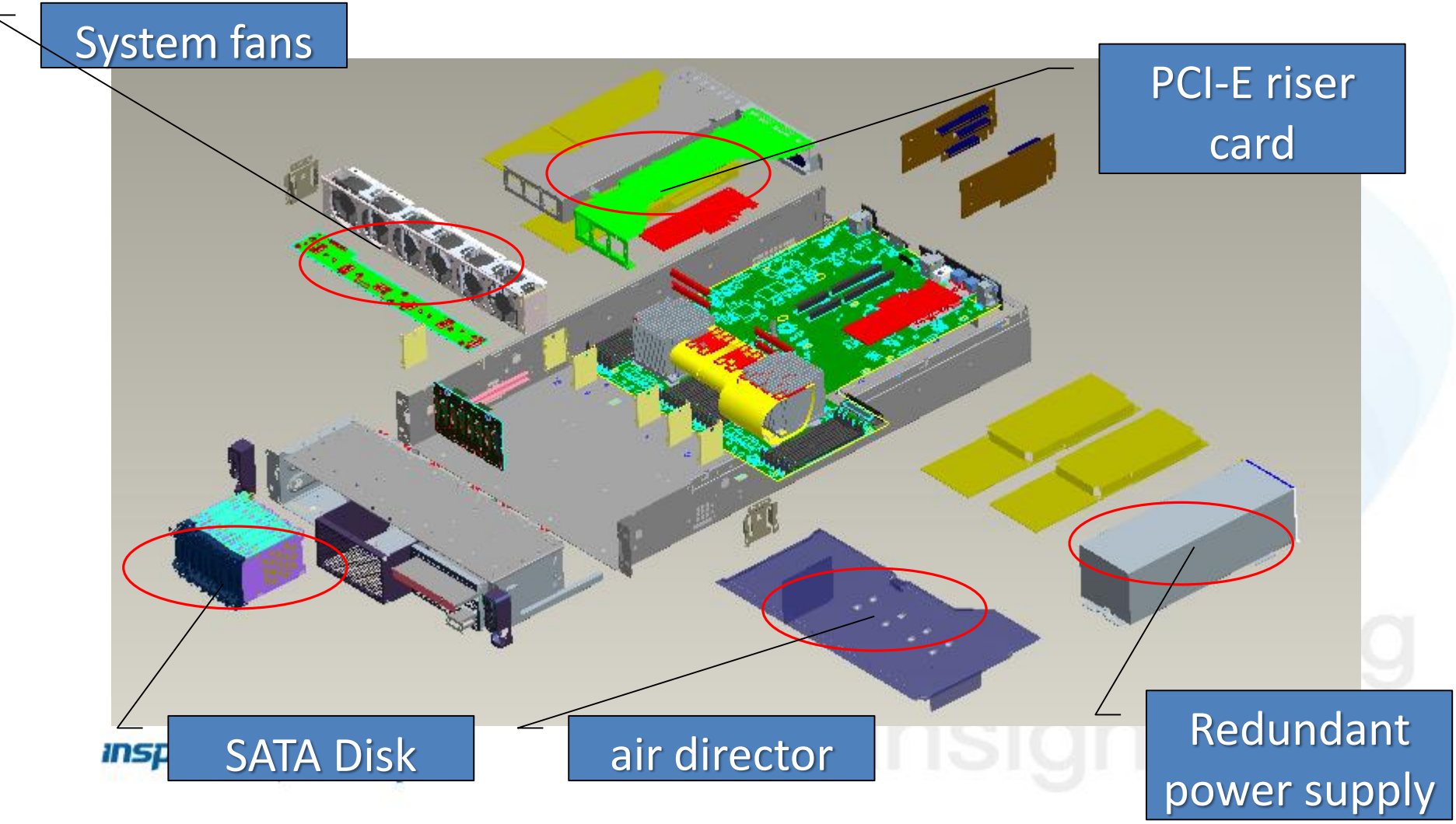
GPU

Xeon Phi

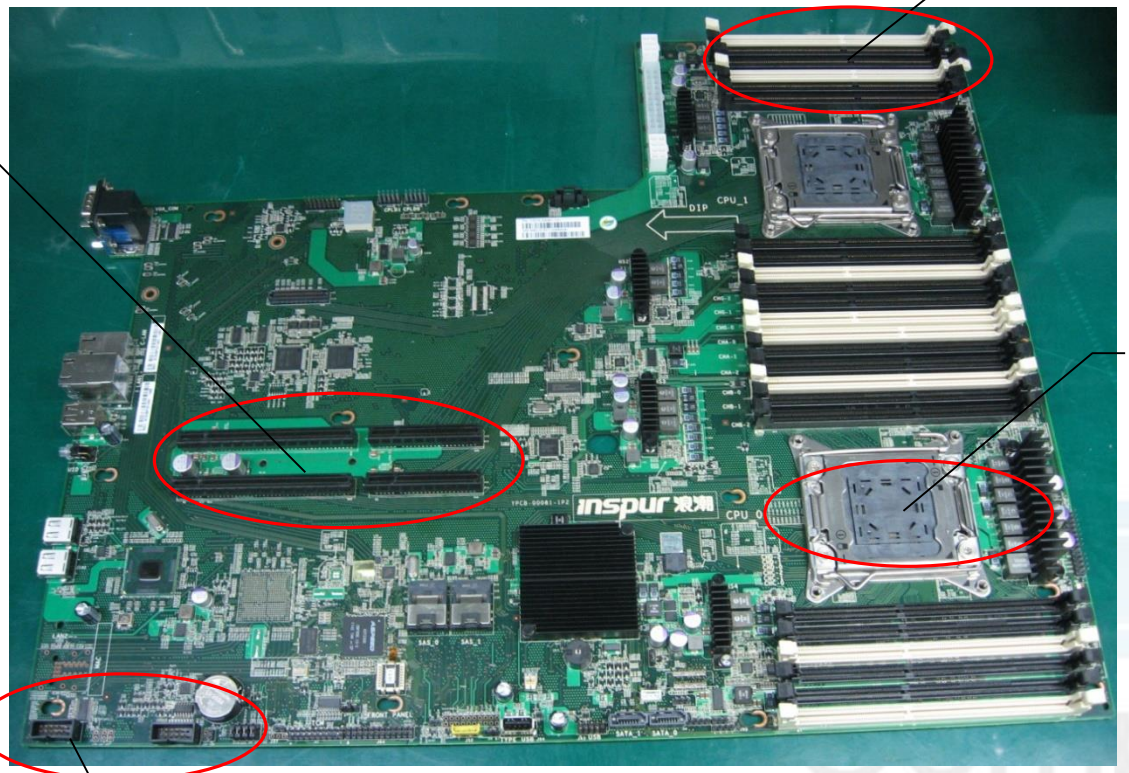
技术要求之集群构建

构建集群的一般原则

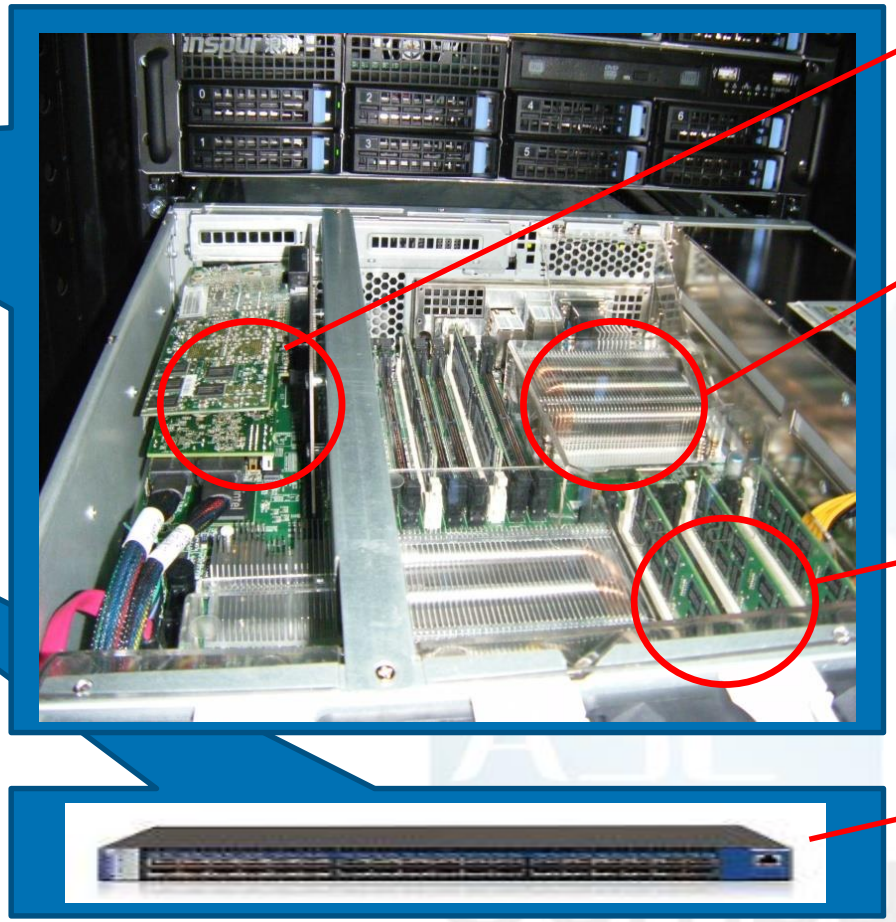
了解Inspur NF5280M4结构



了解服务器主板



功耗预估



GPU or MIC
270w

CPU 120w

Memory
7.5w

IB switch
130w

Total runtime power
≤ 3000 W

Measure the runtime max power when running HPL

系统与应用

- 并行环境（队伍自行准备）
  - OS Linux发行版
  - 网络环境（驱动等）
  - Compiler and MPI
  - File system
  - And so on

MPI简介

MPI(Message Passing Interface)

- MPI 是一个库，而不是一门语言；
- MPI 是一种消息传递编程模型，并成为这种编程模型的代表和事实上的标准；
- MPI 是一种标准或规范的代表，而不特指某一个对它的具体实现；
- 目标: 是提供一个实际可用的、可移植的、高效的和灵活的消息传递接口标准

MPI的编译与运行

- MPI的编译：
  - –mpif77 –o mpi_progmpi_prog.f　　Fortran77
  - –mpicc –o mpi_progmpi_proc.c　　C
  - –mpif90 –o mpi_progmpi_prof.f90　　Fortran90
  - –mpiCC –o mpi_progmpi_prof.C　　C++
- MPI的运行：
  - mpirun –machinefile filename –np N <program name>
  - Machinefile 指定节点机的配置文件 np 指定运行几个进程

High Performance Linpack简介

- **HPL is a software package that solves a (random) dense linear system in double precision (64 bits) arithmetic on HPC computers.**
- **As a yardstick of performance we are using the best performance as measured by the LINPACK Benchmark. LINPACK was chosen because it is widely used and performance numbers are available for almost all relevant systems.**

HPL运行环境

- **Compiler and math library**
  - http://software.intel.com/en-us/intel-composer-xe/
- **MPI**
  - For instance, Intel MPI , OpenMPI
- **IB driver**
  - OFED
- **Make ssh access without password**

编译器

- **Intel Compiler and math library**
  - http://software.intel.com/en-us/intel-composer-xe/
  - Run install.sh
  - Follow the instructions
- **Other Compiler and math library**
  - GCC
  - Goto Blas or …

MPI

- **MPI**
  - Intel MPI -- like Intel Compilers
  - Other MPI - Mpich2 …
- **MPI commands**
  - Mpd &
  - Mpdboot -n X -f your-hostfile
  - Mpiexec -n X your-binary
  - Mpdallexit
  - Or just mpirun -n X -f your-hostfile

HPL.dat

[chewbacca@node001 em64t]$ cat HPL.dat

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University o
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
417536        Ns
1            # of NBs
224           NBs
0            PMAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
16           Ps
64           Qs
```

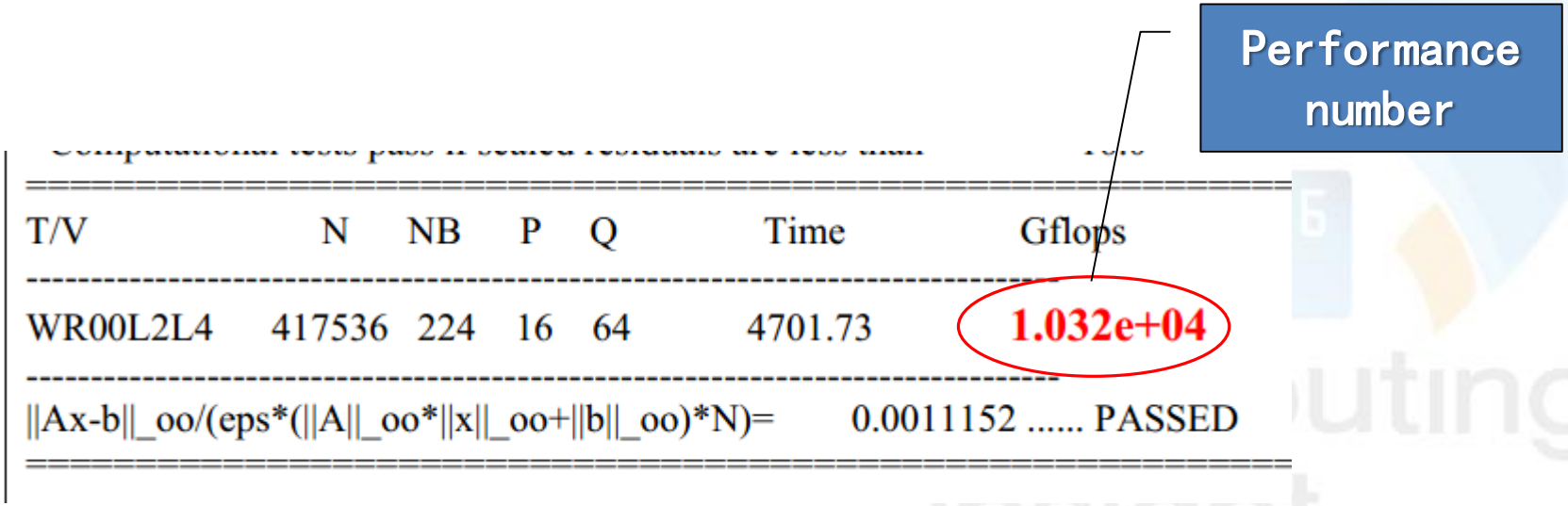
N = Sqrt(0.8x1024x1024x1024x Mem/8)

P x Q = process

P < Q

HPL output file

- GFlops=CPU freq x FP op per sec x core number x CPU number



集群效率

HPL效率

实测节点数	实测峰值（Tflops）	736节点理论峰值（Tflops）	效率（736节点）	740节点理论峰值（Tflops）	效率（740节点）
736	90.34	103.511	87.28%	104.073	86.80%

HPCG

HPCG指标更容易反映出有限元法和流体分析等实际应用的性能。要求处理器的运算性能、内存容量、带宽以及互连性能之间取得平衡。

参数设定：

bin/hpcg.dat文件中的第3和4行。

hpcg.dat文件提供了默认的参数。

第3行：指定了每个MPI进程能处理问题的本地维数，默认值是104 104 104。

第4行：指定了时间部分的benchmark允许运行多长时间，单位秒，默认是60s

注：关于第3行的输入参数设置，取值需要是偶数，且连续三次对2取余时，余数都为0。

HPCG

- 编译

cd .. #返回目录hpcg

mkdir build_Linux #建立编译目录

cd build_Linux

../configure Linux #Linux是配置文件Make.Linux的后缀

make

此时，会在bin目录下生成可执行文件xhpcg

- 测试

cd /hpcg-2.4/build_Linux/bin

mpirun –n 48 –machinefile /root/soft/pt2pt.host ./xhpcg #默认是会读取 hpcg.dat文件中的值。

THANKS