1.1 Peer-To-Peer 介绍

最近几年，对等计算(Peer-to-Peer，简称 P2P) 迅速成为计算机界关注的热门话题之一，财富杂志更将 P2P 列为影响 Internet 未来的四项科技之一。
目前,在学术界、工业界对于 P2P 没有一个统一的定义，下面列举几个常用的定义供参考：
定义:1 、Peer-to-peer is a type of Internet network allowing a group of computer users with the same networking program to connect with each other for the purposes of directly accessing files from one another's hard drives.
2 、Peer-to-peer networking (P2P) is an application that runs on a personal computer and shares files with other users across the Internet. P2P networks work by connecting individual computers together to share files instead of having to go through a central server.
3、P2P 是一种分布式网络，网络的参与者共享他们所拥有的一部分硬件资源（处理能力、存储能力、网络连接能力、打印机等），这些共享资源需要由网络提供服务和内容，能被其它对等节点(Peer)直接访问而无需经过中间实体。在此网络中的参与者既是资源（服务和内容）提供者（Server），又是资源（服务和内容）获取者（Client）。虽然上述定义稍有不同，但共同点都是 P2P 打破了传统的 Client/Server (C/S)模式，在网络中的每个结点的地位都是对等的。每个结点既充当服务器，为其他结点提供服务，同时也享用其他结点提供的服务。

P2P 技术的特点体现在以下几个方面。

非中心化（Decentralization）：网络中的资源和服务分散在所有结点上，信息的传输和服务的实现都直接在结点之间进行，可以无需中间环节和服务器的介入，避免了可能的瓶颈。P2P 的非中心化基本特点，带来了其在可扩展性、健壮性等方面的优势。
可扩展性：在 P2P 网络中，随着用户的加入，不仅服务的需求增加了，系统整体的资源和服务能力也在同步地扩充，始终能较容易地满足用户的需要。整个体系是全分布的，不存在瓶颈。理论上其可扩展性几乎可以认为是无限的。
健壮性：P2P 架构天生具有耐攻击、高容错的优点。由于服务是分散在各个结点之间进行的，部分结点或网络遭到破坏对其它部分的影响很小。P2P 网络一般在部分结点失效时能够自动调整整体拓扑，保持其它结点的连通性。P2P 网络通常都是以自组织的方式建立起来的，并允许结点自由地加入和离开。P2P 网络还能够根据网络带宽、结点数、负载等变化不断地做自适应式的调整。
高性能/价格比：性能优势是 P2P 被广泛关注的一个重要原因。随着硬件技术的发展，个人计算机的计算和存储能力以及网络带宽等性能依照摩尔定理高速增长。采用 P2P 架构可以有效地利用互联网中散布的大量普通结点，将计算任务或存储资料分布到所有结点上。利用其中闲置的计算能力或存储空间，达到高性能计算和海量存储的目的。通过利用网络中的大量空闲资源，可以用更低的成本提供更高的计算和存储能力。
隐私保护: 在 P2P 网络中，由于信息的传输分散在各节点之间进行而无需经过某个集中环节，用户的隐私信息被窃听和泄漏的可能性大大缩小。此外，目前解决 Internet 隐私问题主要采用中继转发的技术方法，从而将通信的参与者隐藏在众多的网络实体之中。在传统的一些匿名通信系统中，实现这一机制依赖于某些中继服务器节点。而在 P2P 中，所有参与者都可以提供中继转发的功能，因而大大提高了匿名通讯的灵活性和可靠性，能够为用户提供更好的隐私保护。

负载均衡: P2P 网络环境下由于每个节点既是服务器又是客户机，减少了对传统 C/S 结构服务器计算能力、存储能力的要求，同时因为资源分布在多个节点，更好的实现了整个网络的负载均衡。
与传统的分布式系统相比，P2P 技术具有无可比拟的优势。同时，P2P 技术具有广阔的应用前景。Internt 上各种 P2P 应用软件层出不穷，用户数量急剧增加。2004 年 3 月来自 www.slyck.com 的数据显示，大量 P2P 软件的用户使用数量分布从几十万、几百万到上千万并且急剧增加，并给 Internet 带宽带来巨大冲击。P2P 计算技术正不断应用到军事领域，商业领域，政府信息，通讯等领域。器节点。而在 P2P 中，所有参与者都可以提供中继转发的功能，因而大大提高了匿名通讯的灵活性和可靠性，能够为用户提供更好的隐私保护。

根据具体应用不同，可以把 P2P 分为以下这些类型：
提供文件和其它内容共享的 P2P 网络，例如 Napster、Gnutella、eDonkey、emule、BitTorrent 等；
挖掘 P2P 对等计算能力和存储共享能力，例如 SETI@home、Avaki、Popular Power 等；
基于 P2P 方式的协同处理与服务共享平台，例如 JXTA、Magi、Groove、.NET My Service 等；
即时通讯交流，包括 ICQ、OICQ、Yahoo Messenger 等；
安全的 P2P 通讯与信息共享，例如 Skype、Crowds、Onion Routing 等。

P2P 网络中的拓扑结构研究

拓扑结构是指分布式系统中各个计算单元之间的物理或逻辑的互联关系，结点之间的拓扑结构一直是确定系统类型的重要依据。目前互联网络中广泛使用集中式、层次式等拓扑结构，Internet 本身是世界上最大的非集中式的互联网络，但是九十年代所建立的一些网络应用系统却是完全的集中式的系统,很多 Web 应用都是运行在集中式的服务器系统上。集中式拓扑结构系统目前面临着过量存储负载、Dos 攻击等一些难以解决的问题。
P2P 系统一般要构造一个非集中式的拓扑结构，在构造过程中需要解决系统中所包含的大量结点如何命名、组织以及确定结点的加入/离开方式、出错恢复等问题。
根据拓扑结构的关系可以将 P2P 研究分为 4 种形式：中心化拓扑（Centralized Topology）；全分布式非结构化拓扑（Decentralized Unstructured Topology）；全分布式结构化拓扑（Decentralized Structured Topology，也称作 DHT 网络）和半分布式拓扑（Partially Decentralized Topology）。
其中，

（一）中心化拓扑最大的优点是维护简单发现效率高。

由于资源的发现依赖中心化的目录系统，发现算法灵活高效并能够实现复杂查询。最大的问题与传统客户机/服务器结构类似，容易造成单点故障，访问的“热点”现象和法律等相关问题，这是第一代 P2P 网络采用的结构模式，经典案例就是著名的 MP3 共享软件 Napster。
Napster。
Napster 是最早出现的 P2P 系统之一，并在短期内迅速成长起来。Napster 实质上并非是纯粹的 P2P 系统，它通过一个中央服务器保存所有 Napster 用户上传的音乐文件索引和存放位置

的信息。当某个用户需要某个音乐文件时，首先连接到 Napster 服务器，在服务器进行检索，并由服务器返回存有该文件的用户信息；再由请求者直接连到文件的所有者传输文件。
Napster 首先实现了文件查询与文件传输的分离，有效地节省了中央服务器的带宽消耗，减少了系统的文件传输延时。这种方式最大的隐患在中央服务器上，如果该服务器失效，整个系统都会瘫痪。当用户数量增加到 105 或者更高时，Napster 的系统性能会大大下降。另一个问题在于安全性上，Napster 并没有提供有效的安全机制。
在 Napster 模型中，一群高性能的中央服务器保存着网络中所有活动对等计算机共享资源的目录信息。当需要查询某个文件时，对等机会向一台中央服务器发出文件查询请求。中央服务器进行相应的检索和查询后，会返回符合查询要求的对等机地址信息列表。查询发起对等机接收到应答后，会根据网络流量和延迟等信息进行选择，和合适的对等机建立连接，并开始文件传输。Napster 的工作原理如图 1 所示。
这种对等网络模型存在很多问题，主要表现为：
(1)中央服务器的瘫痪容易导致整个网络的崩馈，可靠性和安全性较低。
(2)随着网络规模的扩大，对中央索引服务器进行维护和更新的费用将急剧增加，所需成本过高。
(3)中央服务器的存在引起共享资源在版权问题上的纠纷，并因此被攻击为非纯粹意义上的 P2P 网络模型。对小型网络而言，集中目录式模型在管理和控制方面占一定优势。但鉴于其存在的种种缺陷，该模型并不适合大型网络应用。

（二）全分布非结构化网络

在重叠网络（overlay）采用了随机图的组织方式，结点度数服从"Power-law"[a][b]规律，从而能够较快发现目的结点，面对网络的动态变化体现了较好的容错能力，因此具有较好的可用性。同时可以支持复杂查询，如带有规则表达式的多关键词查询，模糊查询等，最典型的案例是 Gnutella。
Gnutella 是一个 P2P 文件共享系统，它和 Napster 最大的区别在于 Gnutella 是纯粹的 P2P 系统，没有索引服务器，它采用了基于完全随机图的洪泛（Flooding）发现和随机转发（Random Walker）机制。为了控制搜索消息的传输，通过 TTL (Time To Live)的减值来实现。具体协议参照［Gnutella 协议中文版］
在 Gnutella 分布式对等网络模型 N 中，每一个联网计算机在功能上都是对等的，既是客户机同时又是服务器，所以被称为对等机(Servent，Server+Client 的组合)。
随着联网节点的不断增多，网络规模不断扩大，通过这种洪泛方式定位对等点的方法将造成网络流量急剧增加，从而导致网络中部分低带宽节点因网络资源过载而失效。所以在初期的 Gnutella 网络中，存在比较严重的分区，断链现象。也就是说，一个查询访问只能在网络的很小一部分进行，因此网络的可扩展性不好。所以，解决 Gnutella 网络的可扩展性对该网络的进一步发展至关重要。
由于没有确定拓扑结构的支持，非结构化网络无法保证资源发现的效率。即使需要查找的目的结点存在发现也有可能失败。由于采用 TTL（Time-to-Live）、洪泛（Flooding）、随机漫步或有选择转发算法，因此直径不可控，可扩展性较差。
因此发现的准确性和可扩展性是非结构化网络面临的两个重要问题。目前对此类结构的研究主要集中于改进发现算法和复制策略以提高发现的准确率和性能。
由于非结构化网络将重叠网络认为是一个完全随机图，结点之间的链路没有遵循某些预先定义的拓扑来构建。这些系统一般不提供性能保证，但容错性好，支持复杂的查询，并受结点频繁加入和退出系统的影响小。但是查询的结果可能不完全，查询速度较慢，采用广播查询

的系统对网络带宽的消耗非常大，并由此带来可扩展性差等问题。
另外，由于非结构化系统中的随机搜索造成的不可扩展性，大量的研究集中在如何构造一个高度结构化的系统。目前研究的重点放在了如何有效地查找信息上，最新的成果都是基于 DHT 的分布式发现和路由算法。这些算法都避免了类似 Napster 的中央服务器，也不是像 Gnutella 那样基于广播进行查找，而是通过分布式散列函数，将输入的关键字惟一映射到某个结点上，然后通过某些路由算法同该结点建立连接。
最新的研究成果体现在采用分布式散列表（DHT）[a]的

（三）完全分布式结构化拓扑网络。

分布式散列表（DHT）实际上是一个由广域范围大量结点共同维护的巨大散列表。散列表被分割成不连续的块，每个结点被分配给一个属于自己的散列块，并成为这个散列块的管理者。DHT 的结点既是动态的结点数量也是巨大的，因此非中心化和原子自组织成为两个设计的重要目标。通过加密散列函数，一个对象的名字或关键词被映射为 128 位或 160 位的散列值。一个采用 DHT 的系统内所有结点被映射到一个空间 ，如果散列函数映射一个 位的名字到一个散列值 ，则有 。
分布式散列表起源于 SDDS（Scalable Distribute Data Structures）[a]研究，Gribble 等实现了一个高度可扩展，容错的 SDDS 集群。
最近的研究集中在采用新的拓扑图构建重叠路由网络，以减少路由表容量和路由延时。这些新的拓扑关系的基本原理是在 DHT 表一维空间的基础上引入更多的拓扑结构图来反映底层网络的结构。
DHT 类结构能够自适应结点的动态加入/退出，有着良好的可扩展性、鲁棒性、结点 ID 分配的均匀性和自组织能力。由于重叠网络采用了确定性拓扑结构，DHT 可以提供精确的发现。只要目的结点存在于网络中 DHT 总能发现它，发现的准确性得到了保证，最经典的案例是 Tapestry，Chord,CAN,和 Pastry。
Tapestry 提供了一个分布式容错查找和路由基础平台，在此平台基础之上，可以开发各种 P2P 应用(OceanStore 即是此平台上的一个应用) 。Tapestry 的思想来源于 Plaxton。在 Plaxton 中，结点使用自己所知道的邻近结点表，按照目的 ID 来逐步传递消息。Tapestry 基于 Plaxtion 的思想，加入了容错机制，从而可适应 P2P 的动态变化的特点。OceanStore 是以 Tapestry 为路由和查找基础设施的 P2P 平台。它是一个适合于全球数据存储的 P2P 应用系统。任何用户均可以加入 OceanStore 系统，或者共享自己的存储空间，或者使用该系统中的资源。通过使用复制和缓存技术，OceanStore 可提高查找的效率。最近，Tapstry 为适应 P2P 网络的动态特性，作了很多改进，增加了额外的机制实现了网络的软状态（soft state），并提供了自组织、鲁棒性、可扩展性和动态适应性，当网络高负载且有失效结点时候性能有限降低，消除了对全局信息的依赖、根结点易失效和弹性（resilience）差的问题。
Pastry 是微软研究院提出的可扩展的分布式对象定位和路由协议，可用于构建大规模的 P2P 系统。在 Pastry 中，每个结点分配一个 128 位的结点标识符号(nodeID) ，所有的结点标识符形成了一个环形的 nodeID 空间，范围从 0 到 2128 - 1 ，结点加入系统时通过散列结点 IP 地址在 128 位 nodeID 空间中随机分配。
在 MIT，开展了多个与 P2P 相关的研究项目：Chord，GRID 和 RON。Chord 项目的目标是提供一个适合于 P2P 环境的分布式资源发现服务，它通过使用 DHT 技术使得发现指定对象只需要维护 O(logN)长度的路由表。
在 DHT 技术中，网络结点按照一定的方式分配一个唯一结点标识符(Node ID) ，资源对象通过散列运算产生一个唯一的资源标识符(Object ID) ，且该资源将存储在结点 ID 与之相等

或者相近的结点上。需要查找该资源时，采用同样的方法可定位到存储该资源的结点。因此，Chord 的主要贡献是提出了一个分布式查找协议，该协议可将指定的关键字(Key) 映射到对应的结点(Node) 。从算法来看，Chord 是相容散列算法的变体。MIT 的 GRID 和 RON 项目则提出了在分布式广域网中实施查找资源的系统框架。

AT&T ACIRI 中心的 CAN(Content Addressable Networks) 项目独特之处在于采用多维的标识符空间来实现分布式散列算法。CAN 将所有结点映射到一个 n 维的笛卡尔空间中，并为每个结点尽可能均匀的分配一块区域。CAN 采用的散列函数通过对(key, value) 对中的 key 进行散列运算，得到笛卡尔空间中的一个点，并将(key, value) 对存储在拥有该点所在区域的结点内。CAN 采用的路由算法相当直接和简单，知道目标点的坐标后，就将请求传给当前结点四邻中坐标最接近目标点的结点。CAN 是一个具有良好可扩展性的系统，给定 N 个结点，系统维数为 d，则路由路径长度为 O(n1/d) ，每结点维护的路由表信息和网络规模无关为 O(d) 。

DHT 类结构最大的问题是 DHT 的维护机制较为复杂，尤其是结点频繁加入退出造成的网络波动（Churn）会极大增加 DHT 的维护代价。DHT 所面临的另外一个问题是 DHT 仅支持精确关键词匹配查询，无法支持内容/语义等复杂查询。

（四）半分布式结构（有的文献称作 Hybrid Structure）

吸取了中心化结构和全分布式非结构化拓扑的优点，选择性能较高（处理、存储、带宽等方面性能）的结点作为超级点（英文文献中多称作：SuperNodes, Hubs)，在各个超级点上存储了系统中其他部分结点的信息，发现算法仅在超级点之间转发，超级点再将查询请求转发给适当的叶子结点。半分布式结构也是一个层次式结构，超级点之间构成一个高速转发层，超级点和所负责的普通结点构成若干层次。最典型的案例就是 KaZaa。

KaZaa 是现在全世界流行的几款 p2p 软件之一。根据 CA 公司统计，全球 KaZaa 的下载量超过 2.5 亿次。使用 KaZaa 软件进行文件传输消耗了互联网 40%的带宽。之所以它如此的成功，是因为它结合了 Napster 和 Gnutella 共同的优点。从结构 上来说，它使用了 Gnutella 的全分布式的结构，这样可以是系统更好的扩展，因为它无需中央索引服务器存储文件名，它是自动的把性能好的机器成为 SuperNode，它存储着离它最近的叶子节点的文件信息，这些 SuperNode,再连通起来形成一个 Overlay Network. 由于 SuperNode 的索引功能，使搜索效率大大提高。

半分布式结构的优点是性能、可扩展性较好，较容易管理，但对超级点依赖性大，易于受到攻击，容错性也受到影响。下表比较了 4 种结构的综合性能，比较结果如表 1-1 所示。

表 1：4 种结构的性能比较

比较标准 / 拓扑结构	中心化拓扑	全分布式非结构化拓扑	全分布式结构化拓扑	半分布式拓扑
可扩展性	差	差	好	中
可靠性	差	好	好	中
可维护性	最好	最好	好	中
发现算法效率	最高	中	高	中
复杂查询	支持	支持	不支持	支持

国内的 P2P 研究现状

学术机构研发

北京大学—Maze

Maze 是北京大学网络实验室开发的一个中心控制与对等连接相融合的对等计算文件共享系统，在结构上类似 Napster，对等计算搜索方法类似于 Gnutella。网络上的一台计算机，不论是在内网还是外网，可以通过安装运行 Maze 的客户端软件自由加入和退出 Maze 系统。每个节点可以将自己的一个或多个目录下的文件共享给系统的其他成员，也可以分享其他成员的资源。Maze 支持基于关键字的资源检索，也可以通过好友关系直接获得。

清华大学—Granary

Granary 是清华大学自主开发的对等计算存储服务系统。它以对象格式存储数据。另外，Granary 设计了专门的结点信息收集算法 PeerWindow 的结构化覆盖网络路由协议 Tourist。

华中科技大学—AnySee

AnySee 是华中科大设计研发的视频直播系统。它采用了一对多的服务模式，支持部分 NAT 和防火墙的穿越，提高了视频直播系统的可扩展性；同时，它利用近播原则、分域调度的思想，使用 Landmark 路标算法直接建树的方式构建应用层上的组播树，克服了 ESM 等一对多模式系统由联接图的构造和维护带来的负载影响。

更详细介绍见［中国计算机学会通讯　Page 38-51　郑纬民等　对等计算研究概论］

企业研发产品

广州数联软件技术有限公司-Poco

POCO 是中国最大的 P2P 用户分享平台 ，是有安全、流量控制力的，无中心服务器的第三代 P2P 资源交换平台 ，也是世界范围内少有的盈利的 P2P 平台。目前已经形成了 2600 万海量用户，平均在线 58.5 万，在线峰值突破 71 万，并且全部是宽带用户的用户群。 成为中国地区第一的 P2P 分享平台。[a]

深圳市点石软件有限公司-OP

OP-又称为 Openext Media Desktop，一个网络娱乐内容平台，Napster 的后继者，它可以最直接的方式找到您想要的音乐、影视、软件、游戏、图片、书籍以及各种文档，随时在线共享文件容量数以亿计“十万影视、百万音乐、千万图片”。OP 整合了 Internet Explorer、Windows Media Player、RealOne Player 和 ACDSee ，是国内的网络娱乐内容平台。[a]

基于 P2P 的在线电视直播-PPLive

PPLive 是一款用于互联网上大规模视频直播的共享软件。它使用网状模型，有效解决了当前网络视频点播服务的带宽和负载有限问题，实现用户越多，播放越流畅的特性，整 体服务质量大大提高！（2005 年的超级女声决赛期间，这款软件非常的火爆，同时通过它看湖南卫视的有上万观众）

其他商业软件这里不一一列举，请访问 P2P 门户网站 http://www.ppcn.net/

GGF）。P2P 工作组成立的主要目的是希望加速 P2P 计算基础设施的建立和相应的标准化工作。P2PWG 成立之后，对 P2P 计算中的术语进行了统一，也形成相关的草案，但是在标准化工作方面工作进展缓慢。目前 P2PWG 已经和 GGF 合并，由该论坛管理 P2P 计算相关的

工作。GGF 负责网格计算和 P2P 计算等相关的标准化工作。
从国外公司对 P2P 计算的支持力度来看，Microsoft 公司、Sun 公司和 Intel 公司投入较大。Microsoft 公司成立了 Pastry 项目组，主要负责 P2P 计算技术的研究和开发工作。目前 Microsoft 公司已经发布了基于 Pastry 的软件包 SimPastry/ VisPastry。Rice 大学也在 Pastry 的基础之上发布了 FreePastry 软件包。
在 2000 年 8 月，Intel 公司宣布成立 P2P 工作组，正式开展 P2P 的研究。工作组成立以后，积极与应用开发商合作，开发 P2P 应用平台。2002 年 Intel 发布了. Net 基础架构之上的 Accelerator Kit (P2P 加速工具包) 和 P2P 安全 API 软件包，从而使得微软. NET 开发人员能够迅速地建立 P2P 安全 Web 应用程序。
Sun 公司以 Java 技术为背景，开展了 JXTA 项目。JXTA 是基于 Java 的开源 P2P 平台，任何个人和组织均可以加入该项目。因此，该项目不仅吸引了大批 P2P 研究人员和开发人员，而且已经发布了基于 JXTA 的即时聊天软件包。JXTA 定义了一组核心业务：认证、资源发现和管理。在安全方面，JXTA 加入了加密软件包，允许使用该加密包进行数据加密，从而保证消息的隐私、可认证性和完整性。在 JXTA 核心之上，还定义了包括内容管理、信息搜索以及服务管理在内的各种其它可选 JXTA 服务。在核心服务和可选服务基础上，用户可以开发各种 JXTA 平台上的 P2P 应用。

P2P 实际的应用主要体现在以下几个方面

P2P 分布式存储

P2P 分布式存储系统是一个用于对等网络的数据存储系统，它可以提供高效率的、鲁棒的和负载平衡的文件存取功能。这些研究包括：OceanStore，Farsite 等。其中，基于超级点结构的半分布式 P2P 应用如 Kazza、Edonkey、Morpheus、Bittorrent 等也是属于分布式存储的范畴，并且用户数量急剧增加。

计算能力的共享

加入对等网络的结点除了可以共享存储能力之外，还可以共享 CPU 处理能力。目前已经有了一些基于对等网络的计算能力共享系统。比如 SETI@home。目前 SETI@home 采用的仍然是类似于 Napster 的集中式目录策略。Xenoservers 向真正的对等应用又迈进了一步。这种计算能力共享系统可以用于进行基因数据库检索和密码破解等需要大规模计算能力的应用。

P2P 应用层组播

应用层组播，就是在应用层实现组播功能而不需要网络层的支持。这样就可以避免出现由于网络层迟迟不能部署对组播的支持而使组播应用难以进行的情况。应用层组播需要在参加的应用结点之间实现一个可扩展的，支持容错能力的重叠网络，而基于 DHT 的发现机制正好为应用层组播的实现提供了良好的基础平台。

Internet 间接访问基础结构（Internet Indirection Infrastructure）

为了使 Internet 更好地支持组播、单播和移动等特性，Internet 间接访问基础结构提出了基于汇聚点的通信抽象。在这一结构中，并不把分组直接发向目的结点，而是给每个分组分配一个标识符，而目的结点则根据标识符接收相应的分组。标识符实际上表示的是信息的汇聚点。目的结点把自己想接收的分组的标识符预先通过一个触发器告诉汇聚点，当汇聚点收到分组时，将会根据触发器把分组转发该相应的目的结点。Internet 间接访问基础结构实际上在 Internet 上构成了一个重叠网络，它需要对等网络的路由系统对它提供相应的支持。

P2P 技术从出现到各个领域的应用展开，仅用了几年的时间。从而证明了 P2P 技术具有非常广阔的应用前景。