MTA SZTAKI

Institute for Computer Science and Control
Hungarian Academy of Sciences

European Research Consortium for Informatics and Mathematics
ERCIM

Affordable Supercomputing for Data Mining Applications

TASKS

Computation intensive ← → **Data intensive**

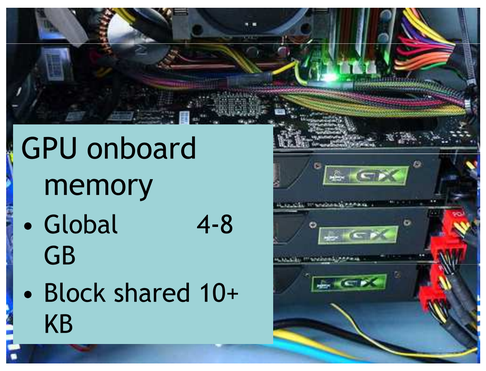
Image, multimedia processing
Expectation Maximization
(k-means, Gauss mixtures, …)
Gene sequencing
Evolutionary trees, …

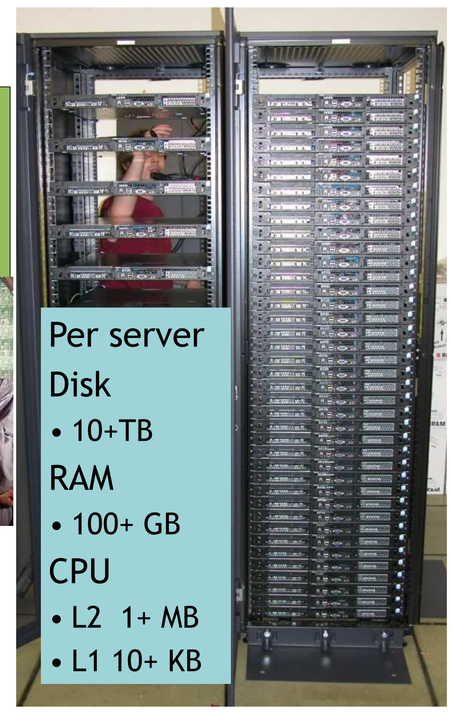
Web processing
- crawl, organization, filtering,
- quality assessment

Transaction log processing
- IT, telco, supermarket, …
- Reporting, fraud, security

ARCHITECTURES

Many-core ← → **Data centers of low-end servers**

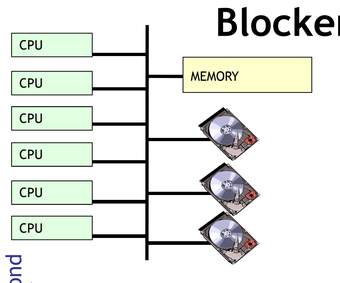




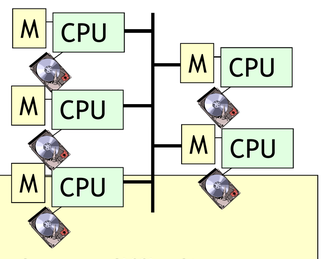
Moore's Law

- **Certain physical limits reached (?)**
- **No growth in single core processors**
- **But gowth in *many core***

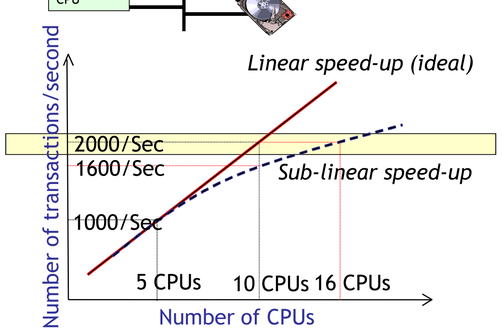
Blockers in Scalability



Read 1 MB sequentially…
- memory 250,000 ns
- network 10,000,000 ns
- disk 30,000,000 ns



☹ Cost
☹ Security
☹ Integrity control more difficult
☹ Lack of standards
☹ Lack of experience
☹ Complexity of management and control
☹ Increased storage requirements
☹ Increased training cost



Connolly, Begg: Database systems: a practical approach to design, implementation, and management], International computer science series, Pearson Education, 2005

„Numbers Everyone Should Know"

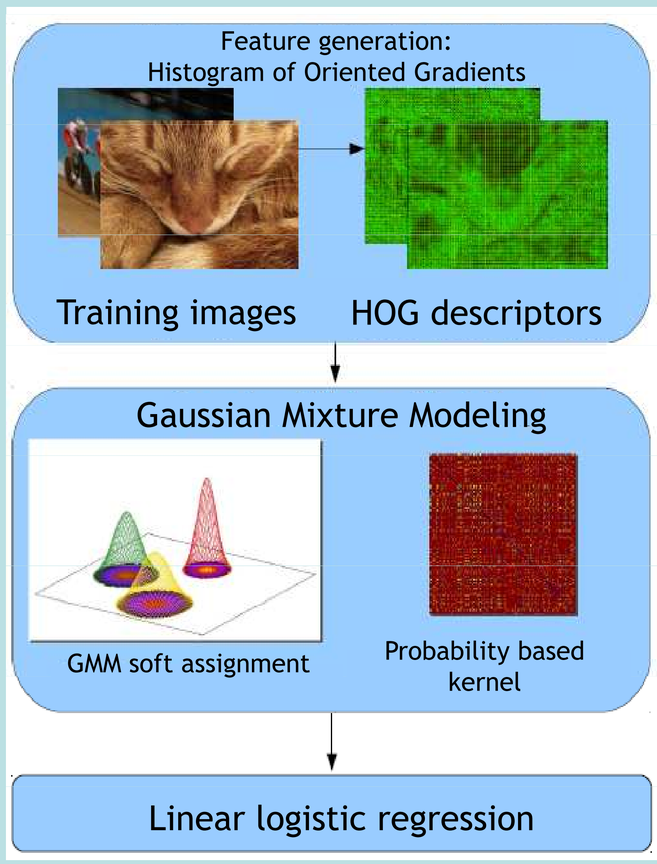
Jeff Dean, Google

Intra-process communication
- Mutex lock/unlock 100 ns
- Read 1 MB sequentially from network 10,000,000 ns

Disk
- Disk seek 10,000,000 ns
- Read 1 MB sequentially from disk 30,000,000 ns

RAM
- L1 cache reference 0.5 ns
- L2 cache reference 7 ns
- Main memory reference 100 ns
- Read 1 MB sequentially from memory 250,000 ns

Image Classification on GPGPU

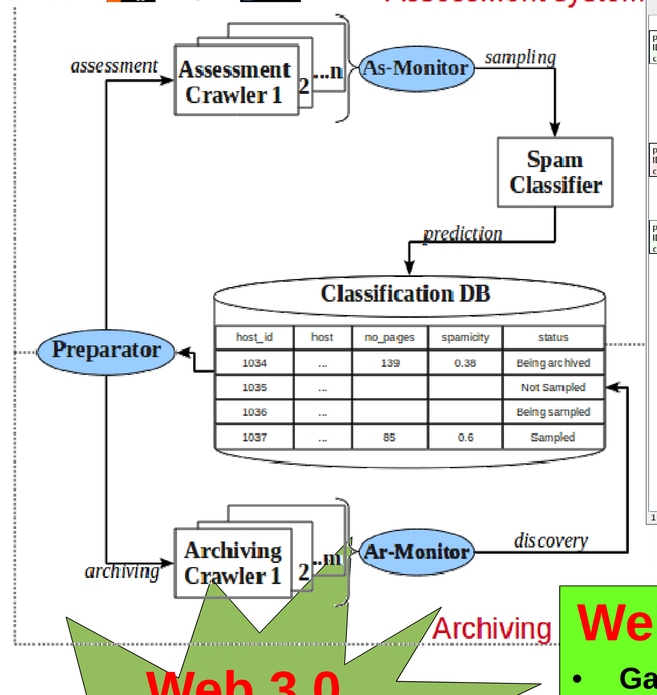


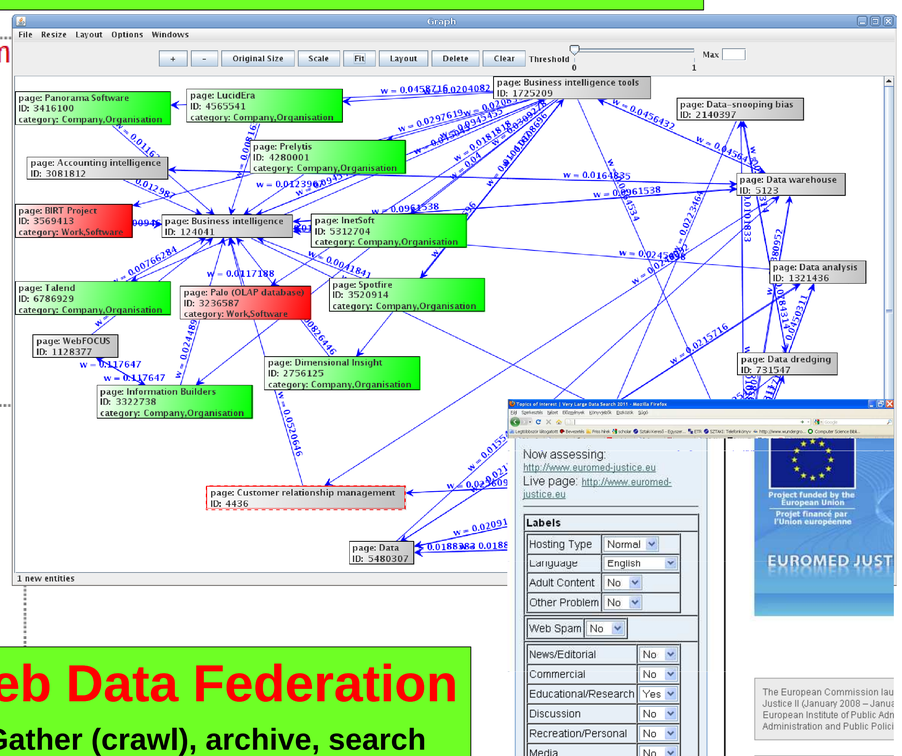
Feature generation:
- Histogram of Oriented Gradients (grayscale)
- Results in 200-400 dimensions
- Lowered with PCA

Modeling: GMM
- 128 Gaussians
- sampled training data
- pdf based $\chi 2$ kernel

Research data continuous growth in volume

Source	Number of Images
CLEF 2006: IAPR TC-12	20,000
CLEF 2008, PASCAL VOC 2007: MIR Flickr	25,000
CLEF 2011, PASCAL VOC 2010: MIR Flickr	1,000,000





Web Data:

The ECML/PKDD Discovery Challenge 2010 on Web Quality

- **LiWA: Living Web Archives**
- **Living Knowledge**
- **Yahoo! Research**
- **Google Research**

Longitudinal Analytics of Web Archive Data in FIRE

LAWA

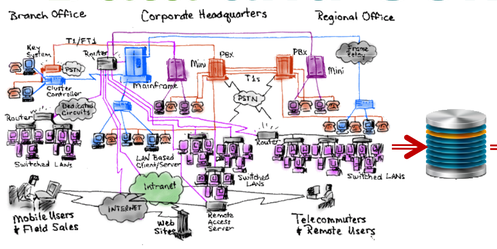
Web 3.0

- **Personalized, „social"**
- **Access by any device**
- **Data Mining aided**
- **Distributed, cloud**
- **Media**

Web Data Federation

- **Gather (crawl), archive, search**
- **Visualize, arrange, navigate**
- **Label, select subcollections**
- **Quality control**

Data and Computation intensive tasks → ARCHITECTURE?

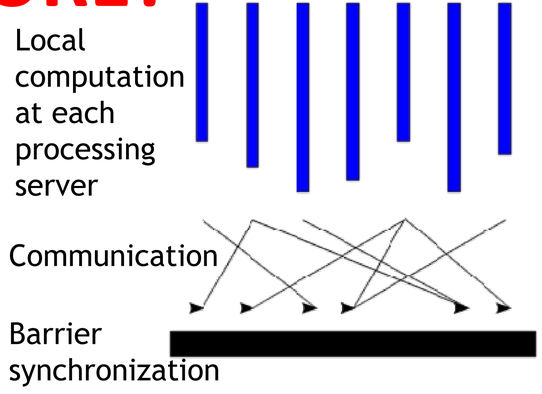


Source	Yearly log records	GBs
AEGON Hungary IT	1+ Bn	10,000+
Leading Hungarian Web portal	20+ Bn	40,000+
ReliaWind Wind power plant contact: zsolt.viharos@sztaki.hu	5 M / turbine	0.1 / turbine

Network Algorithms:

Bulk Synchronous Parallel

Network centrality, ranking
Trust, distrust propagation
Node similarity, friendship prediction
→ Large data, complex network processing

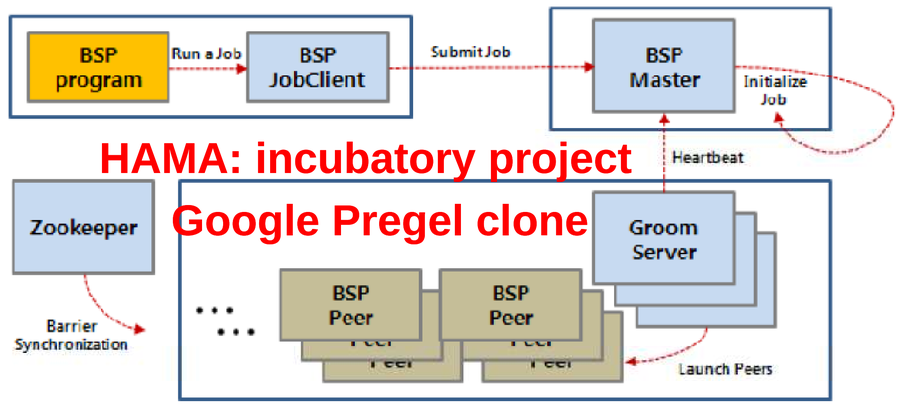


Plethora of Solutions?

hadoop, S3, Sybase IQ, Greenplum, HYPERTABLE, InfiniDB, TERADATA, VERTICA, Sun ORACLE, membase, HIVE, HBASE, INFOBRIGHT, twinFin, Microsoft SQLServer, Redis, CouchDB, DB2, Cassandra, mongoDB

Conclusions

- **New hw/sw architectures for scaling**
- **Interdisciplinary area**
- **Enables novel apps (e.g. Web 3.0)**
- **Computation & Data Intensive problems need breakthrough!**
- **No mature open architecture yet for distributed network analysis (BSP?)**



Data Mining and Web Search Group

Contact: Andras Benczur, benczur@sztaki.hu

LAWA: Longitudinal Analytics of Web Archive Data

SCIIMS: Strategic crime and Immigration management System