Ch4. P2P原理与技术

4.1 P2P网络基本概念

4.2 混合式P2P网络（第一代）

4.3 无结构P2P网络（第二代）

4.4 结构化P2P网络（第三代）

4.5 P2P网络的问题与研究

4.1 P2P网络基本概念

4.1.1 What is P2P ? (Peer-to-Peer)

- 对等(网络，计算)...;端到端...
- 经系统间直接交换来共享计算资源和服务的应用模式
- 以非集中方式使用分布式资源来完成关键任务的一类系统和应用
  - 资源包括计算、存储、带宽、场景（计算机、人和现场）和信息等资源
  - 关键任务可能是分布式计算、数据/内容共享，通信和协同、或平台服务

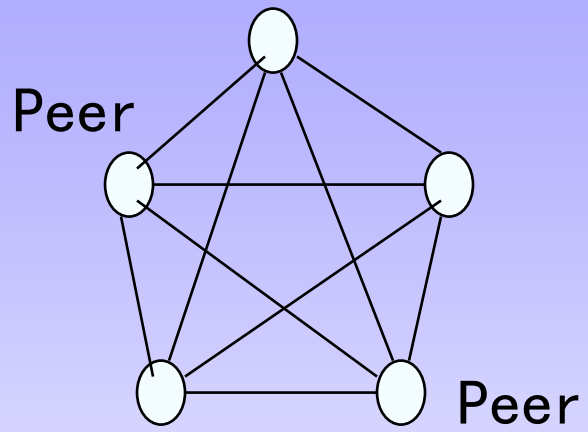
◆ 典型位置： 因特网边界或ad-hoc网内

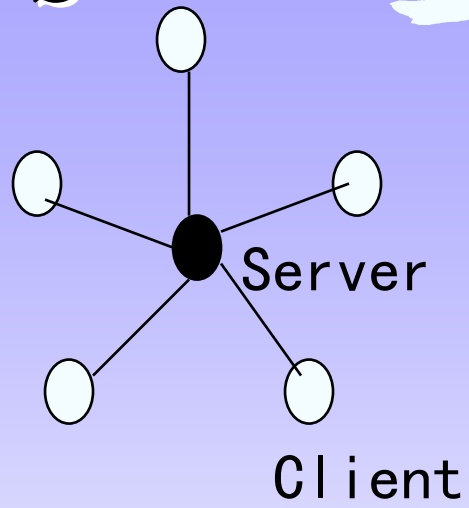
典型定义

- Intel 工作组：通过在系统之间**直接交换**来**共享**计算机**资源**和服务的一种应用模式
- A.Weytsel:在因特网周边以非客户地位使用的设备
- R.I.Granham: 通过3个关键条件定义
  - 具有服务器质量的可运行计算机
  - 具有独立于DNS的**寻址**系统
  - 具有与可变连接合作的能力
- C.Shirky:利用因特网边界的存储/CPU/内容/现场等资源的一种应用。访问这些非集中资源意味着运行在不稳定连接和不可预知IP地址环境下，P2P节点必须运行在DNS系统外边，对中心服务器来说具备有效的或全部的**自治**

- Kindberg:独立生存的的系统
- D.J.Milojicic:给对等组提供或从对等组获得共享
  - 对等端向组给出某些资源，并从组获得某些资源
  - Napster:把音乐供给组内其他人，并从其他人获得音乐
  - 捐赠计算资源用于外星生命的搜索或战胜癌症，获得帮助其他人的满足
- 另一种应用模式选择：
  - 相对集中式、和C/S模式
  - 纯P2P：没有服务器的概念，所有成员都是对等端
- 并不是新的概念
  - 早期分布式系统：如UUCP和交换网络
  - 电话通信
  - 计算机网络中的通信、网络游戏中的诸玩家

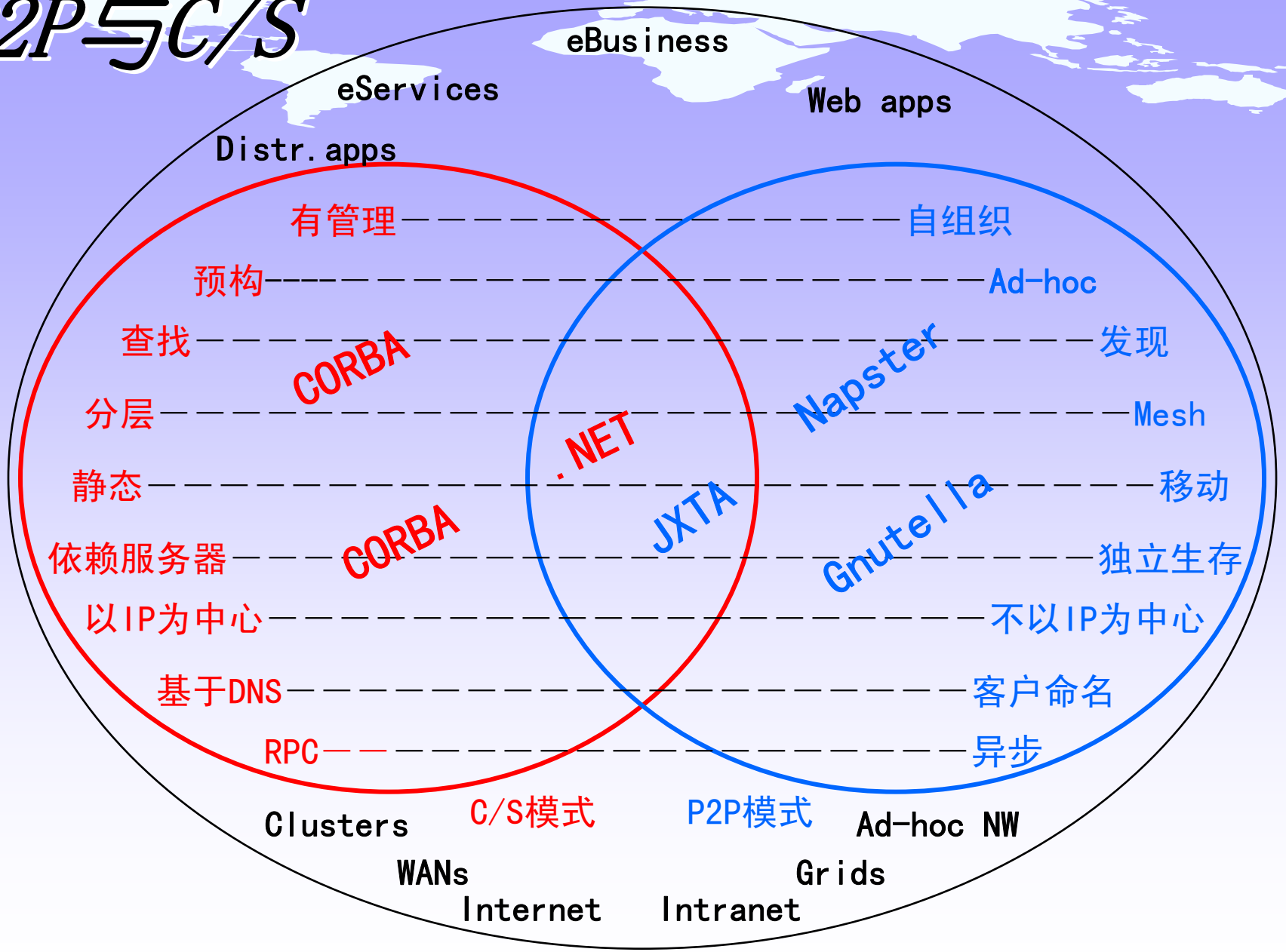
P2P 与 C/S



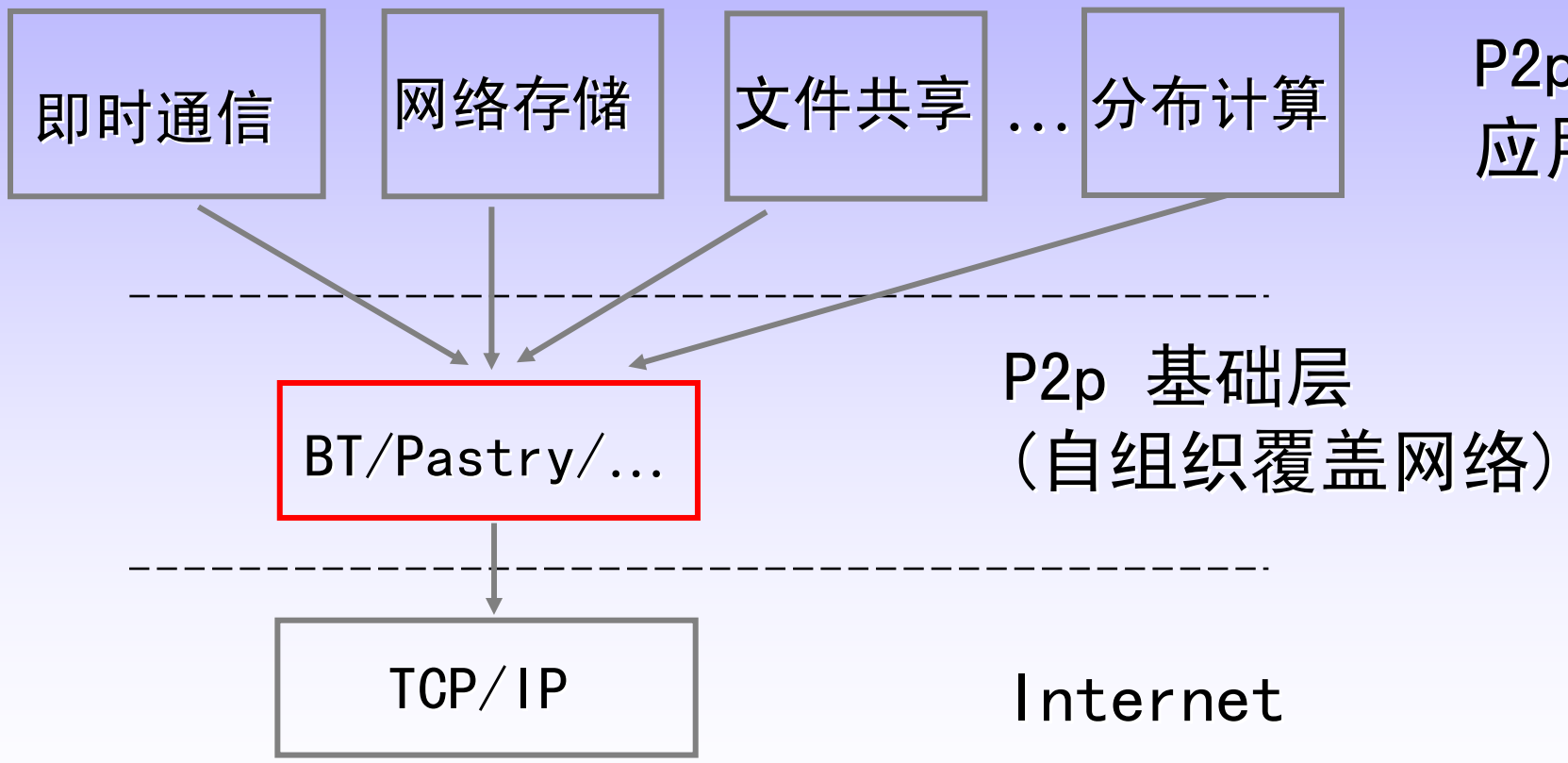


- 二者在结构和构成上有很大的区别
  - 管理能力、构态能力、功能（查找或发现）、组织(分层与网孔)、元素（DNS）和协议(IP)
  - C/S通常是简单的端到端通信，P2P通常要构成自己的应用层网络
- 但又无明显的边界
  - 都能运行在不同的（Internet / Intranet）平台上
  - 都能服务传统或新的应用：eBusiness eServuices ...

P2P与C/S


eBusiness
eServices
Web apps
Distr.apps
有管理
自组织
预构
Ad-hoc
查找
发现
CORBA
Napster
分层
Mesh
.NET
静态
移动
JXTA
CORBA
Gnutella
依赖服务器
独立生存
以IP为中心
不以IP为中心
基于DNS
客户命名
RPC
异步
Clusters
C/S模式
P2P模式
Ad-hoc NW
WANs
Grids
Internet
Intranet

P2P的定位


即时通信
网络存储
文件共享
...
分布计算
P2p
应用层
BT/Pastry/...
P2p 基础层
(自组织覆盖网络)
TCP/IP
Internet

C/S 模式的挑战

单服务器或搜索引擎已不能满足或覆盖日益增长的Web内容需求

- 2×10^{18} Bytes/year Internet上增长.
- 但仅 3×10^{12} Bytes/year 可被公众利用 (0.00015%).
- Google 仅搜索 1.3×10^{8} Web pages.

(Source: IEEE *Internet Computing, 2001*)

C/S

C/S 模式严重限制**可用带宽和服务**的利用

- 流行的服务器和搜索引擎已成为**流量瓶颈**
- 但许多高速网络连接的客户端却**很空闲**
- 客户端的计算能力与信息被忽视

Content Delivery Networks模式

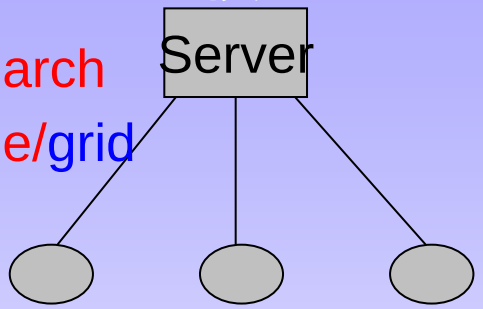
- 服务器在因特网上分散部署（内容重复）
- 分布部署的服务器由总部中心授权控制
- Examples: Internet content distributions by Akamai, Overcast, and FFnet.
- C/S和CDN 模式都有单点失效问题

面向Peer的系统

- 既是客户端consumer也是服务器端producer=Prosumer
- 任何时候都有加入或离开的自由
- 无限的peer diversity：服务能力、存储空间 、网络带宽和服务需求
- 挑战与机遇：开放的广域无中心分布系统

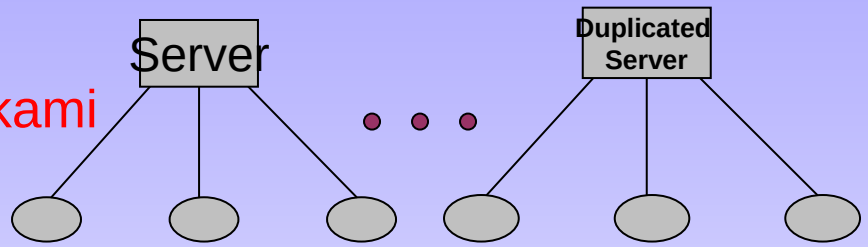
C/S模式

a search engine/grid

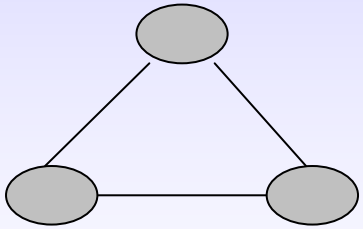


Content Delivery Networks

e.g. Akami



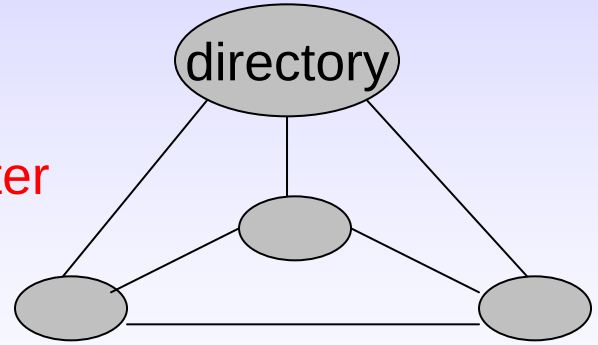
Pure P2P



e.g. Freenet & Gnutella

Hybrid P2P

e.g. Napster



P2P 的目标与优势

- 只要不存在网络的物理断开，目标文件总是可以找到！
- 信息可扩展：往P2P系统加入更多内容将不影响其性能！
- 系统可扩展：加入或离开，将不影响P2P 系统的性能！

P2P 的应用

- File Sharing: document sharing among peers with no or limited central controls.
- Instant Messaging (IM): Immediate voice and file exchanges among peers.
- Distributed Processing: One can widely utilize resources available in other remote peers.

P2P 的目标与吸引力

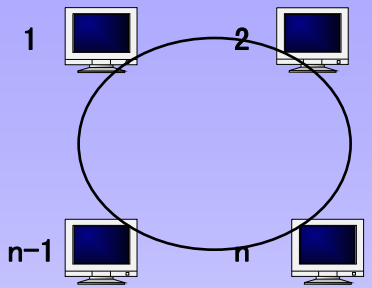
- P2P是一类发挥互联网边缘资源（存储、处理能力、内容、带宽、用户现场）可用性的应用
- 每个参与者（进程）既是客户端也是服务器：
  - 你的PDA可以存放部分音乐目录
  - 你的PC可以存放部分音乐仓库

- 简化地依赖个人设备和子网（去中心服务器）
- 非脆弱的健壮性（无单点故障）
- 柔软性/快速恢复（内建冗余）
- 抗DOS攻击（无中心服务器）
- 更高的可扩展性
- 改进的高峰请求服务（提出需求的设备越多，意味着服务器资源也越多）

P2P的效果

- 巨大的扩展力
  - 通过低成本交互来聚合资源，导致整体大于部分之和
- 低成本的所有权和共享
  - 通过使用现存的基础设施、削减和分布成本达到
- 匿名和隐私：
  - 通过允许对等端在其数据和资源上很大的自治控制达到

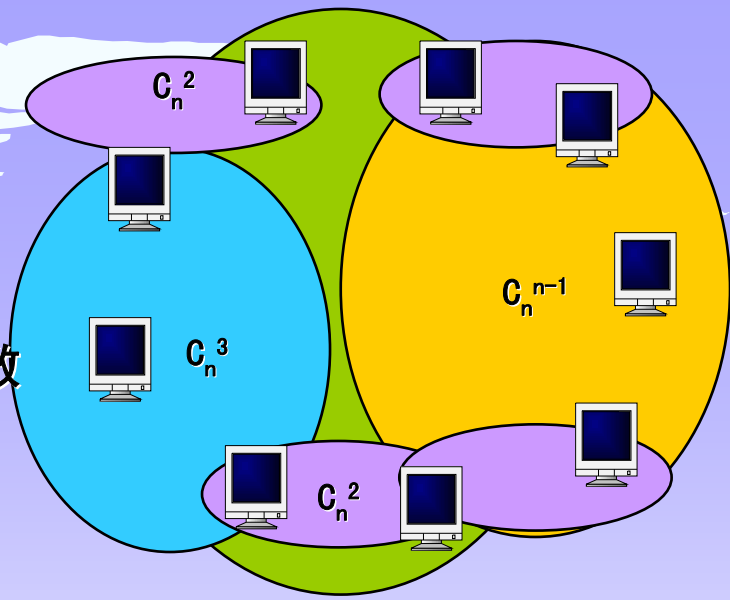
独立子网数四定律



A:Sarnoff 'law :规模是O(1)

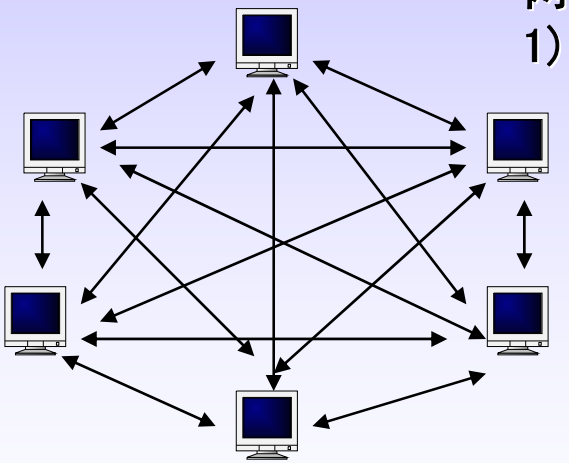
➢Sarnoff 'law：广播网络组数是O(1) = =1

➢Reed 'law：子网络个数 $O(2^n) = Cn2+...Cn^{n-1}+Cn^n = 2^n-n-1$



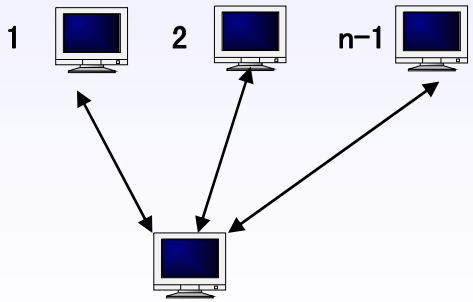
D: Reed 'law：规模是$O(2^n)$

➢Metcalfe 'law：子网络数 $O(n^2) = n(n-1)/2$

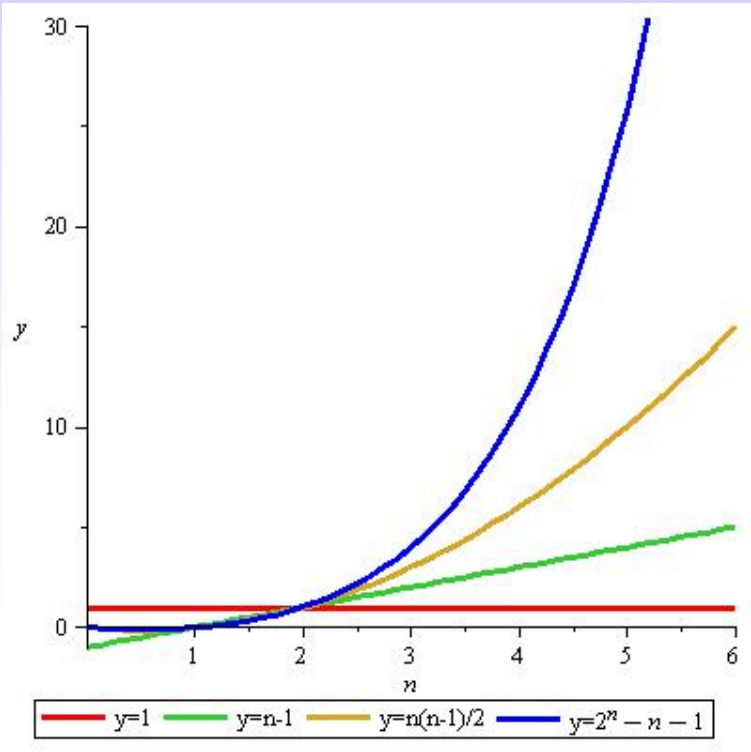


C: Metcalfe 'law :规模是$O(n^2)$

➢Leeying'law：C/S的子网络数 O(n) = n-1

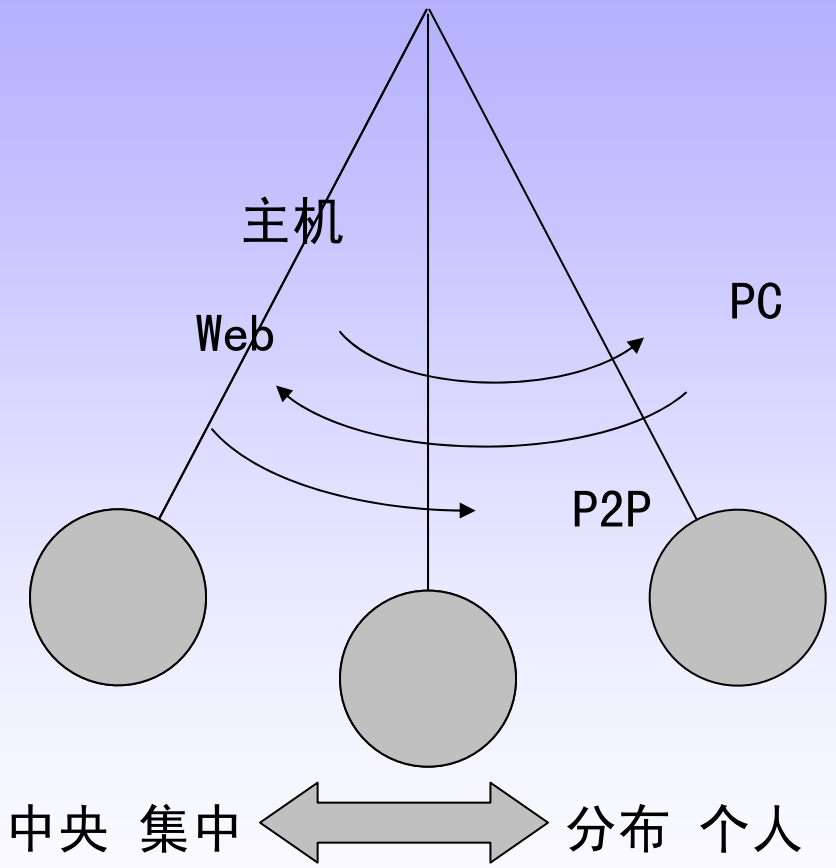


B:leeying 'law :规模是O(n)



2010-10-15

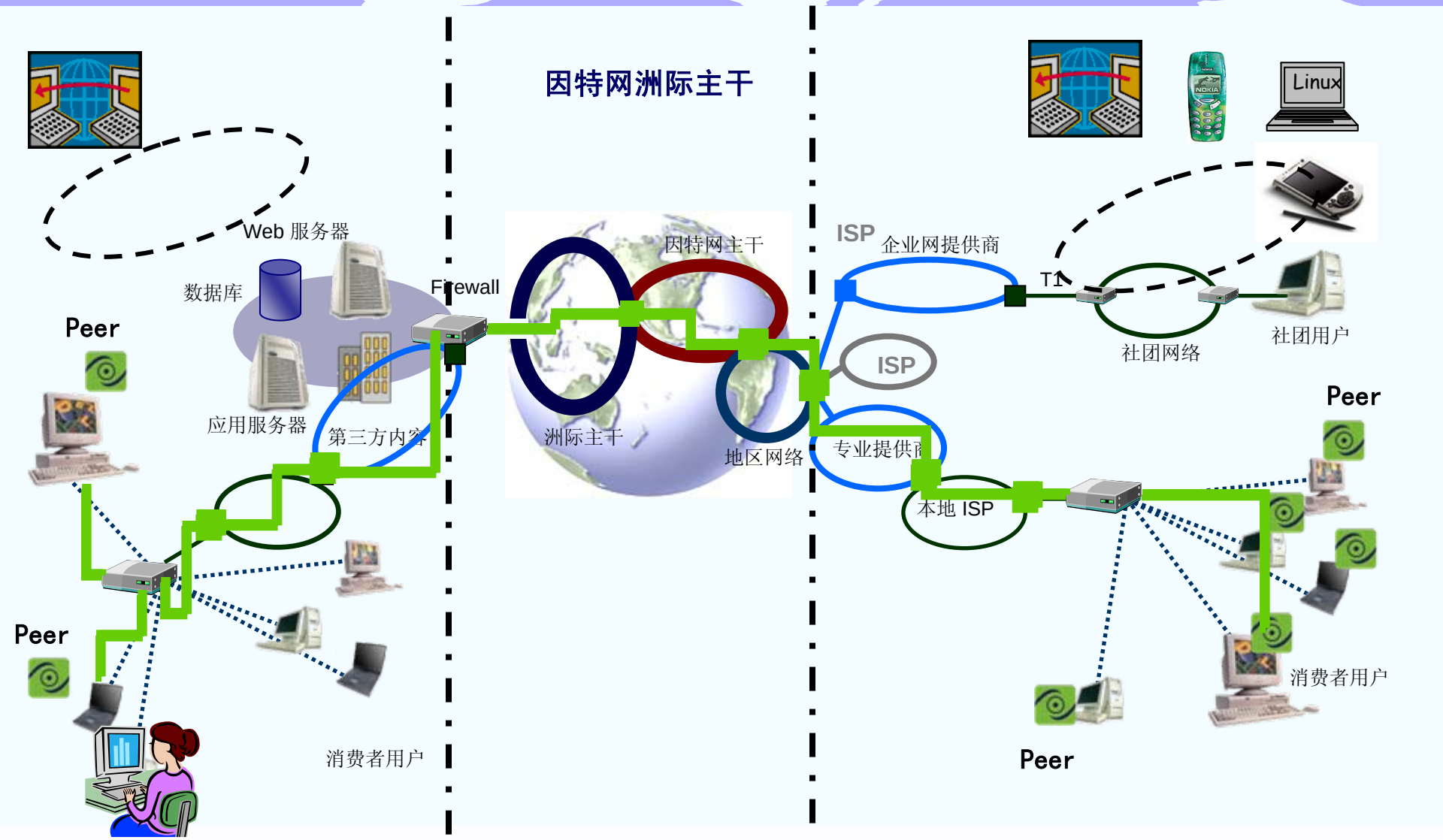
P2P-从集中向分布的演化



- 将每一个 PC 变成服务器
- 适合自组织 ad-hoc组工作
- 推动采用 IPv6,用户直接连接网络
- IPv6提供无服务器DNS
- 开发者的平台

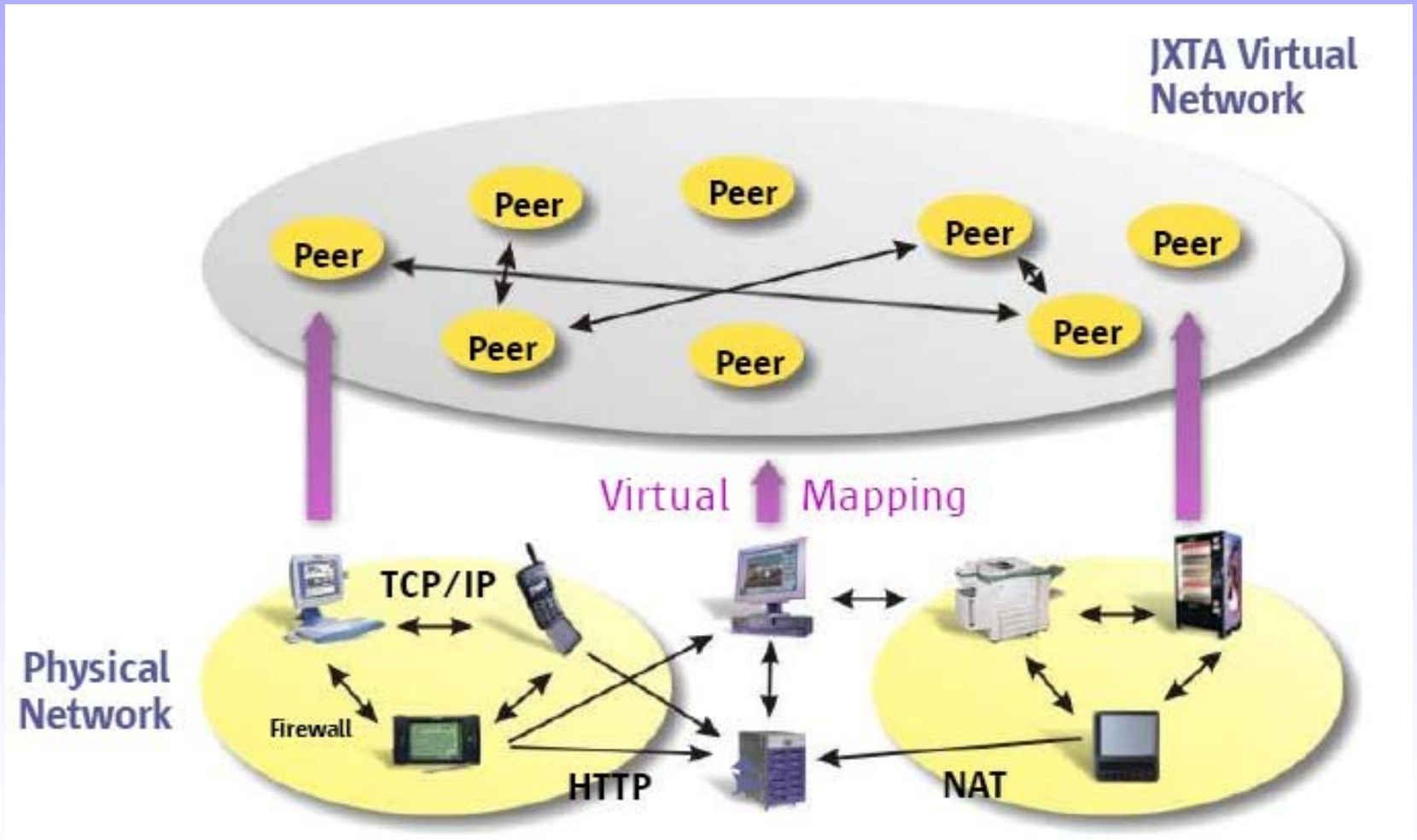
充分发挥互联网无所不在的优势

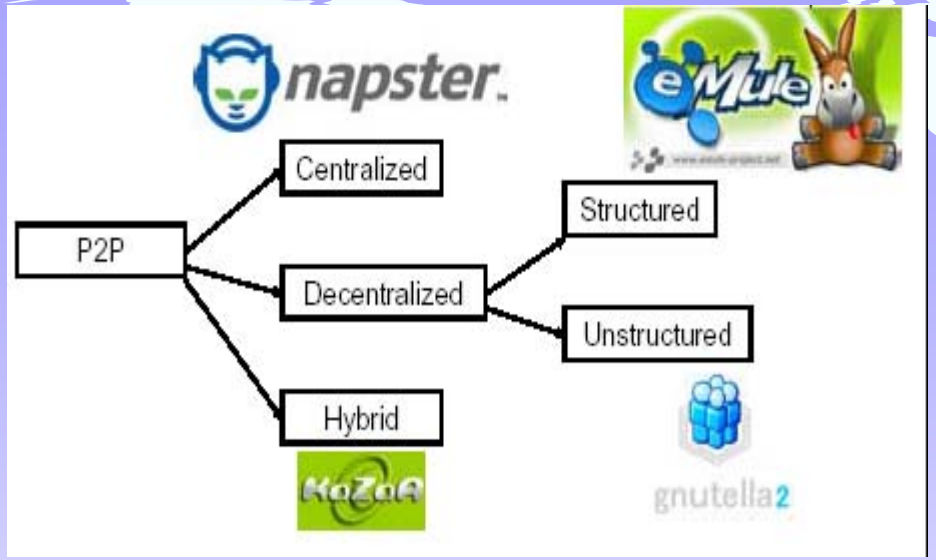
Peer to Peer 过程

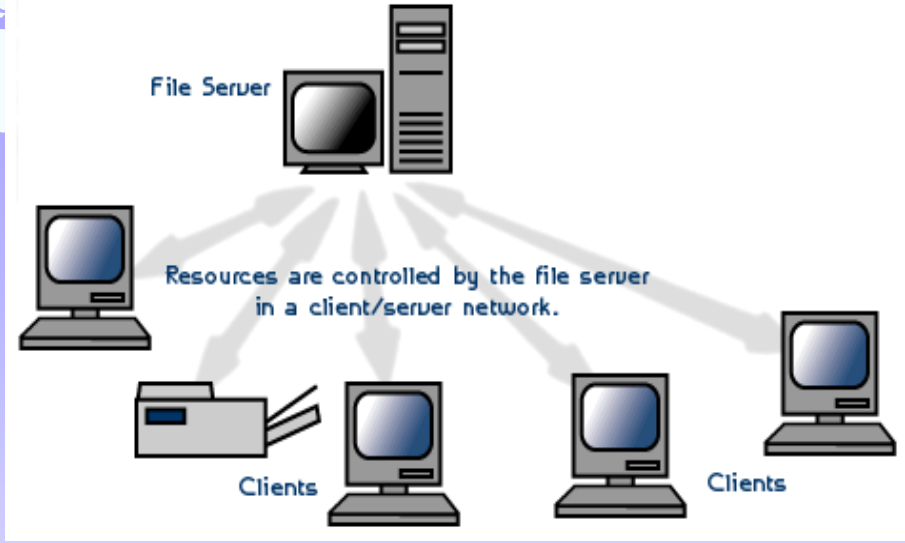


Slide source: www.porivo.com

应用层重叠网络


JXTA Virtual Network
Peer
Peer
Peer
Peer
Peer
Peer
Peer
Peer
Virtual
Mapping
Physical Network
TCP/IP
Firewall
HTTP
NAT





即时通讯软件

skype

流媒体

PPLive

PPStream

下载软件

Maze 天网 Tianwang

KaZaA

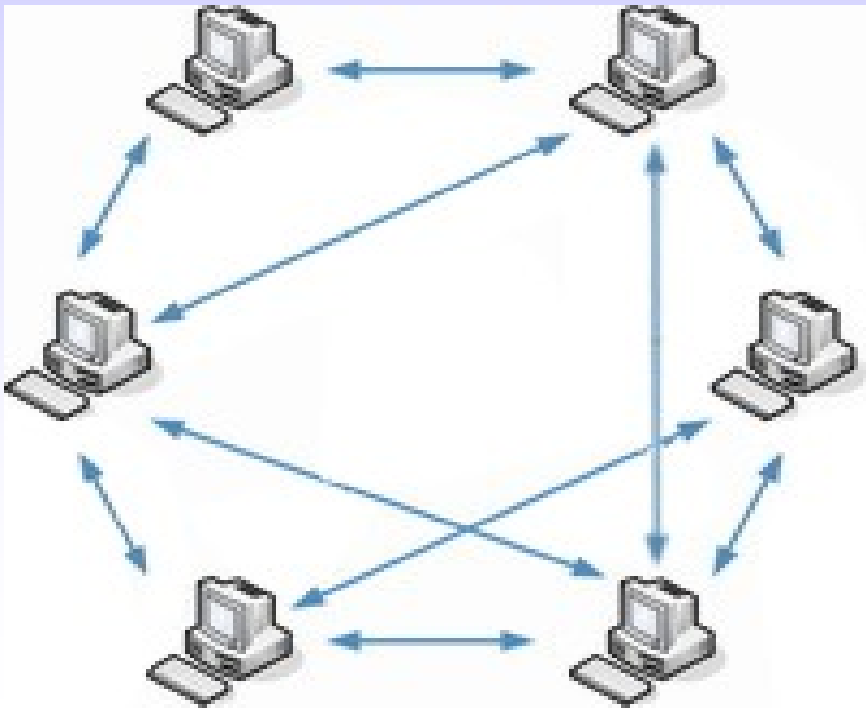
匿名访问

Tor

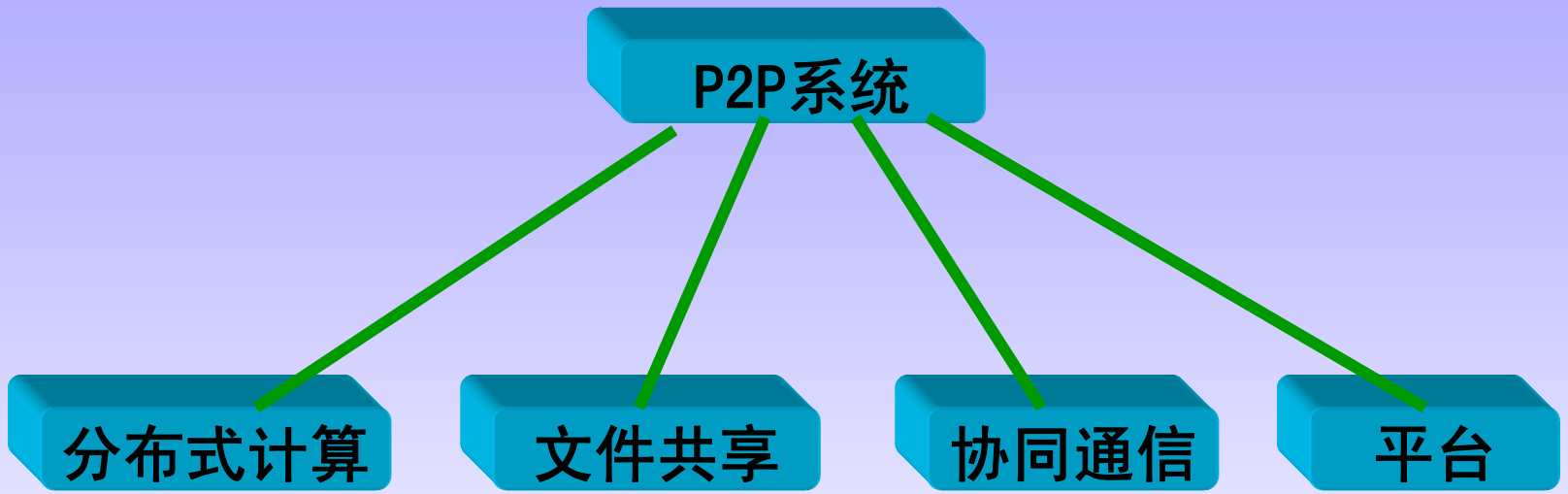
科学研究

SETI HOME

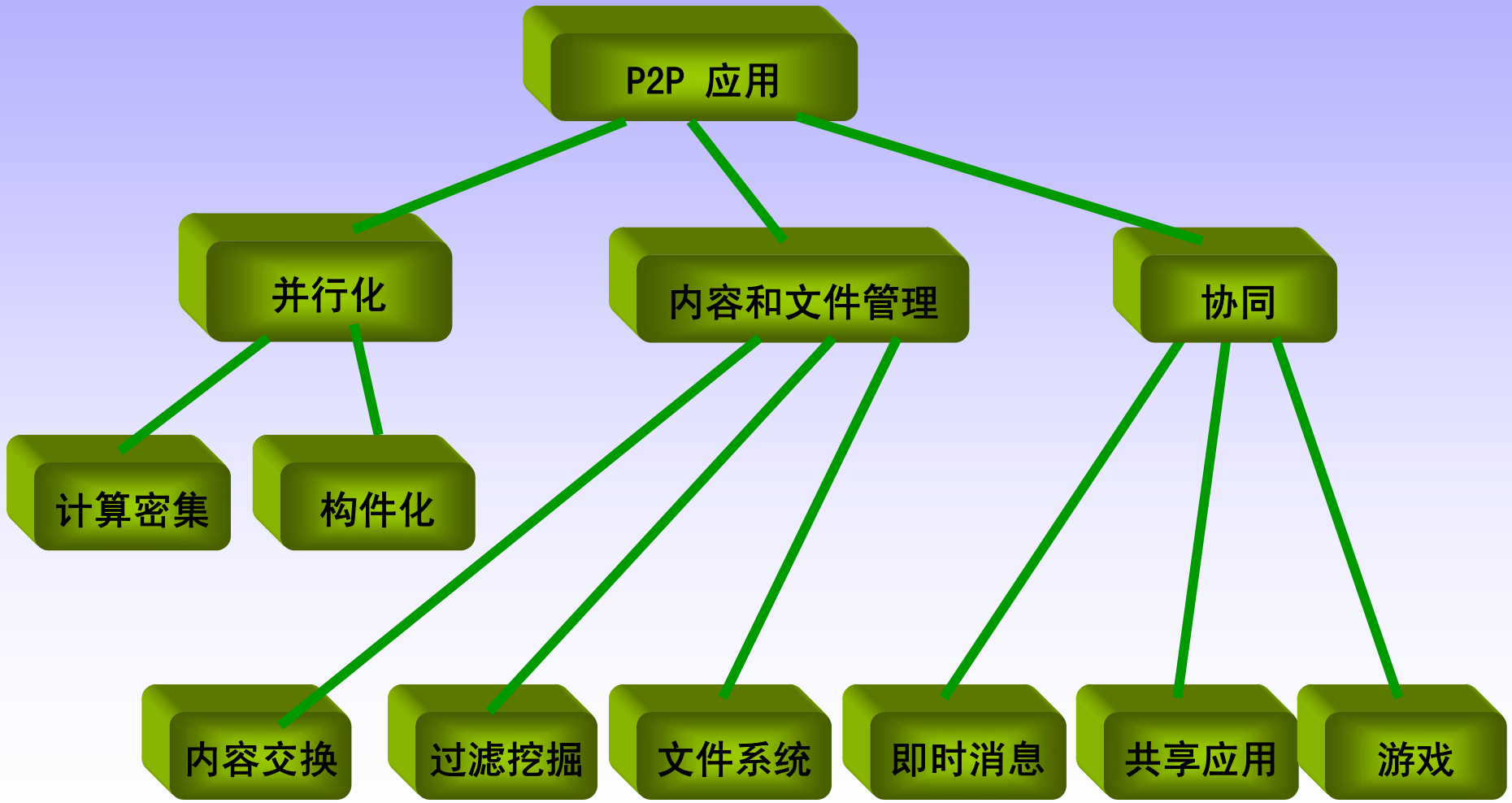
P2P游戏等



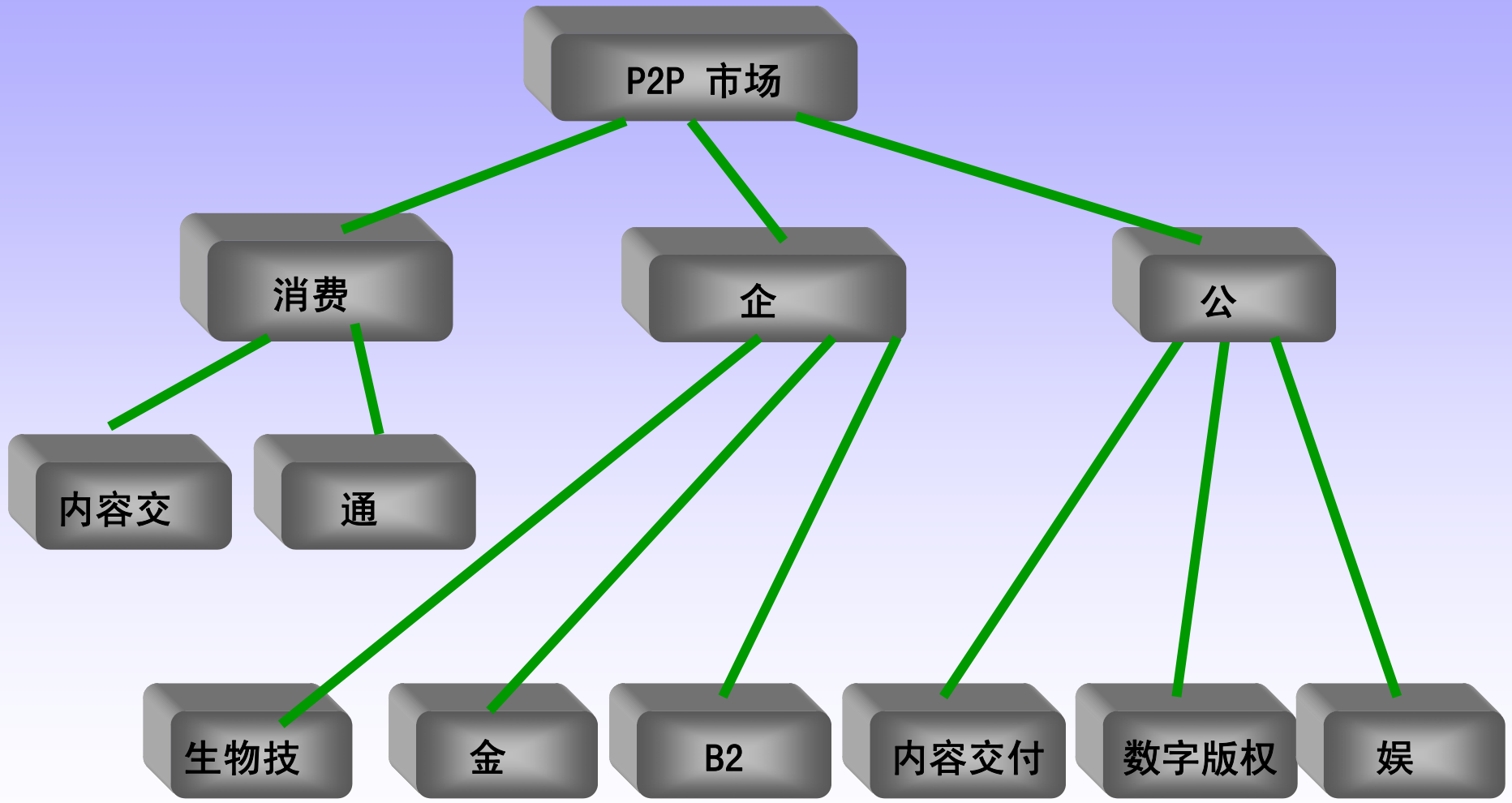
4.1.2 P2P网络的应用



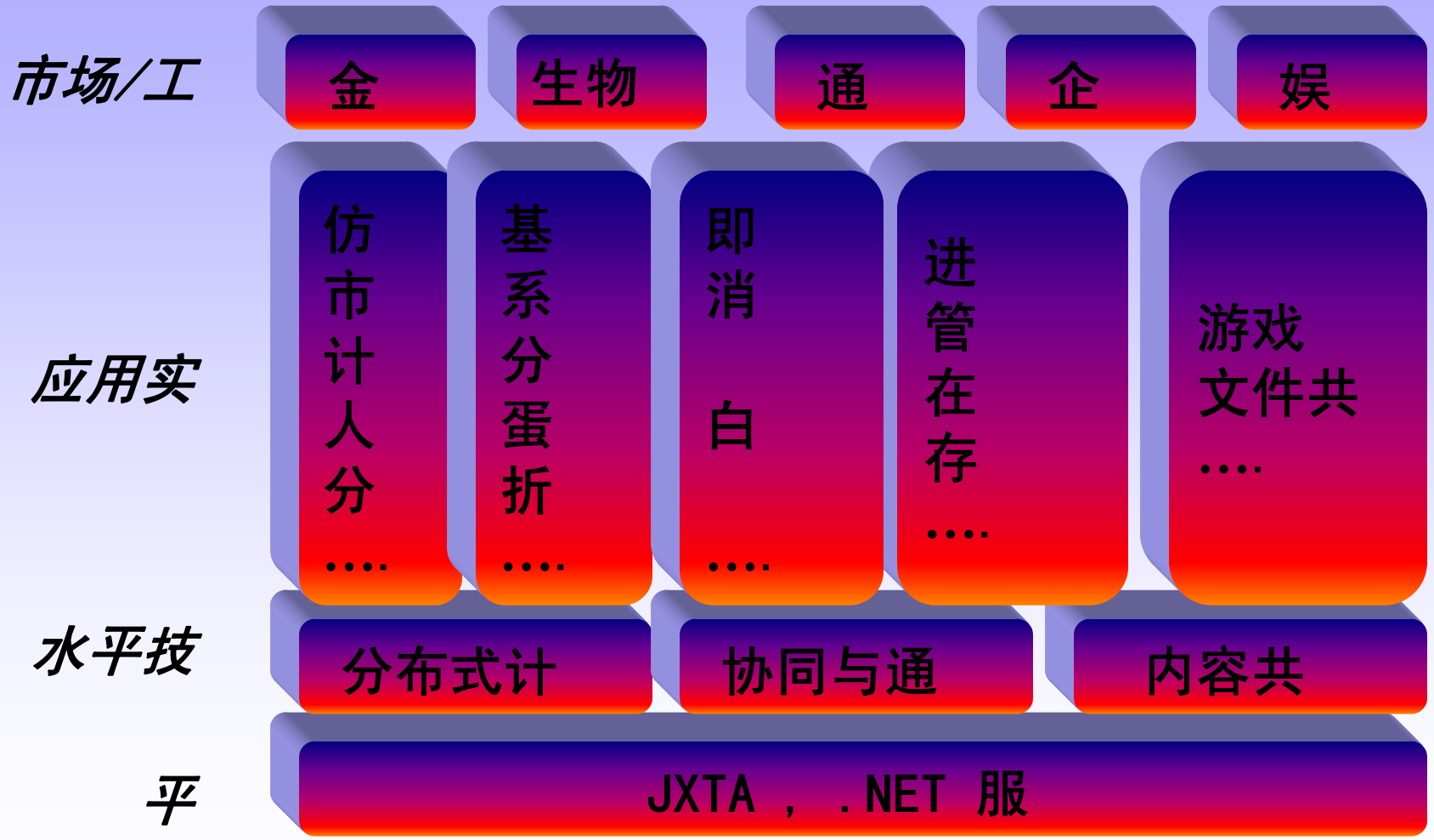
P2P 应用分类



P2P 应用的市场分类



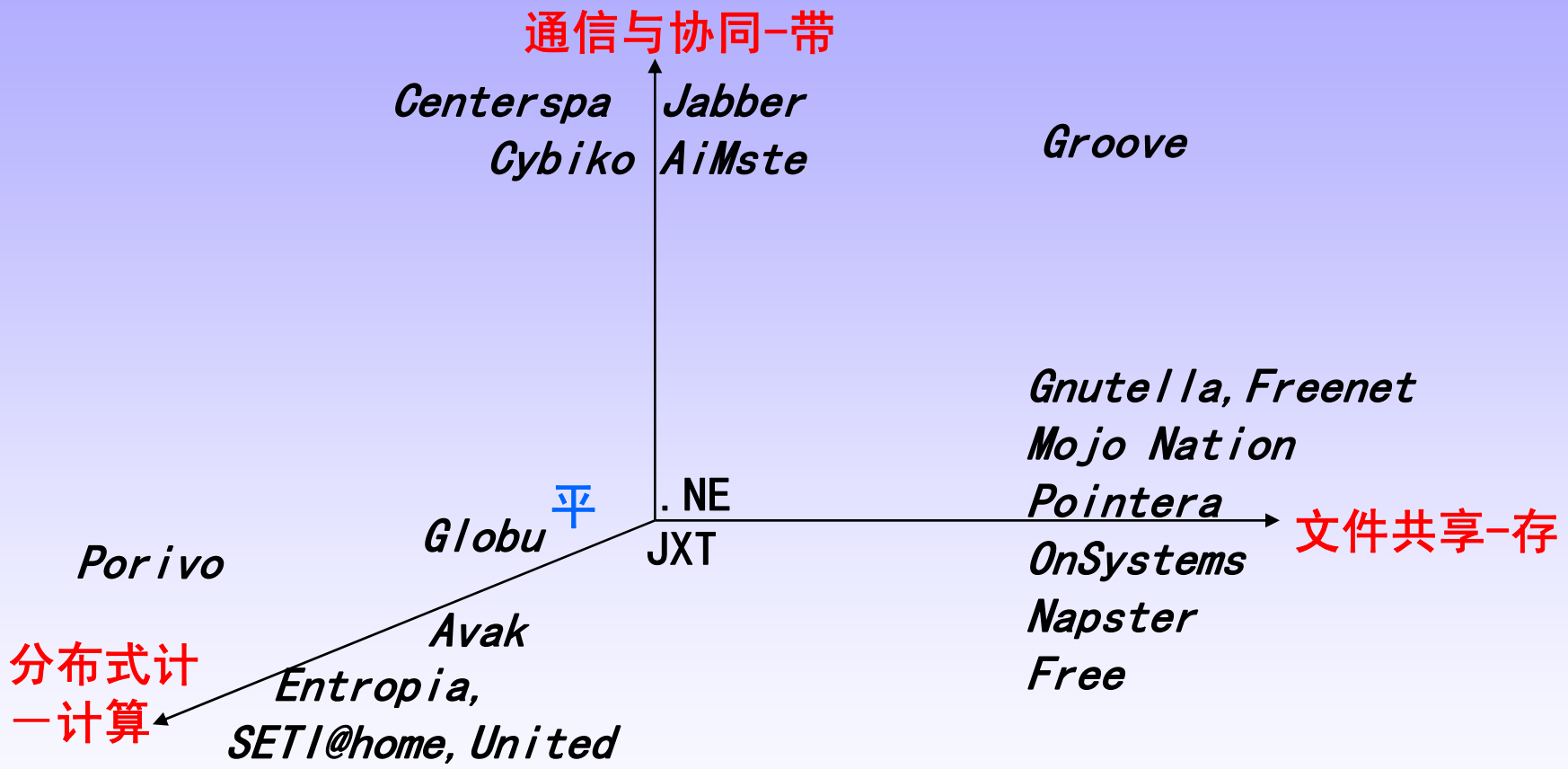
P2P网络应用构成


市场/工
金
生物
通
企
娱
应用实
仿市计人分….
基系分蛋折….
即消白….
进管在存….
游戏文件共….
水平技
分布式计
协同与通
内容共
平
JXTA ， .NET 服

当前应用分类

- 文件共享
  - BitTorrent/Gnutella/E-Mule/E-Donkey/Maze/Kazza
- 处理器共享
  - SET@Home
- 视频直播、点播
  - PPLive/CoolStreaming/PPSteam

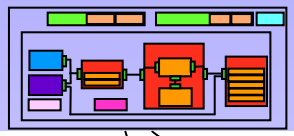
P2P应用的多维视图


通信与协同-带
Centerspa
Cybiko
Jabber
AiMste
Groove
Gnutella,Freenet
Mojo Nation
Pointera
OnSystems
Napster
Free
文件共享-存
平
.NE
JXT
Globu
Porivo
Avak
Entropia,
SETI@home,United
分布式计
一计算

分布式 P2P 仿真

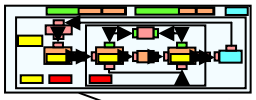
全系统仿真由许多子系统仿真构成

翅膀模块

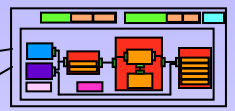


•Lift Capabilities
•Drag Capabilities

机身模块

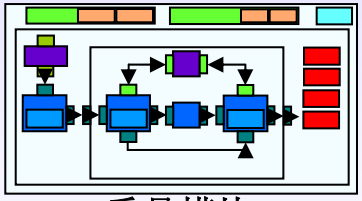


水平尾翼模块

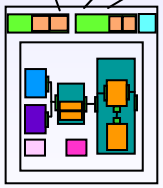


•Deflection capabilities
•Responsiveness

Crew Capabilities
- accuracy
- perception
- stamina
- re-action times
- SOP's



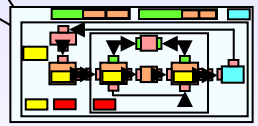
乘员模块



•Braking performance
•Steering capabilities
•Traction
•Dampening capabilities

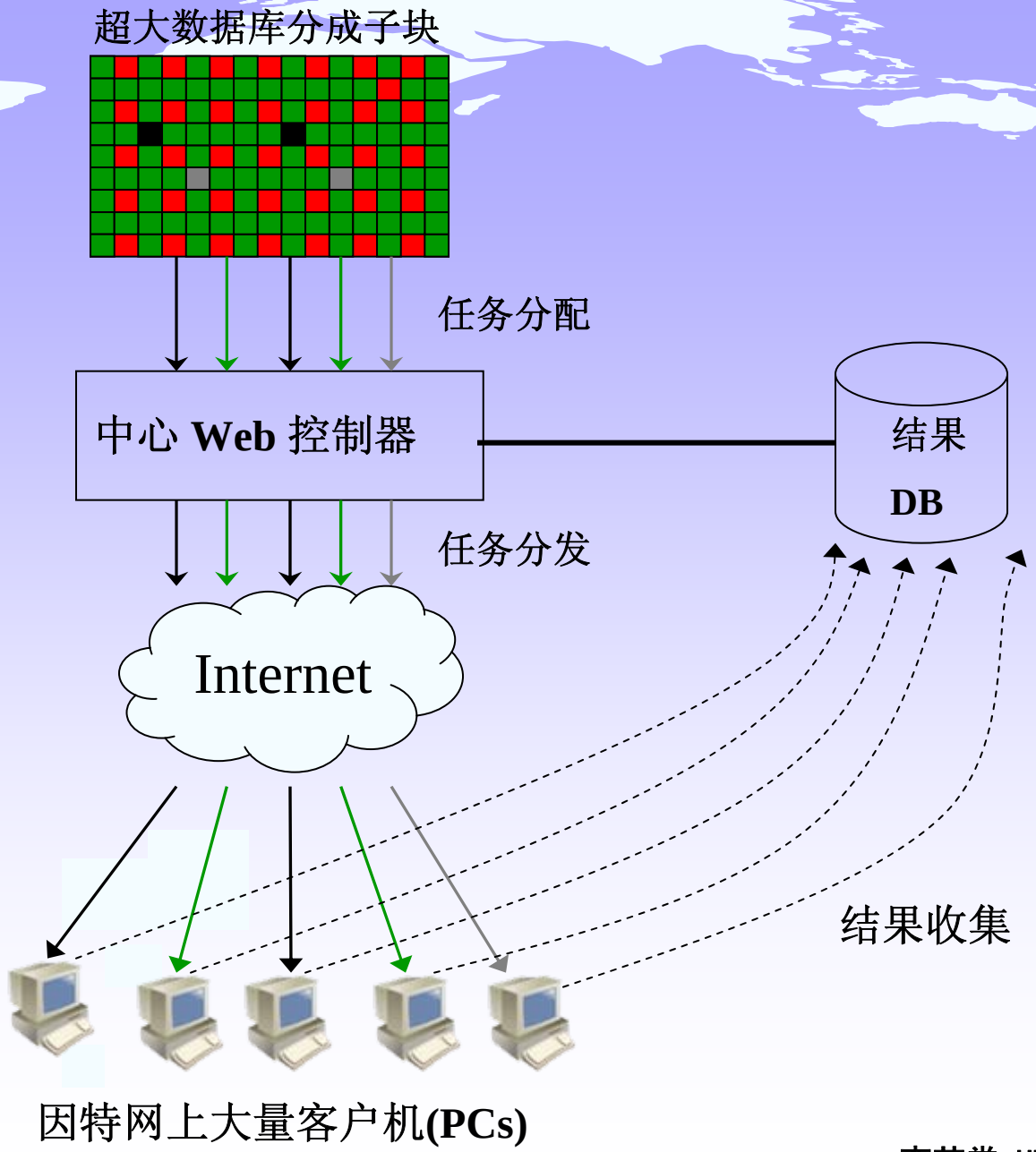
起落架装置模块

引擎模块



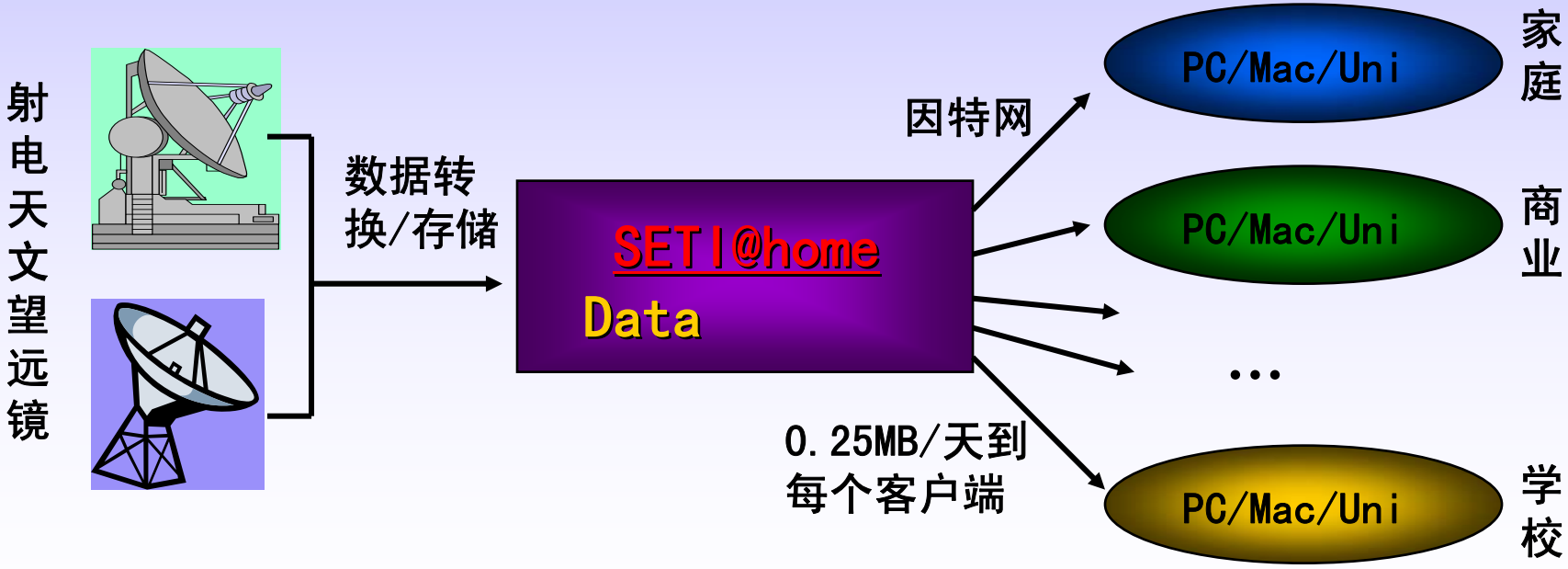
•Thrust performance
•Reverse Thrust performance
•Responsiveness
•Fuel Consumption

Web上的分布式计算


超大数据库分成子块
任务分配
中心 Web 控制器
结果
DB
任务分发
Internet
结果收集
因特网上大量客户机(PCs)

SETI@home

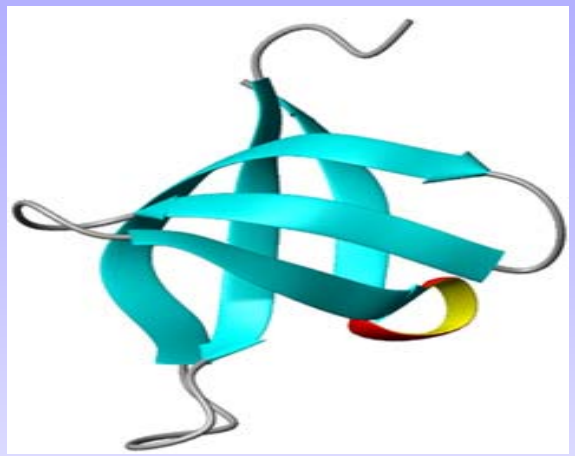
- Search for Extraterrestrial Intelligence
  - 研究工程集合:目的是发现不同文明
  - 工程之一: SETI@home:通过位于Arecibo的巨大天文望远镜接收和收集空间的无线电波发射,并用成千上万的因特网PC来分析

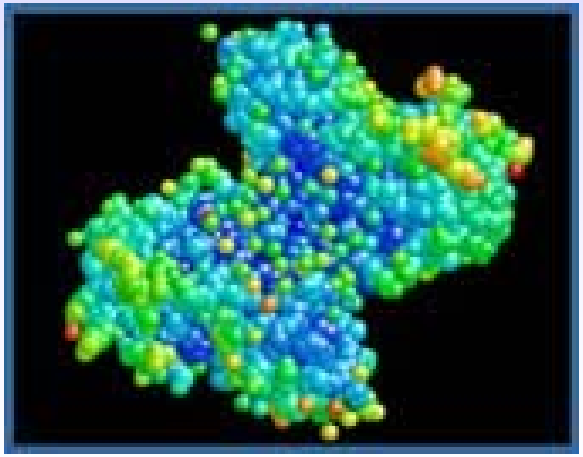


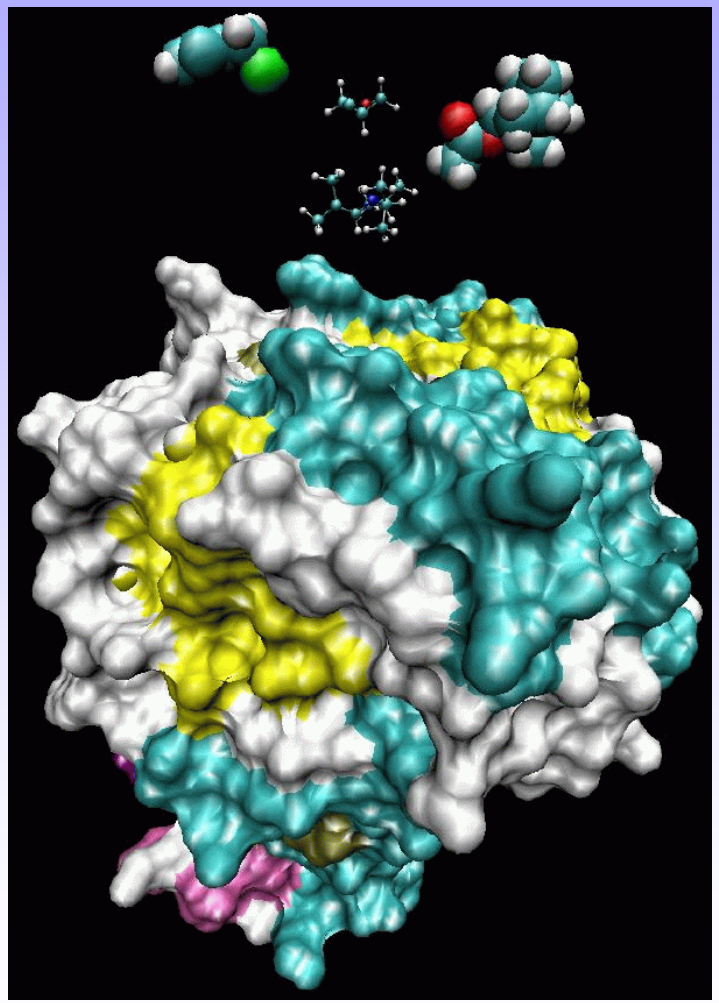
Folding@home/蛋白质折叠和药物设计

http://www.stanford.edu/group/pandegroup/Cosm/

http://members.ud.com/vypc/cancer/







- 虚拟超级计算机 peer-to-peer technology 产生空前大量的计算能力
- 使医疗研究者能加速治疗方法的改进和药物的设计
- 加快癌研究的新发现

4.1.3 P2P关键技术特性

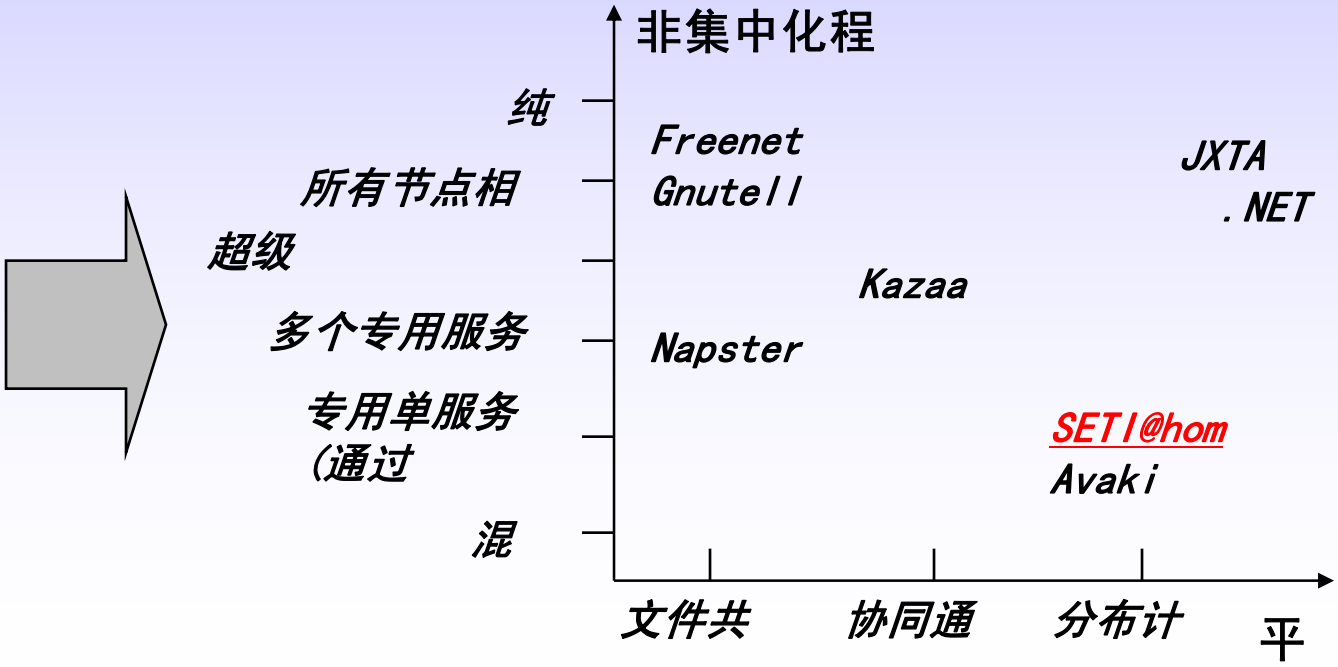
I) 非集中化:置疑 C/S 模式

- 集中化
  - 在访问权限和安全上容易管理
  - 但不可避免导致:低效/瓶颈/资源浪费
  - 尽管硬件性能和成本有了改进,但建立和维护集中化知识库成本高昂,需要人员智能化地建立,保持信息的相关和更新
- 非集中化:更强有力的思想
  - 强调用户端所有权,对数据和资源的控制
  - 每个Peer都是平等的参与者
  - 实现更困难(无全局服务器,看不到全局Peers及其文件)
  - 这也是当前混合模式存在的原因

◆全非集中化文件系统(Gnutella Freenet)

- ➢发现网络是很困难的
- ➢新节点必须知道其他节点
- ➢或由主机列表知道其他Peers的IP地址
- ➢该节点通过和现行网络中至少一个Peer建立连接而加入网络
- ➢从而能发现其他Peers并Cache它们的IP地址在本地

各种P2P系统按非集中化程度分类



II）可扩展性

- 可扩展性受限的主要原因
  - 需要完成大量的集中化操作:如同步与一致
  - 需要维护许多状态
  - 固有的并行性应用展开
  - 用来表示计算的编程模式
- P2P解决可扩展性问题
  - Napster在其服务的高峰用户达到 600万
  - 然SETI@hone2002年止用户 仅接近350万.因为它集中在并行度有限的任务上,依靠因特网上的可用计算力来分析从天文望远镜收集来的数据,搜索外星生命
  - Avaki通过提供分布式对象模型来解决可扩展性问题

- Napster是通过故意保留许多集中化文件操作来实现-达到好的扩展性并不是扩大其它所希望的特点
- Gnutella和Freenet:早期的P2P系统具有Ad-hoc的特点,Peer必须把请求盲目发送到许多其他Peers,促使它们搜索请求的文件
- CAN,Chord,Oceanstore PST:最近的P2P系统
  - 专注在目标键和目标节点间找到一致的映射
  - 每个节点仅维护较少的系统节点信息及其状态,故增加了可扩展性
  - 这些系统设计规模是 数10亿用户,数百万服务器和10^{14}文件
- 未来:带宽和计算能力继续增长,P2P平台能利用这些能力去完成人们感兴趣的应用,结构将更自治可扩展,提供更多的资源,展开更多的应用

III) 匿名

- 目的
  - 重要目的是让人们使用系统时不用关心法律问题和其他节外生枝的问题
  - 进一步目的可能使数字内容的审查制度形同虚设
- 匿名形式
  - 作者:可以不标识文件的作者或创建者
  - 出版者:可以不标识对系统而言的文件发行者
  - 读者:可以不标识文件的读者或其他消费数据者
  - 服务器:可以不标识含有未被标识文件的服务器
  - 文件:服务器并不知道它存储的是什么文件
  - 查询:服务器并不告诉它正用何文件在响应用户的查询

- 必须在通信对之间强迫执行3种匿名，才能达到上述匿名(不管何种匿名形式)
  - 发送者匿名:隐藏发送者的标识符
  - 接收者匿名:隐藏接收者的标识符
  - 相互匿名:隐藏发送者何接收者的标识符，且双方标识符对其他Peers也是隐藏的
- 匿名程度
  - 绝对隐私
  - 不可怀疑:即使攻击者能看到已发送消息的证据，但发送者似乎并不比系统潜在的发送者更像真正的发送者
  - 大概无罪
  - 可能曝光
- 6种不同技术，每个适合不同的匿名方式

- 多播使接收者匿名
- 发送者地址欺骗-使用诸如UDP的无连接协议
  - 发送者伪造其地址，从而达到消息发送者匿名的目的
  - 但并不总是可行的，因为大多数ISP能过虑来自无效IP地址的源发包
- 标识符欺骗-改变通信参与者的标识符
  - Freenet：Peer既可通过自己的Cache，也可通过上游的Peer，传送一件到请求者，但都可说是自己提供的内容
  - 从攻击者的观点看，响应(或提供)者大概是无罪的，但实际响应者(提供)很大可能是其他人
- 隐蔽通道-双方通过某些中间节点，而不是直接通信
- 难管的别名
- 非志愿放置-发布者强行把文档放在某承载主机上，非志愿主机对这些文档不负责任

项目	匿名技术与类型			
	发布者	读者	服务器	文档
Gnutella	组播 隐蔽通道	N/A	N/A	N/A
Freenet	隐蔽通道 标识符欺骗	隐蔽通道	非志愿放置	加密
APFS	隐蔽通道	隐蔽通道	N/A	N/A
FreeHaven	隐蔽通道 (remailer)	隐蔽通道	广播	加密/把文件分成共享部分
Publius	隐蔽通道 (remailer)	N/A	非志愿放置	加密/划分密钥
PAST	N/A	N/A	非志愿放置	加密

IV) 自组织

- 定义
  - 自组织是一个过程，在此一个系统的组织(约束/冗余)自然本能地增加，也就是不通过环境，也不包含其他外部系统来增加控制
- P2P需要自组织
  - 可扩展性：系统数/用户数/负载数等每一个都不可预测，因为需要进行频繁的集中化重构
  - 故障容错 （resilience弹性）：大规模导致故障率增加，这就需要对系统的自维护/自修复
  - 资源的间歇连接：在很长期间内保持完整的预定义构态是很难的，故需要处理Peers连接和断开而引起的变化
  - 所有权成本：管理这些专用设备和/或管理这样复杂波动环境的人需要成本，故管理应该在Peers间分布

- 有许多研究系统和产品都表明是自组织的
- OceanStore
  - 其自组织已应用到基础设施的定位和路由
  - 由于Peers的间歇性及网络延迟带宽的变化，基础设施必须适应其路由和定位支持
- Pastry
  - 通过基于全网容错的节点进/出协议处理自组织
  - 客户端请求保证在少于平均$\log_{16}N$步路由达到
  - 负载平衡；文件副本分布，随机存储
- FastTrack
  - 对自组织分布式网络进行快速搜索和下载
  - 系统中强大计算机可自动变成**超节点**作为搜索Hubs
  - 若有处理能力并满足联网标准，任客户端也可变成，**超节点**
  - 这样分布式网络可取代任何集中化服务

V）所有权成本

- P2P的前提
  - 共享所有权
  - 共享所有权减少了自有系统/内容/和维护它们的代价
  - SETI@home 比当今世界上最快的计算机还快，而且成本只是它的1%
  - Napster 音乐共享的全部理念是基于每个成员把音乐文件贡献到文件池中去，其他文件系统也一样.
- P2P协同/通信/平台
  - 集中化计算机存储信息的削减也减少了所有权和维护成本
  - 美国的无线通信采用了类似的方法—寄生网格：在用户之间共享家庭安装的802.11带宽，在成本上同安装有无线基础设施的公司竞争

VI) Ad-hoc连接

- 分布式并行计算
  - 并不能在所有时间/所有系统上执行
  - 某些系统在所有时间可用/部分时间可用/并不可用
- Ad-hoc系统
  - 对P2P计算，可随进随出，是理想的P2P使用
  - 对P2P内容共享系统，高服务保证通过冗余服务实现，削弱了Ad-hoc的特点
  - 对P2P协同， 用户希望用移动设备连接到因特网，Ad-hoc可通过代理群接收消息，或发送中继来保持通信延迟和断开的透明
- 802.11b，蓝牙和红外支持Ad-hoc，其半径有限，它接入P2P，支持容许突然断开是很重要的

VII) 性能

- P2P系统目标:
  - 聚合分散网络上的存储容量(Napster/Gnutella)和计算周期(SETI@home) 来改进系统的性能
- 影响非集中化性能的三类资源
  - 处理/存储/网络
  - 网络资源中的带宽和时延是主要因素
- 中心协调系统(Napster SETI@home)
  - Peers的协调和仲裁通过中心服务器进行
  - 混合P2P以克服这些脆弱点
- 非集中协调系统(Gnutella Freenet)
  - 用消息传递机制搜索信息和数据
  - 查询搜索的带宽与发送消息数,命中前的Perrs数成正比,

优化性能的关键技术

- 复制(Replication)
  - 把对象/文件的拷贝放在请求Peers附近,最小化连接距离
  - 改变数据时必须保持数据拷贝的一致性
  - OceanStore基于冲突解的更新传播模式支持一致性语义
- 高缓(Cache)
  - 减少获取文件/对象路径的长度,进而Peers间交换消息数
  - 这一减少很有意义-Peers间通信时延是严重的性能瓶颈
  - Freenet:命中文件传播到请求者途中所有节点高缓它
  - 目标是最小化时延,最大化请求吞吐率,很少高缓大数据
- 智能路由和网络组织
  - 社交“小世界”现象,60年美， 明信片均6熟链找到生人
  - 局部搜索策略,代价与网络规模成子-线性增加
  - OceanStore/Pastry网络上积极移动数据提高性能