第11章 机群系统

张晨曦 刘依

www.GotoSchool.net

xzhang2000@sohu.com

目前流行的高性能并行计算机系统结构通常可以分成5类：

- 并行向量处理机（PVP）
- 对称多处理机（SMP）
- 大规模并行处理机（MPP）
- 分布共享存储多处理机（DSM）
- 机群（Cluster）
  - 优势：低廉的价格、极强的灵活性和可缩放性
  - 成为近年来发展势头最为强劲的系统结构

全球Top500中机群计算机的数量和比例

时间	1997.6	1997.11	1998.6	1998.11	1999.6	1999.11	2000.6	2000.11
数量	1	1	1	2	6	7	11	28
比例	0.2%	0.2%	0.2%	0.4%	1.2%	1.4%	2.2%	5.6%
时间	2001.6	2001.11	2002.6	2002.11	2003.6	2003.11	2004.6	2004.11
数量	32	43	81	93	149	208	289	294
比例	6.4%	8.6%	16.2%	18.6%	29.8%	41.6%	57.8%	58.8%
时间	2005.6	2005.11	2006.6	2006.11	2007.6	2007.11	2008.6	
数量	304	361	364	361	374	406	400	
比例	60.8%	72.2%	72.8%	72.2%	74.8%	81.2%	80%	

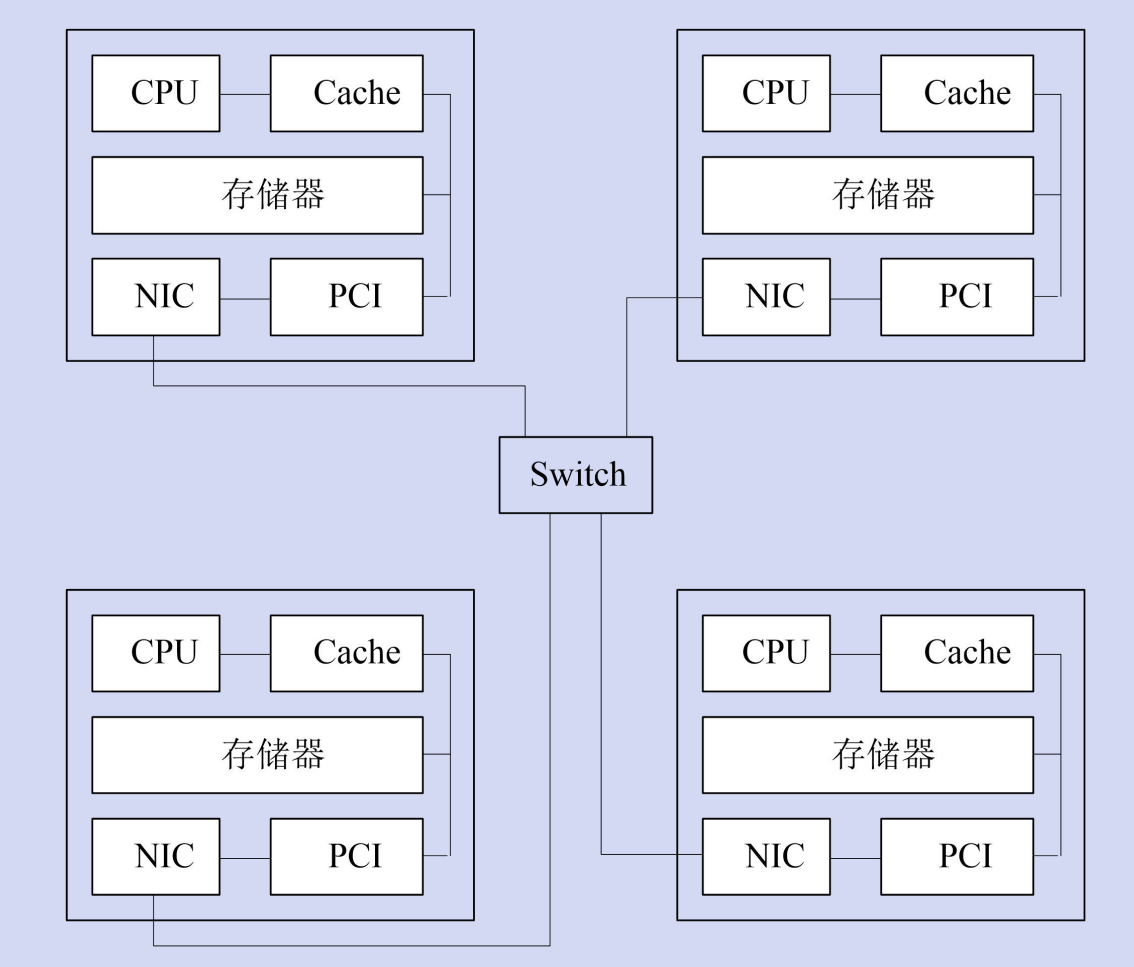
11.1 机群的基本结构

11.1.1 机群的硬件组成

1. 机群

- 一种价格低廉、易于构建、可扩放性极强的并行计算机系统。
- 由多台同构或异构的独立计算机通过高性能网络或局域网互连在一起，协同完成特定的并行计算任务。
- 从用户的角度来看，机群就是一个单一、集中的计算资源。

- 一个简单PC机群的逻辑结构
  - 4台PC机通过交换机连接在一起。
  - NIC表示网络接口，PCI表示I/O总线。
  - 这是一种无共享的结构，大多数机群都采用这种结构。
  - 如果将图中的交换机换为共享磁盘，则可以得到共享磁盘结构的机群系统。



一个包含4个结点的简单PC机群

1. 构成机群的每台计算机都被称为一个结点。
   - 每个结点都是一个完整的系统，拥有本地磁盘和操作系统，可以作为一个单独的计算资源供用户使用。
   - 除了PC机外，机群的结点还可以是工作站，甚至是规模较大的对称多处理机。
   - 结点分类
     - 计算结点
     - 管理登录结点
     - I/O结点

1. 机群的各个结点一般通过商品化网络连接在一起。
2. 网络接口与结点的I/O总线以松散耦合的方式相连。

11.1.2 机群的软件

1. **机群操作系统：**在各结点的操作系统之上建立一层操作系统来管理整个机群。
2. 机群操作系统的功能
   - 提供硬件管理、资源共享以及网络通信
   - 实现单一系统映象

- Single System Image，简称SSI
- 一项重要功能
- 机群的一个重要特征

1. SSI有4重含义
   - 单一系统
   - 单一控制
     - 逻辑上，最终用户或系统用户使用的服务都来自机群中唯一一个位置；
     - 系统管理员通过一个唯一的控制点配置机群的所有软、硬件组件。
   - 对称性：用户可以从任一个结点上获得机群服务。
   - 位置透明：用户不必了解真正提供服务的物理设备的具体位置。

1. 机群系统中的SSI至少应该提供以下三种服务：

- 单一登录
  - 即用户可以通过机群中的任何一个结点登录，而且在整个作业执行过程中只需登录一次，不必因作业被分派到其它结点上执行而重新登录。
- 单一文件系统
  - 在机群系统中，有一些对整个机群所有结点而言都相同的软件，它们没有必要在每一个结点上重复安装。
  - 执行并行作业时要求每个结点都可以访问到这些软件，但它们在整个机群系统中应该只有一个备份。

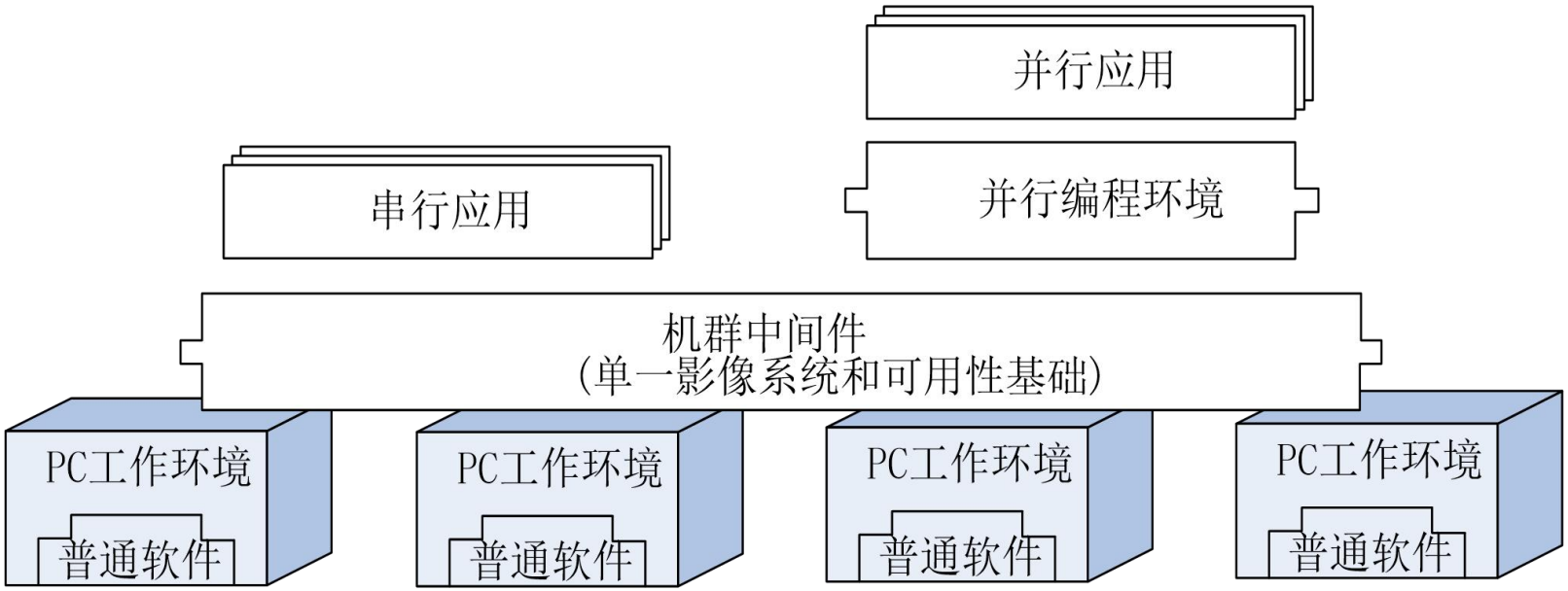
- 单一作业管理系统
  - 用户可以透明地从任一结点提交作业，作业可以以批处理、交互或并行的方式被调度执行。
  - PBS、LSF、Condor和JOSS都是目前比较具有代表性的作业管理系统。

1. 并行编程模型以及相关的并行编程环境

比较流行的并行编程工具：

MPI、PVM、OpenMP、HPF

6. 机群系统的软件框架


并行应用
串行应用
并行编程环境
机群中间件
(单一影像系统和可用性基础)
PC工作环境
普通软件
PC工作环境
普通软件
PC工作环境
普通软件
PC工作环境
普通软件

机群系统的软件框架

11.2 机群的特点

1. 机群系统的优点

- 系统开发周期短
  - 结点采用商品化的PC机、工作站，通过商用网络连接。
  - 系统开发的重点：通信子系统和并行编程环境
- 可靠性高

  每个结点都是独立的PC机或工作站
- 可扩放性强
  - 机群的计算能力随着结点数量的增加而增大
  - 机群结构灵活（结点之间以松耦合方式连接）
  - 机群系统的硬件容易扩充和替换，可以灵活配置。

- 性能价格比高
- 用户编程方便

1. 机群的迅猛发展还得益于微处理器技术、网络技术和并行程序设计技术的进步。
   - 微处理器技术的进步使得微处理器的性能不断提高，价格不断下降；
   - 机群系统更容易融合到已有的网络系统中，而且随着网络技术的进步和高性能通信协议的引入，机群结点间的通信带宽进一步提高，通信延迟进一步缩短，逐步缓解了由于结点松散耦合引起的机群系统通信瓶颈问题。

- 随着PVM、MPI、HPF、OpenMP等并行编程模型的应用与成熟，使得在机群系统上开发并行应用更加方便。

1. 机群的不足之处

由于机群由多台完整的计算机组成，它的维护相当于要同时去管理多个计算机系统，因此维护工作量较大，维护费用也较高。

11.3 机群的分类

1. 根据组成机群的各个结点和网络是否相同，分为：

   同构、异构

2. 根据结点是PC还是工作站，分为：

   PC机群、工作站机群

3. 以机群系统的使用目的为依据，分为：

   高可用性机群、负载均衡机群以及高性能机群

   （最常用的分类方法 ）

1. 高可用性机群

- 主要目的：当系统中某些结点出现故障的情况下，仍能继续对外提供服务。
- 采用冗余机制
  - 当系统中某个结点由于软、硬件故障而失效时，该结点上的任务将在最短的时间内被迁移到机群内另一个具有相同功能与结构的结点上继续执行。
  - 对于用户而言，系统可以一直为其提供服务。
- 适用于Web服务器、医学监测仪、银行POS系统等要求持续提供服务的应用。

1. 负载均衡机群

- 主要目的：提供与结点个数成正比的负载能力
- 要求：机群能够根据系统中各个结点的负载情况实时地进行任务分配。
- 专门设置了一个重要的监控结点，负责监控其余每个工作结点的负载和状态，并根据监控结果将任务分派到不同的结点上。
- 适合大规模网络应用

  如Web服务器或FTP服务器、大工作量的串行或批处理作业（如数据分析）

- 负载均衡机群适用于提供静态数据的服务；而高可用性机群既适用于提供静态数据的服务，又适用于提供动态数据的服务。

1. 高性能计算机群

- 主要目的：降低高性能计算的成本
- 通过高速的商用互连网络，将数十台乃至上千台PC机或工作站连接在一起，可以提供接近甚至超过传统并行计算机系统的计算能力，但其价格却仅是具有相同计算能力的传统并行计算机系统的几十分之一。

1. 按照构建方式将机群分为：

（一种比较常用的分类方法）

- 专用机群
  - 吞吐率较高，响应时间较短。
  - 专用机群的结点往往是同构的，一般采用集中控制，由一个（或一组）管理员统一管理，而且用户一般需要通过一台终端机来访问它。
- 企业机群
  - 各结点之间一般通过标准的LAN或WAN互连
  - 通信开销较大、延迟较长
  - 企业机群的各个结点一般是异构的

11.4 典型机群系统简介

11.4.1 Berkeley NOW

美国加州大学Berkeley分校开发

具有很多优点：

- 采用商用千兆以太网和主动消息通信协议支持有效的通信。
- 通过用户级整合机群软件GLUNIX提供单一系统映象、资源管理和可用性，开发了一种新的无服务器网络文件系统xFS，以支持可扩放性和单一文件层次的高可用性。

1. 主动消息

- 实现低开销通信的一种异步通信机制
- 基本思想
  - 在消息头部控制信息中携带一个用户级子例程（称作消息处理程序）的地址。
  - 当消息头到达目的结点时，调用消息处理程序通过网络获取剩下的数据，并把它们集成到正在进行的计算中。
  - 主动消息相当高效和灵活，以至于各种系统都逐渐地用它作为基本的通信机制。

1. GLUNIX

- 运行在工作站标准UNIX上的一个软件层，属于自包含软件。
- 主要思想
  - 机群操作系统应由底层和高层组成；
  - 底层是执行在核模式下的结点商用操作系统，高层是能提供机群所需的一些功能的用户级操作系统。
  - 特别地，这一软件层能够提供机群内结点的单一系统映象，使得所有的处理器、存储器、网络容量和磁盘带宽均可以被分配给串行和并行应用。

1. 无服务器文件系统xFS

- 一个无服务器的分布式文件系统；
- 将文件服务的功能分布到机群的所有结点上，以提供低延迟高带宽的文件系统服务功能；
- 主要采用廉价冗余磁盘阵列、协同文件缓存和分布式管理等技术。

11.4.2 Beowulf

1. 目标
   - 1GFlops的计算处理能力和10GB的存储容量
   - 价格不能过高
2. 一个具有16个结点的机群

   （Thomas Sterling与Don Becker二人构建）
   - 硬件：Intel的DX4处理器以及10Mb/s的以太网
   - 软件：基于Linux系统以及其它一些GNU软件

- 将这个系统命名为Beowulf
  - 这种基于COTS（Commodity Off The Shelf）思想的技术也迅速由NASA传播到其它科研机构。
  - 这类机群被称为Beowulf机群。

（Beowulf Class Cluster Computers）

1. Beowulf并不是一套具体的软件包或是一种新的网络拓扑结构，它只是一种思想。

在达到既定目标的前提下，把注意力集中在获取更高的性能价格比上。

11.4.3 LAMP

1. 使用低成本、小配置（2～8个处理器）的SMP来构建机群系统逐渐成为主流。
   - 这种结构的系统被统称为CLUMPs
     （CLUster of MultiProcessors）
   - 由于SMP结点内部与SMP结点之间通信能力往往不一致，CLUMPs一般使用专门的通信协议和通信算法。

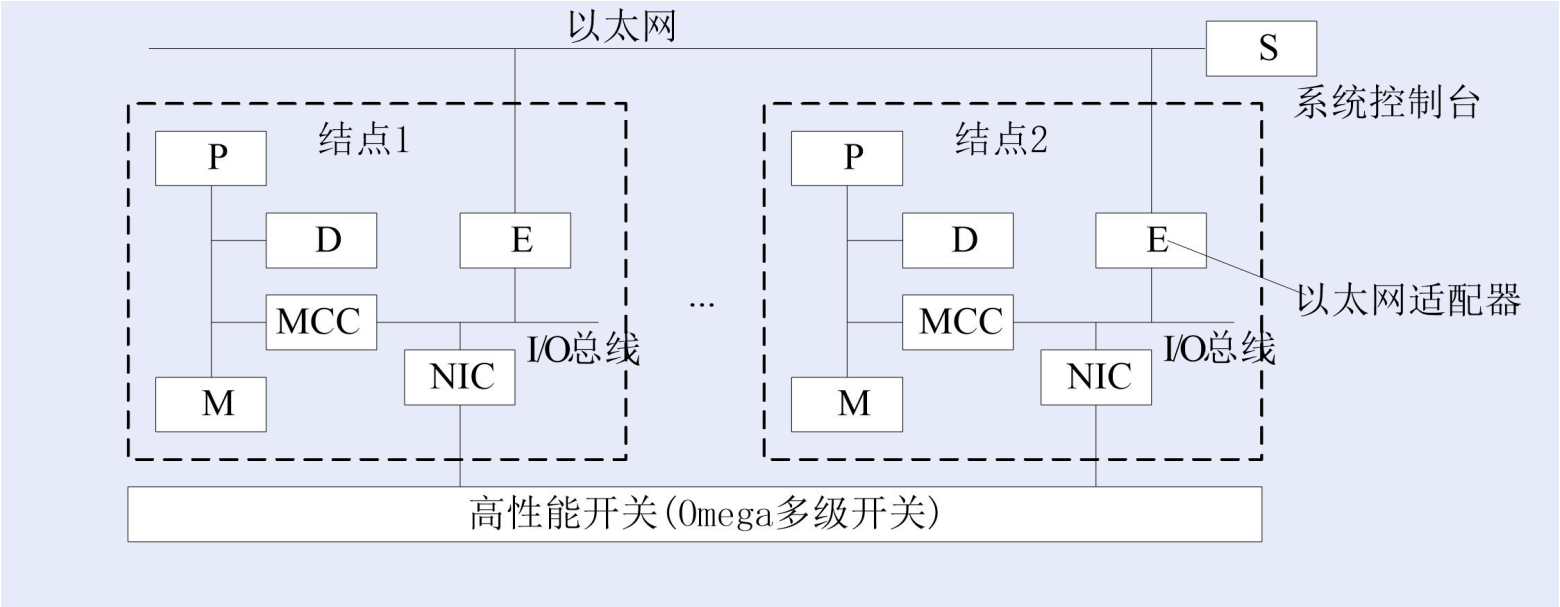
1. LAMP（Local Area MultiProcessor）

- 由NEC实验室构建，基于Pentium Pro PC机、SMP机群
- 共有16个结点

  每个结点包含两个Pentium Pro 200MHz的CPU以及256MB内存。
- 操作系统使用了支持SMP的Linux 2.0.34内核版本，提供MPICH 1.1.0并行程序开发环境。
- 同一个SMP结点内的两个CPU之间采用基于共享存储器的消息传递机制进行通信，而结点间通信则通过Myrinet完成。

11.4.4 IBM SP2

1. 深蓝：采用30个RS/6000工作站（带有专门设计的480片国际象棋芯片）的IBM SP2机群
2. 异步的MIMD，具有分布式存储器系统结构。



- 结点：一台RS/6000工作站，带有自己的存储器和本地磁盘。
- 结点中采用的处理器：一台6流出的超标量处理机

  每个时钟周期可以执行6条指令，包括2条读数写数指令，2条浮点乘或加指令，1条变址增量指令和1条分支指令。
- 每个结点配有一套完整的AIX操作系统（IBM的UNIX）
- 结点间的互连网络接口是松散耦合的，通过结点本身的I/O微通道（MCC）接到网络上，而不是通过本身的存储器总线。

- SP2的结点数可以从2个到512个不等，除了每个结点采用RS/6000工作站外，整个SP2系统还需要配置另外一台RS/6000工作站作为系统控制台 。

1. SP2的结点可分为3类：宽结点、细结点、细2结点
   - 它们都有：
     - 1个指令Cache
     - 1个数据Cache
     - 1个分支指令和转移控制部件
     - 2个整数部件
     - 2个浮点部件

- 但它们在存储器容量、数据宽度和I/O总线插槽个数上有所不同。例如：
  - 在存储器容量方面

    宽结点：64～2048MB

    细结点和细2结点： 64～512MB
  - 在存储器总线的宽度方面

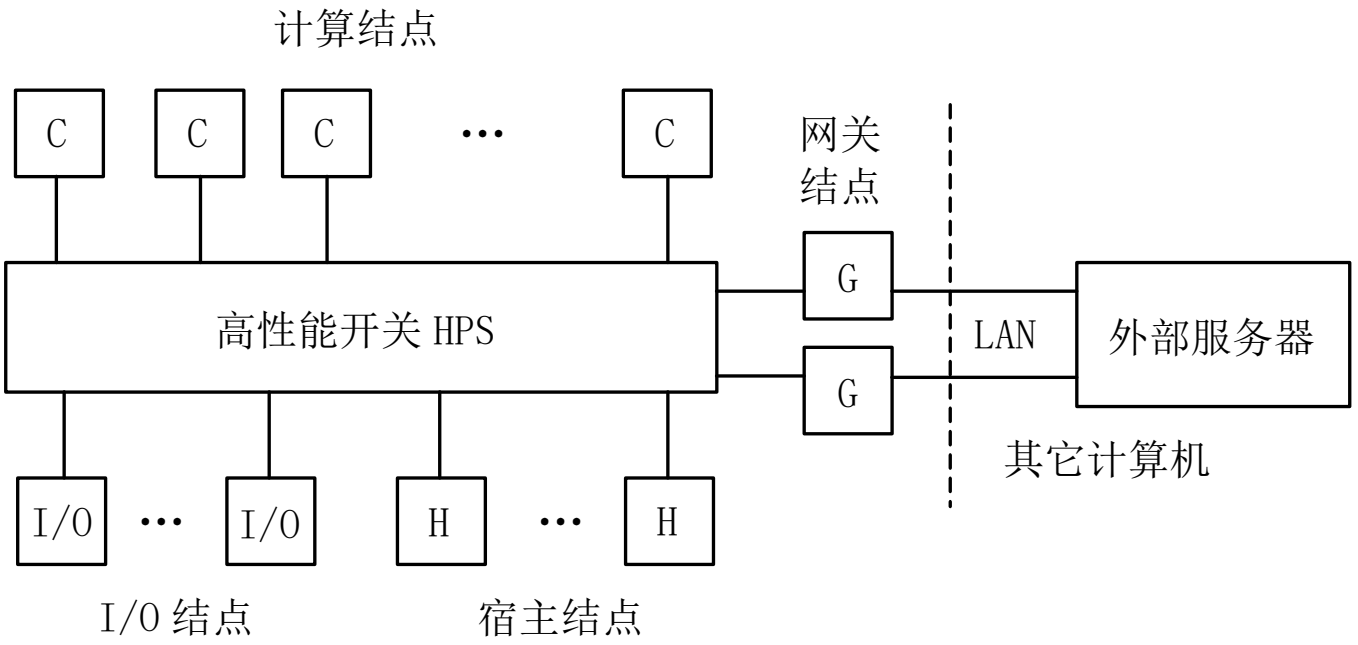
    宽结点：256位

    细2结点：128位

    细结点：64位

- SP2的结点通过网络接口开关NIC接到HPS，IBM将其称为开关适配器。

1. SP2的I/O子系统的总体结构



1. SP2系统软件的核心：AIX操作系统
2. SP2中设置了一个专门的系统控制台用以管理整个系统，系统管理人员可以通过这个系统控制台从单一地点对整个系统进行管理。