南京邮电大学
Nanjing University of Posts and Telecommunications

新型网络计算技术

徐小龙

南京邮电大学 教授/博士生导师

南京邮电大学

个人信息

姓　　名:	徐小龙	性　　别:	男	导师类型:	博士生导师
技术职称:	教授	电子邮箱:	xuxl@njupt.edu.cn		
博士招生学科:	(0810Z2)信息网络				
学术型硕士招生学科:	(083500)软件工程				
专业型硕士招生类别(领域):	(085400)电子信息				

个人简介:

徐小龙，教授，博士生导师。2008年获得“通信与信息系统”专业博士学位，2011年获得国家留学基金委资助赴英国从事博士后研究，一直从事信息网络、分布式计算与信息安全等技术领域的教学和科研工作。现为ACM会员、IEEE会员、中国计算机学会高级会员、中国电子学会青年科学家、江苏省计算机学会“计算机安全专委会”常务委员、人工智能专委会委员、江苏省电子学会“信息安全专委会”委员、江苏省计算机学会“计算机与通信专委会”副主任委员兼秘书长。 入选江苏省“333高层次人才培养工程，江苏省“六大人才高峰”高层次人才、江苏省“青蓝工程”优秀青年骨干教师，获得南京邮电大学课堂教学卓越教师和教学名师称号，成为国家级、省级教学名师培育人选，并被评为江苏省优秀计算机科技工作者。作为项目组负责人已经顺利完成了多项科研项目，包括国家自然科学基金项目、国家863项目、教育部博士点基金、教育部专项课题、中国博士后基金、江苏省科技计划项目，以及华为科研基金和中兴科研基金等横向科研项目等；作为第一作者在SCI 收录的期刊上发表论文18篇，作为通信作者发表53篇论文，独立编写出版2本学术专著《云计

http://dsfc.njupt.edu.cn/dsgl/nocontrol/college/dsfcxq.htm?dsJbxxId=9B9D05C52A802DCFE050007F01006EFE

主讲情况

1. 南京邮电大学计算机学院教授、博士生导师
2. “通信与信息系统专业”博士
3. “电子科学与技术”博士后流动站博士后
4. University of the West of Scotland 博士后
5. 江苏省高层次创新创业人才
6. 江苏省“333高层次人才培养工程”高层次人才
7. 江苏省“六大人才高峰”高层次人才
8. 中国计算机学会高级会员
9. IEEE member
10. 江苏省计算机学会“人工智能专委会”委员
11. 江苏省大数据专家委员会委员
12. 江苏省计算机学会“计算机与通信专委会”秘书长

主讲情况

相关领域承担科研项目情况:

（1）国家自然科学基金项目，面向绿色云计算的节能型资源整合和任务调度关键技术的研究，项目负责人

（2）国家自然科学基金项目，基于安全Agent的可信云计算与对等计算融合模型及关键技术研究，项目负责人

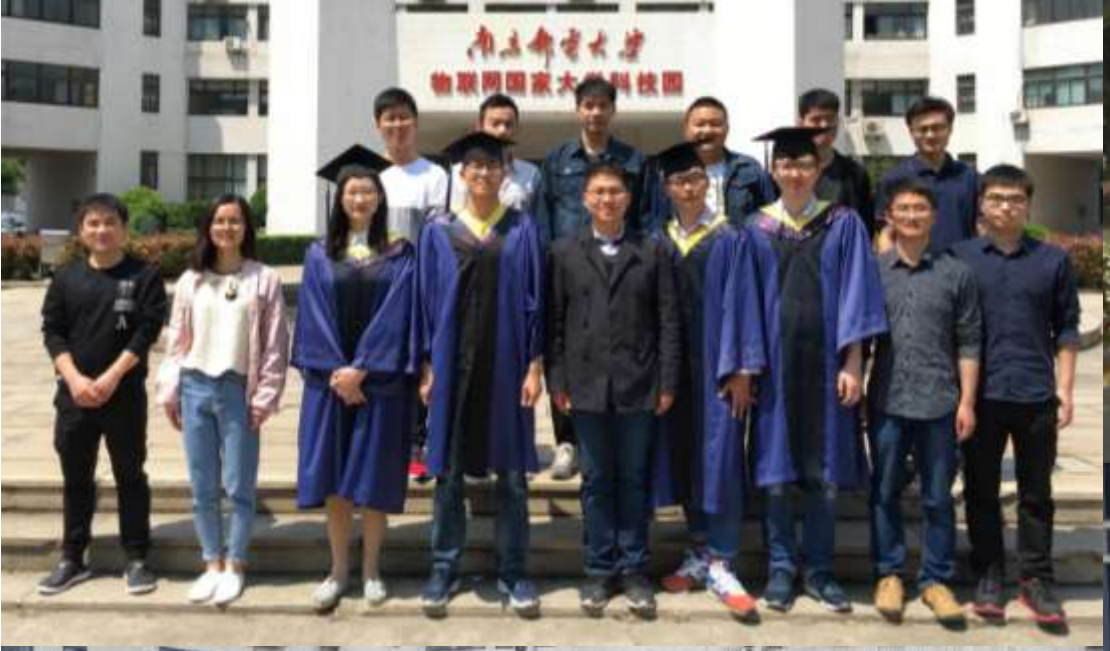
（3）江苏省科技计划项目，基于大数据的灾害管理与应急处理关键技术的研究与应用示范，项目负责人

（4） 校企合作项目，高端制造业大数据分析及商业智能支撑系统构建，项目负责人

（4）江苏省自然科学基金项目，公共云计算环境中可信虚拟私有模型及其关键技术的研究，项目负责人

（5）江苏省博士后科研资助计划项目（江苏省人力资源和社会保障厅），开放计算环境中基于安全Agent的可信虚拟私有云模型的研究，项目负责人

（6）江苏省高校自然科学研究计划资助项目，基于移动云计算的创新型文献共享和交流平台及关键技术，项目组负责人




荣誉证书
HONORARY CERTIFICATE


IOT
2016年全国大学生物联网设计竞赛
华东分赛区决赛暨颁奖典礼
主办:教育部高等学校计算机类专业教学指导委员会
承办:上海交通大学电子信息与电气工程学院软件学院
协办:德州仪器半导体技术(上海)有限公司
金牌赞助商:
银牌赞助商:
华东赞助商:


IOT


Experiments


Welcome to HPCC/SmartCity/DSS 2019
The 21st International Conference on High Performance Computing and Communications
The 17th International Conference on Smart City
The 5th International Conference on Data Science
Organized by


Why we use EEG?
More objective
Timely response
High accuracy

纲要

一 计算技术发展

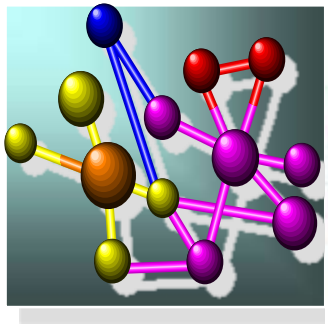
二 新型网络计算

三 主流开发平台

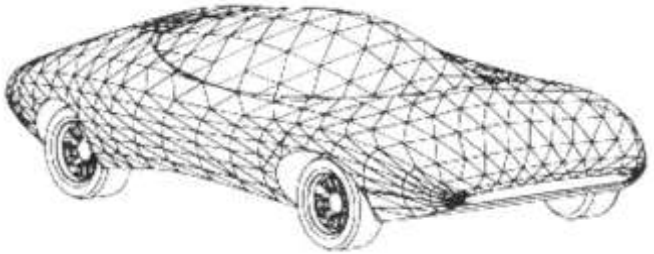
四 其它新型技术

计算技术发展回顾

对性能的需求

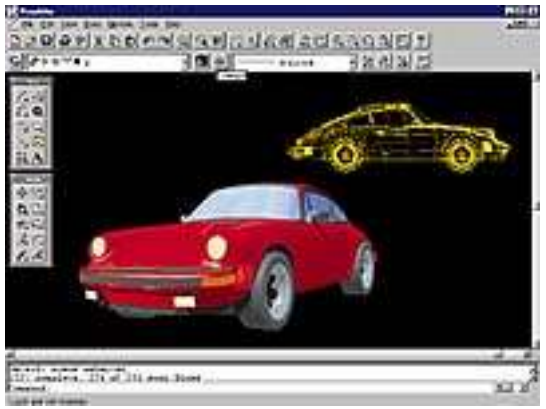


生命科学



CAD/CAM

数字生物学

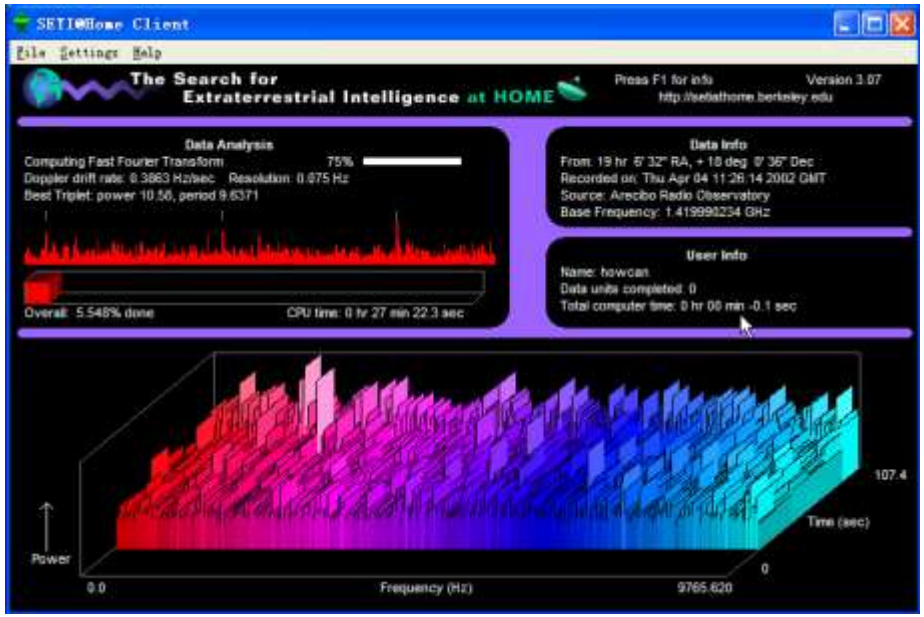


计算技术发展回顾

遥感 天文学

计算技术发展回顾


SETI@home
The Search for Extraterrestrial Intelligence


SETI@Home Client
File Settings Help
The Search for Extraterrestrial Intelligence at HOME
Press F1 for info
Version 3.07
http://setiathome.berkeley.edu
Data Analysis
Computing Fast Fourier Transform 75%
Doppler drift rate: 0.3883 Hz/sec Resolution: 0.075 Hz
Best Triplet: power 10.56, period 9.6371
Overall: 5.548% done
CPU time: 0 hr 27 min 22.3 sec
Data Info
From: 19 hr 6' 32" RA, + 18 deg 9' 36" Dec
Recorded on: Thu Apr 04 11:26:14 2002 GMT
Source: Arecibo Radio Observatory
Base Frequency: 1.419990234 GHz
User Info
Name: howcan
Data units completed: 0
Total computer time: 0 hr 00 min -0.1 sec
Power
0.0
Frequency (Hz)
9765.620
107.4
Time (sec)
0

计算技术发展回顾

军事

计算技术发展回顾

- 主机时代　50S----80S　集中计算
  IBM360/370系统，分时系统本质仍是集中计算
- 个人机时代 80S---90S　分散计算
  Wintel 模式
  - 开放的架构
  - 芯片/存储/IO的快速发展
  - 用户界面和软件的创新
  - 互联网的兴起，早期的 LAN仅实现信息共享，计算模式仍是分散的

计算技术发展回顾

- 网络时代　90S----　网络计算
  Internet的迅速发展使网络计算成为主流技术和计算范式
- Internet的发展
  - 起始阶段（70S~90S)
    - 物理层+TCP/IP
    - 小范围
  - Web阶段（90S~现在）
    - HTTP+Browser =〉呈现信息的窗口
    - 技术＝>文化
    - Client-Server，Browser-Server
  - 智能网络（现在～）

网络计算的概念与特征

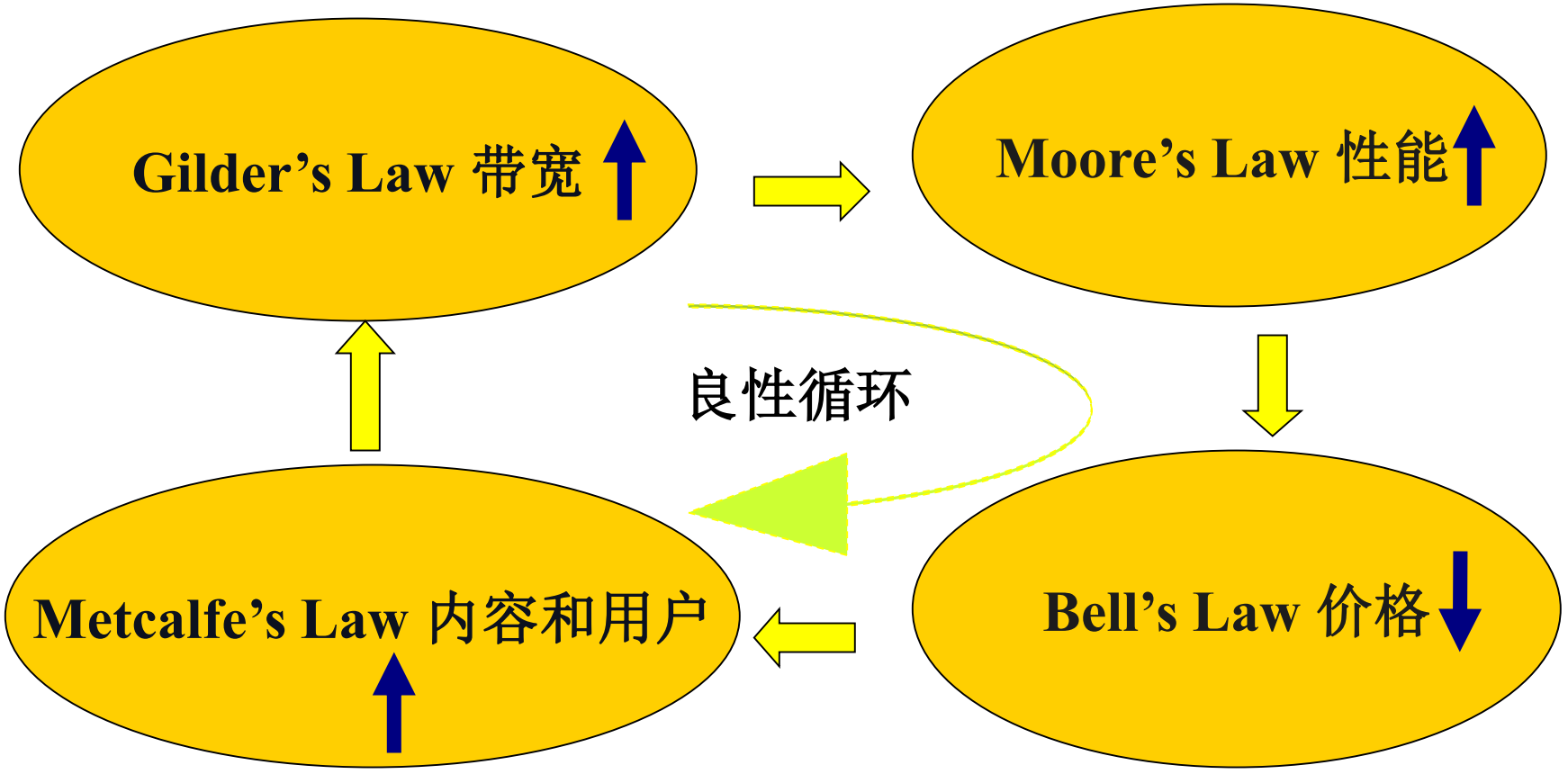
- 网络计算系统的基本概念
  - 网络计算系统是把分布在网络（Internet、物联网、移动网络）上的多个局部自治的异构计算系统进行有机集成，实现广泛的资源共享和协同工作的系统
  - 网络资源：计算资源、信息资源、软件资源、社会资源；资源共享是在资源汇聚的基础上
  - 网络计算具有广泛的应用：科学计算、信息服务、事务处理、数据交换等
  - 网络计算技术要解决的主要问题是网络资源如何广泛共享、如何有效聚合、如何充分释放
  - 技术上是计算和网络通信技术的融合

网络计算的概念与特征

- 网络计算环境的基本特征
  - 成长性：网络资源不断膨胀和变化，节点连接的开放性和动态性
  - 自治性：节点高度自治，无统一控制的“真”分布性，缺乏有效的协同能力。
  - 异构性：设备、软件、人员的多重异构性，网络连接环境的多样性，使用方式的个性化

  - 体现了网络计算系统的技术难度
  - 体现了要解决的关键问题

网络四大定律


Gilder's Law 带宽
Moore's Law 性能
良性循环
Metcalfe's Law 内容和用户
Bell's Law 价格

纲要

一 计算技术发展

二 新型网络计算

三 主流开发平台

四 其它新型技术

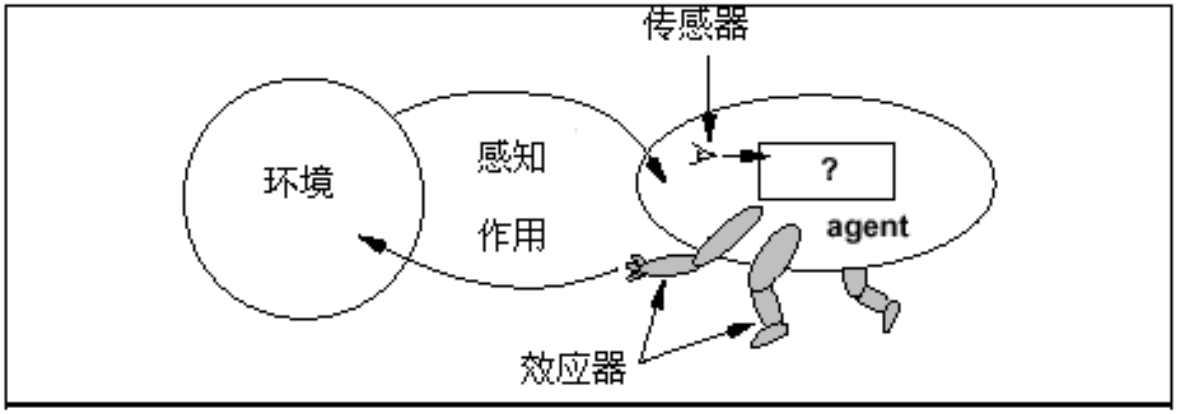
技术1：分布式人工智能

- **基于网络的分布式人工智能：**
- ✓ 分布式人工智能（Distributed Artificial Intelligence, DAI）思想的本质是采用人工智能等技术，研究一组分散的，松散耦合的智能结构如何在分布式环境下实现专家群体间高效率地相互协作联合求解，解决多种协作策略、方案、意见下的冲突和矛盾。
- ✓ 开放分布式网络环境下多点协同工作系统中，每个节点上都有相对独立的智能个体，它们具有自主性，并能根据具有的知识信念以及周围发生的事件进行推理、规划与通信。
- ✓ 多个智能个体之间彼此在逻辑上相互独立，通过共享知识、任务和中间结果，协同在工作中形成问题解决方案。

技术1：分布式人工智能

- 智能Agent
- ✓ FIPA（Foundation for Intelligent Physical Agent）的定义

Agent是存在于某一环境中的实体，能够感知环境，接收来自环境的消息，并做出反应，进而能够反作用于环境。Agent可以是软件，也可以是需要软件控制的硬件。



技术1：分布式人工智能

- 智能Agent

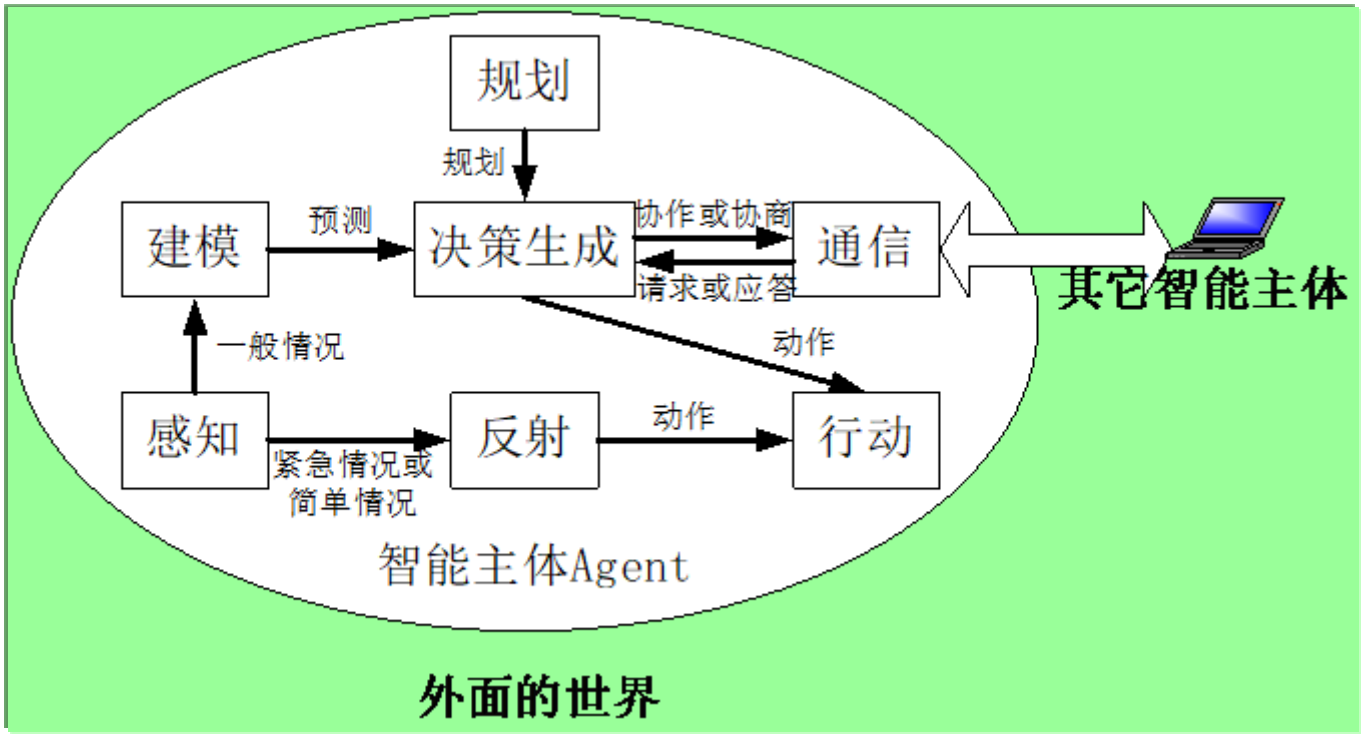
  - 自主性(Autonomy)
  - 主动性(Activity)
  - 反应性(Reactivity)
  - 移动性(Mobility)
  - 社会性(Sociality)
  - 智能性(Intelligence)

Agent可在其所处环境

Agent 具有一定程度上的智能，包括从预定义规则到自学习人工智能推理机等一系列的能力。

技术1：分布式人工智能

- 智能Agent



技术1：分布式人工智能

- 智能Agent

Agent	知觉的对象	动作	目标	环境
无人 驾驶系统	摄像机 里程计 速度计 GPS 声波定位仪 麦克风	操控方向 加速 刹车 与乘客交谈	安全 迅速 合法 理想的路线 利润最大化	道路 交通设施 行人 乘客

技术2：高效能云计算技术

1 云计算技术来由

1. 云计算借用了量子物理中的“电子云”（Electron Cloud），强调说明信息处理的弥漫性、无所不在的分布性和社会性特征。
2. 云计算技术可将计算任务分布在大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算能力、存储空间和信息服务，一般具备以下3个典型特征：

（1）硬件基础设施架构在大规模的廉价服务器集群之上；

（2）应用程序与底层服务协作开发，最大限度地利用资源；

（3）通过多个廉价服务器之间的冗余，使用软件获得高可用性。

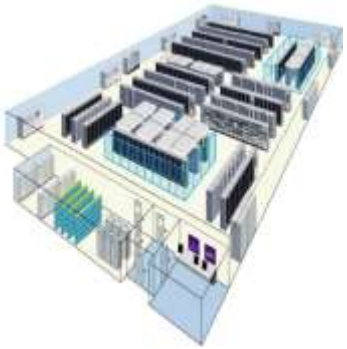
技术2：高效能云计算技术

1 云计算技术来由

1. Gartner从2009年开始发布的《IT行业十大战略性技术报告》中连续6年将云计算技术列为十大战略技术之一；
2. 在发布的《2014年十大战略性技术趋势报告》中，有三项技术与云计算相关：
   （1）混合云和混合IT（Hybrid Cloud and IT as Service Broker）
   （2）云+端联合架构（Cloud/Client Architecture）
   （3）个人云（Personal Cloud）
3. 在发布的《2015年十大战略性技术趋势报告》中，再次提出“云+端计算”，基于云计算实现内容与应用程序状态在多重设备间同步，以及解决跨设备的应用程序可移植性。

技术2：高效能云计算技术

2 云计算数据中心



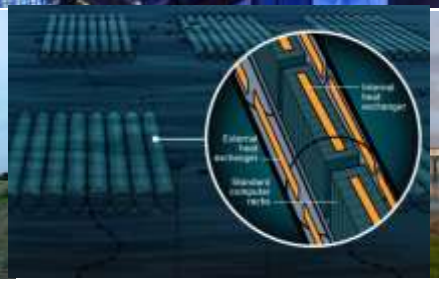
1. 维基百科给出的定义是“数据中心是一整套复杂的设施。它不仅仅包括计算机系统和其它与之配套的设备(例如通信和存储系统),还包含冗余的数据通信连接、环境控制设备、监控设备以及各种安全装置”。
2. 谷歌的《The Datacenter as a Computer》一书中,将数据中心解释为“多功能的建筑物,能容纳多个服务器 以及通信设备。这些设备被放置在一起是因为它们具有相同的对环境的要求以及物理安全上的需求,并且这样放置便于维护”,而“并不仅仅是一些服务器的集合”

技术2：高效能云计算技术

2 云计算数据中心








技术2：高效能云计算技术

2 云计算数据中心

技术2：高效能云计算技术

3 能源消耗现状

- 目前全球数据中心每年的耗电量相当于30座核电站发电量；
- Google的云数据中心每年消耗的电能高达近23亿千瓦时；

《工人日报》

我国数据中心耗电量远高国际水平

3月23日，工信部联合国家能源局、国家机关事务管理局，印发《国家绿色数据中心试点工作方案》（以下简称方案）。方案披露：我国数据中心发展迅猛，总量已超过40万个，年耗电量超过全社会用电量的1.5%，其中大多数数据中心的PUE（平均电能使用效率）仍普遍大于2.2，与国际先进水平相比有较大差距。

根据工信部披露的信息，随着信息化快速发展，全球数据中心建设步伐明显加快，总量已超过300万个，耗电量占全球总耗电量的比例为1.1%～1.5%，其高能耗问题已引起各国政府的高度重视。

与此同时，数据中心产生大量的温室气体排放，消耗大量的水资源，其设备废弃后造成较大污染，给资源和环境带来巨大挑战。

技术2：高效能云计算技术

3 能源消耗现状

- 阿尔法狗用了1920个CPU，256个GPU，它的运算效率大概是3千万亿次，但是它的能耗大概是500千瓦
- 李世石大概是0.1千瓦

技术2：高效能云计算技术

3 能源消耗现状

- 电源使用效率（Power Usage Effectiveness，PUE）值和（数据中心基础设施效率（Data Center Infrastructure Effectiveness，DCiE）。

$$PUE = \frac{\text{数据中心消耗的所有能源}}{\text{IT负载消耗的能源}}$$

- PUE值越接近于1，表示一个数据中心的绿色化程度越高。

PUE = 1 + 制冷能耗因子 + 供电能耗因子 + 其它能耗因子

技术2：高效能云计算技术

3 能源消耗现状

- 计算与存储、网络互联等IT设备产生的能耗。作为数据中心主体，IT设备产生的能耗通常占云数据中心总能耗的最大比例；
- 基于水冷、风冷等的温控设备产生的能耗。这部分能耗占云数据中心总能耗的比例有时甚至比数据中心主体设备产生的能耗还大；
- 电源供应设备等其它配套设备产生的能耗。这部分能耗一般占云数据中心总能耗较小的比例。

技术2：高效能云计算技术

4 绿色节能技术

低功耗硬件

1. 中央处理器（Central Process Unit, CPU）芯片制造商如Intel、AMD等不断采用新的制造工艺，来降低CPU能耗。
2. 固态硬盘取代机械硬盘可以很大程度上降低存储硬件的功耗。机械硬盘从待机状态切换到工作状态，需要进行电机加速；移动磁头臂需要的瞬时电流达到硬盘正常工作电流的两倍以上。而固态硬盘的启动电流几乎和工作电流一样，因此无需进行额外的电源功率设计；固态硬盘只需极短时间就能从待机状态切换到工作状态，所以可频繁将固态硬盘切换到待机状态，而不会增加额外的电力消耗，从而有效节能。
3. Google对服务器主板进行修改，并利用蓄电池来提高能源的利用率。

技术2：高效能云计算技术

4 绿色节能技术

关闭/休眠技术

1. 关闭/休眠技术也是常用的实现分布式系统节能的技术，该技术通过关闭或休眠空闲节点的方式来降低空闲能耗。
2. 通过休眠空闲的节点来减少能耗，并假定休眠后节点的能耗为0，且不考虑休眠节点存储的副本，但事实上在实际应用中必须考虑这些问题。
3. 关闭/休眠技术的缺点是当前活动节点不满足需求时，重启节点需要很长时间，这会导致系统的响应时间变长，影响用户体验。需要准确设定或预测关闭/休眠主机或关键部件的时机。
4. 对于拥有大量计算资源的云数据中心而言，关闭/休眠技术需要解决的难题是在已知单位时间任务的到达量的前提下，准确设定需要关闭/休眠多少主机，以及关闭哪些主机等问题。

技术2：高效能云计算技术

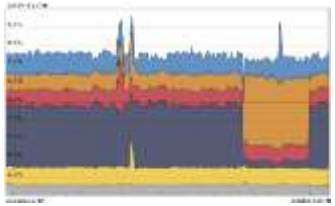
4 绿色节能技术

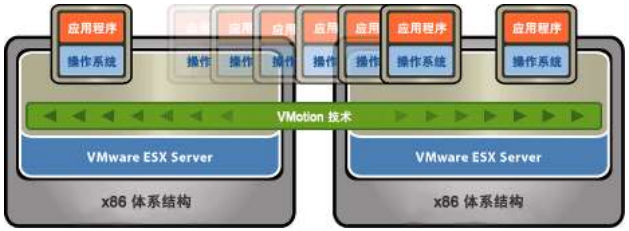
动态电压调节

1. 根据系统实时负载的大小调节系统部件功耗的大小，在降低能耗的同时保证性能。
2. 根据CMOS电路动态功率公式可以得出结论：动态功率与电压的平方是成正比的。因此，如果想要降低处理器的动态功率，可以采取降低处理器电压的方式。
3. 将DVS技术应用于云计算系统时，则需要考虑以下问题：伴随着电压的下降，处理器的性能也会随之下降；任务到达系统的时间是不确定的，所以到达任务的类型很难预测；即使能够预测任务的类型，适合该任务的处理器电压也很难确定；DVS主要用来降低处理器的能耗，但用以优化整个计算机或整个云计算系统的能耗就比较局限。

技术2：高效能云计算技术

4 绿色节能技术





虚拟化技术

- 提供最高的整合率
- 安全地提高利用率
- **减少 80% 的能源消耗**

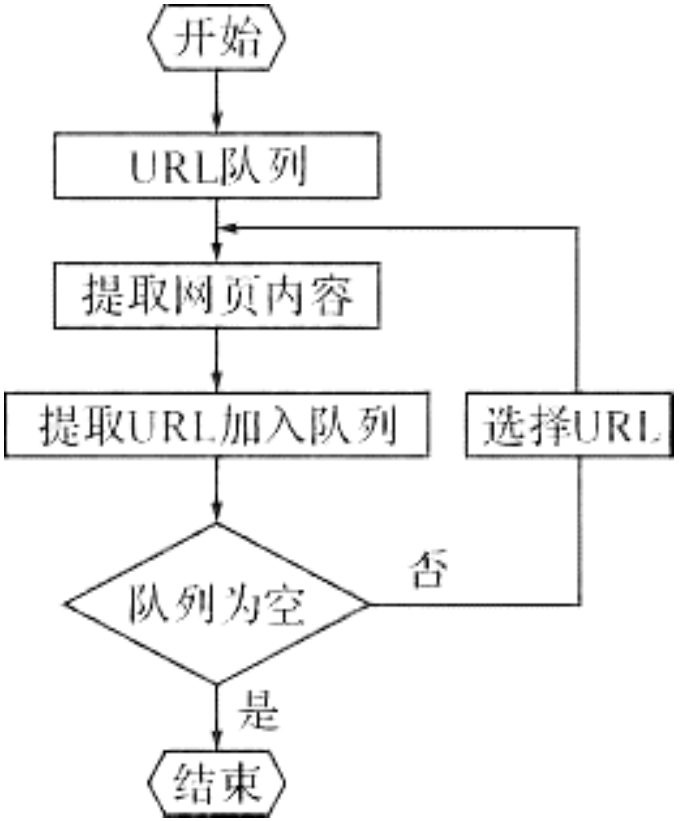
- 动态迁移虚拟服务器
- 实时关闭不需要的服务器
- 动态迁移存储
- **减少 25% 的能源消耗**

- 合理调度资源和任务
- 合理部署数据
- 避免存储相似的数据映像
- **减少 70% 的能源消耗**

技术3：精准数据采集技术

❖ 网络爬虫：

- ✓ 网络爬虫（Crawler）又被称为网页蜘蛛，网络机器人，网络爬虫是一个自动提取网页的程序，它为搜索引擎从Internet网上下载网页，是搜索引擎的重要组成。
- ✓ 网络爬虫使用多线程技术，让爬虫具备更强大的抓取能力。网络爬虫还要完成信息提取任务，基于抓取回来的网页提取出来新闻、电子图书、行业信息等。



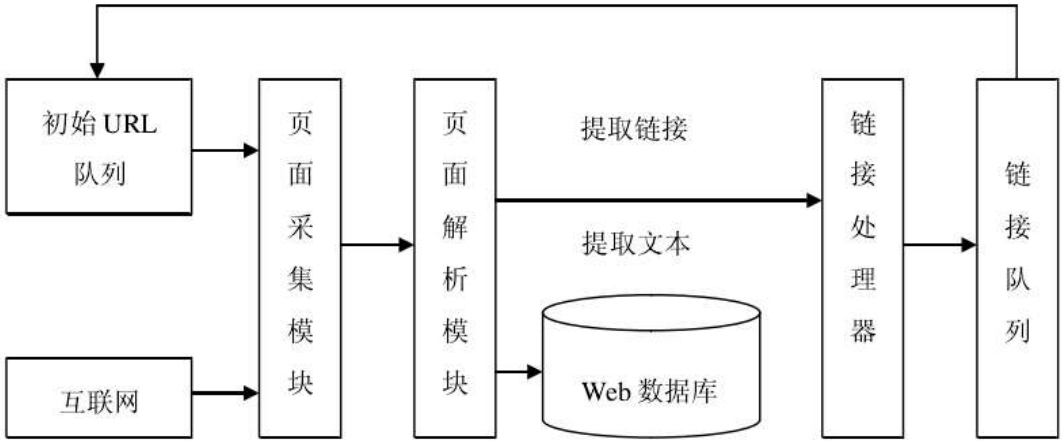
技术3：精准数据采集技术

- **网络爬虫：**
  - ✓ fresh bot——主要考虑网页的时新性
  - ✓ deep crawl bot——针对更新不那么频繁的网页

技术3：精准数据采集技术

❖ 通用爬虫：

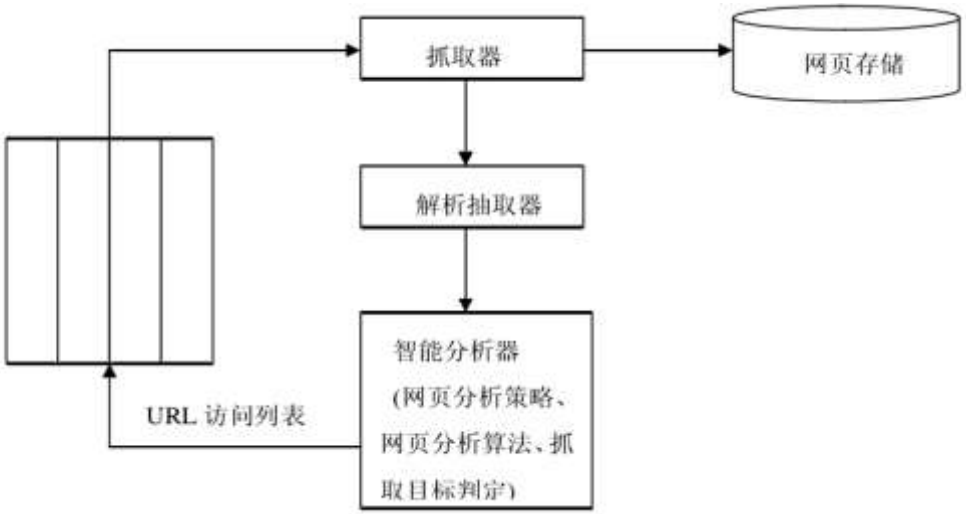
- ✓ 通用网络爬虫从种子链接开始，不断抓取URL网页，将这些URL全部放入到一个有序的待提取的URL队列里。
- ✓ Web信息提取器从这个队列里按顺序取出URL，通过Web上的协议，获取URL所指向的页面，然后从这些页面中分析提取出新的URL，并将它们放到等待提取的URL队列里。



技术3：精准数据采集技术

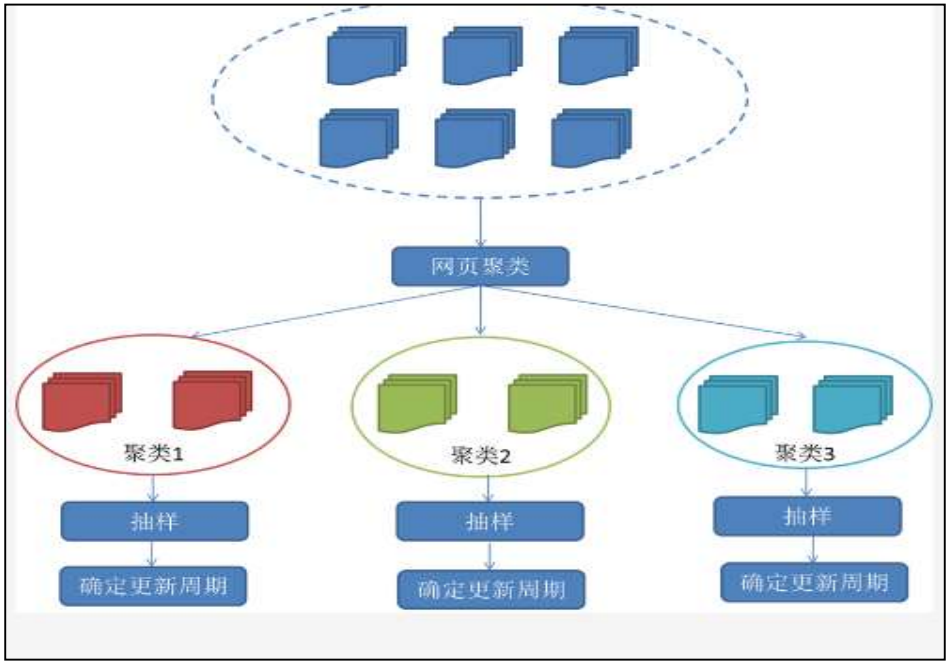
❖ 聚焦爬虫：

- ✓ 聚焦爬虫根据一定的网页分析算法，过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。
- ✓ 根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。



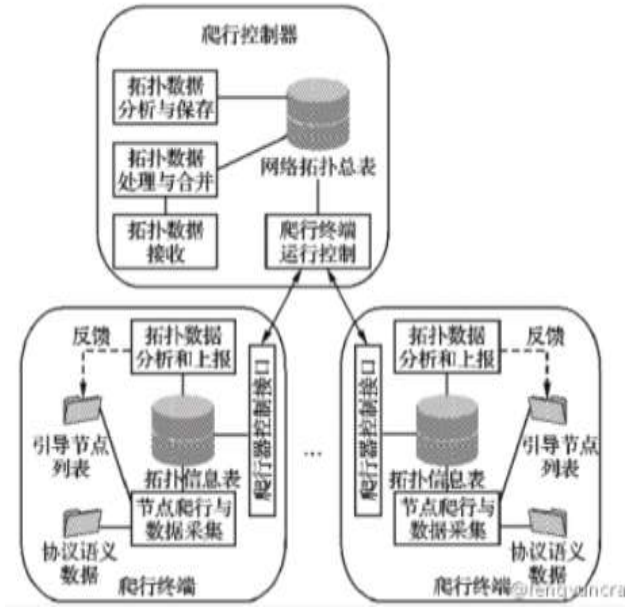
技术3：精准数据采集技术

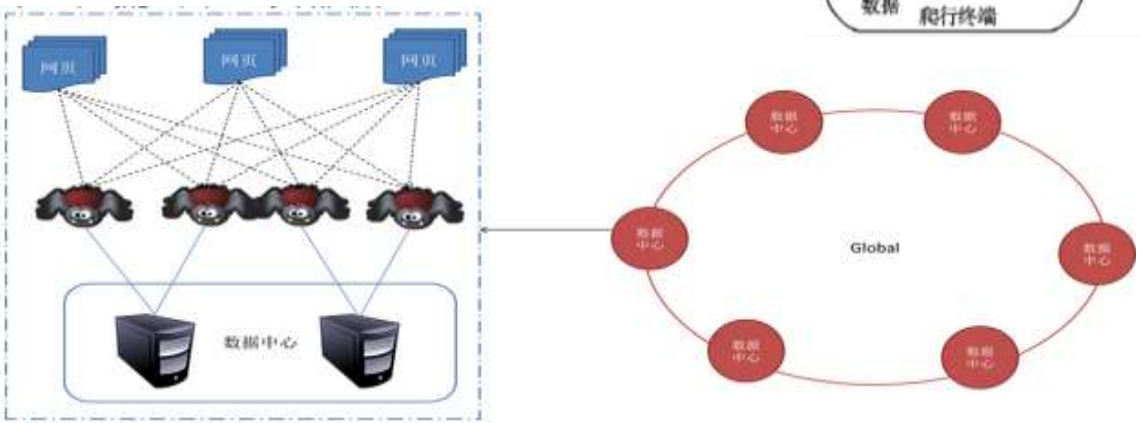
- **爬取策略：**
  - ✓ 网页的抓取策略分为深度优先、广度优先和最佳优先。
  - ✓ 网页更新策略分为历史参考策略、用户体验策略和聚类抽样策略。



技术3：精准数据采集技术

- **分布式爬取架构：**
  - ✓ 主从结构，爬行控制器和终端
  - ✓ 控制器控制(master)全部爬行器同步和终止命令
  - ✓ 终端(slave)负责信息的采集，
  - ✓ 将拓扑信息反馈控制器

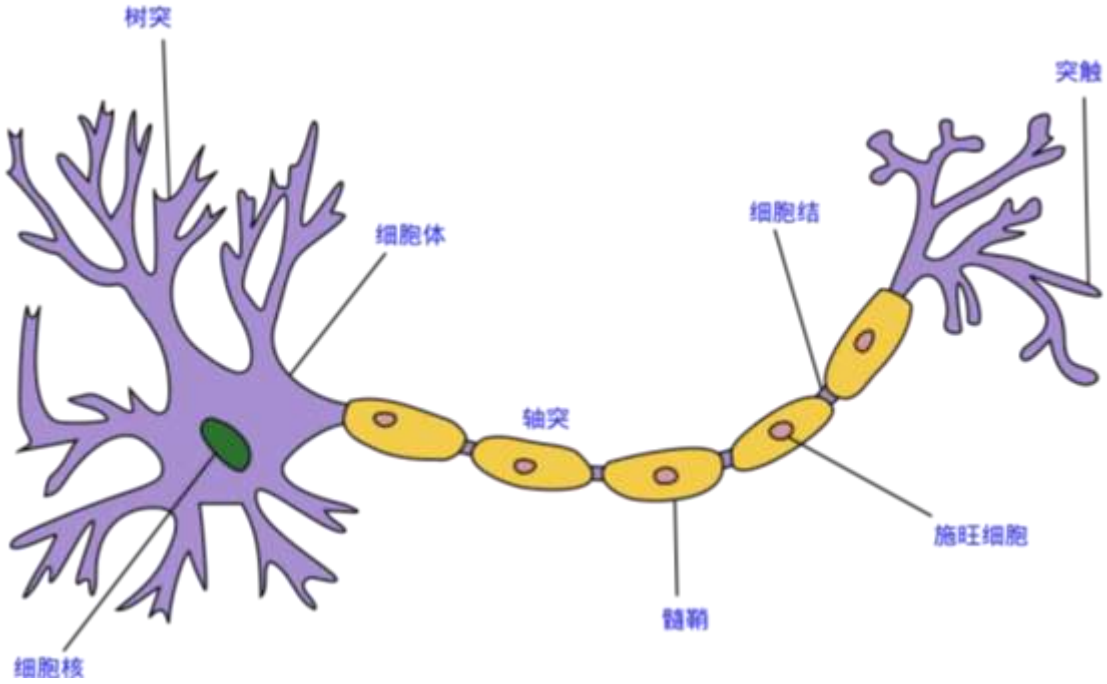


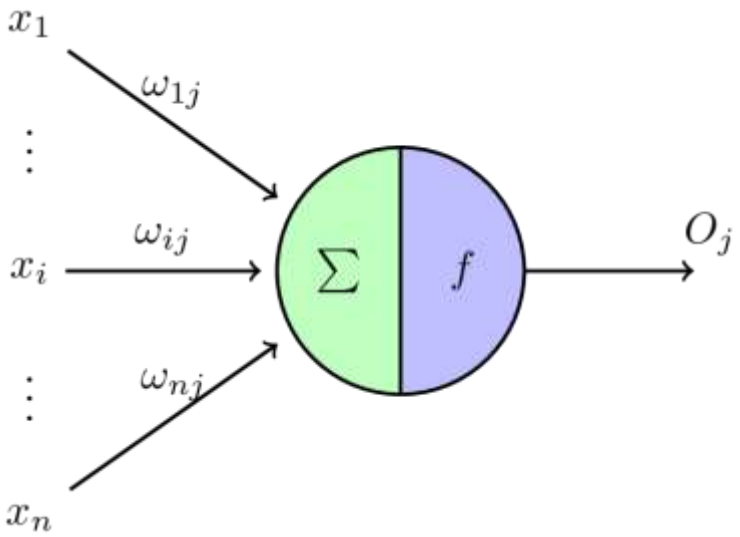


技术4：深度学习与强化学习技术

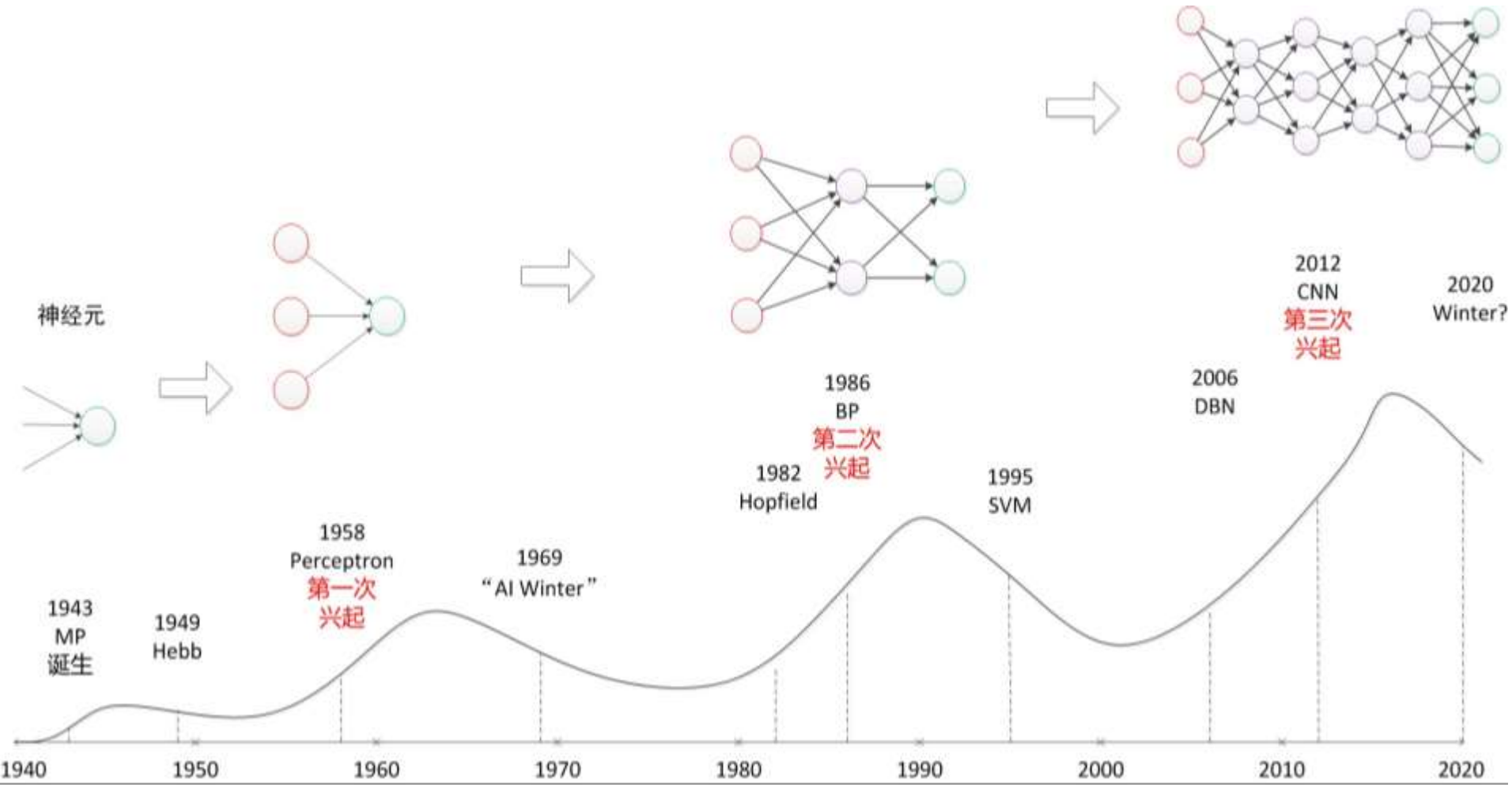
神经元结构的特点：

- 每个神经元都是一个多输入单输出的信息处理单元；
- 神经元输入分兴奋性输入和抑制性输入两种类型；
- 神经元具有空间整合特性和阈值特性；
- 神经元输入与输出间有固定的时滞，主要取决于突触延搁。





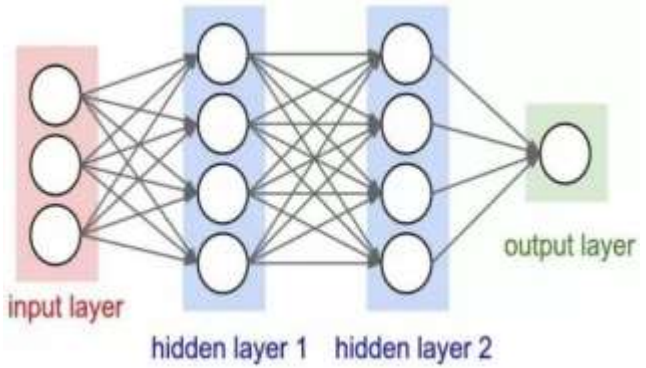
技术4：深度学习与强化学习技术

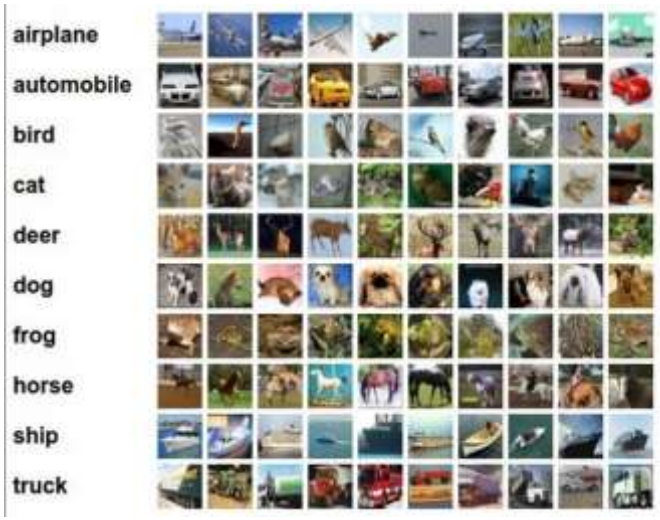


技术4：深度学习与强化学习技术

神经网络的发展阶段总结：

- 感知器：感知器利用两层神经网络进行了简单的分类问题，但只能完成最简单的线性分类问题，对于异或分类问题都无法完成时，神经网络陷入低谷。
- **BP算法**：两层神经网络无法胜任非线性分类问题时，可考虑增加神经网络层数。随着网络层数的增加，权重参数如何训练的问题无法解决，直到**BP**反向传播算法的提出，解决了这一问题，从而发生了神经网络的又一次兴起。**SVM**等算法的通用性和可计算性都优于神经网络，神经网络进入到了第二次低谷。
- 深度学习：深度学习的**CNN**卷积神经网络的提出，将它用于图像分类等问题当中。

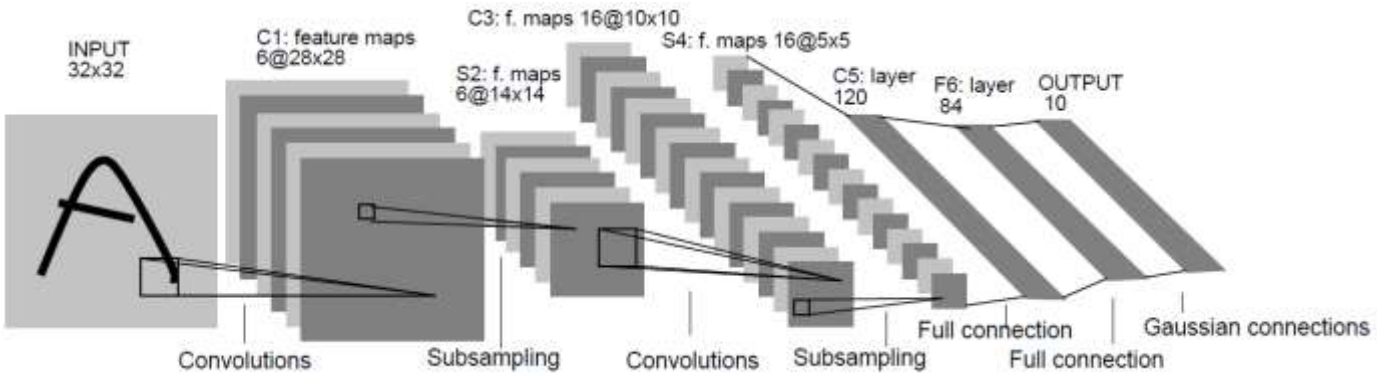




技术4：深度学习与强化学习技术

典型深度学习模型：

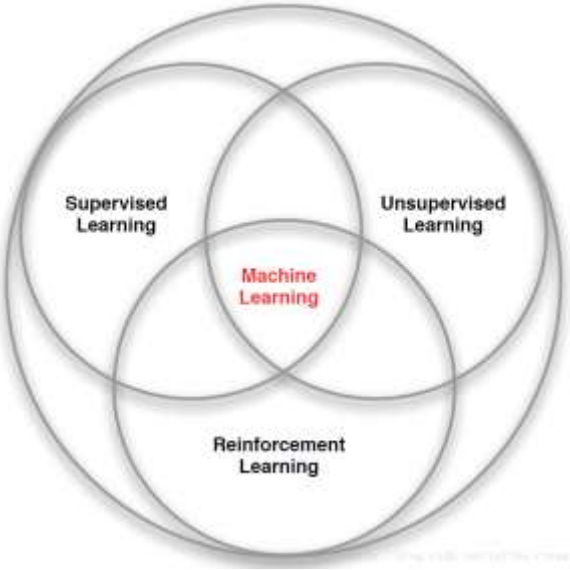
- **CNN**
  常应用于计算机的图像识别、视频分析、自然语言处理、药物发现等
- **RNN**
  用于处理序列数据，被应用在语言分析、机器翻译，语音识别等
- **DNN**
- **ResNet**
- **LSTM**
- 双向**RNN**
- 双向**LSTM**

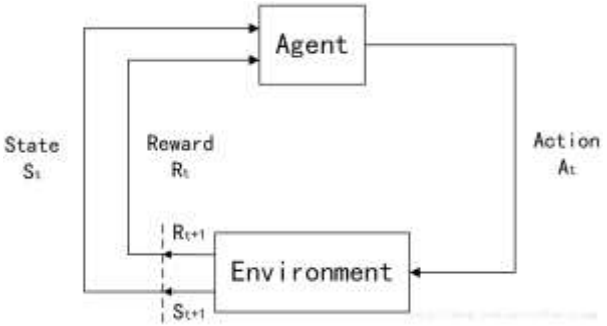


技术4：深度学习与强化学习技术

强化学习（Reinforcement Learning）：

- **Reinforcement learning is learning what to do, how to map situations to actions, so as to maximize a numerical reward signal.**
- 增强学习关注的是智能体如何在环境中采取一系列行为，从而获得最大的累积回报。
- 通过增强学习，一个智能体（**agent**）应该知道在什么状态下应该采取什么行为。从环境状态到动作的映射称为策略。

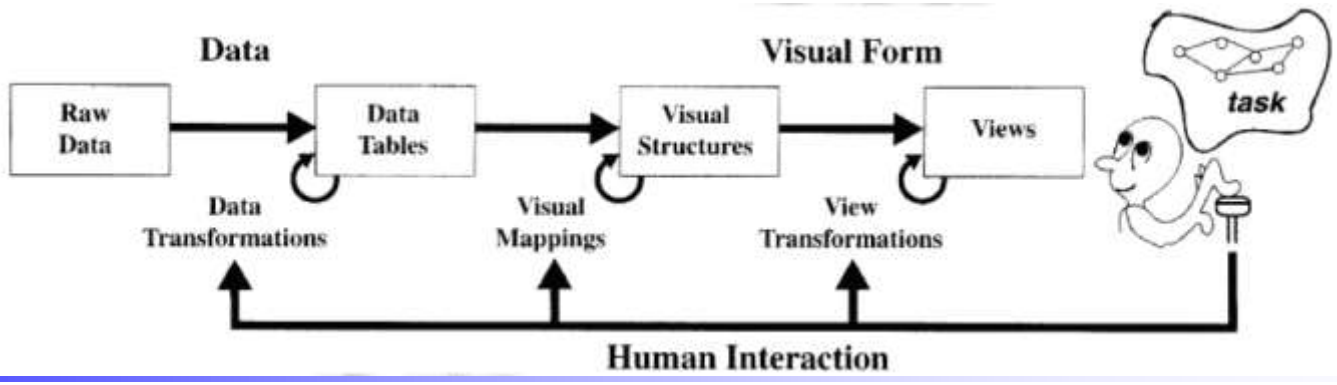




技术5：数据可视化技术

信息可视化参考模型：

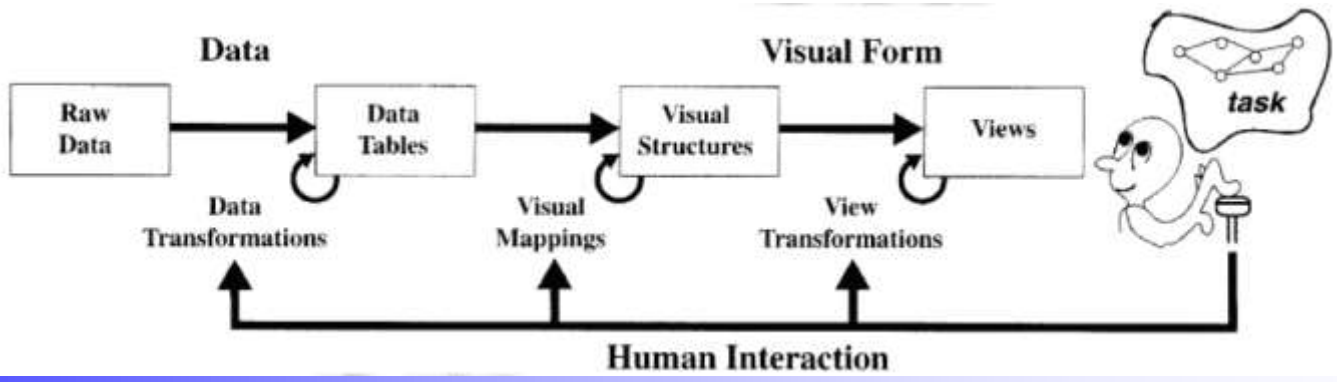
- 从原始数据到可视化形式再到人的感知认知系统的可调节的一系列转换过程
- 数据变换将原始数据转换为数据表形式
- 可视化映射将数据表映射为可视化结构,由空间基、标记、以及标记的图形属性等可视化表征组成
- 视图变换则将可视化结构根据位置、比例、大小等参数设置显示在输出设备上
- 用户根据任务需要,通过交互操作来控制上述3 种变换或映射.



技术5：数据可视化技术

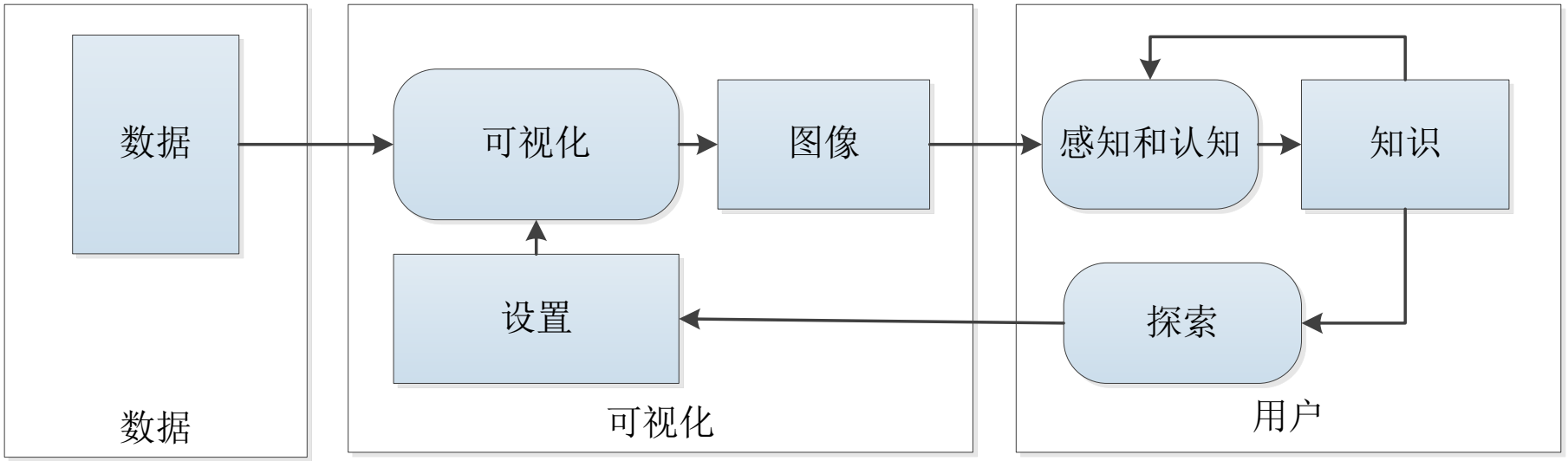
可视化映射基本要求：

- 真实的表示并保持了数据的原貌,并且只有数据表中的数据才能映射至可视化结构;
- 可视化映射形成的可视化表征或隐喻是易于被用户感知和理解的,同时又能够充分地表达数据中的相似性、趋势性、差别性等特征,即具有丰富的表达能力.
- 如何创造新型并且有效的可视化表征以达到一眼洞穿的效果,一直是该领域追求的目标和难点,在大数据时代仍然是信息可视化领域的关键所在.



技术5：数据可视化技术

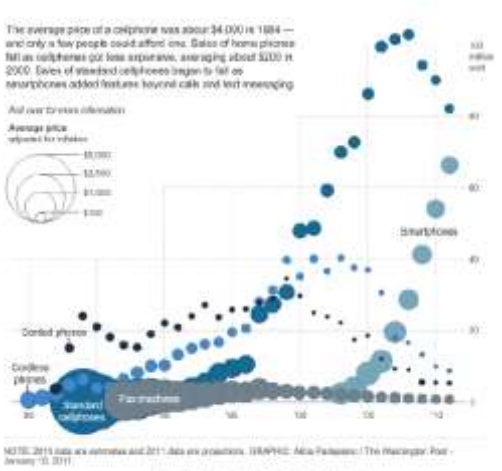
用户参与的可视化：



技术5：数据可视化技术

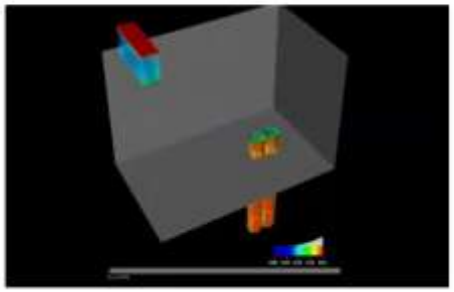
可视化的发展历程：

- 最初，可视化技术被大量应用于统计学领域，用来绘制统计图表，比如圆环图、柱状图和饼图、直方图、时间序列图、等高线图、散点图等
- 后来，又逐步应用于地理信息系统、数据挖掘分析、商务智能工具等，有效促进了人类对不同类型数据的分析与理解.

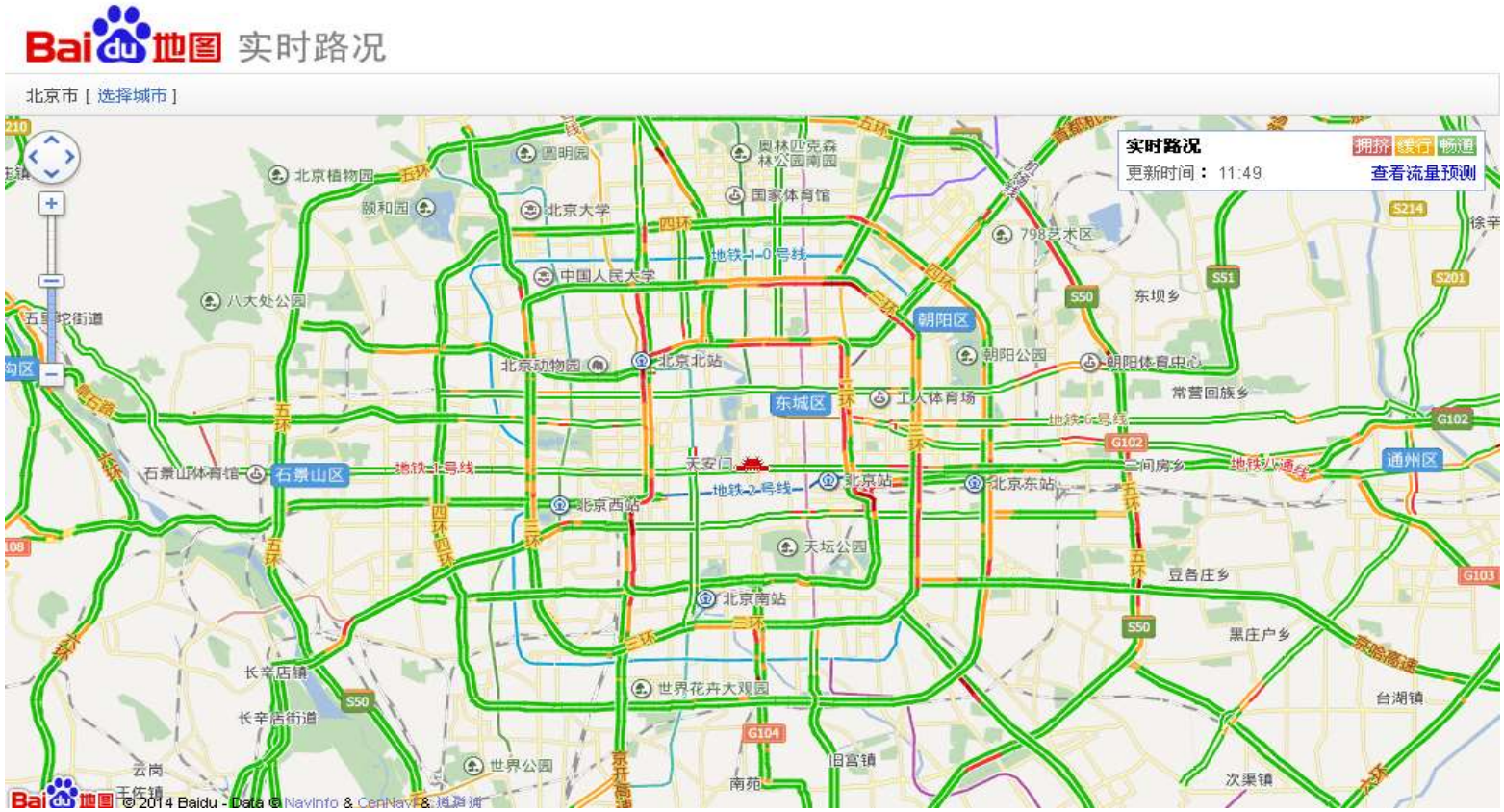


电子产品价格与销量图
圆点大小表示价格

利用涡量可视化核反应堆冷却液体的流场
– 颜色代表液体温度，温度越高颜色越红。

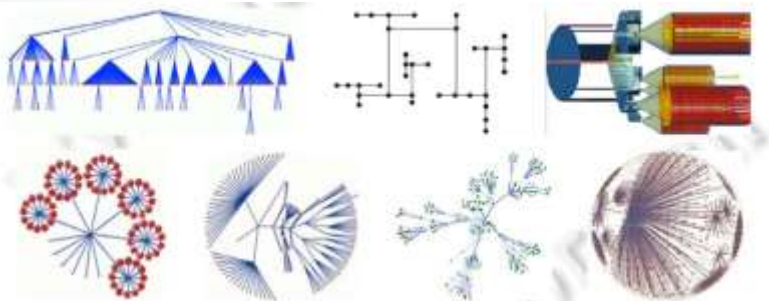


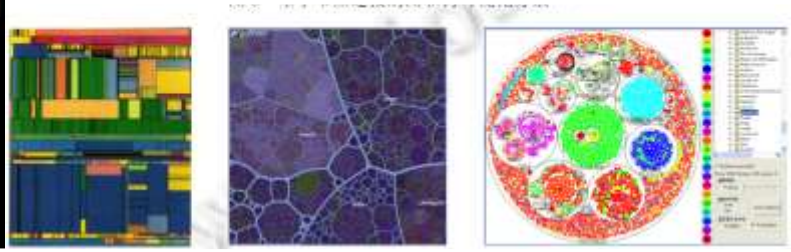
技术5：数据可视化技术


Baidu 地图 实时路况
北京市 [选择城市]
实时路况
拥挤 缓行 畅通
更新时间： 11:49
查看流量预测
北京植物园
颐和园
北京大学
国家体育馆
798艺术区
中国人民大学
八大处公园
朝阳区
东坝乡
北京动物园
北京北站
朝阳公园
朝阳体育中心
常营回族乡
东城区
工人体育场
石景山体育馆
石景山区
天安门
北京站
北京东站
北京西站
通州区
天坛公园
豆各庄乡
北京南站
黑庄户乡
长辛店镇
世界花卉大观园
台湖镇
长辛店街道
世界公园
旧宫镇
南苑
次渠镇
© 2014 Baidu - Data © NavInfo & CenNavi &

技术5：数据可视化技术


周杰伦
姚明
宋柯
余政仪
张靓颖
李承鹏
高雯丽
高文丽
李宇春
徐晨星
李升敏
杨娉璇
何洁
满春
梅丽娜
熊文
林丽
陈礼雪
潘鲸
李炳辰
于振义

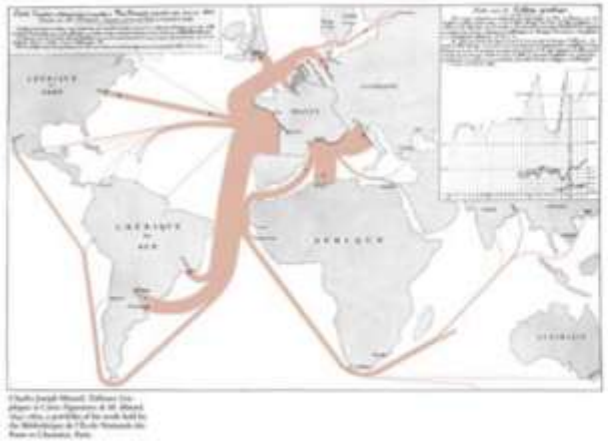




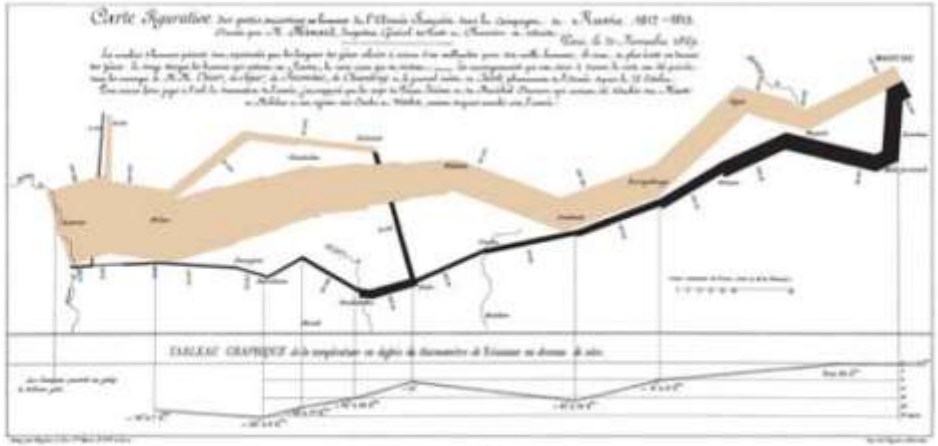

wel
goedkoffie
Vandaag
president
Vriendsnieuwe
Obama
weer
verhaal
Meeus
boek
gaat
mensen
personages
album
even
Barack
experiment


data
visualization
information
approach
analysis
design
visualizations
techniques
large
user
visual
layout
graph

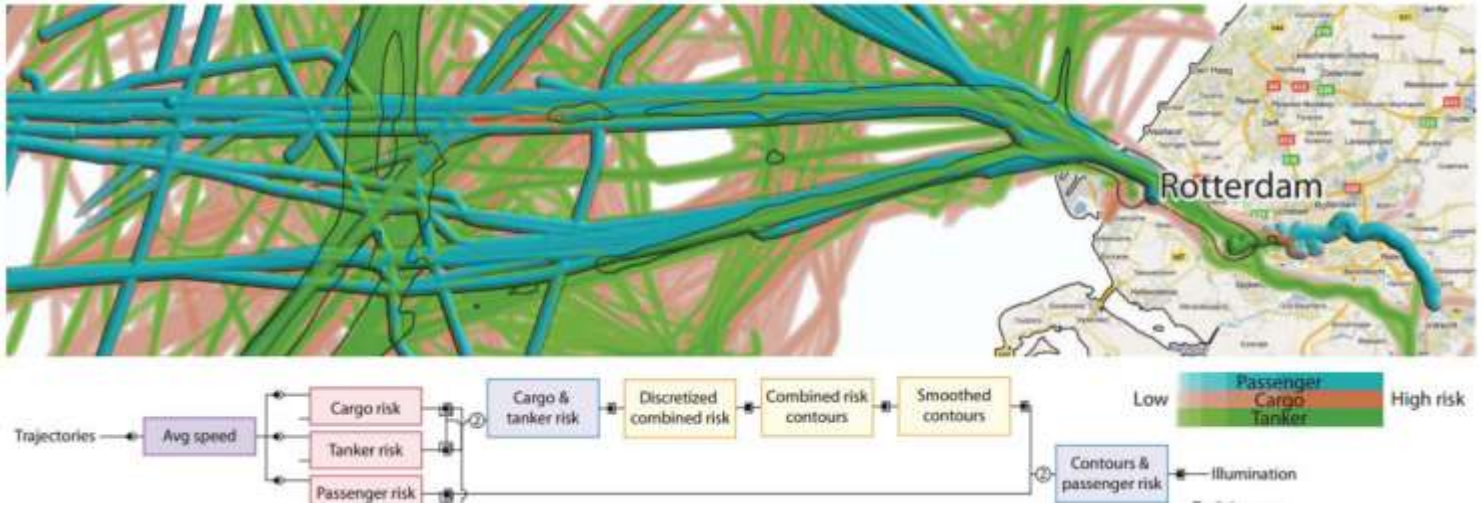
技术5：数据可视化技术



(a) 法国 1864 年红酒出口



(b) 拿破仑 1812 年进攻俄罗斯



技术5：数据可视化技术

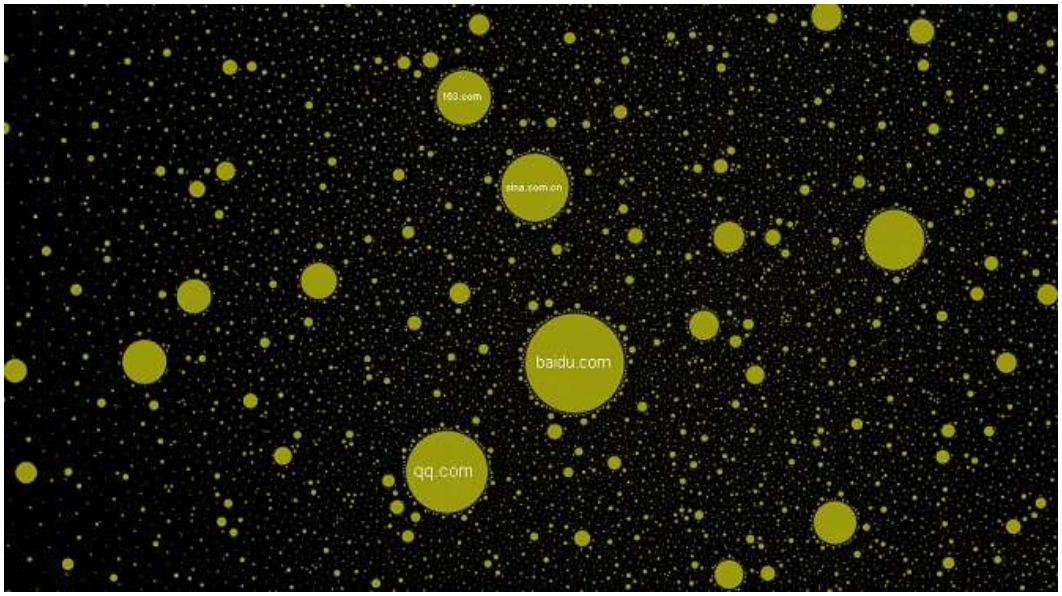
- **全球黑客活动**

安全供应商Norse打造了一张能够反映全球范围内黑客攻击频率的地图，利用Norse 的“蜜罐”攻击陷阱显示出所有实时渗透攻击活动。

技术5：数据可视化技术

- **互联网地图**

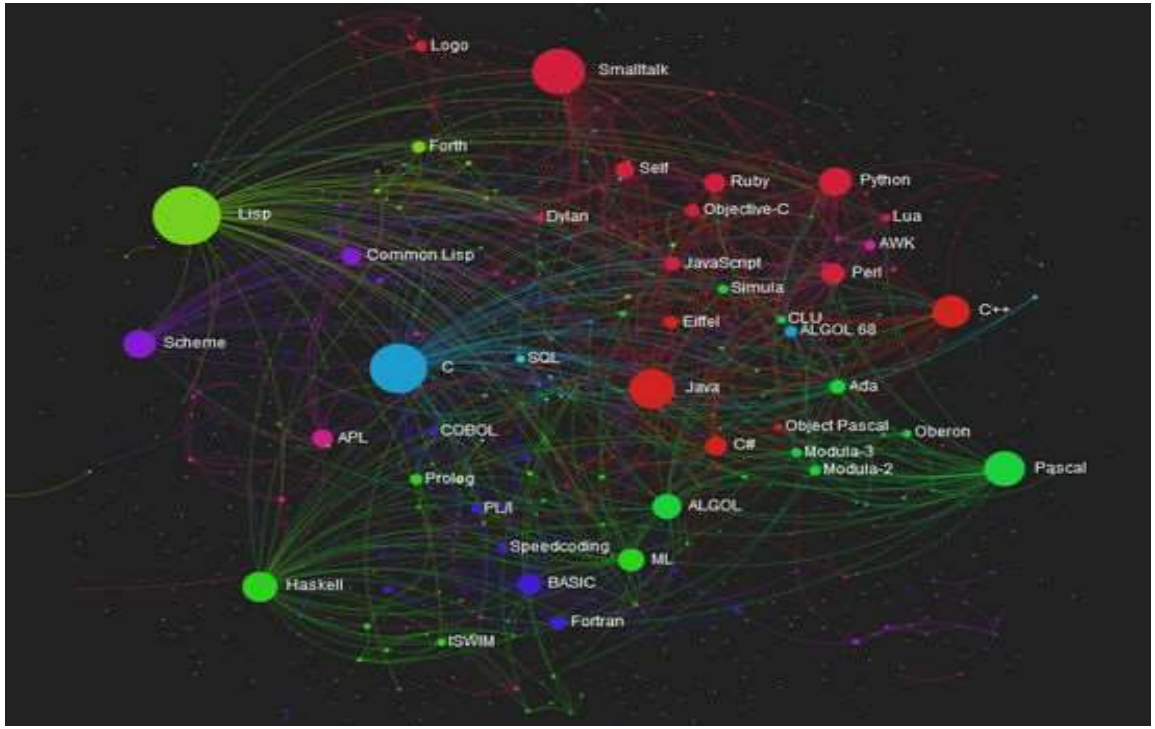
俄罗斯工程师 **Ruslan Enikeev** 将全球 **196** 个国家的 **35** 万个网站数据整合起来，并根据 **200** 多万个网站链接将这些“星球”通过关系链联系起来，每一个“星球”的大小根据其网站流量来决定，而“星球”之间的距离远近则根据链接出现的频率、强度和用户跳转时创建的链接来确定



技术5：数据可视化技术

- **编程语言之间的影响力关系图**

Ramio Gómez利用来自Freebase上的编程语言维护表里的数据，绘制了编程语言之间的影响力关系图



技术5：数据可视化技术

- **全国人口迁徙图**

央视与百度合作，启用百度地图定位可视化大数据播报春节期间全国人口迁徙情况

技术5：数据可视化技术

- **美国股市图**
  - 周期时间
    - 按周期进行堆叠排列

2008年10月金融危机爆发前后美国股市的激烈状况

2006年-2009年美国道琼斯股票指数
横轴表示月份，纵轴表示年份。

纲要

一 计算技术发展

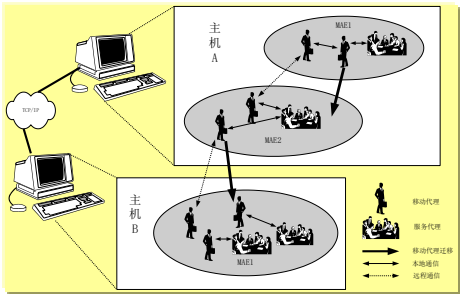
二 新型网络计算

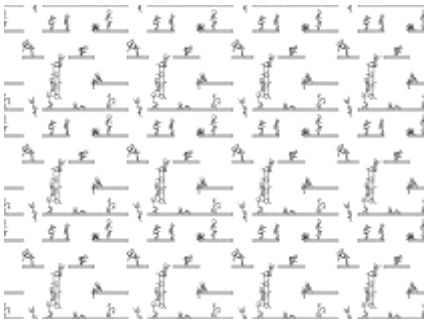
三 主流开发平台

四 其它新型技术

平台1：多智能体平台

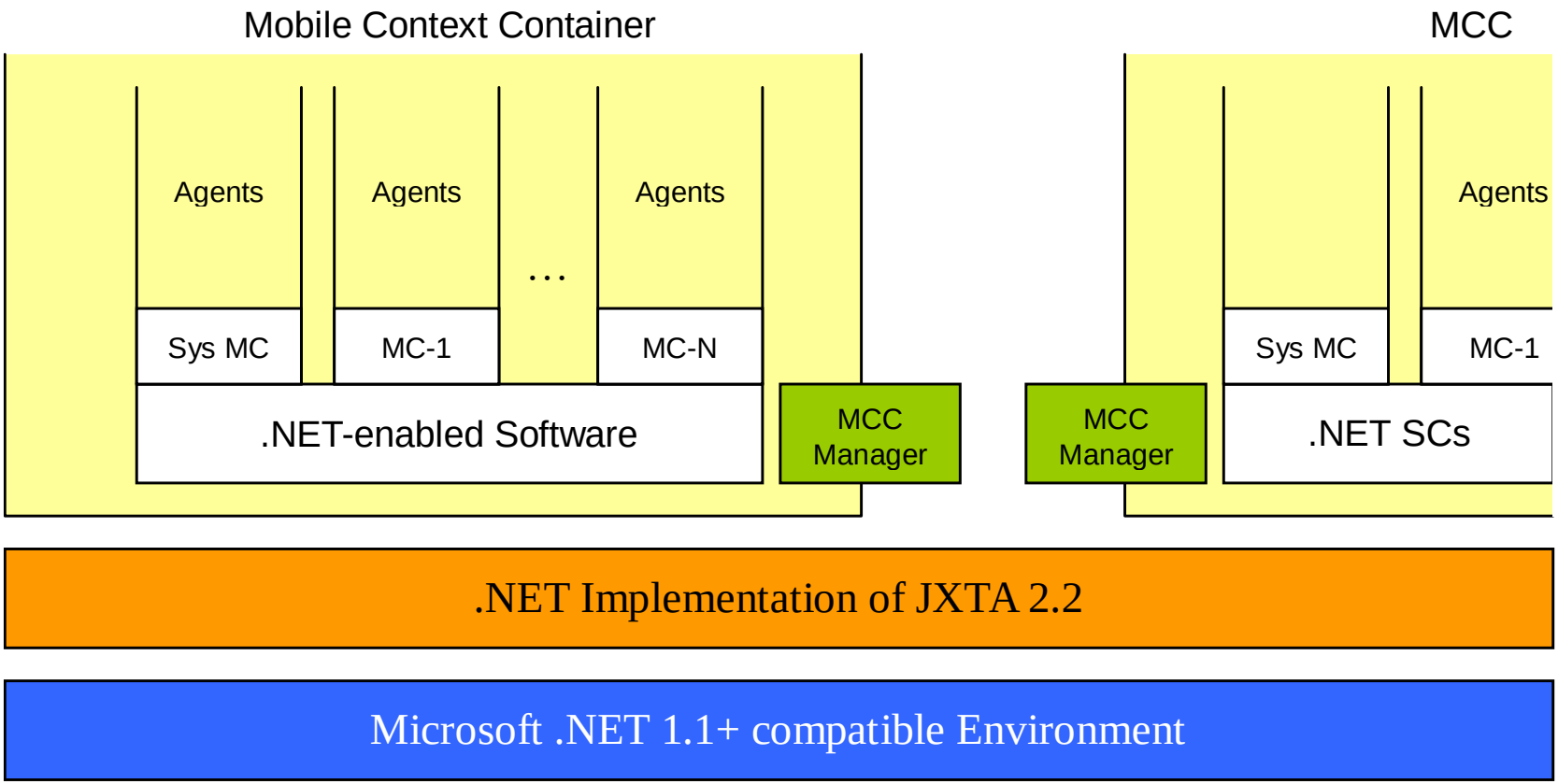
- 美国西北大学网络学习和计算机建模中心的NetLogo。
- 麻省理工学院多媒体实验室的StarLogo
- 芝加哥大学社会科学计算实验室开发研制的Repast
- 爱荷华州立大学的McFadzean、Stewart和Tesfatsion开发的TNG Lab
- 意大利都灵大学Pietro Terna开发的企业仿真项目jES
- 美国布鲁金斯研究所Miles T. Parker开发的Ascape
- 美国桑塔费研究所的Swarm
- 德国IKV++公司的Grasshopper
- 意大利电信实验室的JADE



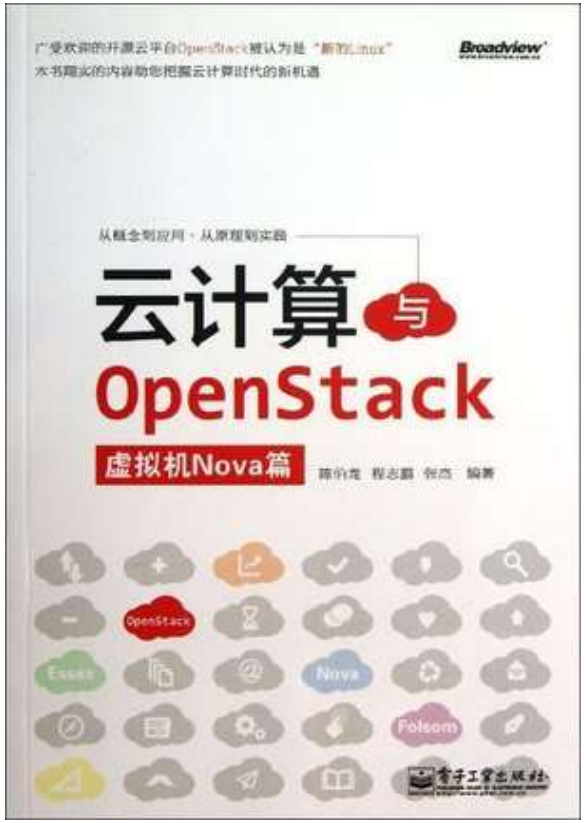


平台1：多智能体平台

- MAP



平台2：开源云平台



云计算与OpenStack
陈伯龙、程志鹏、张杰著
电子工业出版社

OpenStack实战指南
黄凯、毛伟杰、顾俊杰著
机械工业出版社

OpenStack部署实践
张子凡 著
人民邮电出版社

平台2：开源云平台

- OpenStack是由Rackspace和NASA（美国航空航天局）共同发起的开源项目
- 源代码来自于NASA的Nova和Rackspace分布式云存储Swift项目。

平台2：开源云平台

OpenStack是“山寨”的亚马逊AWS。
OpenStack定位是AWS的开源实现。许多组件与AWS基本对应
Nova EC2
Swift S3
Cinder EBS云硬盘
Keystone IAM认证

平台2：开源云平台

OpenStack社区拥有超过150家企业及1500位开发者

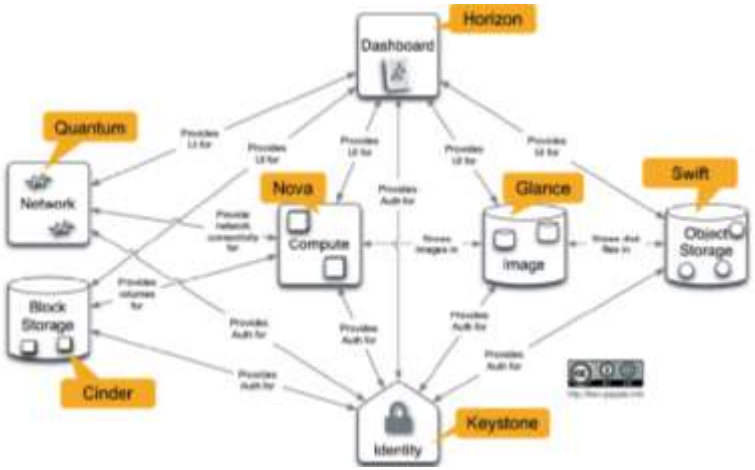
平台2：开源云平台

OpenStack中国应用单位

平台2：开源云平台

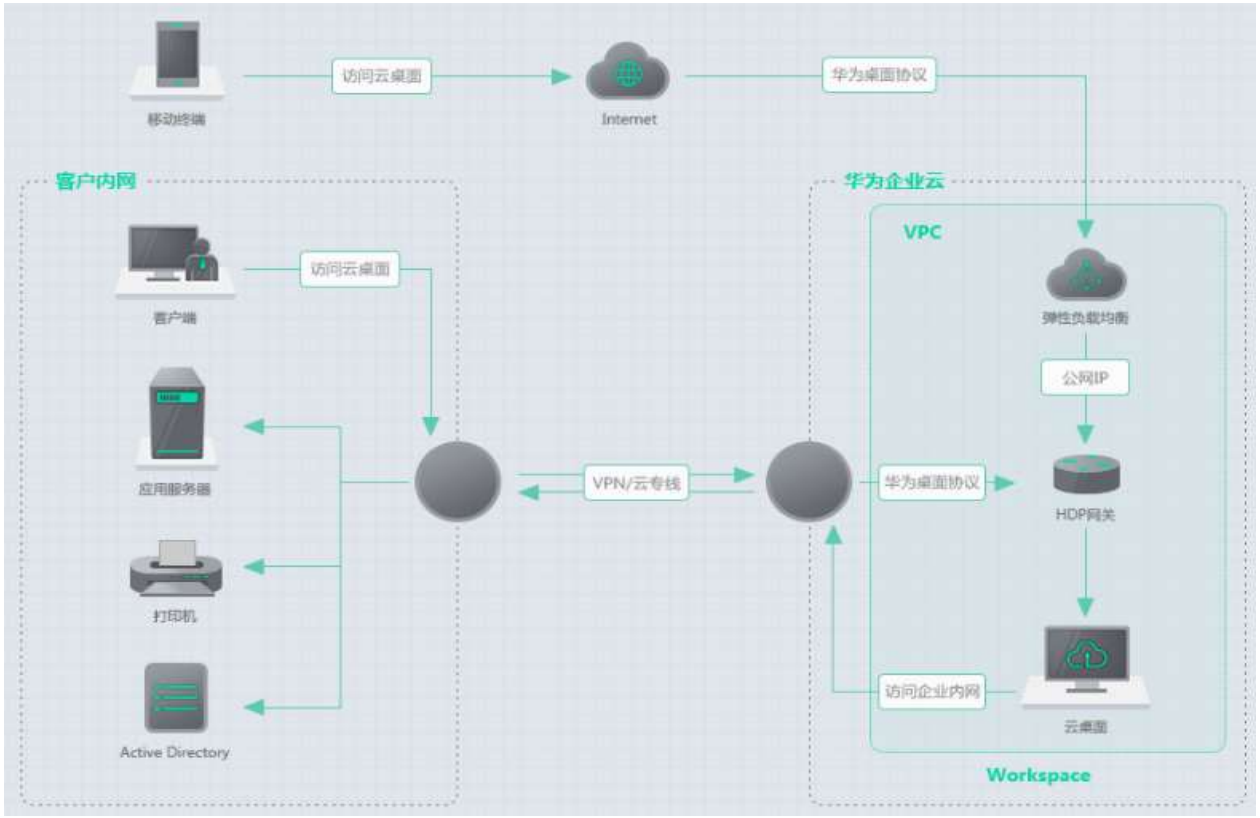
OpenStack组件

- **OpenStack是一个开源软件集合。**
- **OpenStack核心项目：**
  - **Nova计算服务（Compute as a Service）**
  - **Neutron网络服务（Networking as a Service）**
  - **Swift对象存储服务（Object Storage as a Service）**
  - **Cinder块存储服务（Block Storage as a Service ）**
  - **Glance镜像服务（Image as a Service）**
  - **Keystone认证服务（Identity as a Service）**
  - **Horizon仪表盘服务（Dashboard as a Service）**
- **OpenStack管理数据中心的资源：**
  - **计算资源**
  - **网络资源**
  - **存储资源**



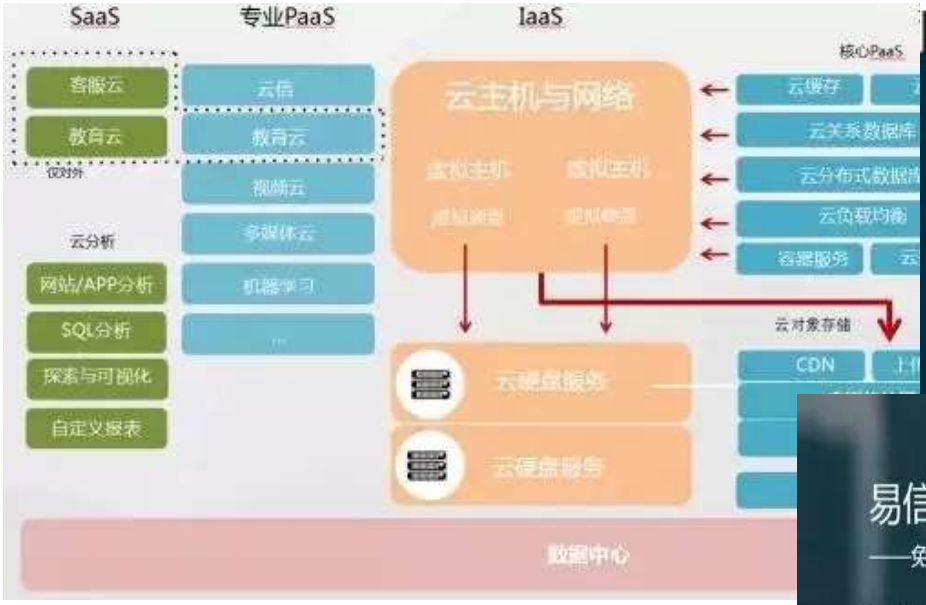
平台2：开源云平台

基于OpenStack的华为云桌面



平台2：开源云平台

基于OpenStack的网易云



平台3：批量大数据处理平台

Hadoop

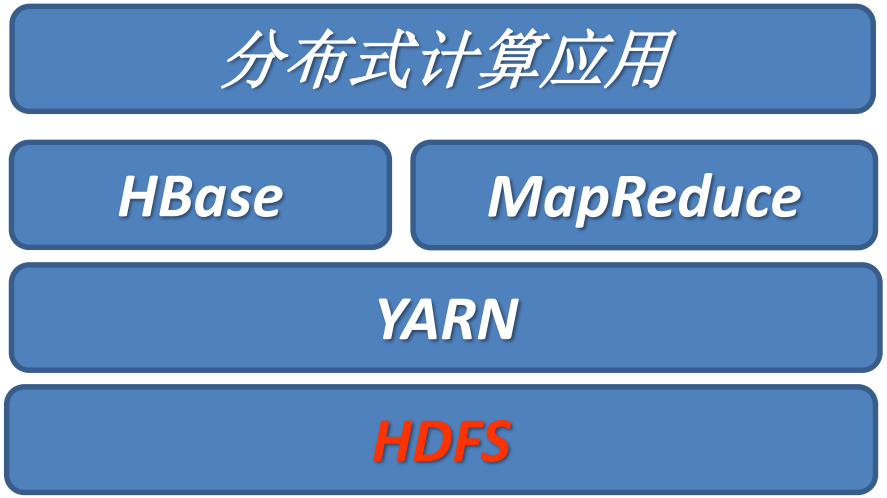
- Apache基金会下开源分布式架构系统，部署在大规模服务器集群上，基于数据中心提供可信赖的计算能力和存储能力。
- 广泛应用于包括Baidu, FaceBook,中国移动、网易、淘宝、腾讯、金山和华为等公司，通常情况下这些机群包括数以千计的服务器和数以万计的CPU。

The Apache Software Foundation
http://www.apache.org/

平台3：批量大数据处理平台

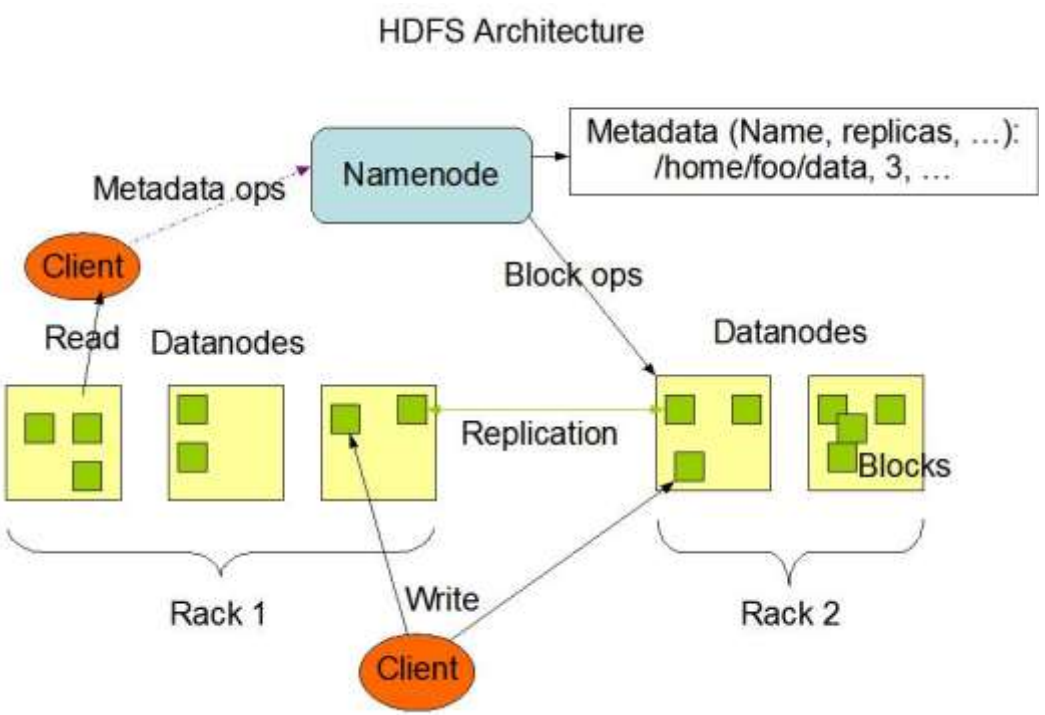
Hadoop

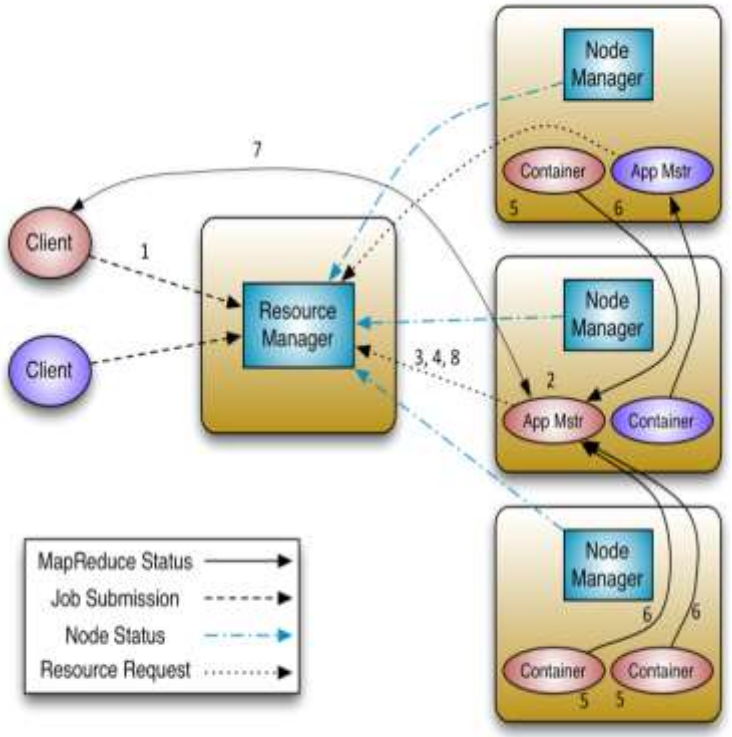
- HDFS 分布式文件系统
- MapReduce 并行计算模型
- Hbase 数据库
- YARN 分布式资源管理与调度系统（2.x）



平台3：批量大数据处理平台

Hadoop





平台4：基于内存计算的大数据平台

Spark

- 2009年诞生于加州大学伯克利分校AMPLab
- 2010年开源
- 2013年进入Apache孵化器
- 基于内存计算的大数据并行计算框架
- 提高了在大数据环境下数据处理的实时性
- 保证了高容错性和高可伸缩性
- 可部署在可大规模廉价服务器集群上

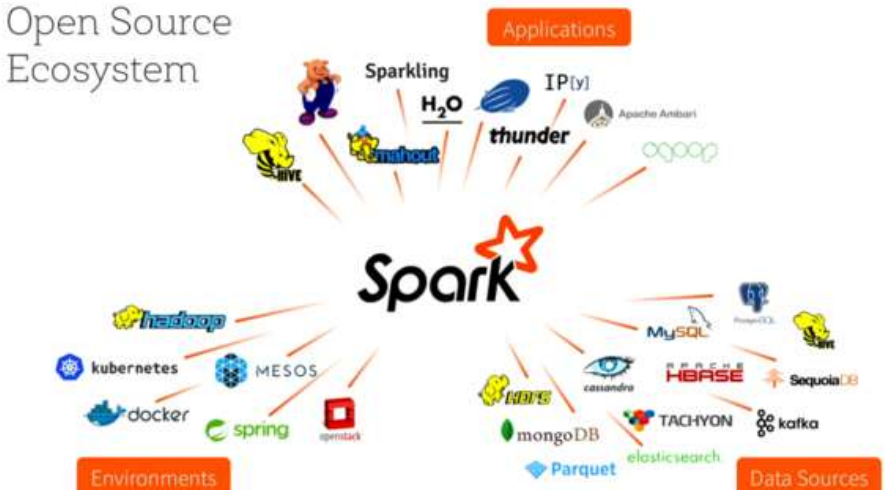
Matai zaharia

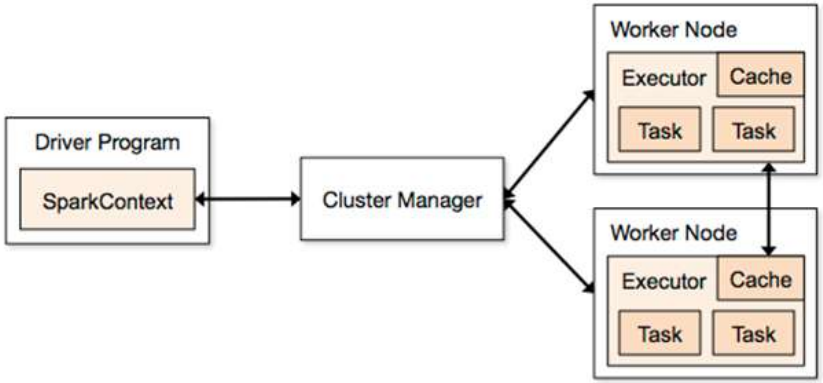
Spark Lightning-fast cluster computing

平台4：基于内存计算的大数据平台

Spark

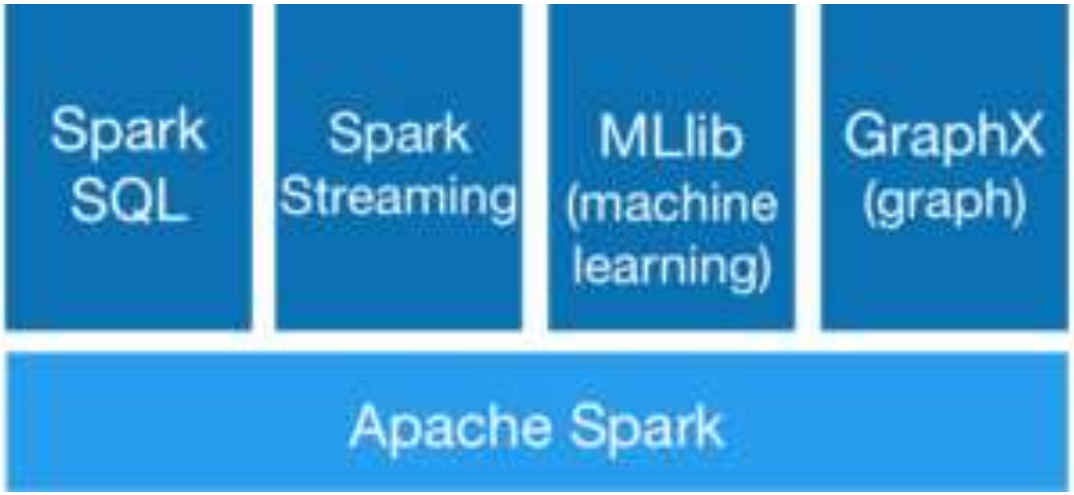
- Spark全面兼容Hadoop的数据持久层，比如HBASE、HDFS、HIVE。从而让把计算任务从原来的MapReduce计算任务迁移到Spark中更加简单。
- 目前Spark的工业应用在国内已经大范围落地。包括BAT在内的一众互联网公司都建立了自己的Spark集群

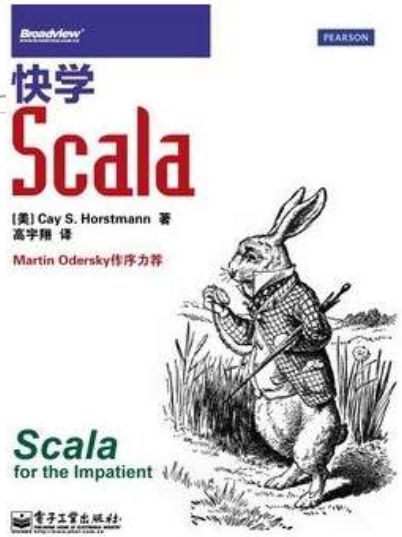




平台4：基于内存计算的大数据平台

Spark





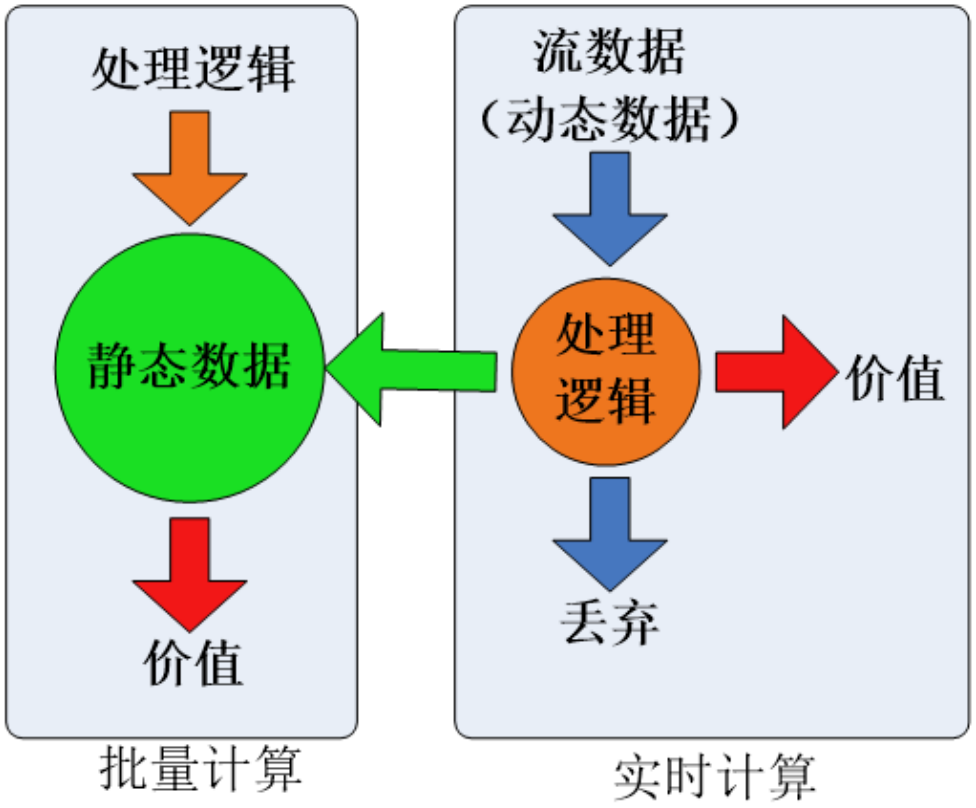
平台5：基于流计算的实时大数据平台

静态数据和流数据

- 近年来，在Web应用、网络监控、传感监测等领域，兴起了一种新的数据密集型应用——流数据，即数据以大量、快速、时变的流形式持续到达
- 流数据具有如下特征：
  - 数据快速持续到达，潜在大小也许是无穷无尽的
  - 数据来源众多，格式复杂
  - 数据量大，但是不十分关注存储，一旦经过处理，要么被丢弃，要么被归档存储
  - 注重数据的整体价值，不过分关注个别数据
  - 数据顺序颠倒，或者不完整，系统无法控制将要处理的新到达的数据元素的顺序

平台5：基于流计算的实时大数据平台

静态数据和流数据



平台5：基于流计算的实时大数据平台

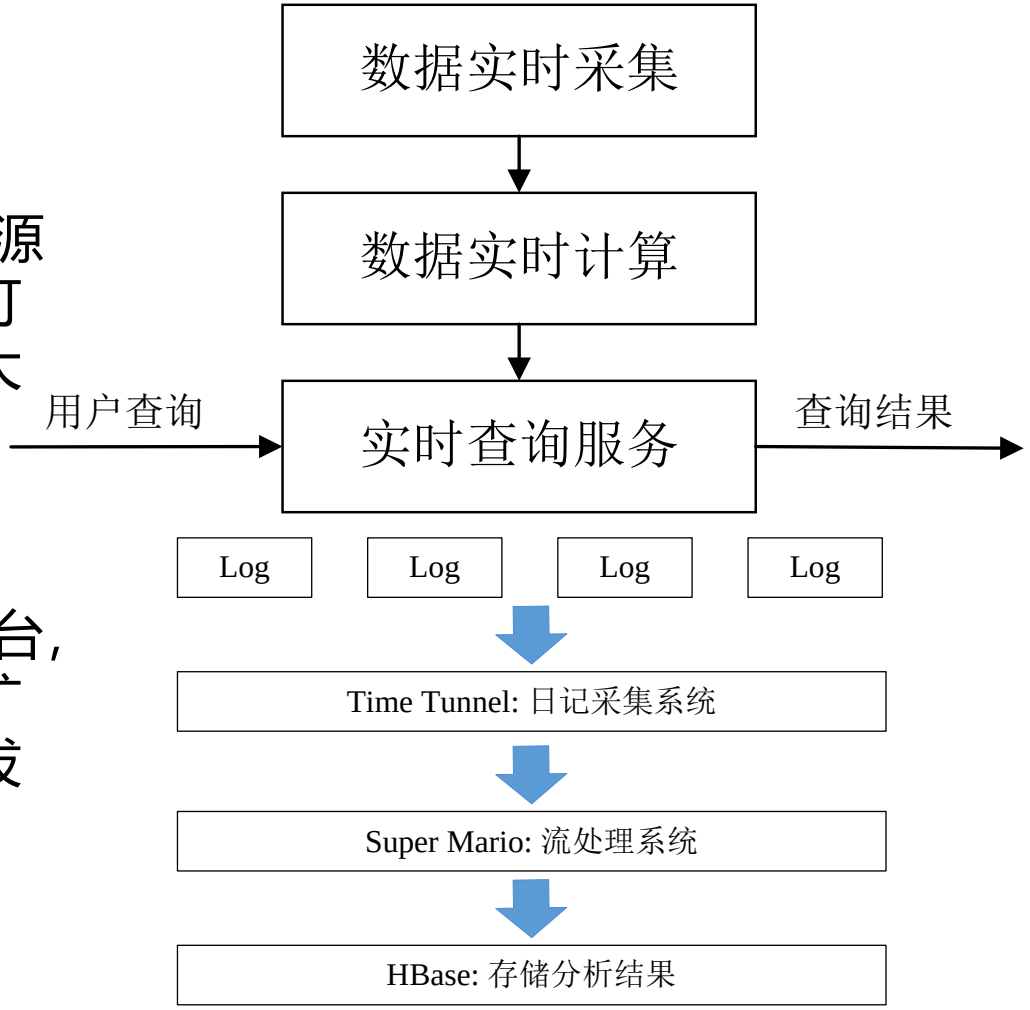
流计算

- 流计算秉承一个基本理念，即**数据的价值随着时间的流逝而降低**。因此，当事件出现时就应该立即进行处理，而不是缓存起来进行批量处理。为了及时处理流数据，就需要一个低延迟、可扩展、高可靠的处理引擎
- 对于一个流计算系统来说，它应达到如下需求：
  - 高性能：处理大数据的基本要求，如每秒处理几十万条数据
  - 海量式：支持TB级甚至是PB级的数据规模
  - 实时性：保证较低的延迟时间，达到秒/毫秒级别
  - 分布式：支持分布式架构
  - 易用性：能够快速进行开发和部署
  - 可靠性：能可靠地处理流数据

平台5：基于流计算的实时大数据平台

流计算

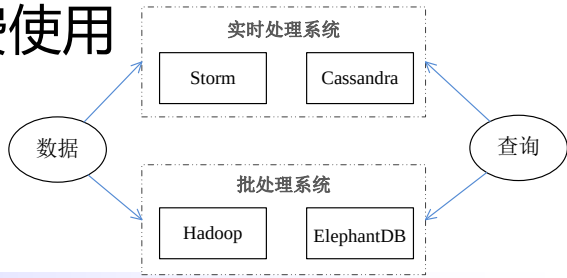
- 开源流计算框架：
  - Twitter Storm：免费、开源的分布式实时计算系统，可简单、高效、可靠地处理大量的流数据
  - Yahoo! S4（Simple Scalable Streaming System）：开源流计算平台，是通用的、分布式的、可扩展的、分区容错的、可插拔的流式系统



平台5：基于流计算的实时大数据平台

Storm

- Storm具有以下主要特点：
  - 整合性：Storm可方便地与队列系统和数据库系统进行整合
  - 简易的API：Storm的API在使用上即简单又方便
  - 可扩展性：Storm的并行特性使其可以运行在分布式集群中
  - 容错性：Storm可自动进行故障节点的重启、任务的重新分配
  - 可靠的消息处理：Storm保证每个消息都能完整处理
  - 支持各种编程语言：Storm支持使用各种编程语言来定义任务
  - 快速部署：Storm可以快速进行部署和使用
  - 免费、开源：Storm是一款开源框架，可以免费使用



平台5：基于流计算的实时大数据平台

Storm

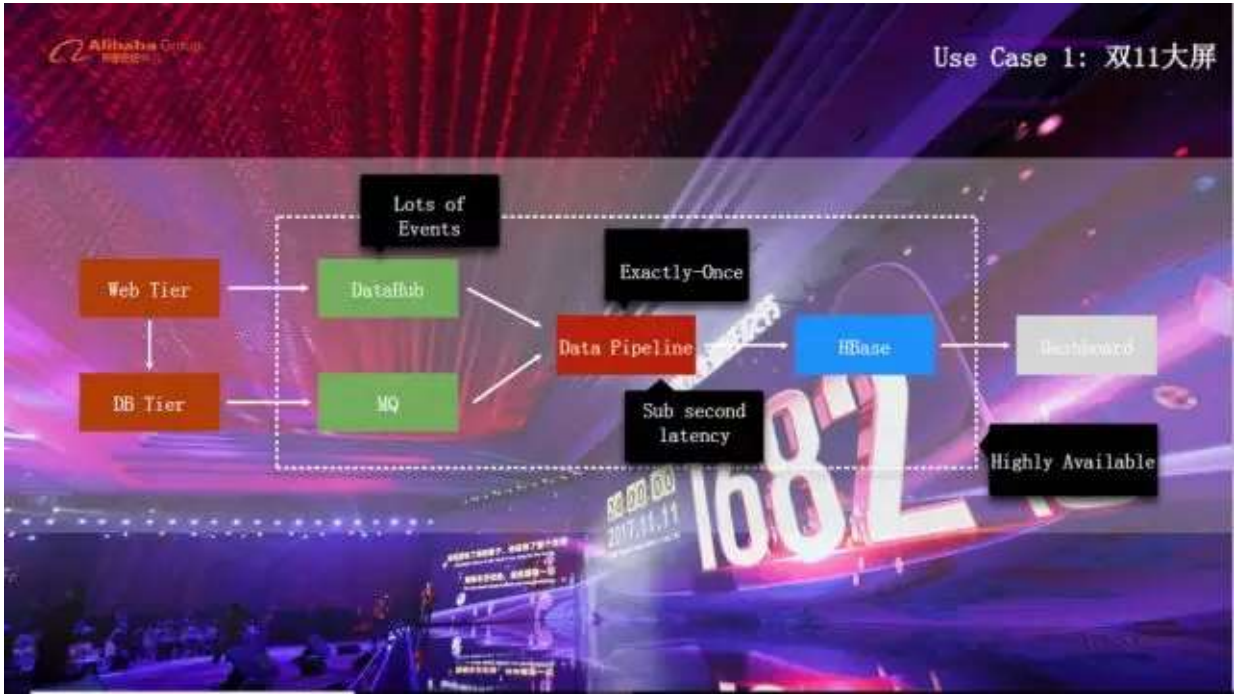
使用Storm的公司和项目

平台5：基于流计算的实时大数据平台

阿里实时计算平台

每年双11阿里都会聚合有价值的数据展现给媒体，GMV大屏是其中之一。

整个GMV大屏是非常典型的实时计算，每条交易数据经过聚合展现在大屏之上。



平台5：基于流计算的实时大数据平台

饿了么实时计算平台

用户的每一步有价值的操作(包括：搜索、点击、浏览、购买、收藏等)，都将实时、智能地影响搜索结果排序，从而显著提升用户搜索体验、搜索转化率。

平台5：基于流计算的实时大数据平台

斗鱼实时计算平台

平台5：基于流计算的实时大数据平台

滴滴实时计算平台

对实际线上业务的各种业务变化，比如订单呼叫量、订单应答量、订单成交量、实时应答率、平均接驾距离等，要将这些数据实时的反馈到运营团队的**dashboard**上。

平台5：基于流计算的实时大数据平台

滴滴实时计算平台

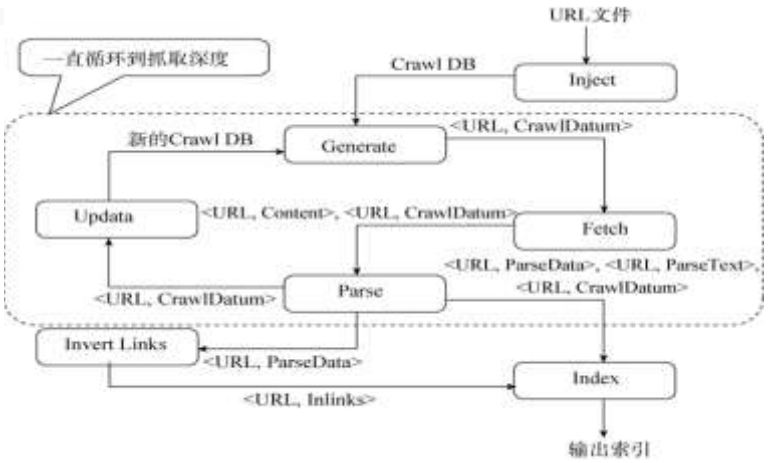
对实际线上业务的各种业务变化，比如订单呼叫量、订单应答量、订单成交量、实时应答率、平均接驾距离等，要将这些数据实时的反馈到运营团队的**dashboard**上。

平台6：网络大数据采集工具

	平台	开发语言	优点	缺点	社区活跃程度
Larbin	Linux	C++	性能好，稳定	没有删除功能，排重会误判	★★★★★
Nutch	Windows/linux	java	Nutch 和 Lucene ，Hadoop结合的很好	不太稳定	★★★★★
Heritrix	Windows/linux	Java	高度可扩展性，性能优秀，对抓取的高度控制性，功能齐全	对中文支持不够，没有很好的容错性以及回复机制	★★★★★
WebSPHINX	Windows/linux	Java	采集效率高，接口清晰，易于扩展	不再被维护了	★★★☆☆
Mercator	Windows/linux	Java	可伸缩、可扩展	资料少	★★☆☆☆
PolyBot	Linux	Python/c++	可配置性好	缺点就是直观性太差	★★☆☆☆

平台6：网络大数据采集工具

Nutch



开发语言：**Java**

http://lucene.apache.org/nutch/

简介：

分为两个部分：网络爬虫和查询。

1.网络爬虫的主要作用是从网络上抓取网页数据并建立索引；

2.查询则主要是利用这些索引来检索用户所提交的关键词并产生和返回查找结果。两大部分之间的交汇点是索引，耦合度相对较低。

两个优点：

1.基于**Lucene** 的高效索引和检索功能；

2.基于**Apache** 开源项目**Hadoop** 实现类似**Google** 的分布式文件系统，它大量使用**Google** 的**MapReduce** 机制。

平台6：网络大数据采集工具

Scrapy

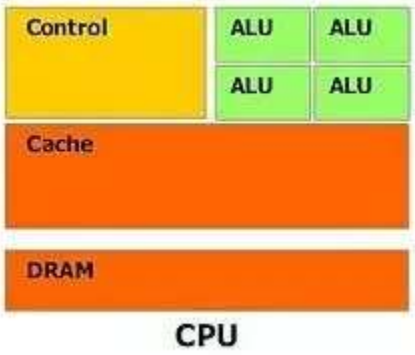
开发语言：**Python**

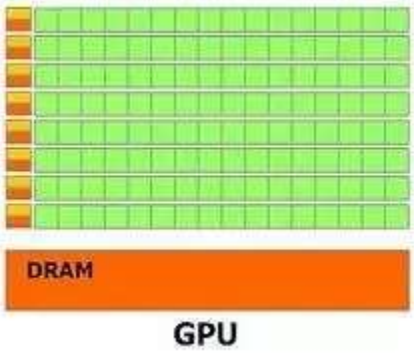
简介：

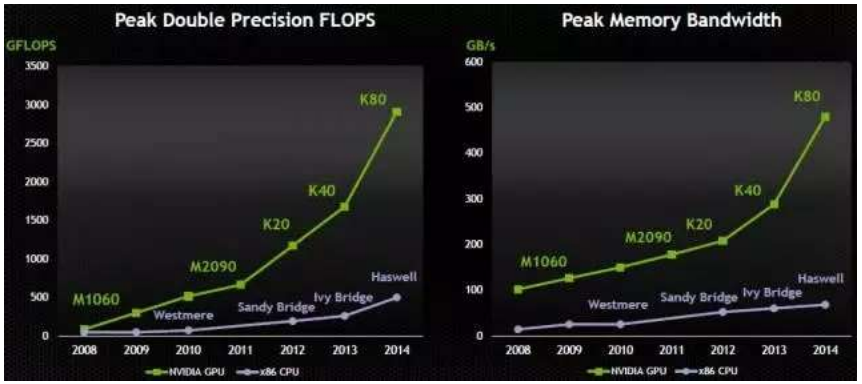
- 快速,高层次的屏幕抓取和web抓取框架，用于抓取web站点并从页面中提取结构化的数据。
- 用途广泛，可以用于数据挖掘、监测和自动化测试。
- 吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。
- 提供了多种类型爬虫的基类，如BaseSpider、sitemap爬虫等。
- 最新版本提供了web2.0爬虫的支持。

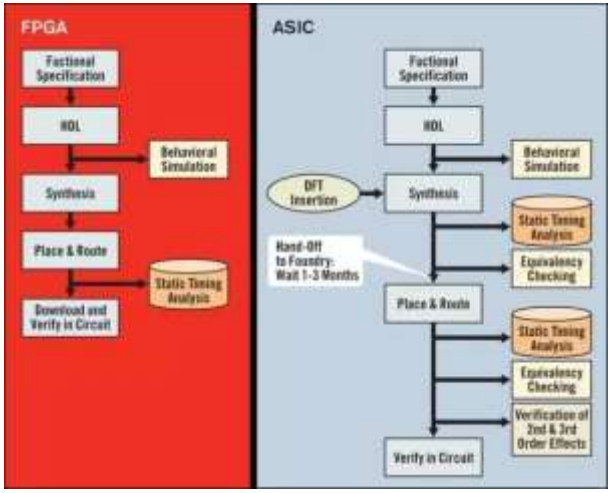
平台7：深度学习开发平台

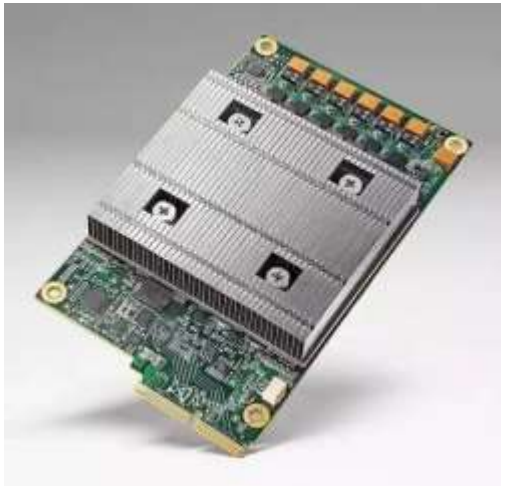
处理芯片











平台7：深度学习开发平台

CPU vs GPU

	# Cores	Clock Speed	Memory	Price
CPU (Intel Core i7-7700k)	4 (8 threads with hyperthreading)	4.4 GHz	Shared with system	$339
CPU (Intel Core i7-6950X)	10 (20 threads with hyperthreading)	3.5 GHz	Shared with system	$1723
GPU (NVIDIA Titan Xp)	3840	1.6 GHz	12 GB GDDR5X	$1200
GPU (NVIDIA GTX 1070)	1920	1.68 GHz	8 GB GDDR5	$399

平台7：深度学习开发平台

Caffe (UC Berkeley) → Caffe2 (Facebook)

Torch (NYU / Facebook) → PyTorch (Facebook)

Theano (U Montreal) → TensorFlow (Google)

Paddle (Baidu)

CNTK (Microsoft)

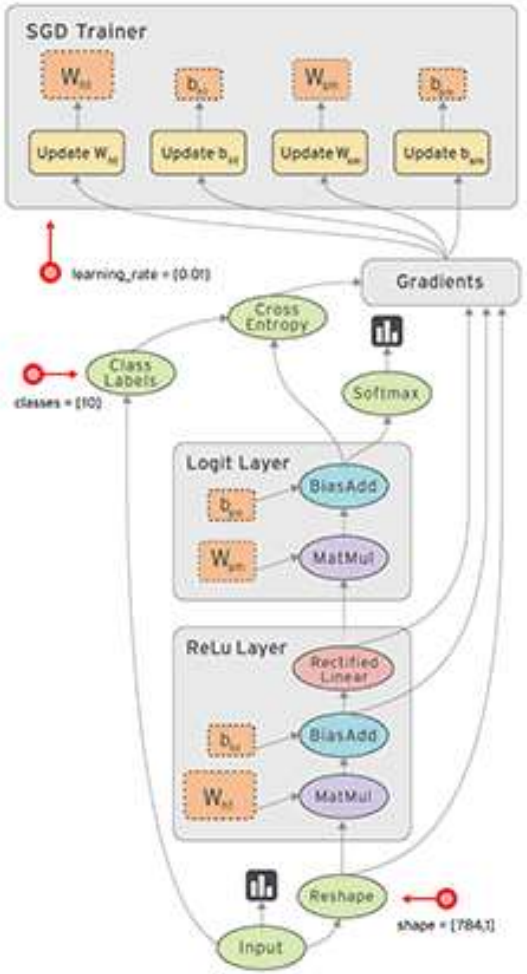
MXNet (Amazon)

Developed by U Washington, CMU, MIT, Hong Kong U, etc but main framework of choice at AWS

平台7：深度学习开发平台

TensorFlow

- 谷歌基于DistBelief进行研发的第二代人工智能学习系统。Tensor（张量）意味着N维数组，Flow（流）意味着基于数据流图的计算，TensorFlow为张量从流图的一端流动到另一端计算过程。
- TensorFlow将复杂的数据结构传输至人工智能神经网中进行分析和处理。
- TensorFlow被用于语音识别或图像识别等多项机器学习和深度学习领域，对2011年开发的深度学习基础架构DistBelief进行了各方面的改进。
- TensorFlow可在小到一部智能手机、大到数千台数据中心服务器的各种设备上运行。



平台7：深度学习开发平台

TensorFlow

环境

- Windows 10 64bit
- GPU：NVIDIA GeForce GTX 1060

Visual Studio 2015

- 安装cuda需要使用VC++编译
- 安装选项中只选择了Visual C++

CUDA 8.0

- 安装CUDA时需要关闭VS2015，并选择自定义安装，取消勾选driver，可以安装成功
- 安装完成后，在cmd中输入nvcc –V可以查看安装的CUDA版本，测试是否安装成功
- 安装完成后，打开VS2015，新建NVDIA项目，直接回建一个示例程序，可以跑通示例就ok

纲要

一 计算技术发展

二 新型网络计算

三 主流开发平台

四 其它新型技术

其它新型技术

- 雾计算与边缘计算
- 对等计算
- 位置计算
- 区块链
- 可信网络计算
- ……

xuxl@njupt.edu.cn

谢 谢 ！