AMAX 基础知识培训

AMAX China Group

Author：Mike Huang
Version：V1.0

什么是集群

集群硬件

集群软件

集群相关服务

AMAX在集群方面的优势

什么是集群？

集群，是指一批独立的设备（计算机等）构成的一个整体。该整体内部的各个独立设备之前存在着较强的相互通信与协作。

我们常说的集群，实际上是集群这个大范围中的一个小部分：高性能计算集群。注意，我们的以后所说的集群，全称是高性能计算集群。

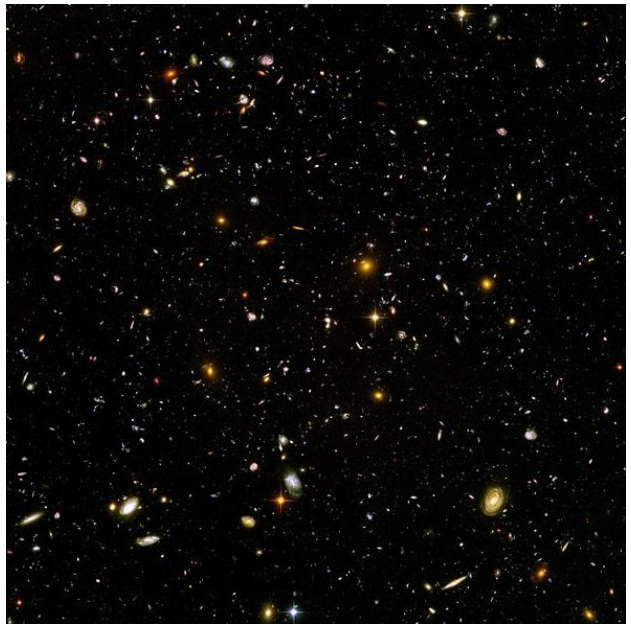
高性能计算集群，是由一批配置一样，或者相近的计算机组成的一个集合。这个集合的是为了提供比单台计算机更强大的计算性能。这类集群多用于高校、科研院所等。除了追求高性能，这类集群还有一个比较大的特点是使用超高速网络来负载MPI数据。

为什么要有集群？

一台计算机，再强大，终归是一台机器。无论是内存数量还是CPU核心数都是有一定的极限。

那么，面对一个需要上PB内存，几十万CPU核心的计算课题，我们该怎么办？答案就是使用集群。

宇宙中所有星球的质量之和是多少？



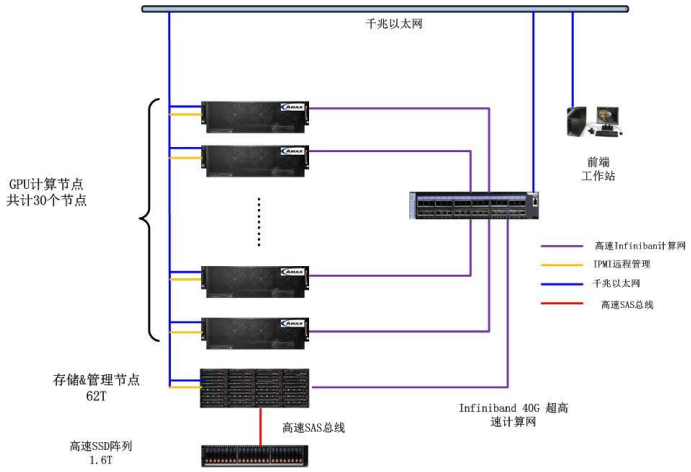
集群网络与一般网络

这里的一般网络，是指公司里面，或者数据中心里面的一批计算机通过网络连接一起。

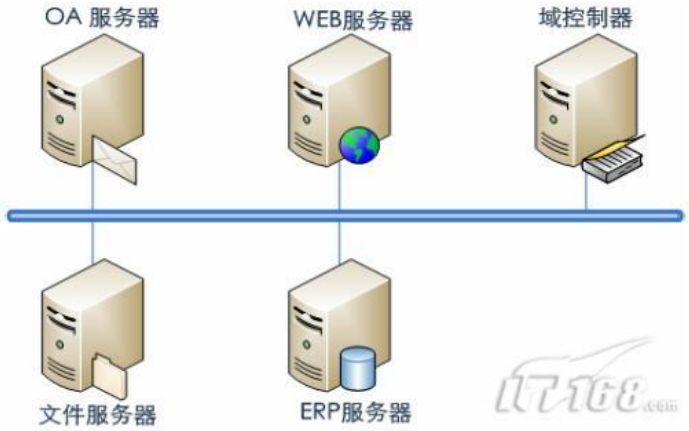
同样是一批计算机通过网络连接在一起，那么，他们与集群有何区别?

他们最本质区别是：

- 集群是很多台机器在一起做一件事情，解决同一个计算问题
- 一般网络的很多台机器各干各的，解决不同的需求。



VS



集群与网格计算

网格计算，也就是大家常说的分布式计算。

集群与网格计算在概念上类似，都是是指一批计算机为了解决同一个计算量较大的计算难题而奋斗。但是，他们还是有很大的区别。这个区别就在于每个计算机所承担的计算任务之间的关联度（耦合度）。

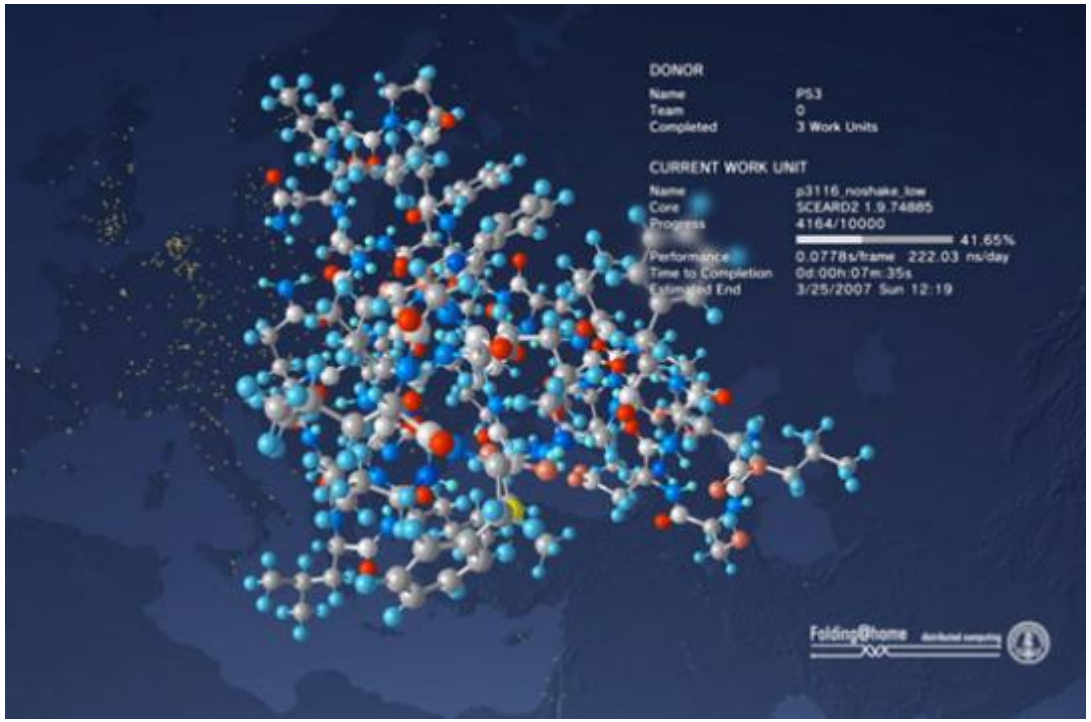
网格计算，往往将一个大的计算拆分成许多不相关的子部分，分发给成员计算机。各成员计算机上跑的任务是没有关联的。各个计算机之间没有联系。就像搞地下工作的卧底一样，只和上级联系，不横向联络。

然而，世界上绝大部分的计算任务不是由许多决然没有联系的子任务。现实的情况是，虽然可以拆分成很许多相对独立的子任务，但是这些任务之间存在着或强或弱的联系。这样，就需要利用到集群。集群的高速互联网络和MPI可以帮助这些子任务之间更快更好地协作。

Folding @ Home

Folding@home是网格计算的一个典型例子。

该系统是一个研究蛋白质折叠、误折、聚合及由此引起的相关疾病的分布式计算工程。最开始F@H仅支持CPU，后来加入了对PS3游戏机的支持，但同样是使用内置的CELL处理器做运算。F@H因ATI的加入为GPU计算翻开了新的一页，当然F@H加入了对NVIDIA DX10 GPU的支持



集群与大型机

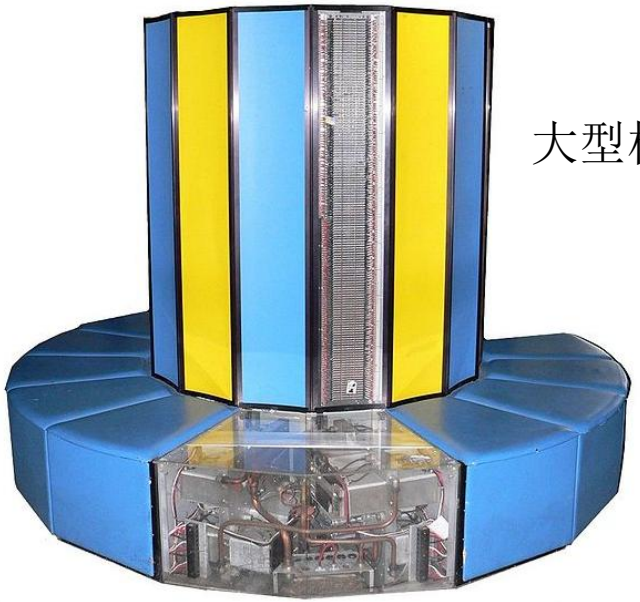
早期的超级计算机都是大型机。随着集群的兴起，大型机已经基本上被集群所取代。现在世界排名靠前的超级计算都是集群而非大型机。

大型机采用SMP（多处理器构架）构架，特点是一台机器内集成了很多路CPU和大容量内存。正是由于这种特性，大型机早已遇到发展的极限，因为很难在一个机箱里装入上百颗CPU。

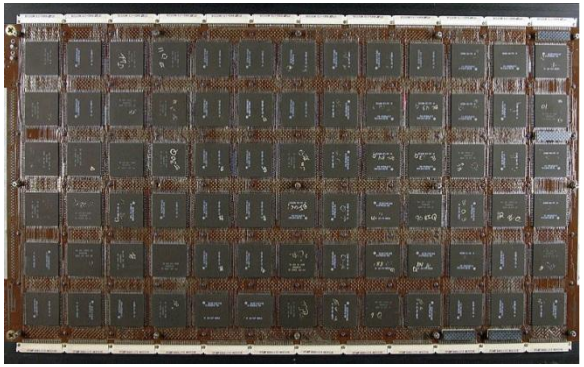
本质上，大型机是一台计算机，运行一个操作系统。集群包含很多台计算机，每个节点有自己的操作系统。

大型机采用共享内存，所以各个进程之间可以直接进行信息交互，不需要借助MPI通信；集群采用分布式内存，各节点上的进程如需要互相通信，必须借助MPI。

大型机



大型机构架

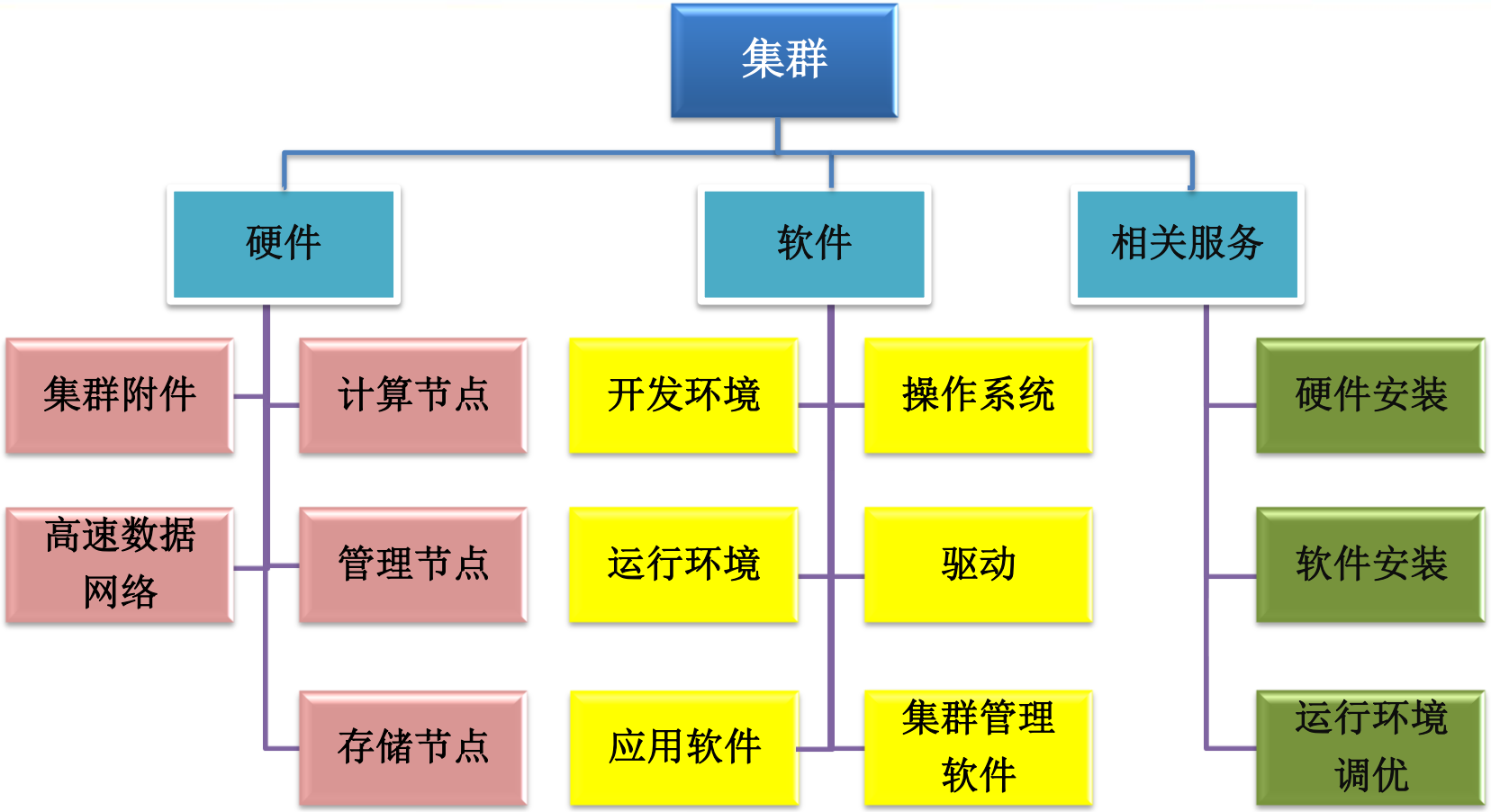


大型机处理器板

云计算 VS 网格计算 VS HPC

	云计算（现今）	大型机	网格计算	HPC集群
针对领域	承载各种应用	承载大规模计算	承载大规模计算	承载大规模计算
计算类型	并行、串行	并行	并行	并行
构架	异构	同构	异构	同构
网络连接速度	高	无要求	低	极高
计算耦合度	N/A	强、弱耦合计算	弱耦合计算	强、弱耦合计算
开发平台	统一、开放	封闭	专属、封闭	统一、开放
应用对应关系	一对多	一对多	一对一	一对多
典型应用	MS Azure	Cray 的大型机	Folder@home	AMAX SuperG

集群的构成


集群
硬件
软件
相关服务
集群附件
计算节点
高速数据网络
管理节点
存储节点
开发环境
操作系统
运行环境
驱动
应用软件
集群管理软件
硬件安装
软件安装
运行环境调优

什么是集群

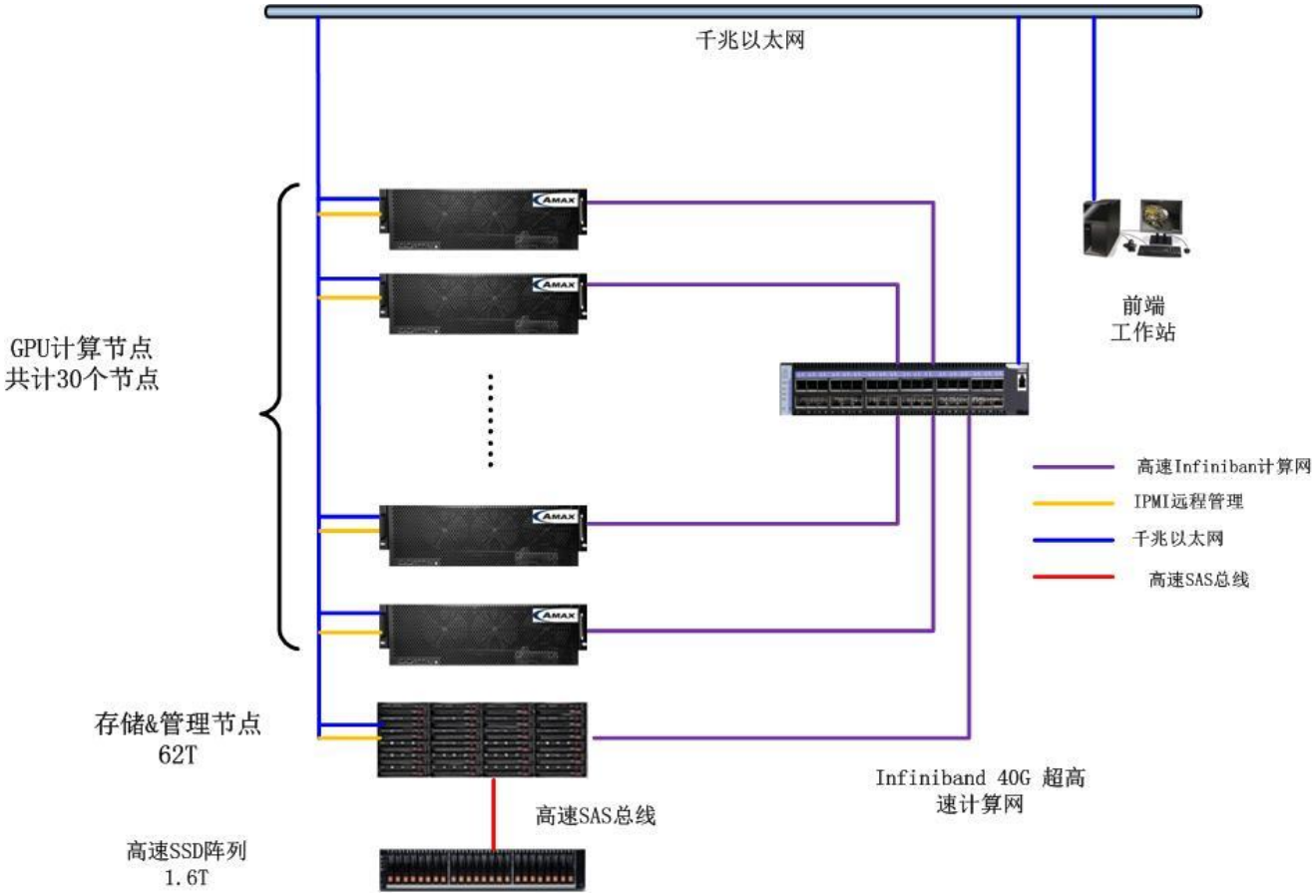
集群硬件

集群软件

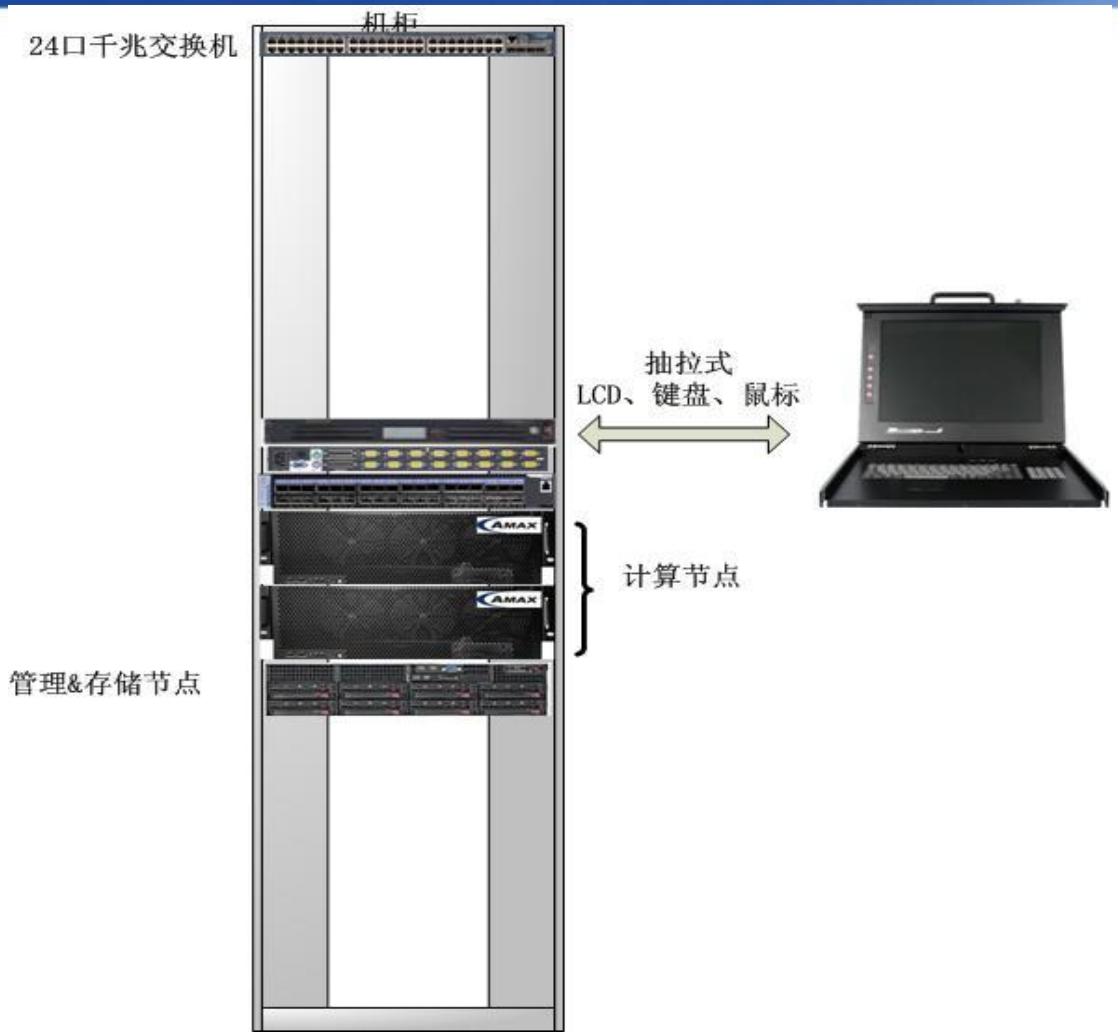
集群相关服务

AMAX在集群方面的优势

GPU集群方案拓扑图样例



GPU集群方案机柜图样例



计算节点

管理节点

存储节点

高速数据网络

集群附件

计算节点

计算节点，是一个集群里面负责计算的部分，也直接影响着一个集群的性能。

如果计算节点是CPU机型，那么集群就是一个CPU计算集群；如果计算节点是GPU机型，那么集群就是一个GPU计算集群；如果计算节点既有GPU，又有CPU机型，那么集群就是一个混合计算集群。

计算节点的计算密度一般较高（单位U里面的GPU或者CPU数量）。计算节点往往需要2种高速网络连接：一种是用来负载MPI的计算网络，另一种是用来负载存储数据的存储网络。但是，我们大多数情况会把这两种网络都放在同一个物理网络上面：Infiniband网络。

计算节点往往都是机架式的，因为比较节省空间。但是我们也有些案例使用塔式计算机作为集群。

计算节点对于可靠性往往没有较高需求，因为大多数的软件支持计算节点故障转移。

计算节点往往对于存储的容量和速度都没有较高的需求，因为往往计算的源数据和目标数据都存储在专业的存储节点上。而计算过程产生的临时数据大多临时存储在内存上。

PSC-6N

- 塔式与4U机架式互转
- 最大支持2路Westmere 6核 CPU
- 最大支持96GB内存(8G单根)
- 最大支持24GB显存
- 支持最多4片高端Quadro或Tesla
- 2800W高效能电源（功耗低于1400W时为冗余电源）

产品优势：
✓成熟稳定
✓扩展性好（插满4片卡还可以插3片PCIE卡片）
✓可上机架
✓硬盘支持热拔插
✓拥有12根内存插槽

产品劣势：
✓噪音大
✓插满4片GPU后剩余的PCIE插槽的带宽为4X，安装Infiniband或者RAID卡的时候会出现性能瓶颈
✓市面上有很多厂家都有这款机器

适用客户:
1.需要插满4片卡，并且还需要安装别的PCIE设备的客户
2.需要部署4片GPU的集群的客户

AMAX Xn-1113M

ServMax Xn-1113M

- **1U机架式机箱**
- **支持2个高性能M2090或M2075**
- **最大支持2路Westmere 6核 CPU**
- **最大支持96GB内存(8G单根)**
- **最大支持12GB显存**
- **最大支持6TB硬盘**
- **可选20Gb或40Gb Infiniband**
- **支持IPMI与KVM-over-IP**

产品优势：

✓ 机型成熟稳定

✓ 支持2路M2090，高达1024处理单元,提供高达1330Gflops的双精度计算能力

✓ 支持热拔插硬盘

✓ 散热设计较好

产品劣势：

✓单机GPU卡片数较少

✓用来安装Infiniband卡的PCIE插槽带宽为4X,存在瓶颈。

适用客户:

1.适合需要CPU数量与GPU数量1:1 或者2:1的客户，如Anasys等

Xn-4101G

- ➢4U机架式机箱
- ➢最大支持8个C2075或者M2090
- ➢最大支持2路Westmere 6核 CPU
- ➢最大支持144GB内存
- ➢最大支持48GB显存
- ➢可选20Gb或40Gb Infiniband
- ➢最大支持4块2.5寸硬盘或者2块3.5寸硬盘

产品优势：

✓4U服务器 最大支持8路GPU，媲美2台Nvidia S2050

✓支持M2090

✓散热设计优秀

产品劣势：

✓插满8片GPU后无法安装其他PCIE卡片

✓仅支持4块2.5寸硬盘或者2块3.5寸硬盘

适用客户：

1. 对单机GPU运算有求高的客户

单机完美支持8片卡具有很强的计算能力可以最为卖点之一 ，4101g是市场上唯一的单机8卡的服务器。即使有其他的厂商也拥有同样的Tyan的barebone,但是他们并不能保证支持8片卡。因为我们使用的BIOS的独有的，经过严格的测试，可以完美兼容8片卡。而其他供应商使用的公版BIOS不能够完美支持8片卡。这是我们AMAX的独家优势

2. 超高计算能力集群的客户

Xn-2207

- 2U机架式机箱内嵌4台独立的计算机（节点）
- 每节点最大支持2路Westmere 6核 CPU
- 每节点最大支持96GB内存(8G单根)
- 每节点最大支持3个热拔插硬盘
- 每节点集成40Gb InfiniBand网卡
- 每节点拥有一个半高PCIE插槽

产品优势:
√最佳性价比
√高密度 (2U空间集成4节点和12个3.5" 硬盘位)
√易维护 (可热插拔的HDD、电源、风扇和节点)
√高可用性 (每节点支持3块硬盘位，可做RAID1, 可选冗余电源)
√可选RAID配置

产品劣势：
✓无

适用客户:
1.适合做高性能CPU集群的计算节点
2.大规模虚拟化，云计算

H2207

- 2U机架式机箱内嵌4台独立的计算机（节点）
- 每节点最大支持2路AMD 6200 16核 CPU
- 每节点最大支持128GB内存(8G单根)
- 每节点最大支持3个热拔插硬盘
- 每节点集成40Gb InfiniBand网卡
- 每节点拥有一个半高PCIE插槽

产品优势:
√最佳性价比，远超Intel平台
√高密度 (2U空间集成4节点和12个3.5" 硬盘位)
√易维护 (可热插拔的HDD、电源、风扇和节点)
√高可用性 (每节点支持3块硬盘位，可做RAID1, 可选冗余电源)

产品劣势：
✓不支持SAS硬盘和RAID 5
✓客户对AMD CPU接受度不高

适用客户:
1.适合做高性能CPU集群的计算节点
2.大规模虚拟化，云计算

计算节点

管理节点

存储节点

高速数据网络

集群附件

管理节点

管理节点，是一个集群里面的管理者。

集群管理软件，一般都是安装在管理节点上。软件运行环境，编译环境，以及应用软件也都在管理节点上进行安装和调试，完成后，经过配置的操作系统会被推送到各个计算节点。

管理节点，对于CPU和内存一般没有太大的要求。管理节点对于可靠性具有极高需求，因为一旦管理节点故障，整个集群将陷入瘫痪。对于要求不是非常高的集群，我们一般采取配置冗余风扇，冗余电源和冗余硬盘的方式保证其可靠性。对于要求较高的集群，需要配置双管理节点，做节点冗余。

对于规模不大的集群，我们常将管理节点和存储节点放在一台服务器上，以节省成本。

管理节点和存储节点可以放在一台服务器上

Xn-1201

- 1U机架式机箱
- 最大支持2路Westmere 6核 CPU
- 最大支持96GB内存(8G单根)
- 最大支持8TB硬盘
- 可选硬件RAID
- 可选冗余电源
- 可选40Gb Infiniband
- 支持IPMI与KVM-over-IP

产品优势：

- ✓ 可选冗余电源
- ✓ 冗余风扇
- ✓ 拥有两个PCIE扩展槽

产品劣势：

- ✓无

适用客户:

1.适合作为入门级服务器

2.适合做集群的管理节点

管理节点产品介绍

- 2U机架式机箱
- 最大支持2路Westmere 6核 CPU
- 最大支持96GB内存(8G单根)
- 最大支持16TB硬盘
- 可选硬件RAID
- 可选冗余电源
- 可选40Gb Infiniband
- 支持IPMI与KVM-over-IP

产品优势：

✓ 可选冗余电源
✓ 冗余风扇
✓ 拥有5个PCIE扩展槽，可选半高或者全高档案
✓ 扩展性强

产品劣势：

✓无

适用客户:

1.适合作为入门级服务器
2.适合做集群的管理节点

GPU
计算节点

管理节点

存储节点

高速数据
网络

集群附件

存储节点

存储节点，是一个集群里面的负责数据存储的节点。

原则上，计算数据是不存储在计算节点上的。除了速度和可靠性之外的考量，将数据存储在存储节点上有利于实现计算节点的故障转移。

存储节点，对于CPU和内存一般没有太大的要求。存储节点对于可靠性具有极高需求，因为一旦存储节点故障，可能导致数据丢失。对于要求不是非常高的集群，我们一般采取配置冗余风扇，冗余电源和冗余硬盘的方式保证其可靠性。对于要求较高的集群，需要配置双管理节点，做冗余。

对于规模不大的集群，我们常将存储节点和管理节点放在一台服务器上，以节省成本。

存储节点需要高速网络。常见的网络有FC（光纤通道），千兆、万兆以太网和Infiniband。我们通常使用Infiniband网络。该网络性价比较高。

Xn-4202

- 4U机架式机箱
- 最大支持2路Westmere 6核CPU
- 最大支持96GB内存(8G单根)
- 拥有36个热拔插硬盘位
- 硬件RAID卡，板载512M缓存，可选RAID卡保护电池
- 支持SSD Cache技术
- 可选冗余电源
- 可选40Gb IB，万兆以太网等
- 支持IPMI与KVM-over-IP

产品劣势：
✓无

产品优势:
✓业界最高密度：4U36盘位
✓拥有强大的处理能力，从容应对严苛的高IO环境
✓支持各种存储操作系统，如OpenE，Windows Storage 2008， Gluster等

适用客户:
1.适合做高性能集群的存储节点
2.适合打造NAS，并行存储

J4502

- **4U 60盘位存储机箱**
- **包含JBOD SAS 2.0 I/O 控制器**
- **四口SFF-8088 6Gb/sec SAS 2.0主机和扩展接口，支持高达2400 MB/S 带宽**
- **通过其他的SFF-8088 SAS接口以SAS菊链方式可连接更多的J4502 JBOD存储**

热插拔I / O控制器和高效的双冗余电源/冷却（APC）模块

使用APC模块的的冗余系统散热风扇

产品优势：

1.密度超高，4U60盘位，业界第一

2.AMAX独占机型

3.冗余电源，冗余数据接口

适用客户:

1.适合对存储容量要求较高，且不迷信名牌的用户

产品劣势：

1.噪音巨大

2.没有硬盘指示灯

3.机箱较长，仅能方在1米的标准机柜内，深度低于1米的机柜很难放进去。

4.背板挂掉所有硬盘都不工作了（但是背板一般不会坏，而且所有JBOD都存在这种问题）

5.只是JBOD，需连接CPU服务器才可使用

GPU
计算节点

管理节点

存储节点

高速数据
网络

集群附件

Infiniband

Infiniband是由Intel等大多厂商联合提出，并开发的。初始的目的是为了取代PCI和PCIX，但是这个目的彻底地失败了，因为现在已经是PCIE的天下。所以Infiniband联盟里的大厂纷纷抛弃了Infiniband。时至今日，只有两家公司还坚守着Infiniband：Mellanox和Qlogic。

虽然，Infiniband完败给了PCIE，但是在HPC领域，Infiniban却取得了巨大的成功。现在主流的HPC集群，基本上都是采用Infiniband。

虽然Intel等大厂，已经抛弃了Infiniband，但是在开源社区，Mellanox等厂家的支持下，Infiniband的生态系统已经建立起来了。很多周边的协议，也已经出来，比如IP-over-Infiniband， FC-over-Infiniband等。另外，业界现在还出现了40Gb转10GbE的网关产品。

现有的Infiniband 有DDR 20Gb， QDR 40Gb， FDR 56Gb三种。但是20Gb已经濒临淘汰。56Gb，由于PCIE的带宽限制，实际速度没有提升。40Gb是现在的主流。

几种网络互联方式的比较

	8Gb FC	1GbE	10GbE	40Gb IB GPU	40Gb IB GPU（GPU Direct）	40Gb IB CPU
理论带宽 Gbps	8	1	10	40	40	40
传输延迟微秒(μs）	100	500	25	10	3	2

可以看到Infiniband最大优势不在于带宽，而是在于低延迟。事实上，对于HPC应用，很难出现成GB的数据传输，相反，绝大多数的都是琐碎的小数据包（一个浮点变量，一个小的数组等）的传输。而这些数据传输的延迟，则会极大地影响到整个HPC集群的性能表现。

Infiniband系统的构成

① Infiniband交换机

② Infiniband卡

③ Infiniband 连接线

④ Infiniband 驱动

⑤ Infiniband MPI

Infiniband 交换机

单体式

模块化

我们一般选择Mellanox出品的Infiniband交换机。Mellanox 是Infiniband 行业的领导者。

由于Infiniband的带宽很大，为了达到Non-blocking（无阻塞），不宜使用多个交换机级联得方式。

单体式IB交换机可选18端口，36端口。

如果需要的口数超过36口，就需要选购模块化的IB交换机。模块化交换机一般起步是36口，可以通过增加模块来增加端口。模块一般是18口或者36口。

Infiniband 网卡

- Connect X3

1. 双口 56Gb FDR
2. 56Gb+10GbE

- Connect X3

1. 单口 40Gb QDR

- 板载IB卡

1. 单口 40Gb QDR

Infiniband连接线

- 按接口类型分

1. CX4 接口（淘汰，连接10Gb，20Gb）
2. QSFP接口（连接40Gb，56Gb以及部分20Gb）

CX4
接口

- 按线缆类型分

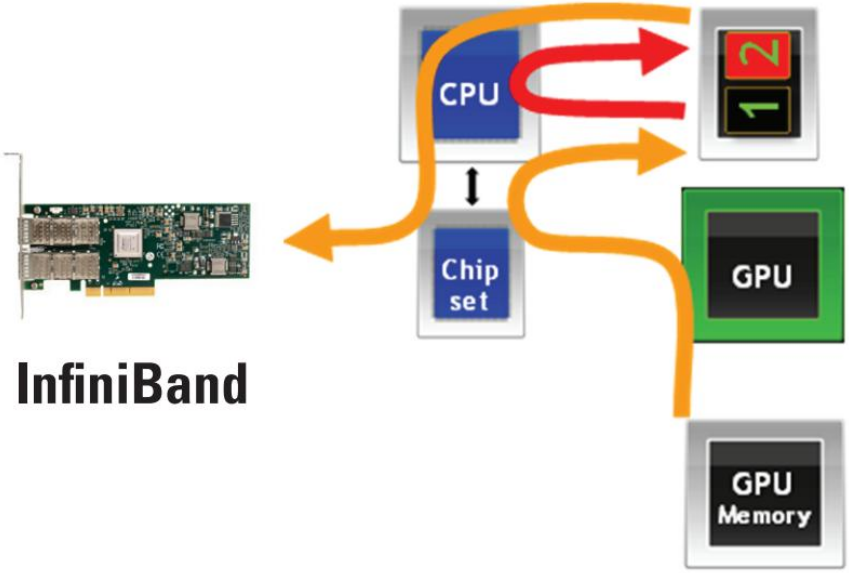
1. 铜缆（不超过10M）
2. 光缆（10M及以上）

- 按长度分

1. 1m，2m，3m~10m

QSFP
接口

GPU Direct （1代）



传统的RDMA技术

- RDMA是infiniband相对于以太网的一大优势，它允许Infiniband卡直接访问内存地址。这在CPU时代无疑是很先进的。

- 然而在GPU的时代，除了内存之外，又出现了显存。因此Infiniband卡必须通过2次读写操作，才可以获得GPU显存的数据。这额外的一次读写操作，无疑大大增加了快捷点GPU数据交互的延迟。

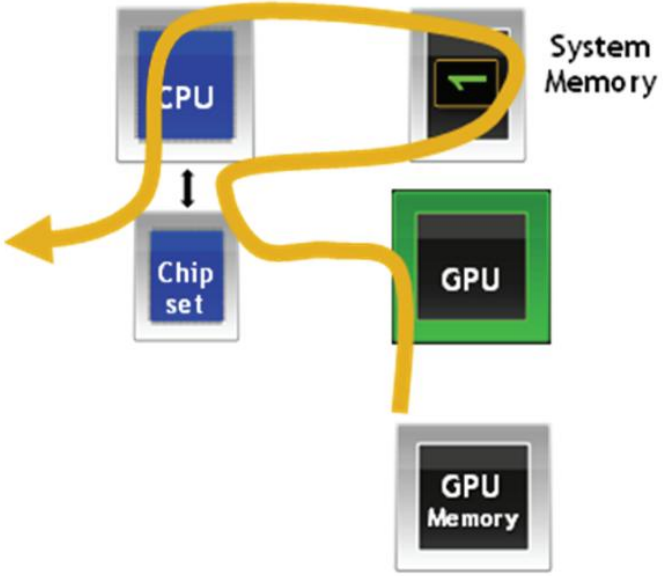
GPU Direct （2代）

- 支持Nvidia最新 GPU Direct技术
- 在此技术的支持下，Infiniband卡只需通过一次内存读写即可获取GPU显存的数据。

- 相对于常规的Infiniband RDMA技术，GPU Direct可以跨节点GPU之间数据交互的延迟降低30%。在跨节点数据交互密集的情况下，该技术可以带来超过的40%的性能提升。

- 该技术仅支持Linux操作系统和Mellanox ConnectX-2或者-3的IB卡

InfiniBand



GPU
计算节点

管理节点

存储节点

高速数据
网络

集群附件

集群附件

机柜

机柜一般是冷轧钢板或合金制作的用来存放计算机和相关控制设备的物件，可以提供对存放设备的保护，屏蔽电磁干扰，有序、整齐地排列设备，方便以后维护设备。机柜一般分为服务器机柜、网络机柜、控制台机柜等。

服务器机柜在机柜的深度、高度、承重等方面均有要求。高度有2.0（42U）米、1.2米（22U）、0.8米（14U）三种高度；宽度为800mm、700mm或600mm三种；常用深度为600mm、800mm、900mm、960mm、1000mm五种。

有些机箱的侧板是可选的。有些则是标配的。比如我们常用的廉价图腾机柜CN46003，就是包含侧板的，但是该机柜的宽度仅有600，走线不是很方便。而另一款高端的机柜CN46006 图腾K3，则需要加钱选配侧板。这款机箱的宽度是800，比较适合走线。

U

1U是一个表示高度的名词 1U=1.75 英寸=4.445厘米

U是一种表示服务器外部尺寸的单位，是unit的缩略语，详细尺寸由作为业界团体的美国电子工业协会（EIA）决定

服务器的宽（48.26cm＝19英寸）与高（4.445cm的倍数）。由于宽为19英寸，所以有时也将满足这一规定的机架称为“19英寸机架”。

1U就是4.445cm，2U则是1U的2倍为8.89cm

也就是说所谓“1U的PC服务器”，就是外形满足EIA规格、厚度为4.445cm的产品。设计为能放置到19英寸机柜的产品一般被称为机架服务器。

1U

2U

3U

KVM

KVM字面的意思是键盘、视频和鼠标。全称应该是键盘、视频和鼠标 切换器。作用就是可以通过一套键盘、视频和鼠标及切换器控制多台计算机。

KVM体系分为传统型和KVM-over-IP两种。一般， 我们的集群里面，传统KVM会作为标配，而KVM-over-IP会作为选配。

传统型KVM一般包含几个组成部分：鼠标，键盘，显示器，KVM切换器，KVM连接线。

KVM-over-IP相对来说比较简单，仅仅需要以太网线连接即可

鼠标，键盘，显示器

专业的KVM系统里面的鼠标，键盘，显示器一般都是会做成1U抽拉式。这样比较专业，不占很多空间。我们一般配的海康SL1701就是这样一款三合一的产品，整合了17寸液晶，键盘，触摸板（鼠标）

正面，拉出后外观

背面，拉出后外观

KVM切换器

KVM切换器是KVM系统里面的核心。KVM切换器往往具有很多的鼠标键盘显示器接口用来连接计算机，以及最少一组二外的鼠标键盘显示器接口用来连接鼠标键盘显示器。

大多数专业级KVM切换器，都采用复合式鼠标键盘显示器接口以节省面板空间

有些四合一机架式鼠标键盘显示器，里面也整合了KVM切换器。

还有些专业级KVM切换器，采用RJ45口连接鼠标键盘显示器，但是这些网口走的不是TCP/IP数据

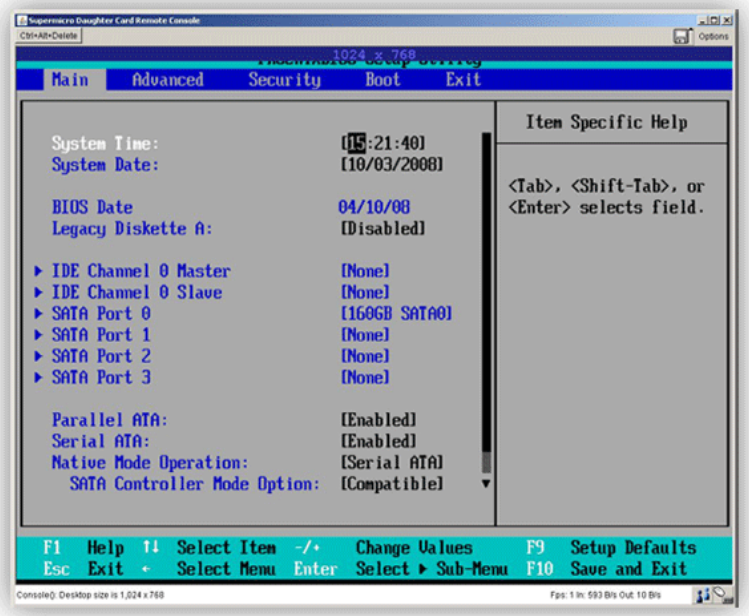
KVM连接线

KVM线是用来连接KVM切换器和被控制的计算机。我们需要注意连接线的两头：连接计算机的那头，有PS2接口和USB的区别，KVM切换器那头有专用接口，PS2&VGA接口，RJ45等多种形式。

我们现在用的SmartView KVM切换器已经包含连接线了

KVM-over-IP

KVM-over-IP是一种通过网络进行KVM管理的一种手段。管理员可以通过IE浏览器从远端办公室，甚至世界各地管理和使用该计算机。相对于各种操作系统里面自带的远程管理功能，如Windows 远程桌面和Linux的SSH， KVM-over-IP可以提供更底层的管理功能，比如安装操作系统，更改BIOS设置等。



AMAX的大多数机器都标配KVM-over-IP功能，仅需多配置一根网线即可实现该功能。

PDU

PDU 是英文 power distribution unit 的缩写，即电源分配单元，说白了，也就是电源插排。

当然，相对于一般家用的多用插排，专业的PDU还是有一些特征：

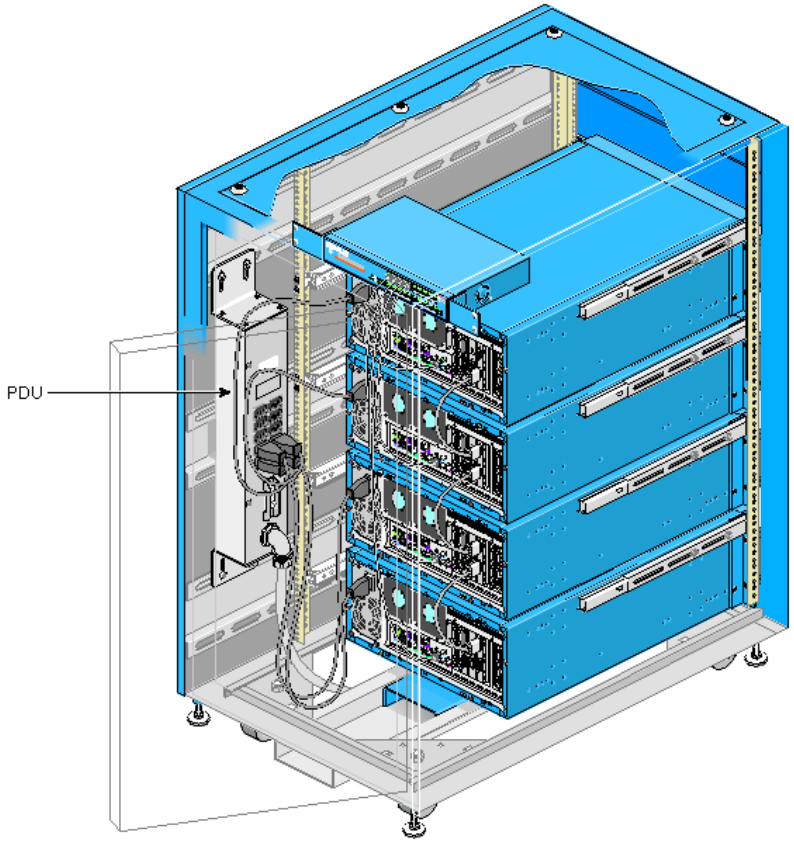
- ✓ 采用1U（横着安装在机柜里）或者0U（竖着，挂在机柜内测）设计。
- ✓ 插口多：8口，16口甚至24口
- ✓ 承载电流大：16A，32A
- ✓ 可选内置保护措施：过载保护，漏电保护，防雷击，稳压滤波等
- ✓ 可选智能控制：IP管理（通过网络控制电源插座的开关）

选择PDU的时候，除了需要注意插口数量外，还需要注意电流是否足够大

PDU

对于国内用的PDU，上面的插口都是国标三相的，有10A和16A（比较少见）。

对于PDU另外一头的插头（插UPS或者市电），则有很大的讲究。其物理型号和电流承载能力。理论上，我们需要先了解客户那里有什么样的插座，我们再配置相应的插头，但是我们往往遇到机器已经进了机房，客户机房的电还没布好。因此，我们一般先为每个PDU报一个电源插座，到时候根据实际情况进行采购

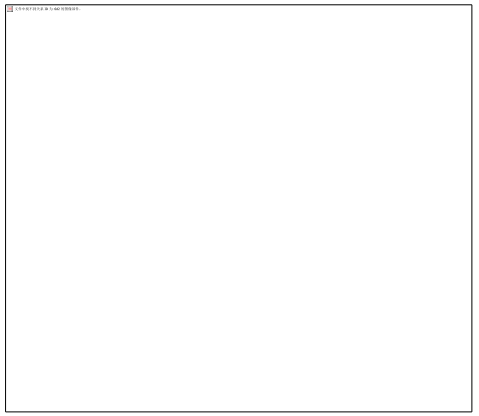


以太网交换机

在集群里面，以太网一般用于负载管理网络和IPMI。因此对于以太网的速度要求不是很高，也允许让更多台交换机通过级联的方式进行扩展

网线

管理网络一般使用六类、超六类线



IPMI网络一般使用五类、超五类即可，当然也可以使用六类线。

什么是集群

集群硬件

集群软件

集群相关服务

AMAX在集群方面的优势

操作系统

驱动

集群管理软件

运行环境

开发环境

应用软件

操作系统

操作系统又名OS。集群，虽然有集群管理软件，而且进行紧密的系统，但是集群里面的每台计算机都有各自的操作系统，而这也是集群与大型机的主要区别之一。

并非所有的操作系统都适合在集群里面使用。事实上，大多数的操作系统，已经被包含在集群管理软件的安装光盘中了。

常见的可用在集群内的操作系统：

- CENTOS. 这个是我们使用的最多的操作系统，基于Redhat的源代码。我们绝大多数的集群管理软件，如Rocks， Bright， Platform都内含了CENTOS。
- Redhat. 与CENTOS基本一样。如果客户指定要求redhat,我们可以使用。注意，该操作系统是需要付费的。
- Suse. Suse Enterprise和Open Suse在一些大型机集群有所使用。
- Widows Server 2008 R2. 2008R2 全系列可以在集群内使用，但是标准版仅支持32GB内存。推荐采用Windows Server 2008 R2企业版，或者Windows Server 2008 R2 HPC版

MS HPC 2008 R2

Red hat Linux Enterprise

OpenSuse

CENTOS

操作系统

驱动

集群管理软件

运行环境

开发环境

应用软件

驱动

驱动，是计算机软件里面的重要组成部分。驱动可以帮助操作系统更好地识别和使用硬件。

驱动是具有很强针对性的，一般一款驱动是针对特定一个键和操作系统

常见的可用在集群内的驱动：

- 主板驱动。 各节点都需要。
- Raid卡驱动。一般在管理节点和存储节点上需要安装。而且一般在安装操作系统的过程中就进行RAID驱动的安装，否则操作系统可能找不到硬盘。
- GPU驱动。一般GPU计算节点需要安装。
- Infiniband驱动。一般计算节点和存储节点需要安装。

操作系统

驱动

集群管理软件

运行环境

开发环境

应用软件

集群管理软件

集群管理软件，是集群的中枢系统。

集群管理软件一般安装在管理节点上，计算节点上的操作系统一般是由管理节点推送过去的。

我们常用的集群管理软件有：Rocks， Rocks+， Platform Cluster Manager，Bright Cluster Manager

集群管理软件一般包含以下几个部分：

- 操作系统
- MPI
- 任务管理与调度
- 集群监管

MPI

MPI是集群运行的核心运行环境。如果没有MPI的话，集群也不称之为集群。

集群管理软件一般会包含一种MPI或多种MPI。

我们常用的软件有：

- OpenMPI, 仅支持Linux
- MPICH，支持Linux和Windows
- MVAPICH，支持Linux和Windows，对Infiniband有优化
- HP MPI(Platform MPI) ，支持Linux和Windows
- Intel MPI，支持Linux和Windows
- MS MPI，仅支持Windows

MPI

MPI,全称Message Passing Interface，译为消息传递接口，是集群应用的核心。前文已经交代，在集群中，不同节点之间的进程，如果需要通信，必须采用MPI，否则，无法进行通信。

MPI是一个库，而不是一门语言。许多人认为，MPI就是一种并行语言，这是错误的。计算语言就像人类的语言，MPI却像人类社会的快递公司。

MPI是一种标准或规范的代表，并不是指一个特定的库或者软件。业界有很多不同的MPI库，比如开源的OpenMPI， 微软的MS MPI，Intel的Intel MPI，HP的HP MPI。这些MPI之间不一定互相兼容，因此，一些著名的集群软件同时支持多种MPI

集群必须安装和配置MPI环境，而且集群中的软件必须能支持这些安装的MPI才可以运行。

有些MPI应用对于网络传输延迟很敏感，所以，在专业的集群内，我们才需要采用高速的网络来负载MPI消息。通常，我们采用Infiniband作为高速网络。

任务管理与调度

任务管理和调度也是集群里面一个主要组成部分。

任务管理和调度模块负责将计算任务分成若干子任务，并分配到相关的计算节点上，另外，该模块还需要负责对各个节点上的计算子任务状态的监控。除此之外，集群往往可以承载多个任务，当这些计算任务对于硬件资源的需求超出集群资源时，该模块还需要对这些任务进行排队。

好的任务管理和调度模块，可以提供各种灵活的排队管理功能，以及众多的，和其他集群应用软件的接口（如果有接口，用户可以在他们的PC机上安装的应用软件直接将需要计算的任务丢到集群里）。

常见的任务管理和调度软件：

- SGE- Sun Grid Engine
- Torque 开源 多与Maui配套使用
- Maui Cluster Scheduler 开源
- Platform LSF 付费
- PBS Professional 付费

集群监管

集群监管，是集群管理软件里面的另一重点。

任务管理和调度模块

常见的任务关和调度软件：

- 计算节点模板配置
- 计算节点操作系统推送
- 添加，管理，删除节点
- 集群状态监控，如节点温度，CPU使用率，内存使用率，温度监控等

常用集群管理软件

- Rocks
- Rocks +
- Bright Cluster Manager
- Platform Cluster Manager
- Microsoft HPC 2008 R2 Suit

集群管理软件 ROCKS

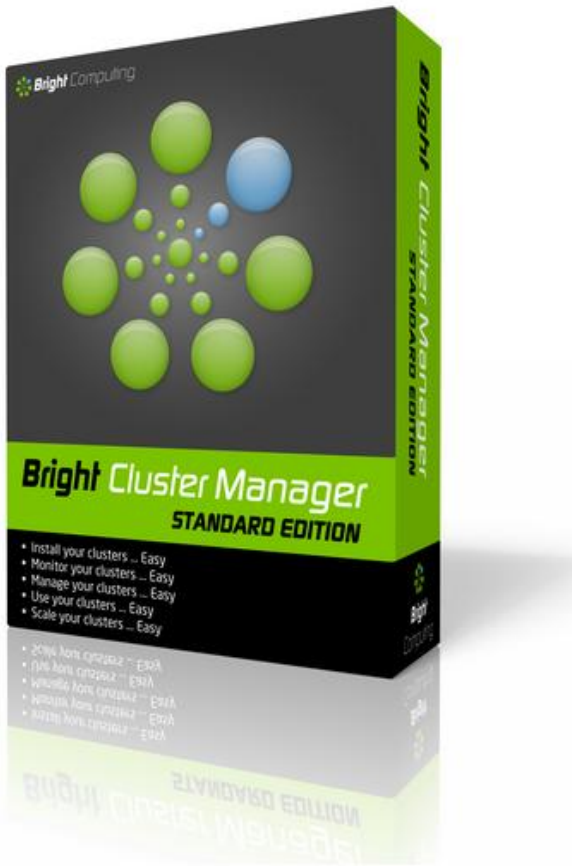
ROCKS 是一个由NPACI(美国高级计算基础设施合作委员会)负责开发的集群管理系统。以性能优秀、成熟和稳定的REDHAT ADVANSCE SERVER为开发基础。

由于其是开源软件，因此我们的SI部门正在开发基于Rocks的AMAX集群管理软件。理论上，客服使用该集群软件，不需要支付任何费用。所以这款集群管理软件也是我们最常使用的集群管理软件之一。

ROCKS的缺点在于其内置的功能和组件较少，比如没有内置GPU的支持。很多功能需要工程师手动安装与配置。另外，ROCKS仅提供有限的用于GPU硬件状况的WEB管理界面，除此之外的各种操作都需要使用命令行进行。因此其使用难度较高。

ROCKS+ 是一款在ROCKS基础上增加了GPU支持等组建的增强版。是商业软件，需要购买授权。但是如果集群节点不超过16个，可以免费使用。

Bright Cluster Manager 集群管理软件



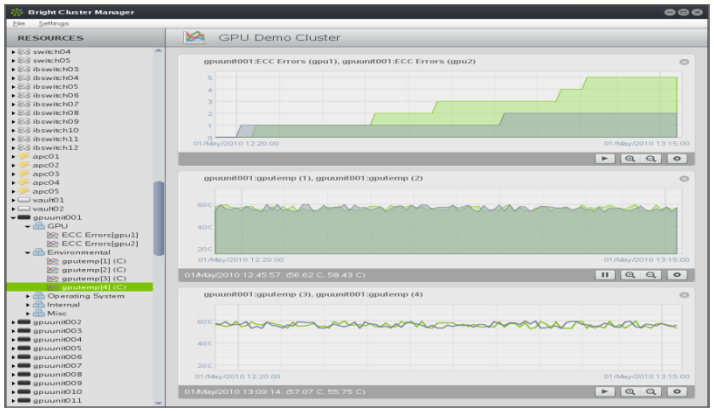
Bright Cluster Manager, 简称BCM，是美国Bright Computing公司开发的一款集群管理软件

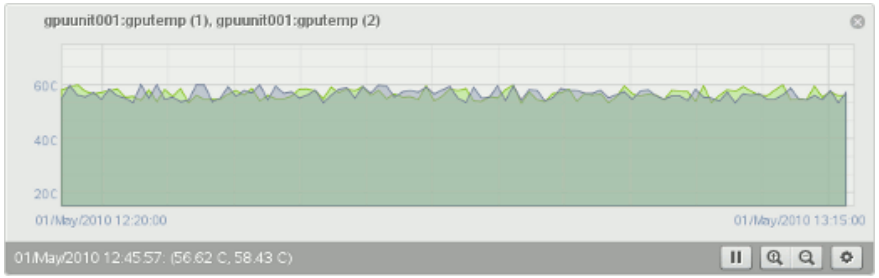
BCM的最大优点在于对GPU的支持非常完善。这表现在两方面：1支持GPU任务的提交与调度；2.可以完善地监控GPU设备的运行状态，并以图形化的方式表现出来

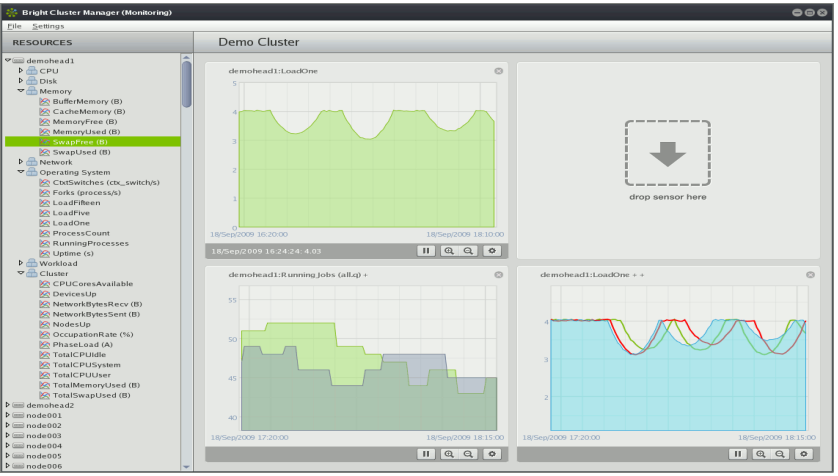
BCM同时支持CPU运算与GPU运算 。BCM的各个功能模块拥有统一的安装和使用界面，简单易用。 仅用一个daemon即可完成所有监管功能，大大降低对系统资源的占用。

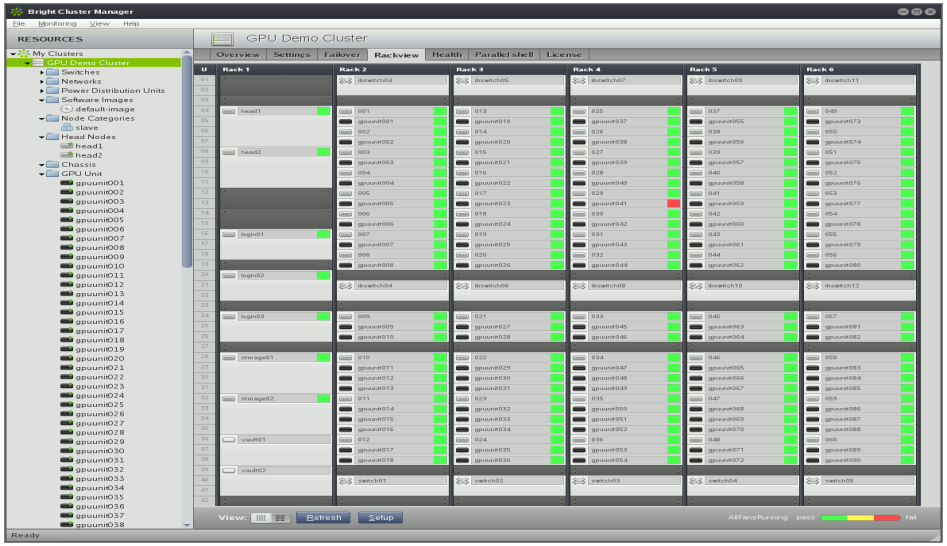
但是BCM不支持Web管理，必须安装客户端才可以使用和管理集群

BCM的图形化管理界面







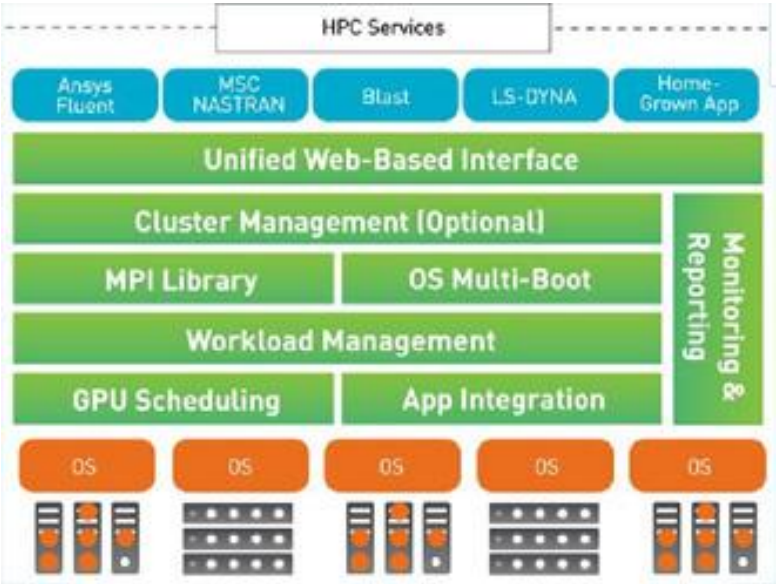


Platform HPC

最为易用与完整的HPC管理软件

Platform HPC是业界最简单、最完整的HPC管理软件。它强大的集群和工作负载管理可以通过最新设计的基于Web的操作界面实现，功能更加强大、操作更加简单。对于需要MPI的应用，强大的商业化MPI库加速和扩展HPC应用，从而缩短解决方案部署时间。

其他厂商的HPC集群解决方案仅是多个工具和接口的结合，并没有进行集成、认证和测试。而Platform HPC则是一体化的产品，具备一整套统一的管理功能，使您能够更加轻松地利用HPC集群强大的功能和可扩展性。

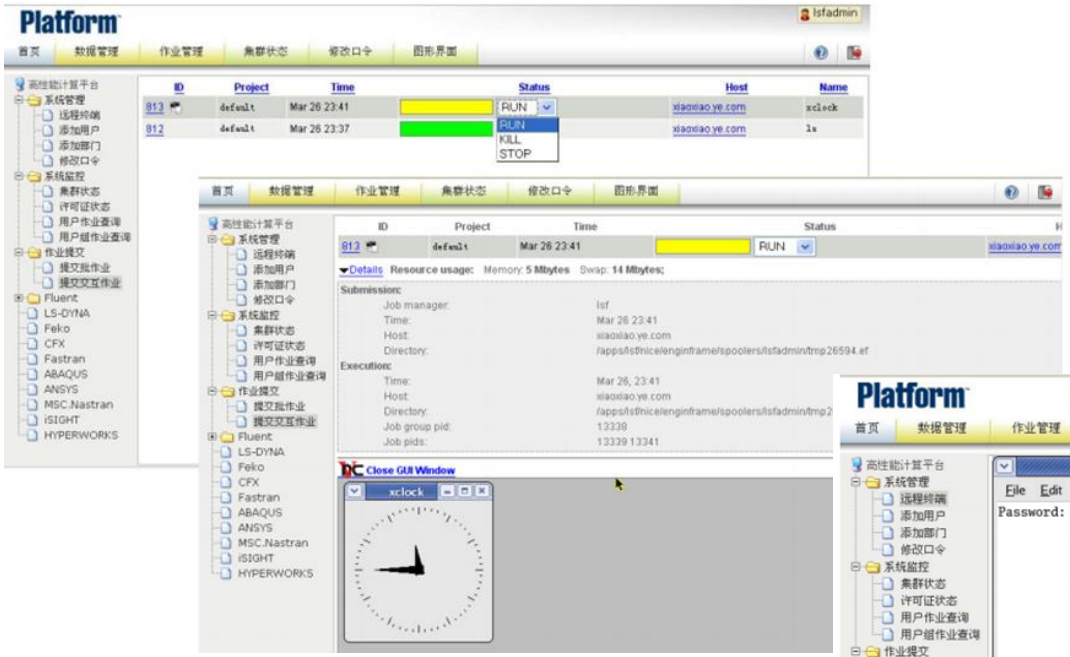


Platform HPC的竞争优势体现在以下方面

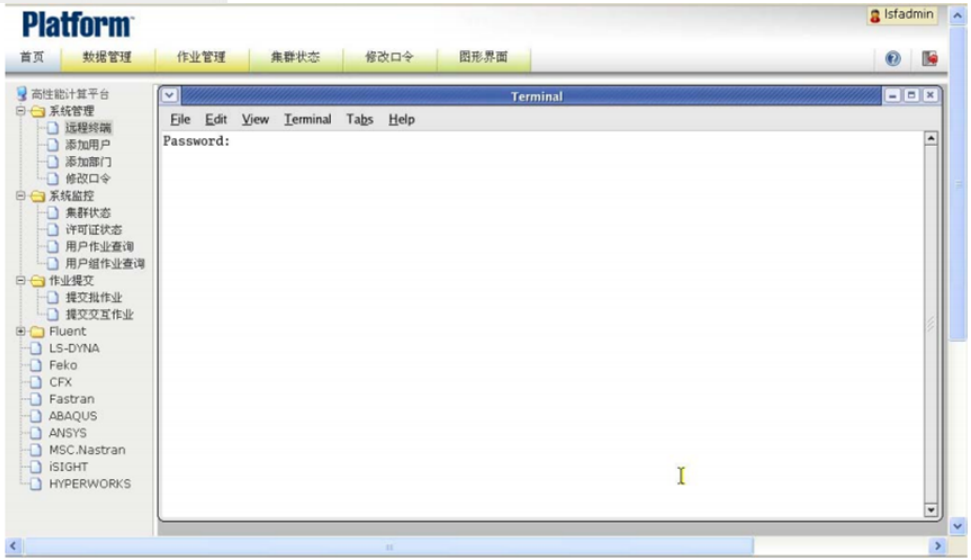
- ✓ 集群的应用支持
- ✓ 完整的、易于使用的集群管理
- ✓ 用户界面友好的、可感知拓扑的工作负载管理
- ✓ 功能强大的工作负载、系统监控和报告
- ✓ 可随处访问的基于Web的中英文的管理界面
- ✓ 动态操作系统的多重启动
- ✓ GPU调度和监控
- ✓ 功能强大的商业MPI库

基于Web使用与管理界面

以用户登录WEB,进行作业提交，作业查看等操作



基于Web的Shell命令介面




操作系统
驱动
集群管理软件
运行环境
开发环境
应用软件

运行环境

运行环境多是指一些驱动，库，和后台驻留程序。绝大多数集群应用软件，都会需要这样，那样的运行环境。

常见的运行环境包括：

- MPI
- CUDA 驱动
- GPU Library

CULA，GPULib，CUDA BLAS，QUDA

- CPU Library

BLAS，GNU Scientific Library，Intel Math Kernel Library

操作系统

驱动

集群管理软件

运行环境

开发环境

应用软件

开发环境

有些用户会在集群环境里自己开发C + MPI，Fortran + MPI，CUDA + MPI的软件。因此就需要有一些比较强大和自动化的开发软件，以提升用户开发的效率。

- Intel® Cluster Studio XE 2012

商业软件，支持Windows和Linux

- Microsoft Visual Studio 2010

商业软件，仅支持Windows

- Caps HMPP

商业软件，仅支持Linux

- PGI CDK Cluster Development Kit

商业软件，支持Windows和Linux

操作系统

驱动

集群管理软件

运行环境

开发环境

应用软件

集群软件

集群应用软件是与大多数用户最息息相关的一部分。大部分用户关心的不是这个集群有多少个节点，用的什么集群管理软件，用的纯CPU还是GPU，用的是什么MPI。用户关心的是他们常用的软件在集群上有没有，性能如何。

集群应用软件，一定是支持MPI的，否则只能在集群的一个节点上运行，而无法使用多节点资源。

常见集群应用软件

- Matlab， Nastran， Anasys, Mathematics, Fluent, NAMD 等等

什么是集群

集群硬件

集群软件

集群相关服务

AMAX在集群方面的优势

服务类型	料号	描述	参考成本	标准	注意事项
集群安装与测试	INSCL2A	In-house Cluster Setup & Testing per extra node by HPC	￥1,100.00	（1台）服务器安装与测试	主要是进行集群管理软件的安装，Infiniband的配置，MPI的安装与配置等。10个点起卖。（不足10个点按10个点算）
	INSCL20	In-house Cluster Setup & Testing upto ten nodes, by HPC	￥11,000.00	（10台）服务器安装与测试，每台2U4Twin的服务器需要4个	
客户现场集群安装与调试	INSCL4A	On-site Cluster Deployment&Testing 1 engineer each extra hr	￥132.00	到客户现场的集群部署服务 每人小时	主要是指在客户现场进行的一些工作，比如和客户实际环境的整合，培训，安装客户所需软件等。 请与超集团队沟通，评估需要多少人工。一般来说不低于3个人天。
	INSCL40	On-site Cluster Deployment&Testing 1 engineer per day by HPC	￥1,100.00	到客户现场的集群部署服务 每人天	
节点上架	LBRNODE	Per Node Cluster Assembly	￥385.00	每个节点的上架，测试。	主要是指集群节点的物理安装，线缆连接等 每台2U4Twin需要4个
机柜安装	LBRCAB2	22U Cluster Cabinet Assembly	￥3,216.40	每个22U机柜的安装	机柜的安装：机柜组装，PDU组装等 每个机柜一个
	LBRCAB4	42U Cluster Cabinet Assembly	￥4,180.00	每个42U机柜的安装	
保修	CW3YB15	Warranty 3 Year Parts & Labor, Server/Workstation B15	￥1,090.00	3年低于一万五	保修有1年、2年、3年的。这里列举的都是3年的。 保修选择标准是：集群附件的成本之和（不包含保修）与各个保修进行比较
	CW3YB35	Warranty 3 Year Parts & Labor, Server/Workstation B35	￥1,657.00	3年低于三万五	
	CW3YB55	Warranty 3 Year Parts & Labor, Server/Workstation B55	￥2,354.00	3年低于五万五	
	CW3YB75	Warranty 3 Year Parts & Labor, Server/Workstation B75	￥2,702.00	3年低于七万五	
	CW3YU75	Warranty 3 Year Parts & Labor, Server/Workstation over75	￥4,460.00	3年高于七万五	

集群相关服务

集群相关服务是能够体现一个公司技术实力的地方。作为集群实施工程师，需要对集群的构架，Linux， Infiniband， MPI， 集群应用软件等具有深刻的了解。另外，作为集群实施工程师需要具有丰富的集群实施经验。因为在集群的实施过程中可能会遇到各种各样的问题，尤其是软件与软件和软件与硬件之间的兼容性问题，所以要求工程师有丰富的经验去解决这些问题。

什么是集群

集群硬件

集群软件

集群相关服务

AMAX在集群方面的优势

单机品质

集群硬件安装品质

集群软件部署品质

成功案例

严格的单机测试

STAGE 5 Quality Testing

Stage 1: 组件测试

对主板，CPU，内存， 硬盘，GPU等各个组件进行测试，剔除不合格的组件。

Stage 2: 系统组装

AMAX的所有机器都严格遵照MPI（生产流程规范）进行生产。每种机型都有自己的MPI。这些MPI中都有明确规定在完成组装的各个步骤中所需要进行的各项测试。

Stage 3: 高温测试

AMAX的所有机器都就将在专用的拷机室之中满负荷运转12到24小时。拷机室中的温度为35度到40度。

Stage 4: 性能测试

AMAX的所有机器都需要通过功能测试。性能测试包含IO性能，运算速度，内存带宽等测试项目。而该测试环节往往为其他一些厂商所忽视。

Stage 5: 最终质量检测

包含对机器外观目测，对机器参数配置的检测，对于机器附件的检测，以及随机挑选一些机器进行第二轮功能测试。当然，对于机器运输包装的检测也是最重质量检测的一部分。

AMAX 生产规范

遵循独特的生产流程规范（MPI）及DataSweep服务，AMAX 重新定义了按单定制生产。

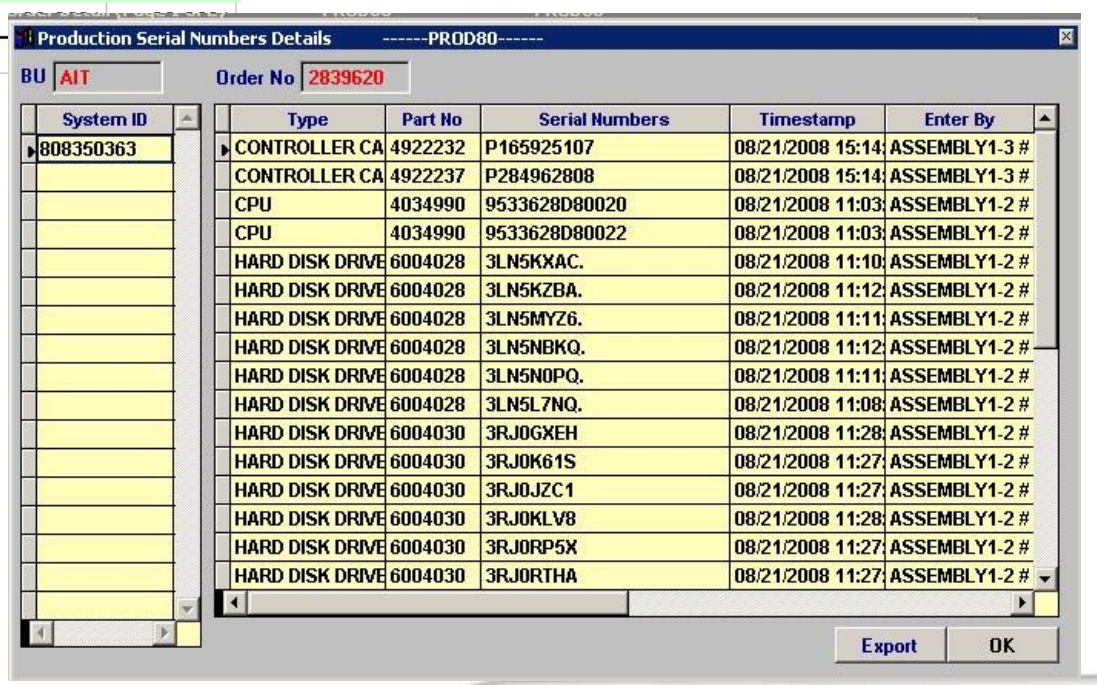
大到机器生产、系统设置，小到线缆和风扇的连接布局，客户都能获得一致正确的产品配置。

AMAX 质量控制

	WK36	WK37	WK38	WK39
Yield Summary				
Mechanical Inspection QC	100.00%	100.00%	100.00%	100.00%
Hi-Pot	100.00%	100.00%	100.00%	100.00%
Burn in	98.00%	98.20%	98.00%	98.10%
Imaging	98.10%	98.10%	98.20%	98.20%
Final QA	100.00%	100.00%	100.00%	100.00%
OBA	100.00%	100.00%	100.00%	100.00%
RTY	96%	96%	96%	96%
RTY-Goal	90.00%	90.00%		
Total Systems Build	181	214		

产品良率分析

所有生产配件的序列号都记录在生产管理系统中。



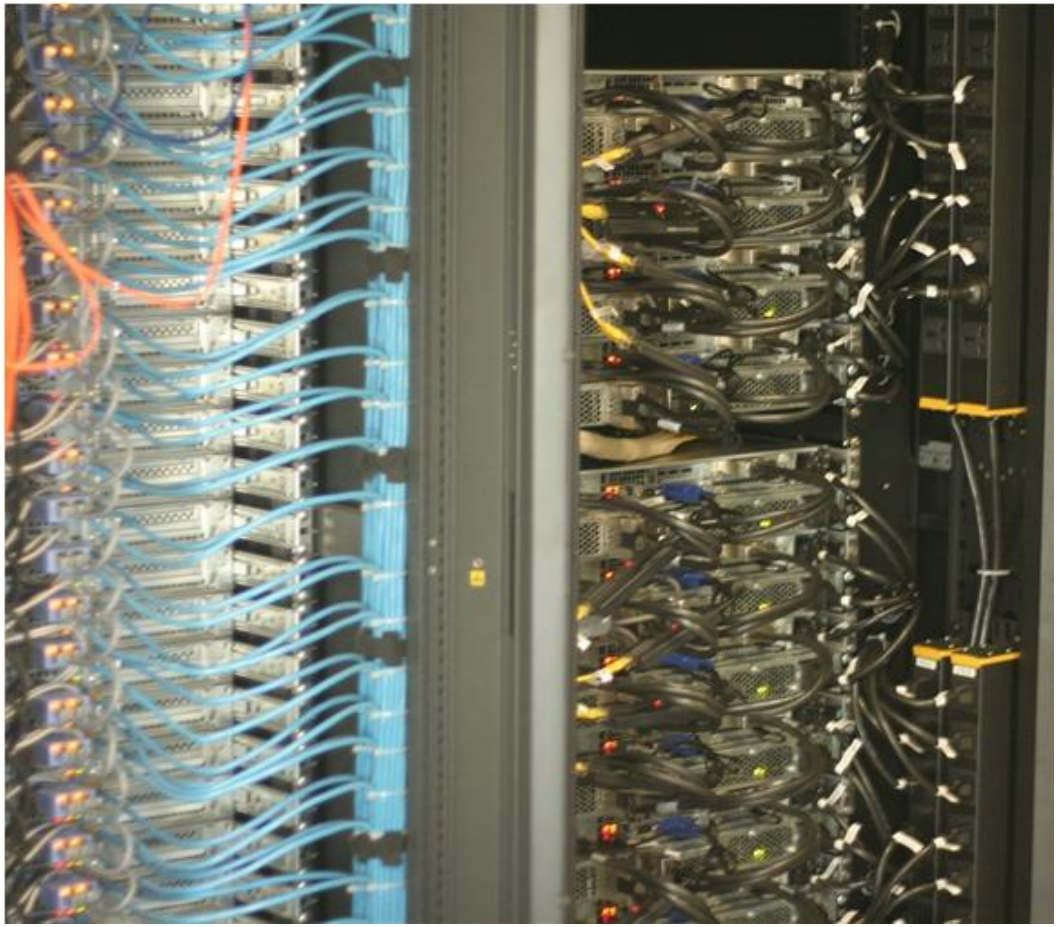
单机品质

集群硬件安装品质

集群软件部署品质

成功案例

集群硬件安装品质



单机品质

集群硬件安装品质

集群软件部署品质

成功案例

集群软件部署品质

每位集群客户都能得到由高级主管负责的专门客户管理团队的服务

工程师技术娴熟且具有丰富的专业知识，他们在协调有序的工作中重新定义个性化客户服务

客户经理

销售工程师

客户服务工程师

技术支持工程师

优秀的项目团队

Aaron Liu

复旦大学物理学硕士。拥有计算机与物理双重背景。熟悉各种常用的科学计算软件。

负责实施过的项目：

清华大学张家港机房CPU集群
国防科技大学CPU集群
重庆大学GPU集群
北京航空航天大学GPU集群
三峡大学CPU集群
南京大学GPU+CPU混合集群

Zack Zhang

拥有丰富集群实施经验，精通Linux，实施过多个高校GPU集群项目。

负责实施过的项目：

恒泰艾普第一数据中心GPU集群
恒泰艾普第二数据中心GPU集群
同济大学海洋学院CPU集群
同济大学海洋学院GPU集群
中国石油大学GPU集群

单机品质

集群硬件安装品质

集群软件部署品质

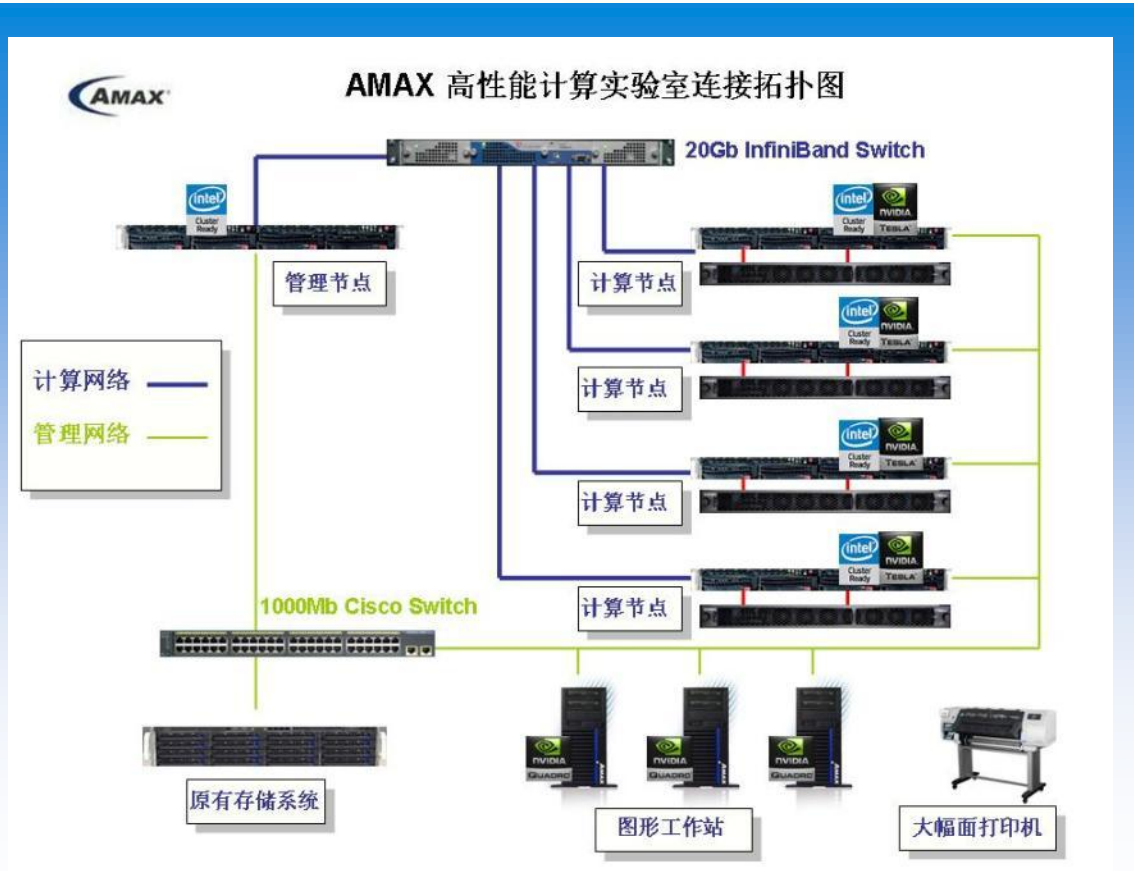
成功案例

AMAX中国成功案例

同济大学海洋学院学院

同济大学GPU集群 2009年9月份，AMAX与同济大学签署并实施了国内高校首例成功应用 Telsa GPU Cluster案例, 利用NVIDIA的CUDA技术产品,和AMAX雄厚的技术实力,为客户整合出一套极具性价比的Cluster解决方案, 极大的提升了客户的运算效率.

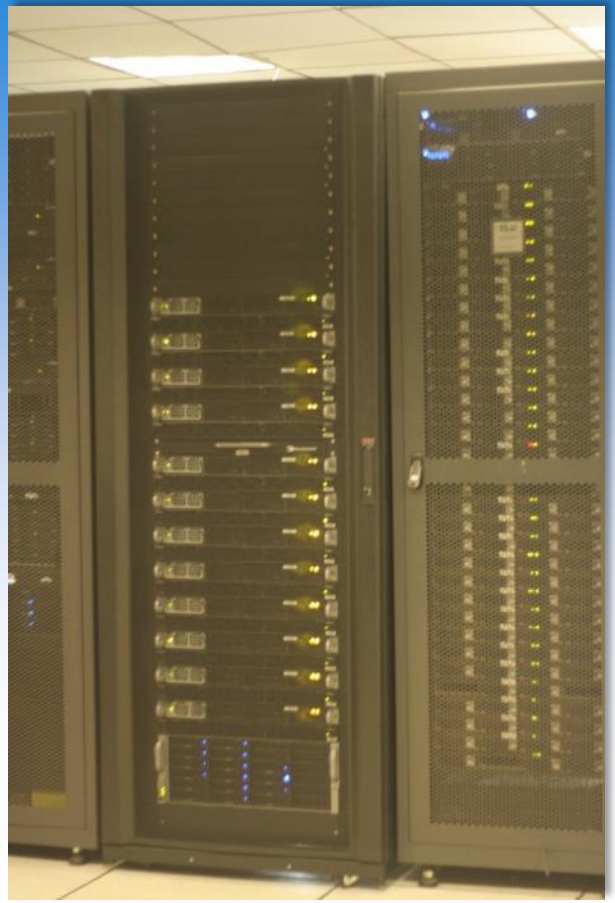
AMAX中国成功案例

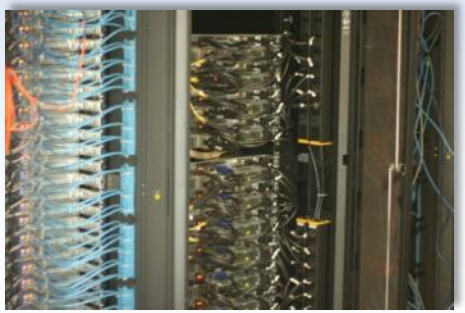


中国石油大学

由于同济大学海洋学院在GPU方面的成功应用，中国石油大学(青岛)于2010年4月请AMAX为其建立了一个GPU拥有16T的计算能力集群，主要加速海洋中的石油与天然气探测。

AMAX中国成功案例

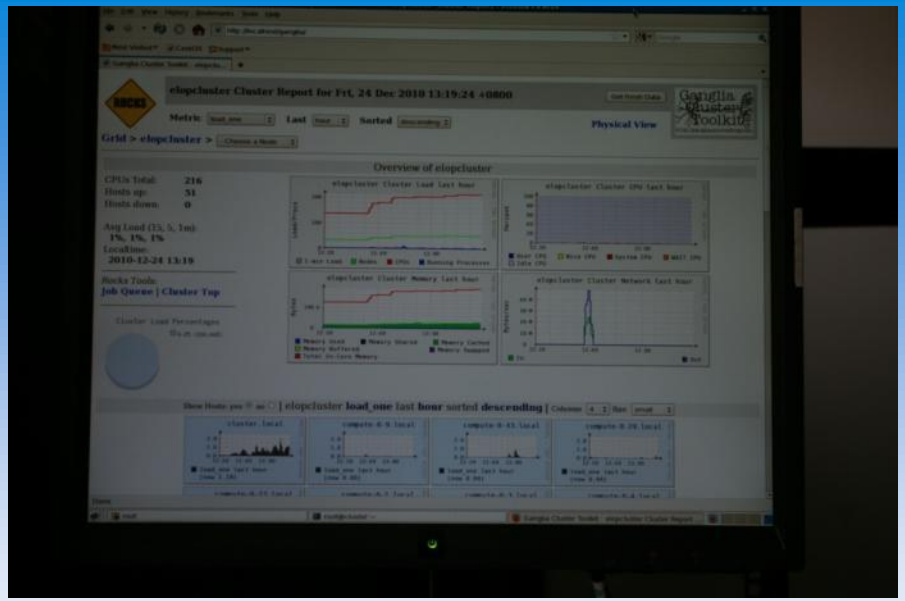


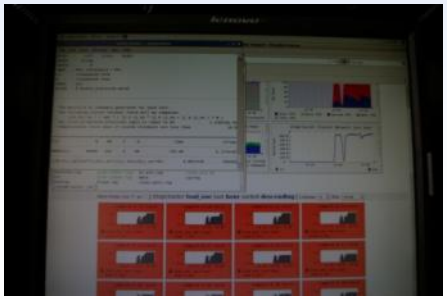


恒泰艾普

恒泰艾普石油天然气技术服务股份有限公司业务包括石油天然气勘探、开发、生产的技术研究、产品研发、技术服务、产品销售。是一家从事石油勘探与开发技术研究、服务与相应软件研发、销售的专业化的高科技公司。AMAX在2010年5月为恒泰爱普提供了一套运算能力为100T的GPU集群用以恒泰爱普在哥伦比亚的石油勘探工作。

AMAX中国成功案例





清华大学

AMAX在2010年12月为清华大学张家港机房建立了一套50节点的低功耗CPU计算机群以供清华的学生学习和联系大规模并行计算软件的开发。

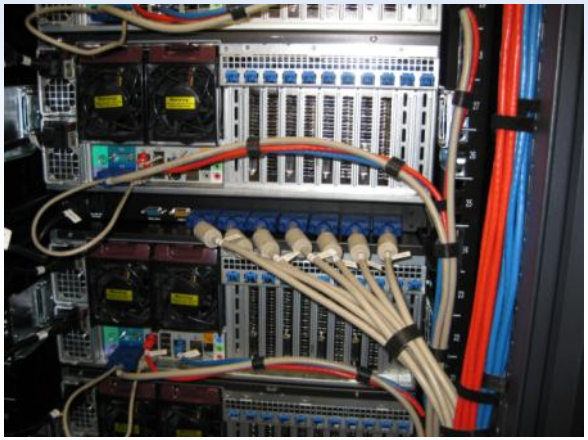
AMAX中国成功案例

国防科技大学

AMAX与2011年1月，帮助国防科技大学建立了一套16个节点的CPU计算机群。

AMAX在该项目中采用了先进的迷你刀片机型，既保证了国防应用领域对于减少单点故障的要求，又大大减少了空间占用。

AMAX中国成功案例

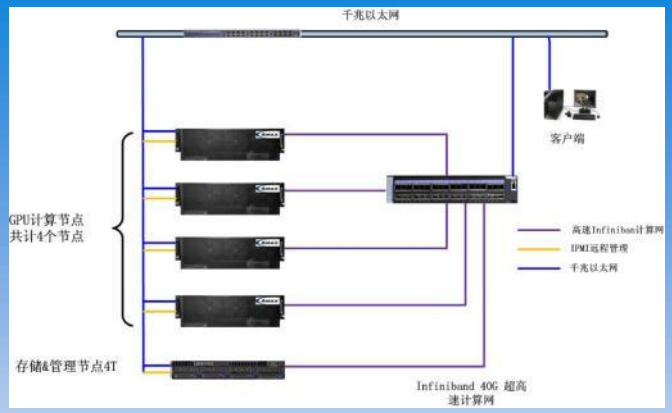


加州理工学院

AMAX与2011年1月，AMAX为美国加州理工学院部署了一个计算计算能力为74T的GPU计算机群。

该项目由两位 AMAX中国 的资深工程师飞赴美国进行实施的。

AMAX中国成功案例

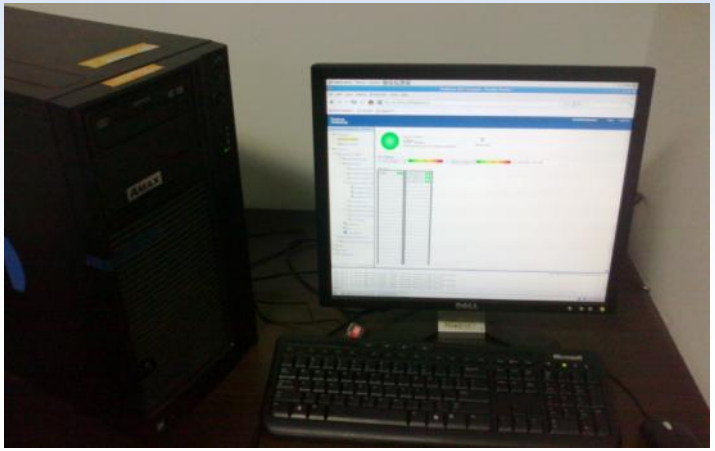


重庆大学现代物理中心

AMAX与2011年3月，为重庆大学现代物理中心打造了一套计算能力24T的GPU计算机群。

该集群是国内首个成功在GPU集群上全面使用Matlab的项目。

AMAX中国成功案例



北京航空航天大学

AMAX与2011年4月，为北京航空航天大学打造了一套特殊的GPU计算机群。

该集群是由四台塔式超级计算机构成。该集群支持Windows与Platform集群双启动。

用户在平时可以使用各自独立地使用计算机，在下班或节假日则切换到集群模式进行高强度计算。

AMAX中国成功案例

三峡大学

AMAX与2011年5月，为三峡大学打造了一套CPU+GPU混合计算机群。

该集群的CPU计算部分采用了AMD Opteron 6176 12核心CPU。该CPU在同等预算的前提下提供了超过Intel CPU 多达35%计算能力。

AMAX在该AMD CPU集群上成功部署了Intel MKL Compiler并运行成功。

AMAX中国成功案例

南京航空航天大学

AMAX与2011年10月，为南航打造了一套CPU+GPU混合计算机群。

该集群的由一套CPU刀片服务器和两台GPU服务器构成。

在该套集群中，AMAX创造性地提供了多组热拔插硬盘。并在不同的硬盘组上安装了2套从物理层面完全隔离的集群环境。

这样，一套硬件既可以服务于敏感的课题研究，又可以为学生提供集群开发实践环境。