互联网超级云计算平台

方坤

谷歌资深软件工程师

题纲

- 计算的规模与极限
- 云计算平台设计挑战
- 谷歌云计算基础技术
- 谷歌云计算新技术

谷歌的云计算理念

- 满足搜索和数据分析的需要
  - 智能，迅速，规模
- 在现有云计算架构上提供公有服务
  - **Gmail, Calendar, Picasa, Docs, Google Apps,...**
  - **App Engine**
- 探索新的云计算技术
  - 云存储
  - 云计算
  - 任务调度
  - 节约能源

高性能计算的极端

应用	用户数	精确度	可靠度	数据量
科学计算	少	极高	低 -- 中等	**Tera**
股市交易	大量	高	极高	**Gega**
基因排序	少	高	高	**Tera -- Peta**
搜索引擎	大量	中等 -- 高	中等	**Peta**

数据规模 vs 算法精度

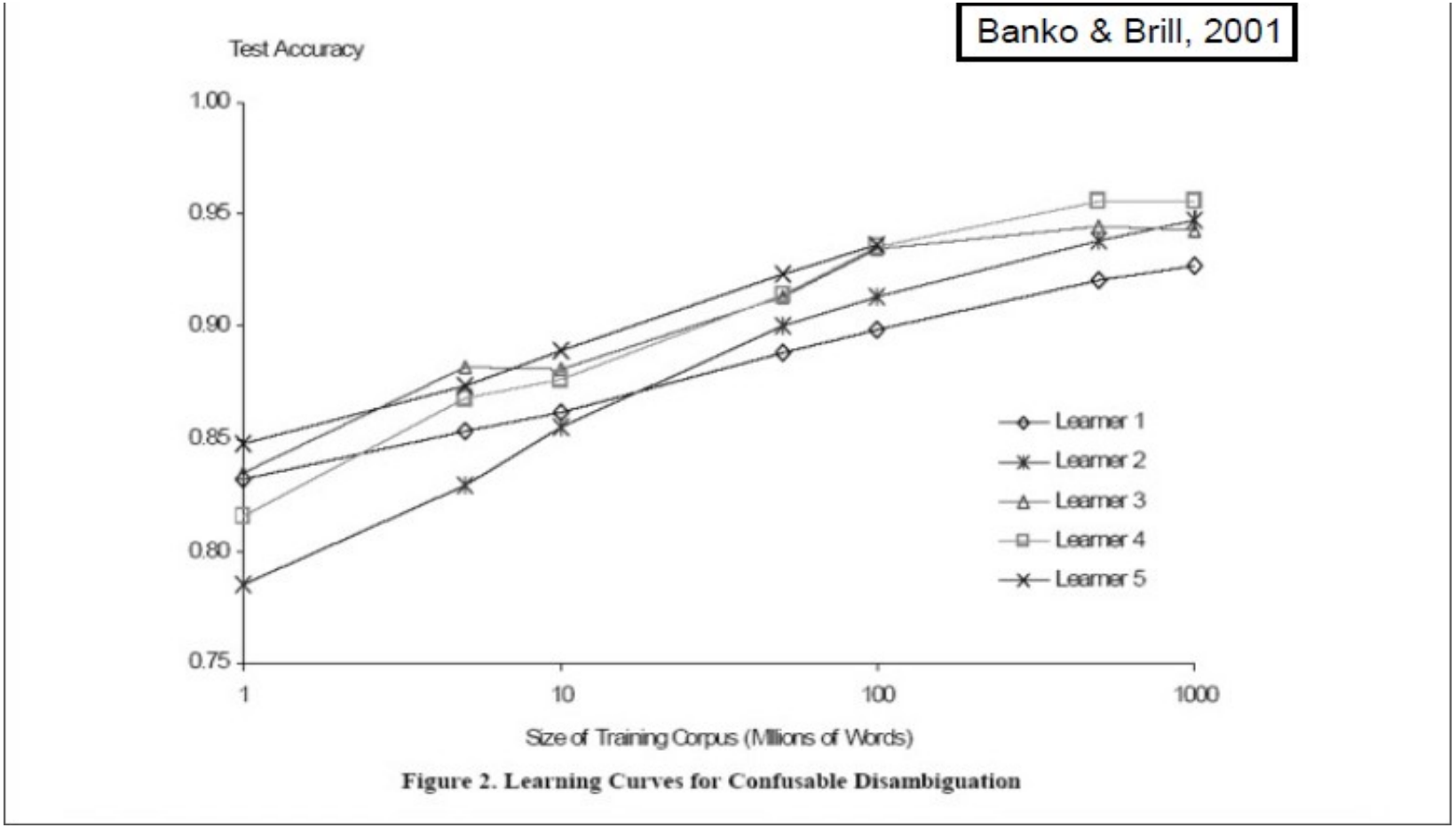


Figure 2. Learning Curves for Confusable Disambiguation

数据规模 **vs** 算法精度

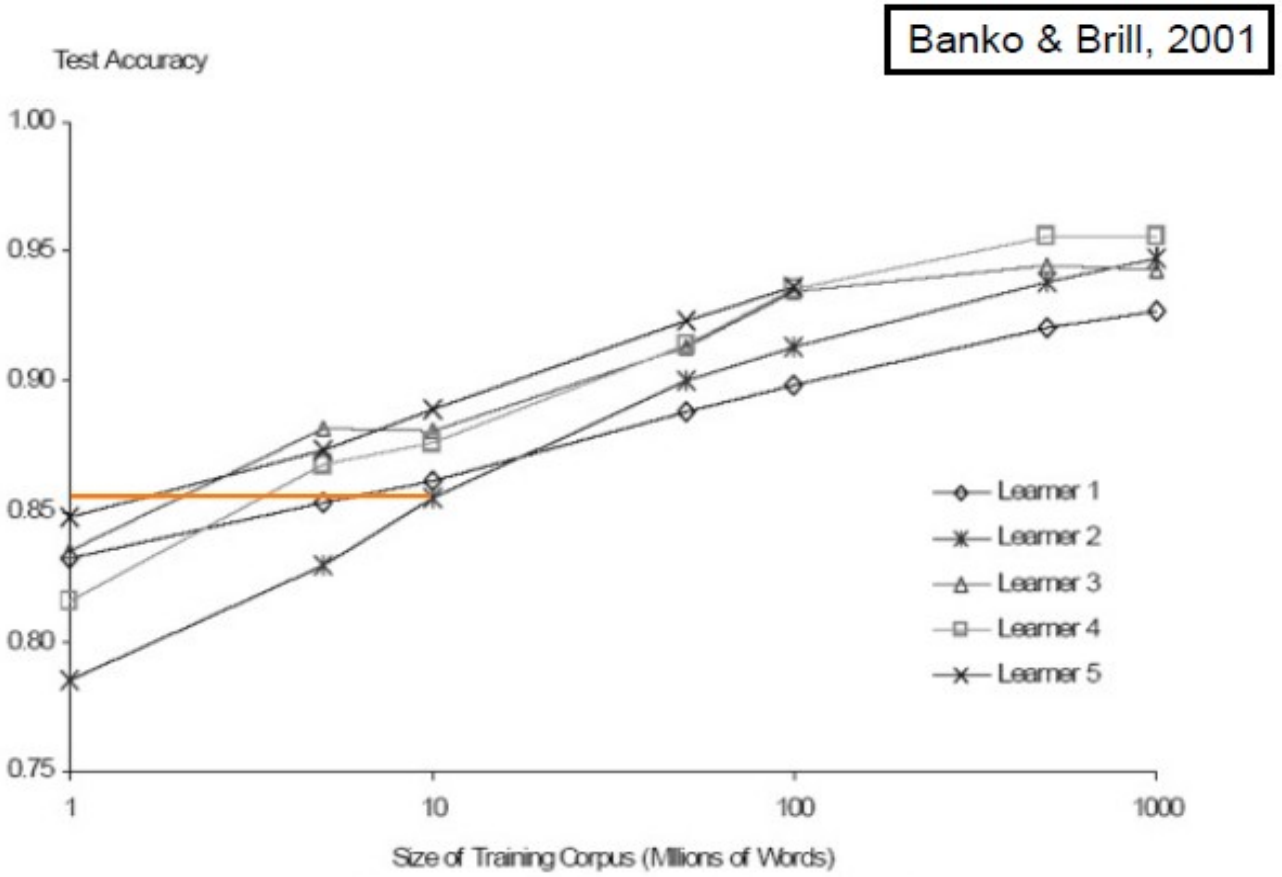


Figure 2. Learning Curves for Confusable Disambiguation

数据规模 vs 算法精度

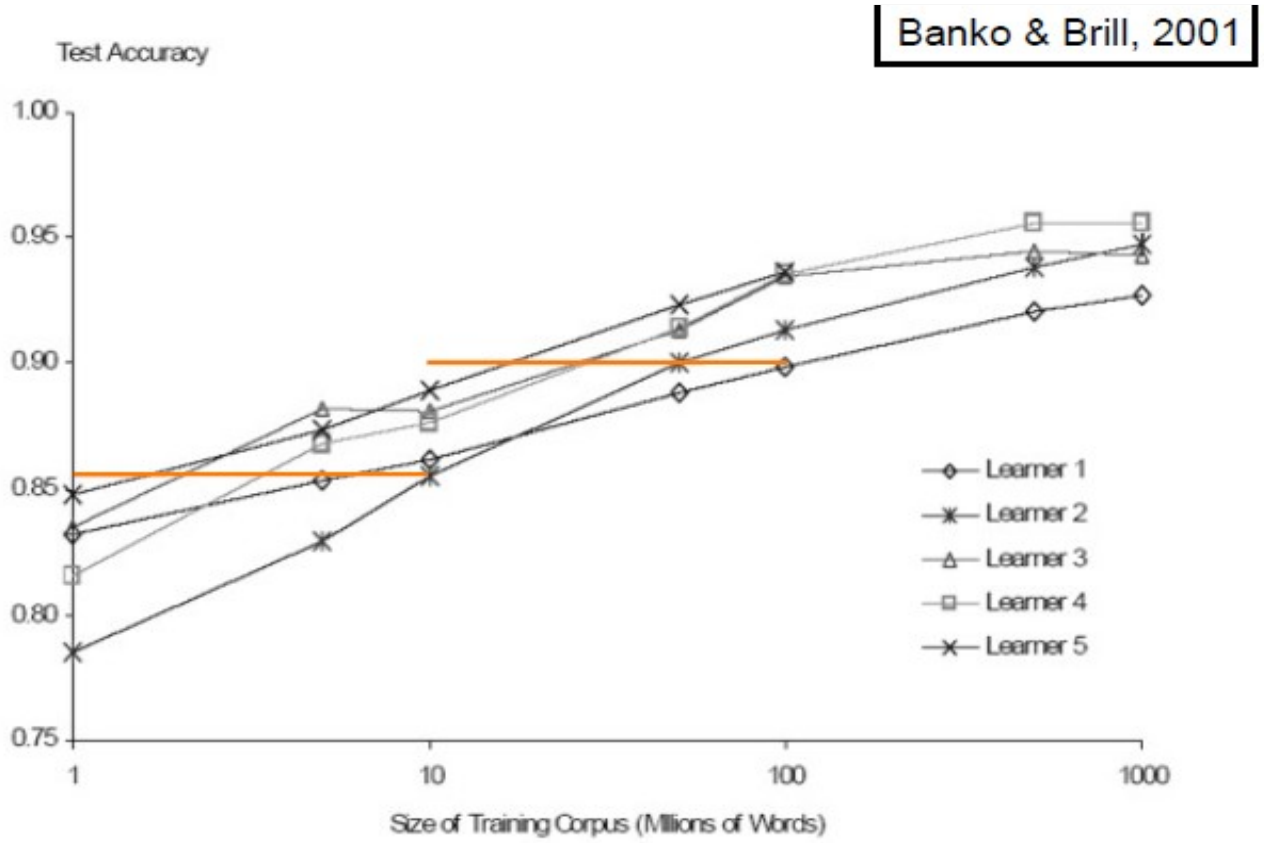


Figure 2. Learning Curves for Confusable Disambiguation

数据规模 vs 算法精度

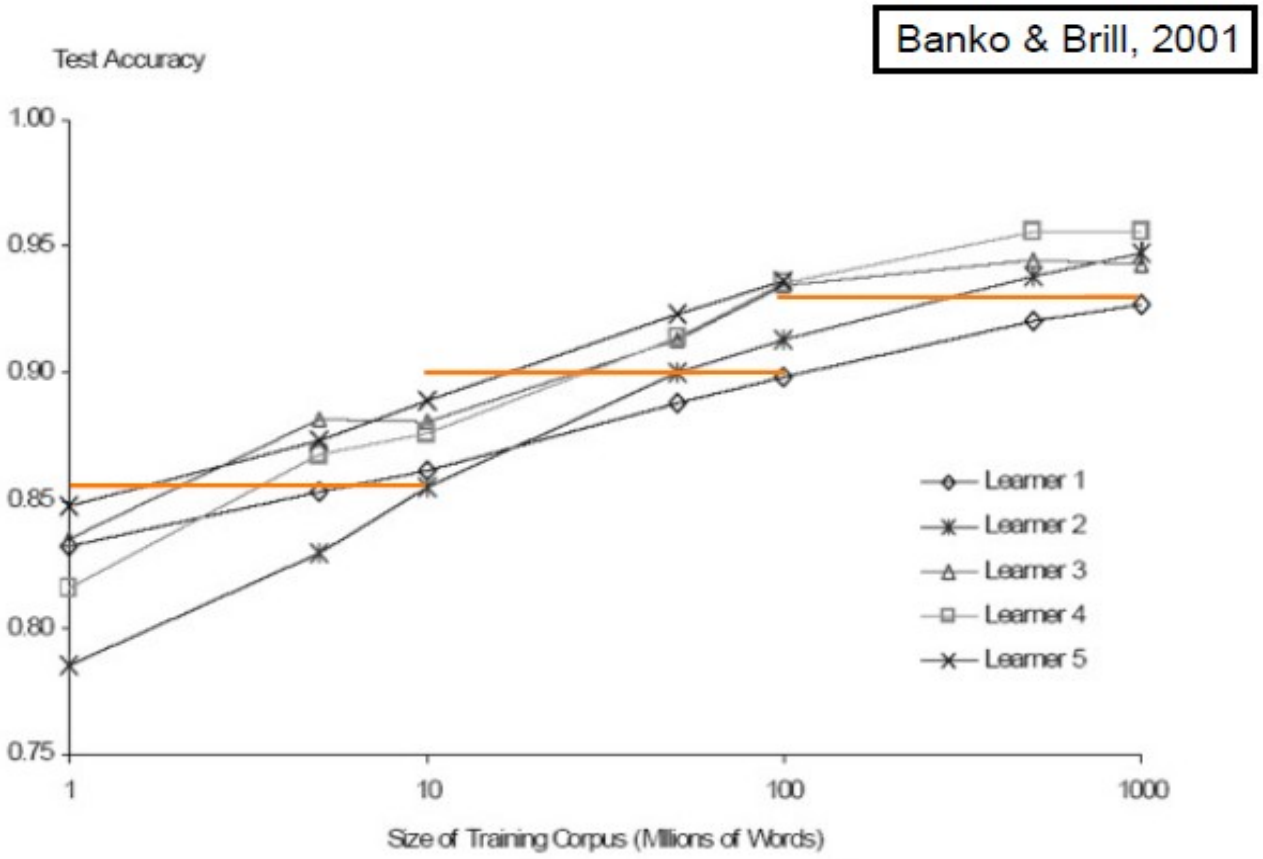


Figure 2. Learning Curves for Confusable Disambiguation

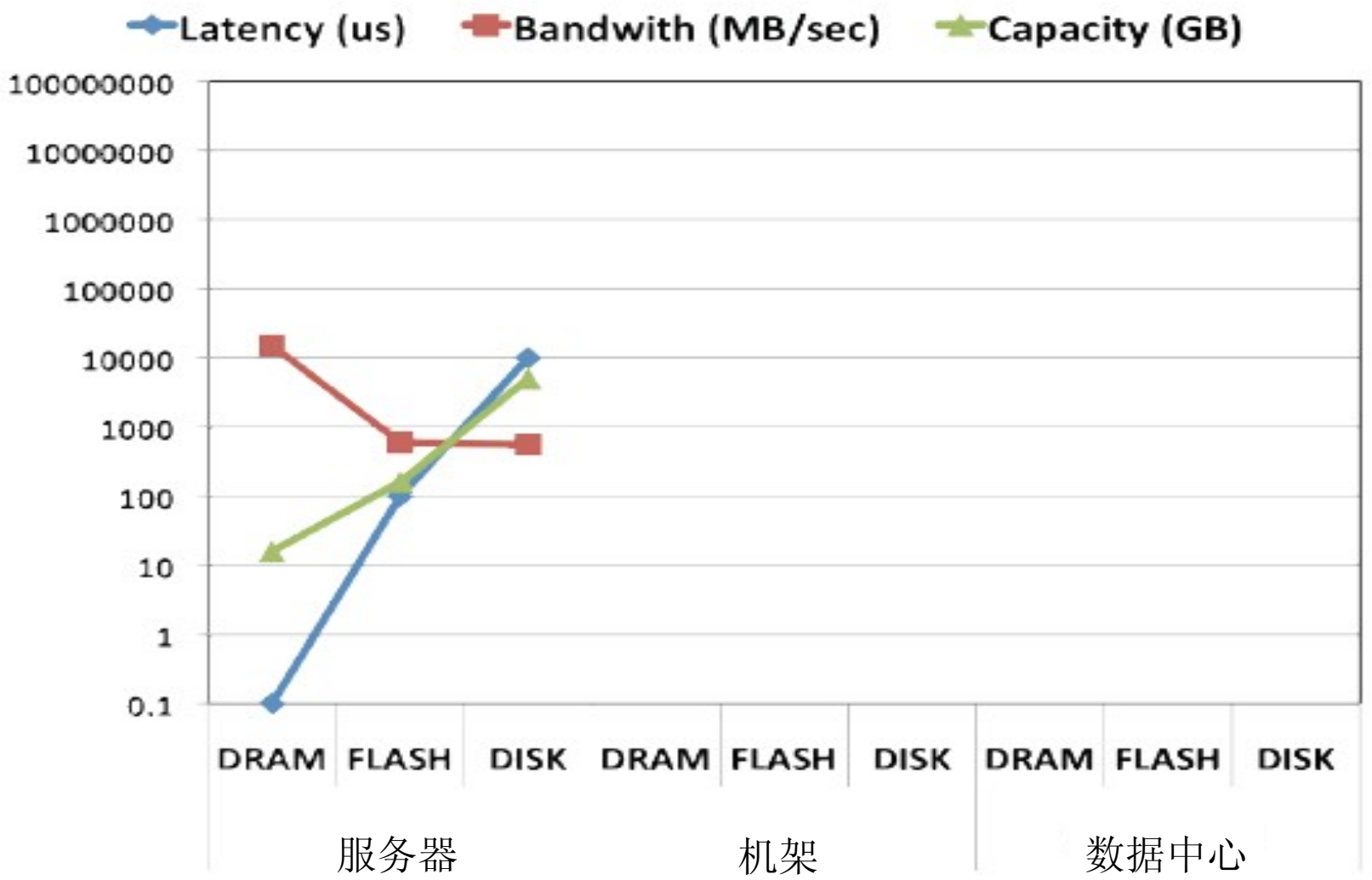
题纲

- 计算的规模与极限
- 云计算平台设计挑战
- 谷歌云计算基础技术
- 谷歌云计算新技术

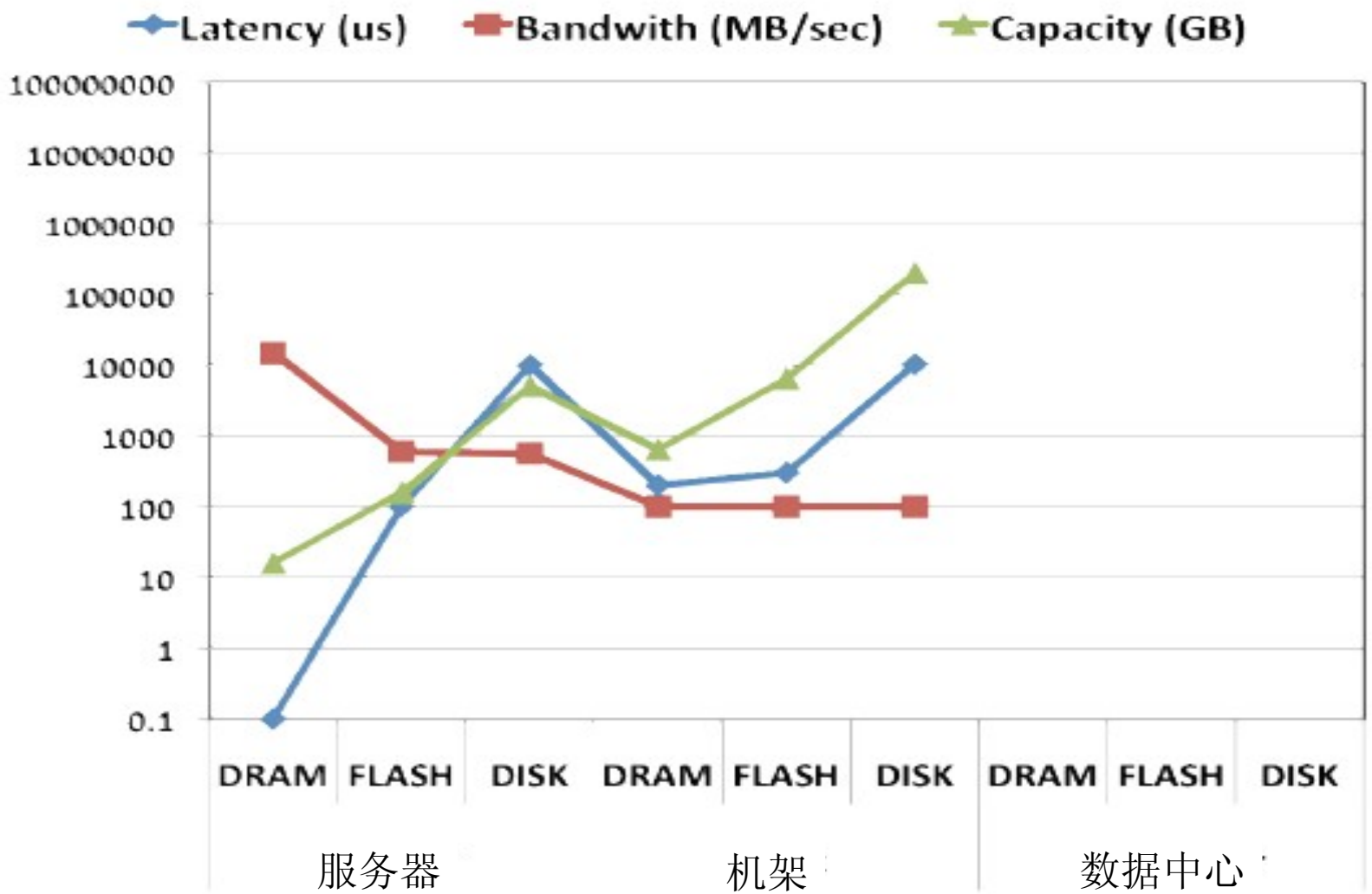
样本平台

- 计算机
  - 16GB DRAM; 160 GB SSD; 5 x 1TB disk
- 计算机机架 (Rack)
  - 40 计算机
  - 48 port Gigabit Ethernet switch
- 数据中心 (Warehouse)
  - 10,000 计算机 (250 racks)
  - 2K port Gigabit Ethernet switch

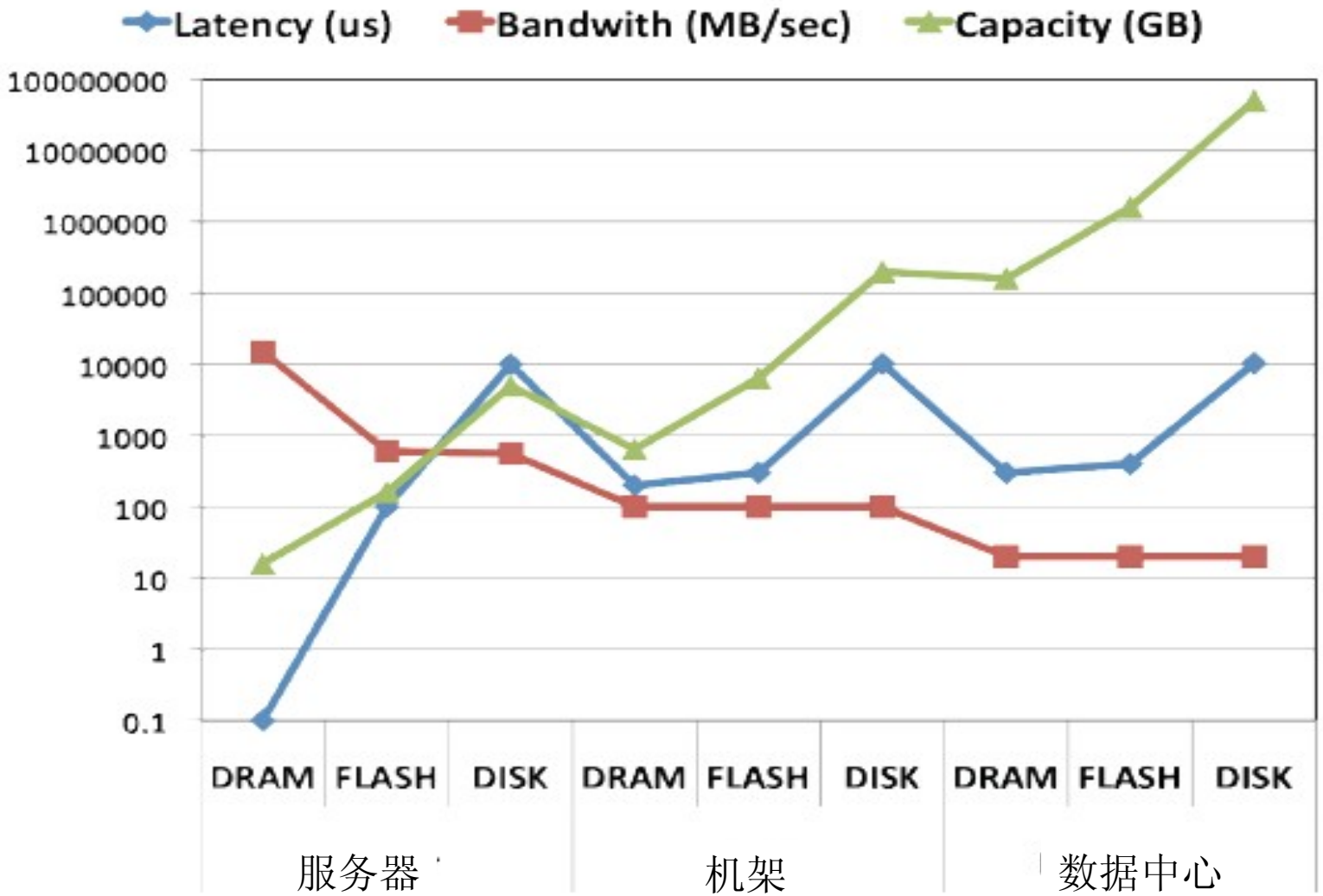
储存 --- 计算机单机


Latency (us)
Bandwith (MB/sec)
Capacity (GB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
0.1
DRAM
FLASH
DISK
DRAM
FLASH
DISK
DRAM
FLASH
DISK
服务器
机架
数据中心

储存 --- 计算机机架


Latency (us)
Bandwith (MB/sec)
Capacity (GB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
0.1
DRAM
FLASH
DISK
DRAM
FLASH
DISK
DRAM
FLASH
DISK
服务器
机架
数据中心

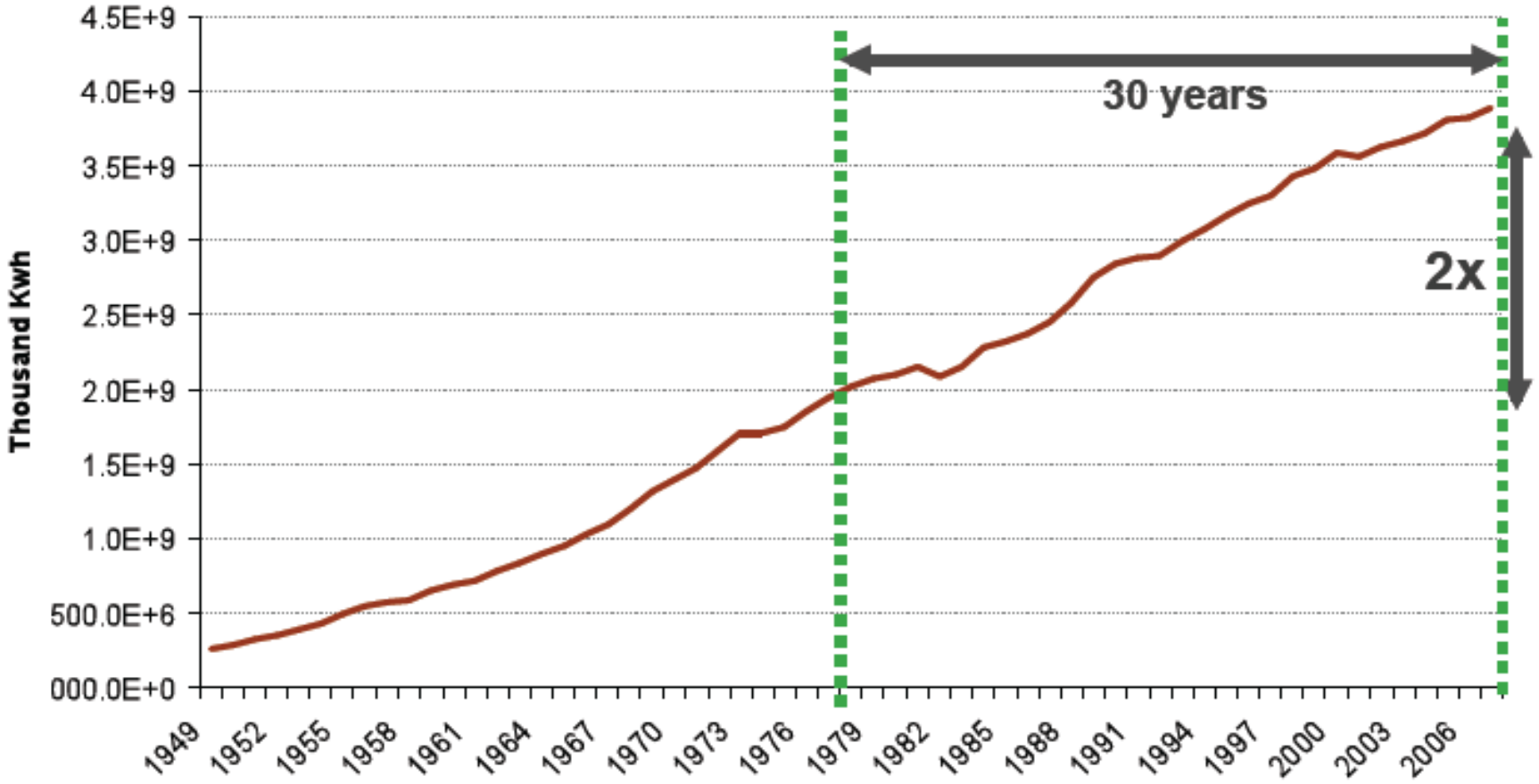
储存 --- 数据中心


Latency (us)
Bandwith (MB/sec)
Capacity (GB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
0.1
DRAM
FLASH
DISK
DRAM
FLASH
DISK
DRAM
FLASH
DISK
服务器
机架
数据中心

设计挑战

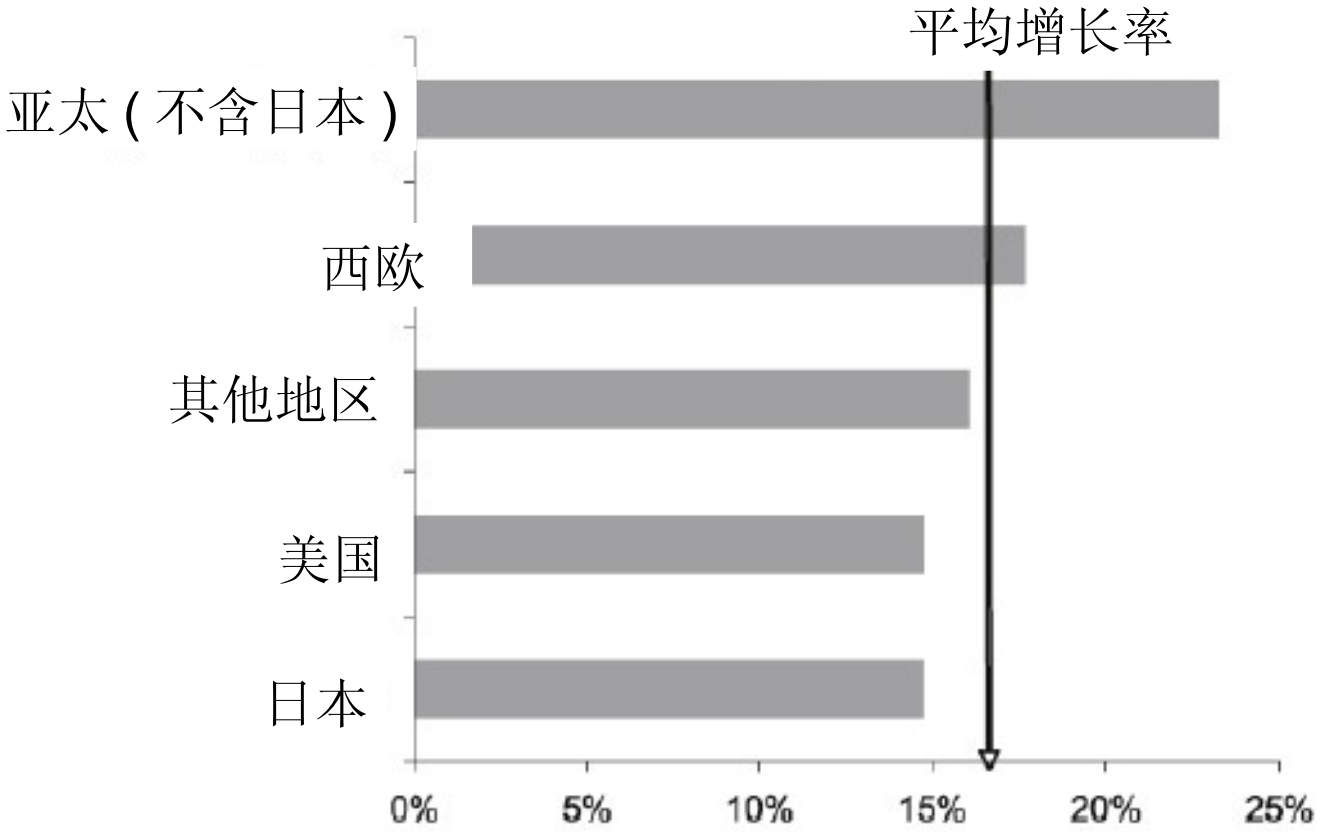
- 节约能源
  - 硬件，软件，数据中心
  - 编码，压缩，传输数据
- 故障恢复
  - 硬件和软件出错
  - 等待运行较慢的机器
- 新程式设计模型
  - 并行计算
  - Flash (SSD); GPU

能源消耗全球增长趋势



能源消耗每 30 年增长一倍

能源消耗 地区增长趋势



2000 – 2005 年平均增长率

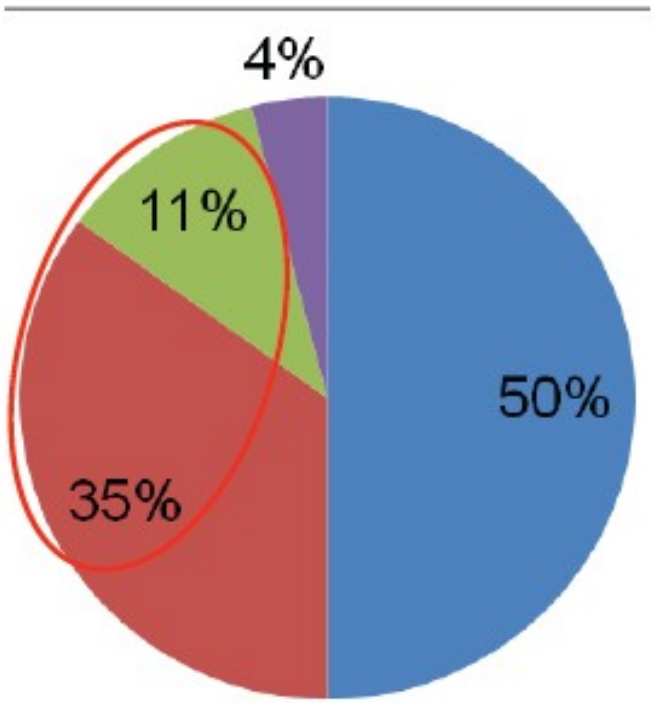
亚太区能源消耗每 6 – 8 年增长一倍

能源消耗 -- 油当量

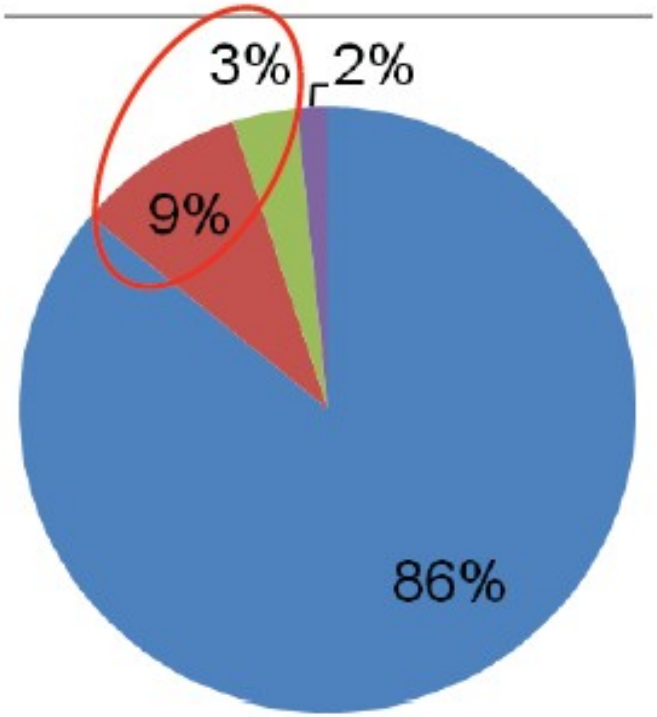
0.2g	Answering one **Google query**
20g	Using a **Laptop** for one hour
75g	Using a **PC & monitor** for one hour
173g	One weekday **newspaper** (physical copy)
209g	Producing a single glass of **orange juice**
280g	Washing one load of **laundry** in an efficient machine
532g	One **beer**

电源使用效率

平均电源使用效率 = 2.0



Google 电源使用效率 = 1.16



IT 冷却 电源分布和备用电源 照明

设计挑战

- 节约能源
  - 硬件，软件，数据中心
  - 编码，压缩，传输数据
- 故障恢复
  - 硬件和软件出错
  - 等待运行较慢的机器
- 新程式设计模型
  - 并行计算
  - Flash (SSD); GPU

故障率

- 99.9% 正常运行时间 = 9 小时故障 / 年
- 10,000 计算机中心
  - 0.25 次断电
  - 3 次路由器故障
  - 1,000 计算机故障
  - 1,000s 硬盘故障
  - etc., etc., etc.

故障后快速修复

- 复制
- 定期检查
- 尽可能：
  - 松散的一致性
  - 近似的答案
  - 不完整的答案

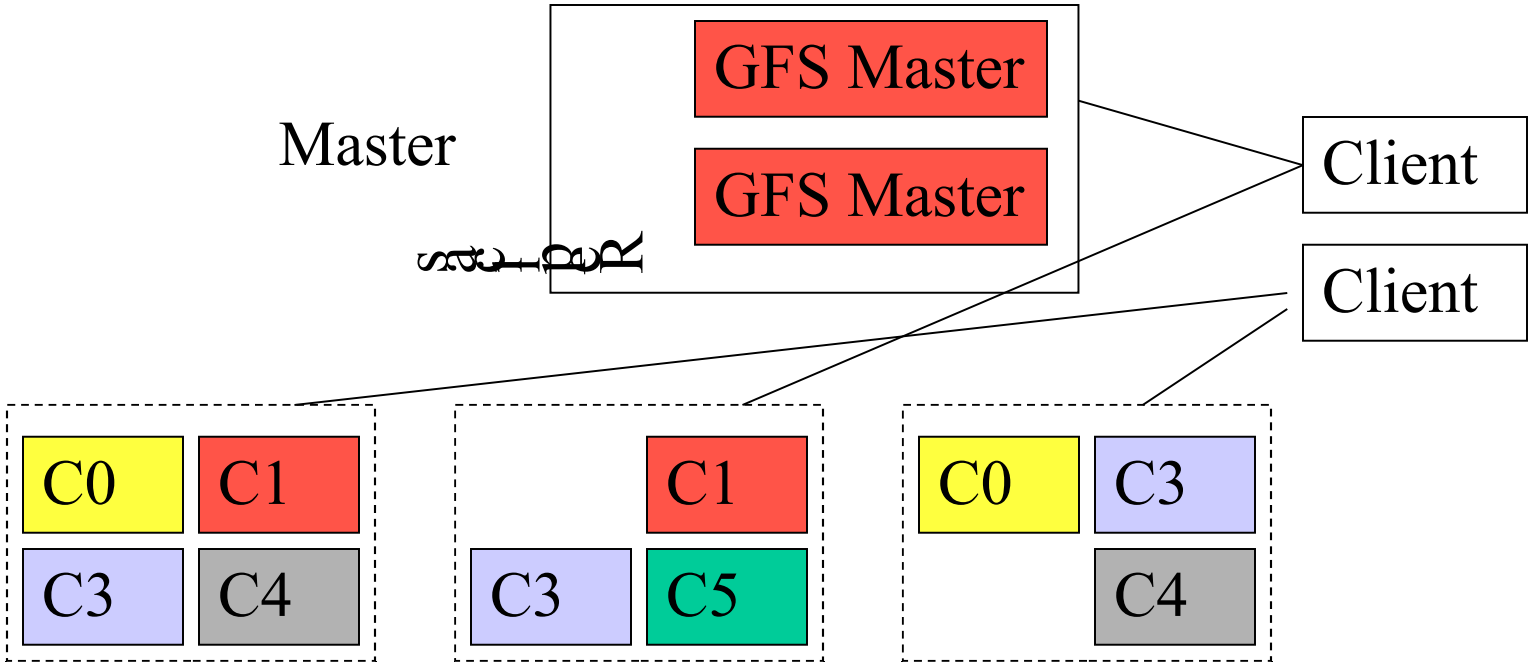
设计挑战

- 节约能源
  - 硬件，软件，数据中心
  - 编码，压缩，传输数据
- 故障恢复
  - 硬件和软件出错
  - 等待运行较慢的机器
- 新程式设计模型
  - 并行计算
  - Flash (SSD); GPU

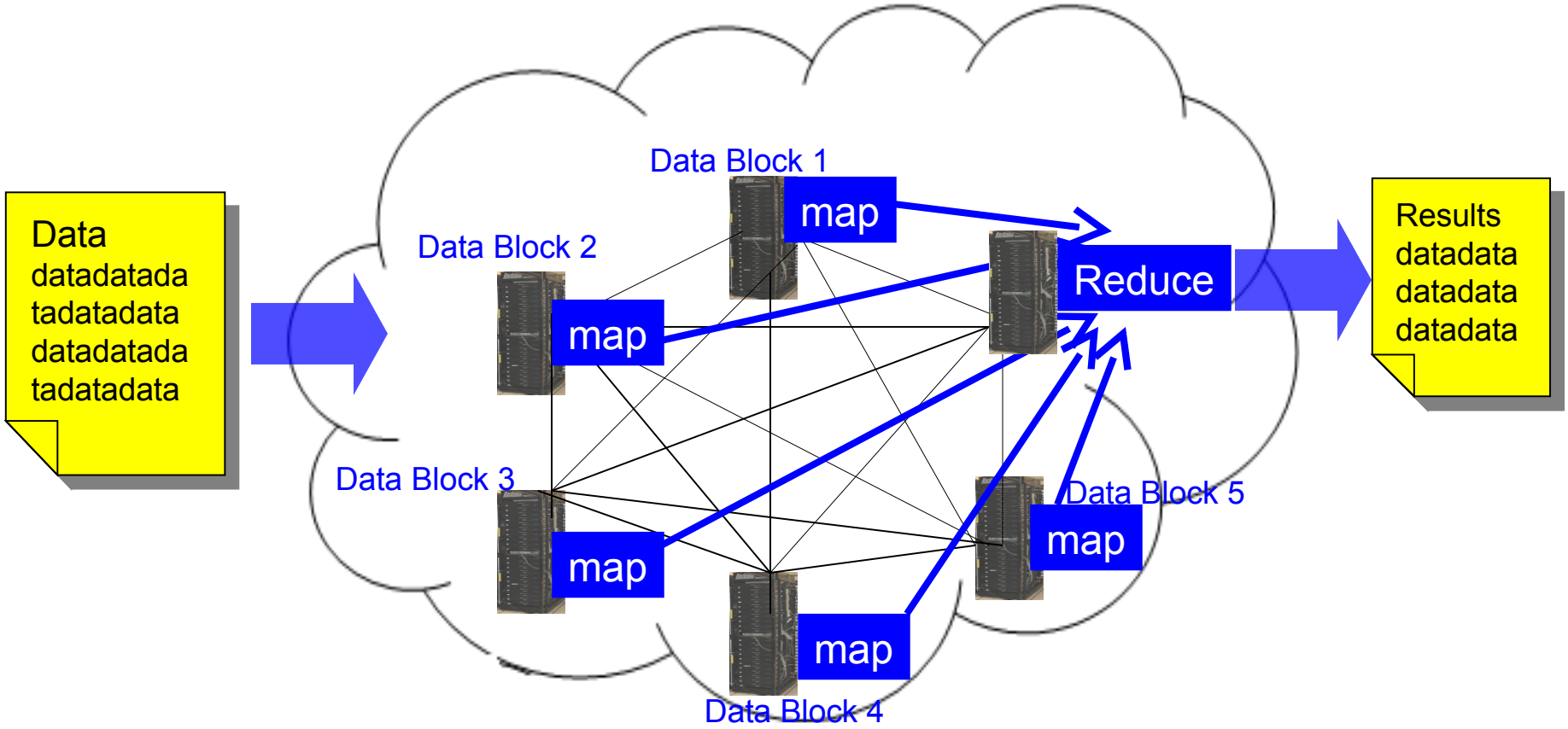
题纲

- 计算的规模与极限
- 云计算平台设计挑战
- 谷歌云计算基础技术
- 谷歌云计算新技术

谷歌文件系统 (GFS)


Master
GFS Master
GFS Master
Client
Client
C0
C1
C3
C4
C1
C3
C5
C0
C3
C4

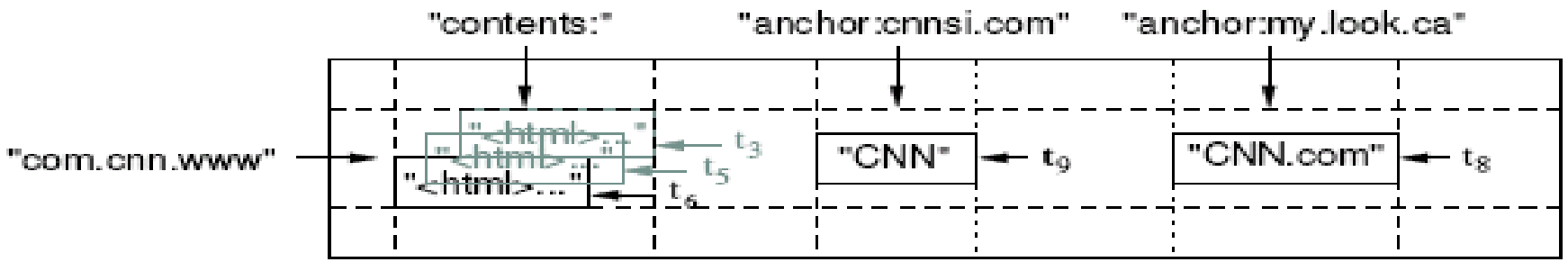
MapReduce



GFS 和 *Mapreduce* 于 *2004* 年发表，并成为开源 *Hadoop* 系统的基础，雅虎，微软和 *Facebook* 在各自的应用里均使用了 *Hadoop* 系统。

谷歌大表格 (Big Table)

- 内存管理



- 举例
  - 行：网页
  - 列：网页具体信息
  - 时间戳：网页信息提取的时间

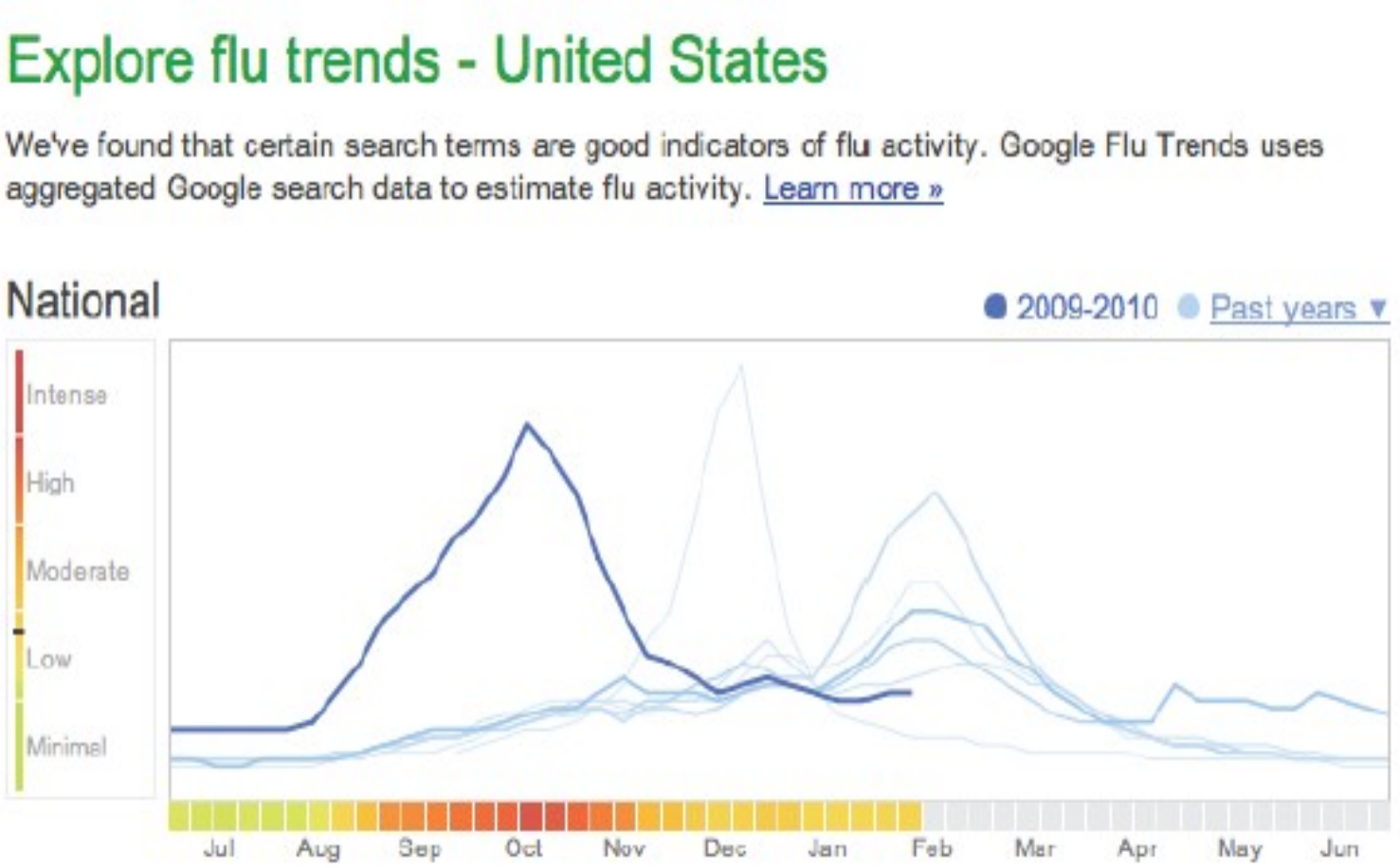
题纲

- 计算的规模与极限
- 云计算平台设计挑战
- 谷歌云计算基础技术
- 谷歌云计算新技术

赢在规模的例证

- 谷歌翻译
- 谷歌语音识别
- 趋势预测

流感趋势图


Explore flu trends - United States
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »
National
2009-2010
Past years
Intense
High
Moderate
Low
Minimal
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun

总结

- 计算的规模与极限
- 云计算平台设计挑战
- 谷歌云计算基础技术
- 谷歌云计算新技术

感谢

- 感谢如下谷歌同事贡献：
  - Peter Norvig
  - Stuart Feldman
  - Edward Chang
  - Xuemei Gu
  - Hai Fang

谢谢！