An Architectural Characterization Study of Data Mining and Bioinformatics Workloads

Berkin Ozisikyilmaz Ramanathan Narayanan Gokhan Memik Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
{boz283, ran310, memik, choudhar}@eecs.northwestern.edu

Joseph Zambreno
Iowa State University
zambreno@iastate.edu

Acknowledgements:
National Science Foundation: NGS CNS-0406341, IIS-0536994/002, CNS-0551639, CCF-0621443, CCF-0546278, NSF/CARP CCF-0444405.
Department of Energy: SDiDAC DEFC02-01ER25485, DE-FG02-05ER25683, DE-FG02-05ER25691.
Intel Corporation

NORTHWESTERN UNIVERSITY

An Explosion of Data

- Recent trends indicate that data collection rates are growing at an exponential pace
- 2003 study – Five *exa*bytes of new information was stored in the previous year[1]
  - Equivalent to 37,000 Libraries of Congress (LoC)
  - 800MB of new data per person

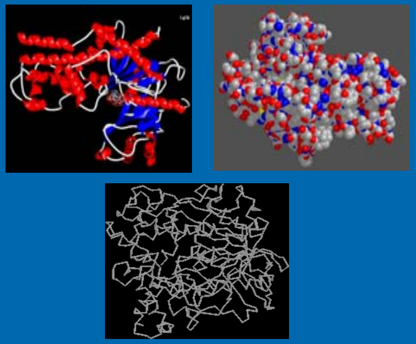
Storage Medium	2002	1999–2000	% Change
Paper	1,634	1,200	36%
Film	420,254	431,690	-3%
Magnetic	5,187,130	2,779,760	87%
Optical	103	81	28%
Total storage:	5,609,121	3,212,731	74.5%

(Upper estimates, expressed in TB)

- [1]Source: “How much information” project, UC-Berkeley

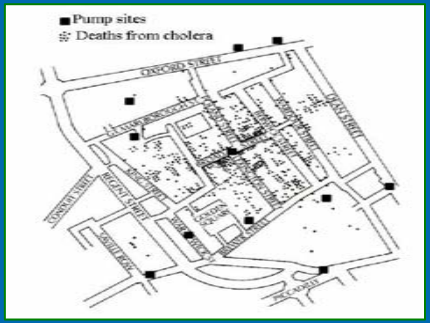
Utilizing Large Data Sets – Mining

- Enormous data growth in both commercial and scientific databases
  - Data mining to extract information from large and complex datasets
- Data scales at a high rate, exceeding Moore's Law
  - Advances in computing capabilities and technological innovation needed to harvest the available wealth of data
- Our contribution
  - Developing an understanding of architectural characteristics of these applications



Biomedical Data

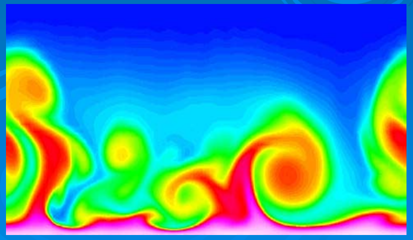
Homeland Security



Geo-spatial intelligence

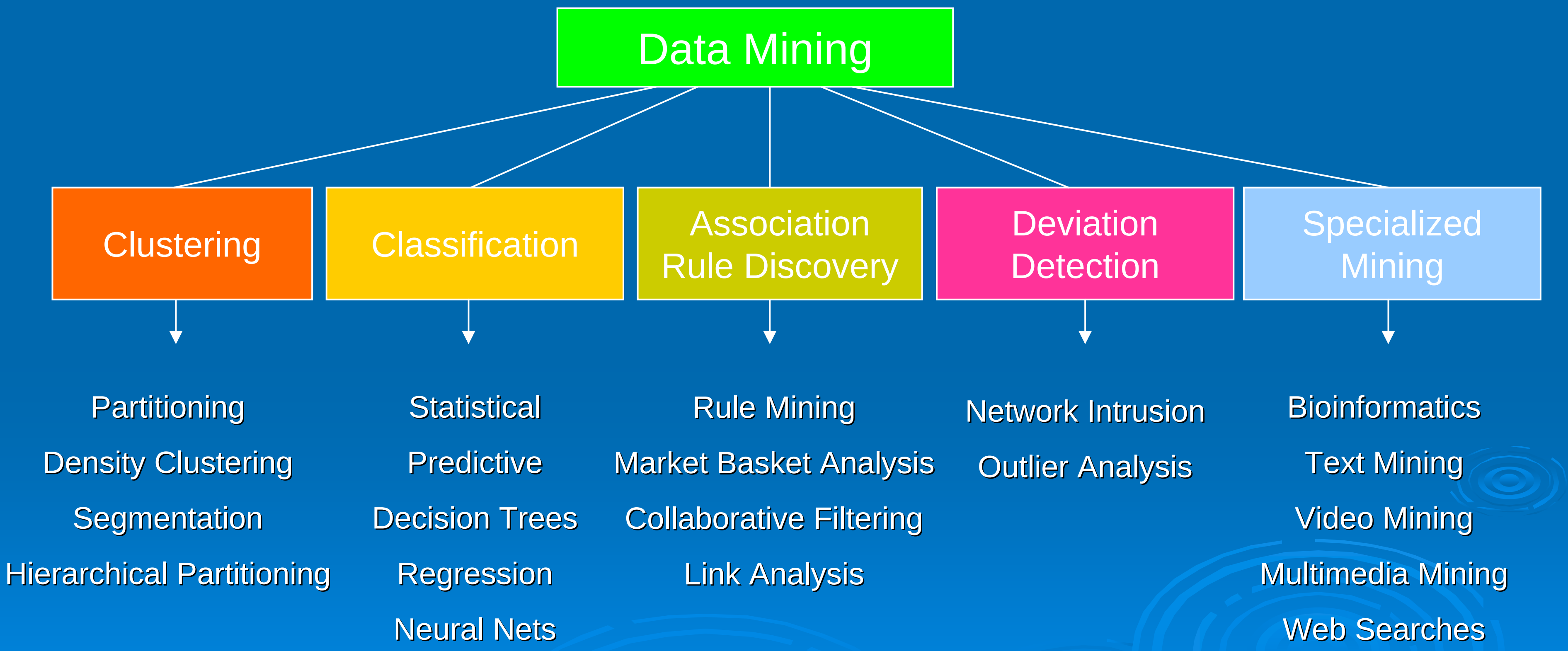
Information Assurance Network Intrusion Detection

Sensor Networks



Computational Simulations

Taxonomy of Data Mining Methods



Characterization Goal

- General purpose processors are targeted for
  - Compute intensive applications (integer and floats)
  - Multimedia applications
  - Database applications
  - …

- **Are data mining applications different from the above?**
- **Why are they different/same?**
  - Identify CHARACTERISTICS that differentiates them if any

Outline

- Introduction / Motivation
- Overview of applications
- Uniqueness of data mining applications
- Performance characterization
  - Execution time
  - Scalability
  - Memory hierarchy behavior
  - Instruction efficiency
- Related work
- Conclusions

MineBench Overview

Non-bioinformatics workload, includes applications from:
- a) Decision trees
- b) Clustering/ Hierarchical Clus.
- c) Utility Mining
- d) Predictive Modeling
- e) Market Basket Analysis

Data taken from:
- a) Image processing
- b) Astrophysics
- c) Grocery chain
- d) Pharmaceutical

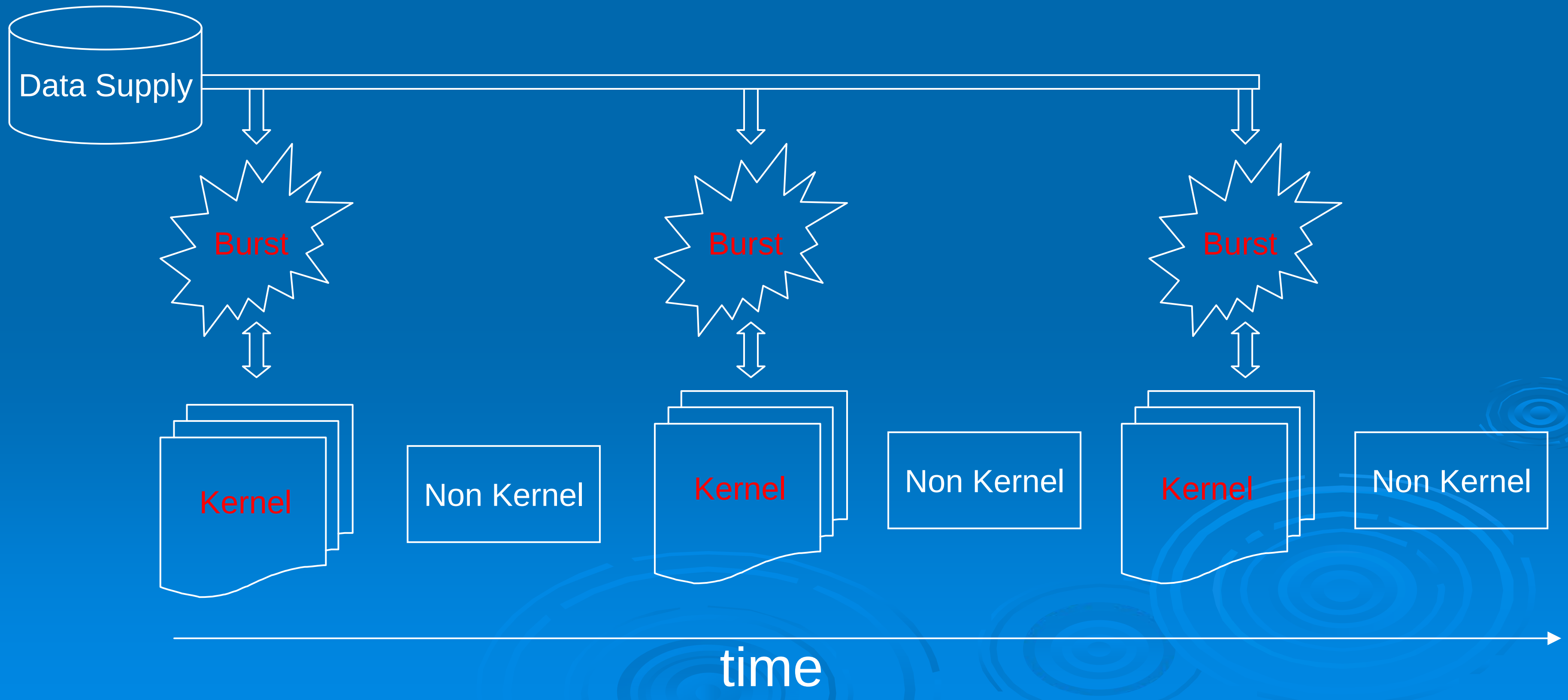
Bioinformatics workload (algorithms used in other fields as well)

Application	Instruction Count (billions)				Binary Size (kB)
	1 Processor	2 Processors	4 Processors	8 Processors	
ScalParC	23.664	24.817	25.550	27.283	154
Naïve Bayesian	23.981	N/A	N/A	N/A	207
K-means	53.776	54.269	59.243	77.026	154
Fuzzy K-means	447.039	450.930	477.659	564.280	154
HOP	30.297	26.920	26.007	26.902	211
BIRCH	15.180	N/A	N/A	N/A	609
Apriori	42.328	42.608	43.720	47.182	847
Eclat	15.643	N/A	N/A	N/A	2169
Utility	13.640	19.902	20.757	22.473	853
SNP	429.703	299.960	267.596	241.680	14016
GeneNet	2,244.470	2,263.410	2,307.663	2,415.428	13636
SEMPHY	2,344.533	2,396.901	1,966.273	2,049.658	7991
Rsearch	1,825.317	1,811.043	1,789.055	1,772.200	676
SVM-RFE	51.370	55.249	63.053	82.385	1336
PLSA	4,460.823	4,526.160	2,080.610	4,001.675	836

*** Ramanathan Narayanan, New Benchmarks Session.**

Data Mining Characteristics

- Multi-phased operations
  - Cyclic data+compute (large) nature



Kernels of the Applications

Kernel Distribution: % of the total execution time

Application	Top 3 Kernels (%)			Sum %
	Kernel 1 (%)	**Kernel 2** (%)	**Kernel 3** (%)	
k-Means	distance (68%)	clustering (21%)	minDist (10%)	99
Fuzzy k-Means	clustering (58%)	distance (39%)	fuzzySum (1%)	98
BIRCH	distance (54%)	variance (22%)	redistribution (10%)	86
HOP	density (39%)	search (30%)	gather (23%)	92
Naïve Bayesian	probCal (49%)	variance (38%)	dataRead (10%)	97
ScalParC	classify (37%)	giniCalc (36%)	compare (24%)	97
Apriori	subset (58%)	dataRead (14%)	increment (8%)	80
Eclat	intersect (39%)	addClass (23%)	invertClass (10%)	72

- Kernels could be prominent/spread across
- Common kernels across applications: distance, variance

Evaluation Framework

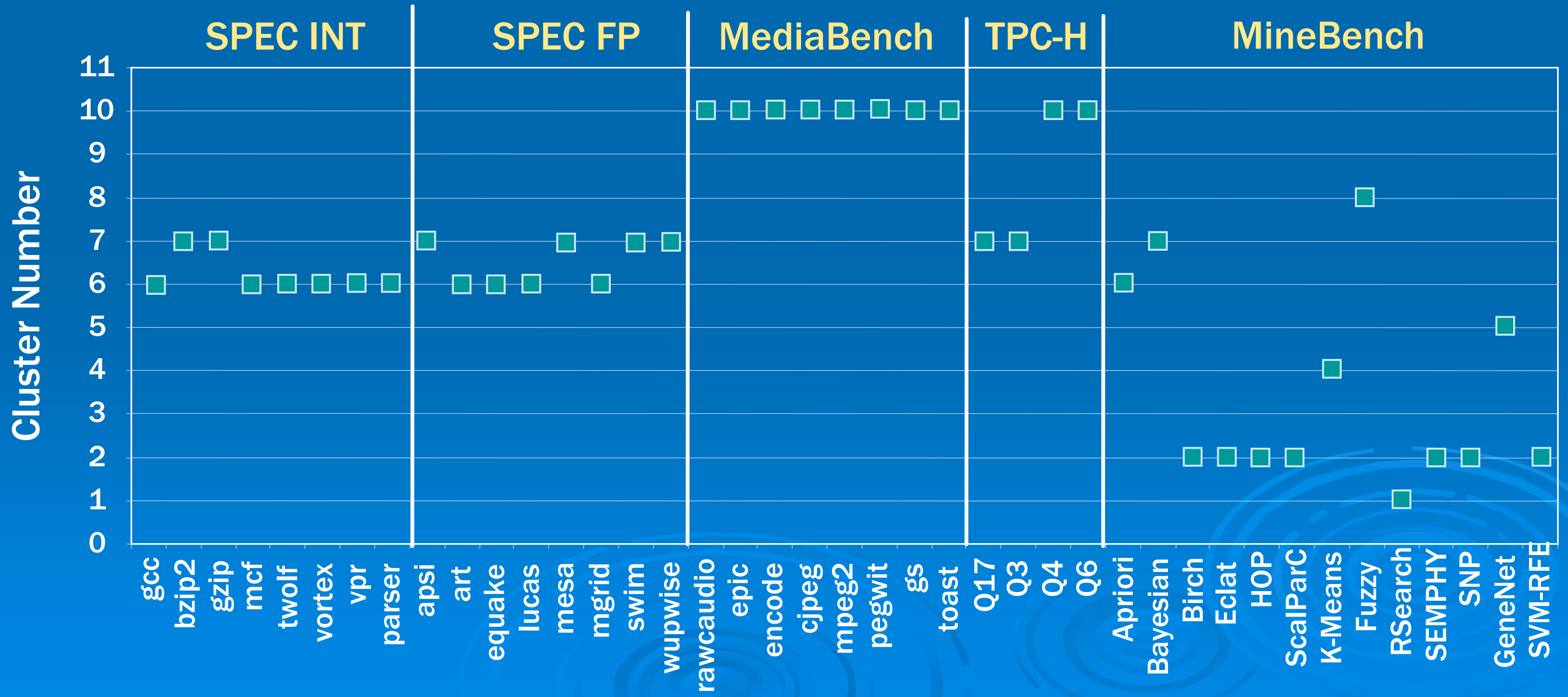
- Target platform:
  - 8-way Shared Memory Parallel (SMP) machine
  - Intel Xeon processors:
    - 700 MHz clock
    - 16 KB non-blocking integrated L1 cache
    - 1024 KB L2 cache
  - 4 GB of shared memory
- Red Hat Advanced Server 2.1
- Intel C++ compiler v7.1
- VTune Performance Analyzer

Metrics of Interest

- Monitored a wide assortment of performance metrics:
  - Cache miss ratios and events
  - Memory statistics
  - Bus usage
  - Branch performance
  - Application execution times
  - Page faults
  - Synchronization & lock overheads
  - Parallelization overheads

Uniqueness of Data Mining Apps

- Performance metrics gathered from VTune were fed into Clementine data mining software
- Data for various benchmark suites run through Kohenen clustering:
  - Other benchmarks tend to fall into one or two clusters
  - Data mining applications span multiple clusters
  - Most importantly, mining apps have their own cluster



Minebench versus Other Benchmarks

Parameter†	Benchmark of Applications				
	SPECINT	SPECFP	MediaBench	TPC-H	MineBench
Data References	0.81	0.55	0.56	0.48	**1.10**
Bus Accesses	0.030	0.034	0.002	0.010	**0.037**
Instruction Decodes	1.17	1.02	1.28	1.08	**0.78**
Resource Related Stalls	0.66	1.04	0.14	0.69	**0.43**
CPI	1.43	1.66	1.16	1.36	**1.54**
ALU Instructions	0.25	0.29	0.27	0.30	**0.31**
L1 Misses	0.023	0.008	0.010	0.029	**0.016**
L2 Misses	0.003	0.003	0.0004	0.002	**0.006**
Branches	0.13	0.03	0.16	0.11	**0.14**
Branch Mispredictions	0.009	0.0008	0.016	0.0006	**0.006**

† The numbers shown here for the parameters are values per instruction

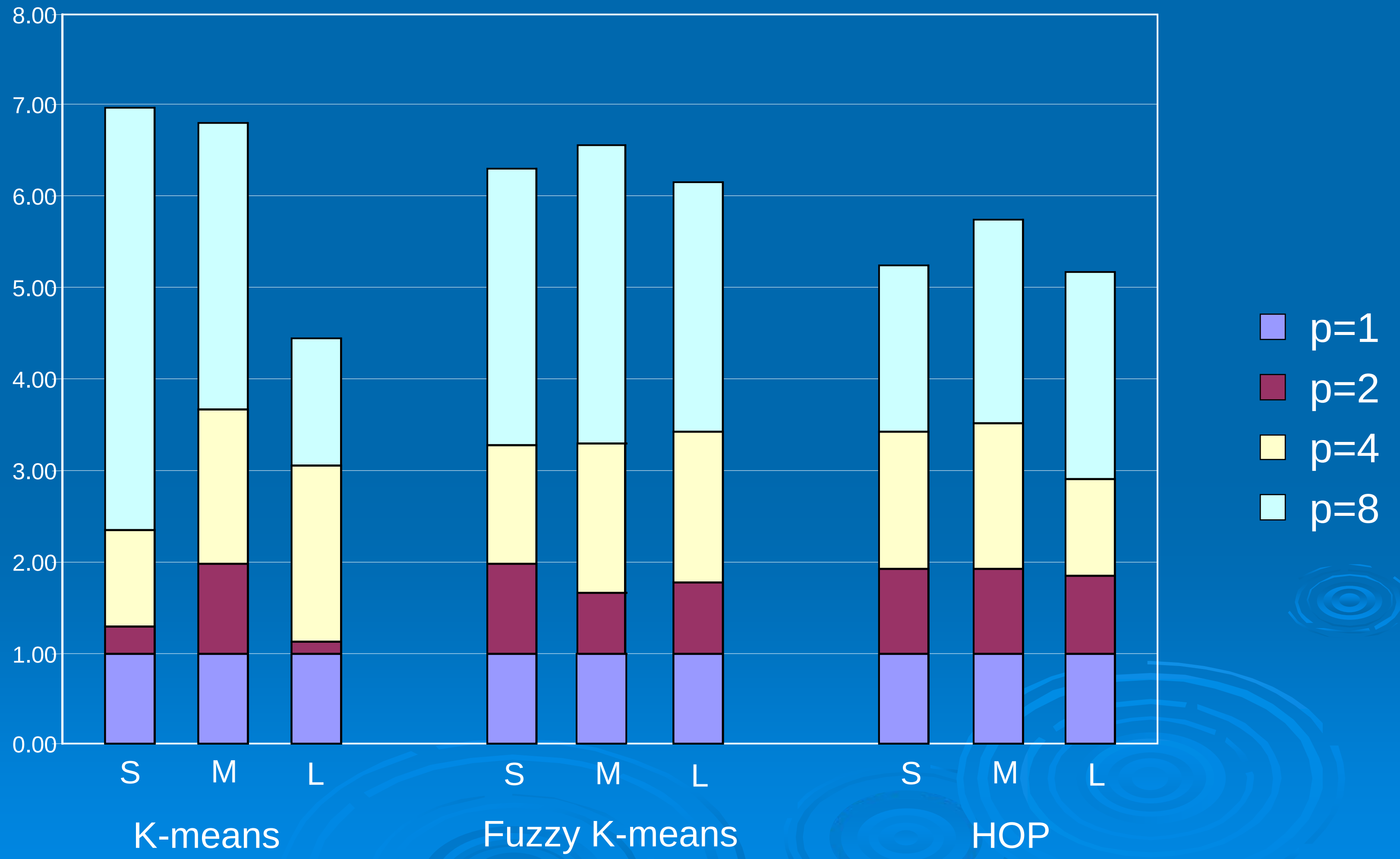
- Key unique attribute: number of data references retired
- Other differentiating attributes:
  - L2 miss rates
  - The ratio of total instruction decodes to the instructions retired
  - The ALU operations per instruction retired

Outline

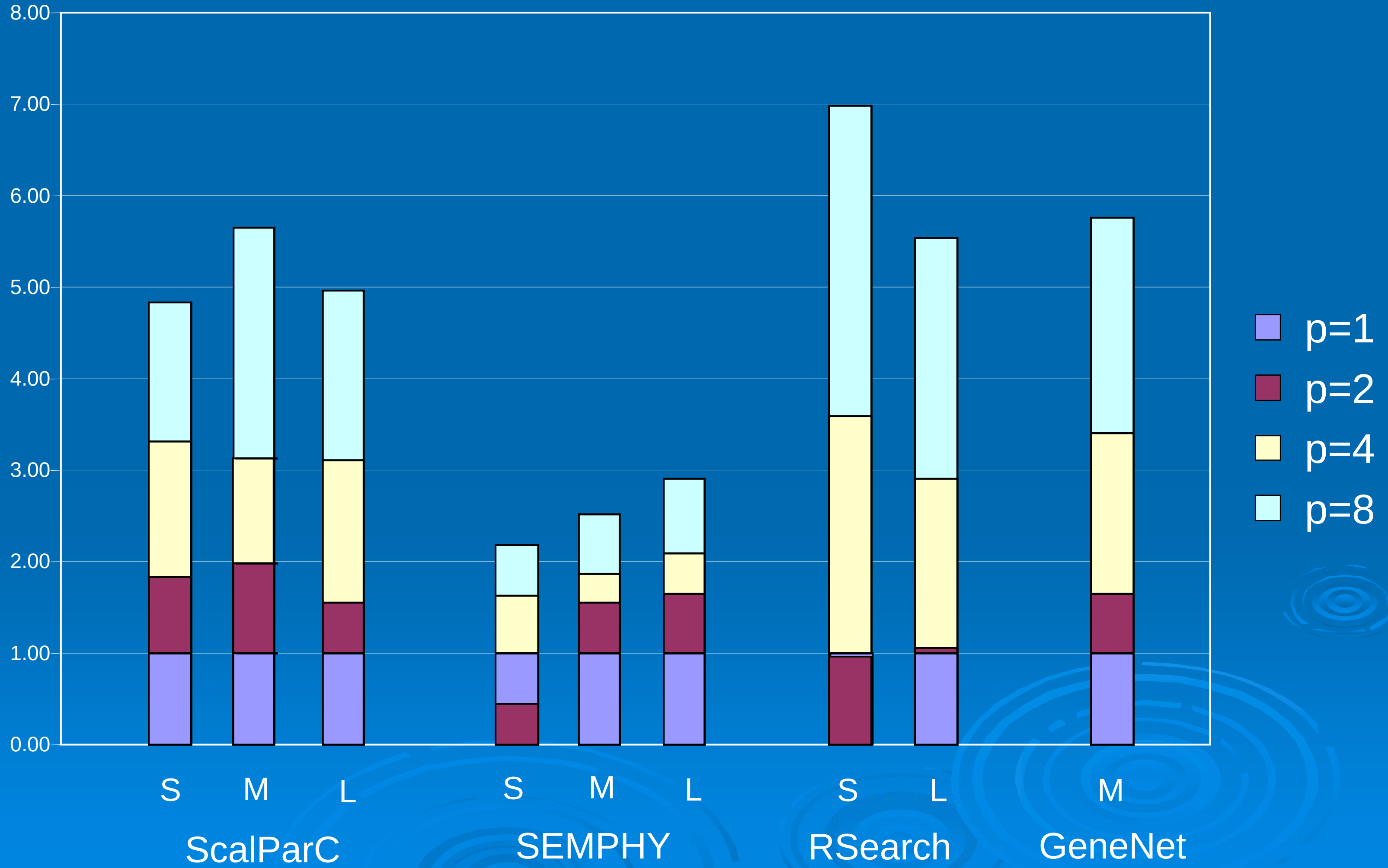
- Introduction / Motivation
- Overview of applications
- Uniqueness of data mining applications
- Performance characterization
  - Execution time
  - Scalability
  - Memory hierarchy behavior
  - Instruction efficiency
- Related work
- Conclusions

Execution Speedups

Clustering

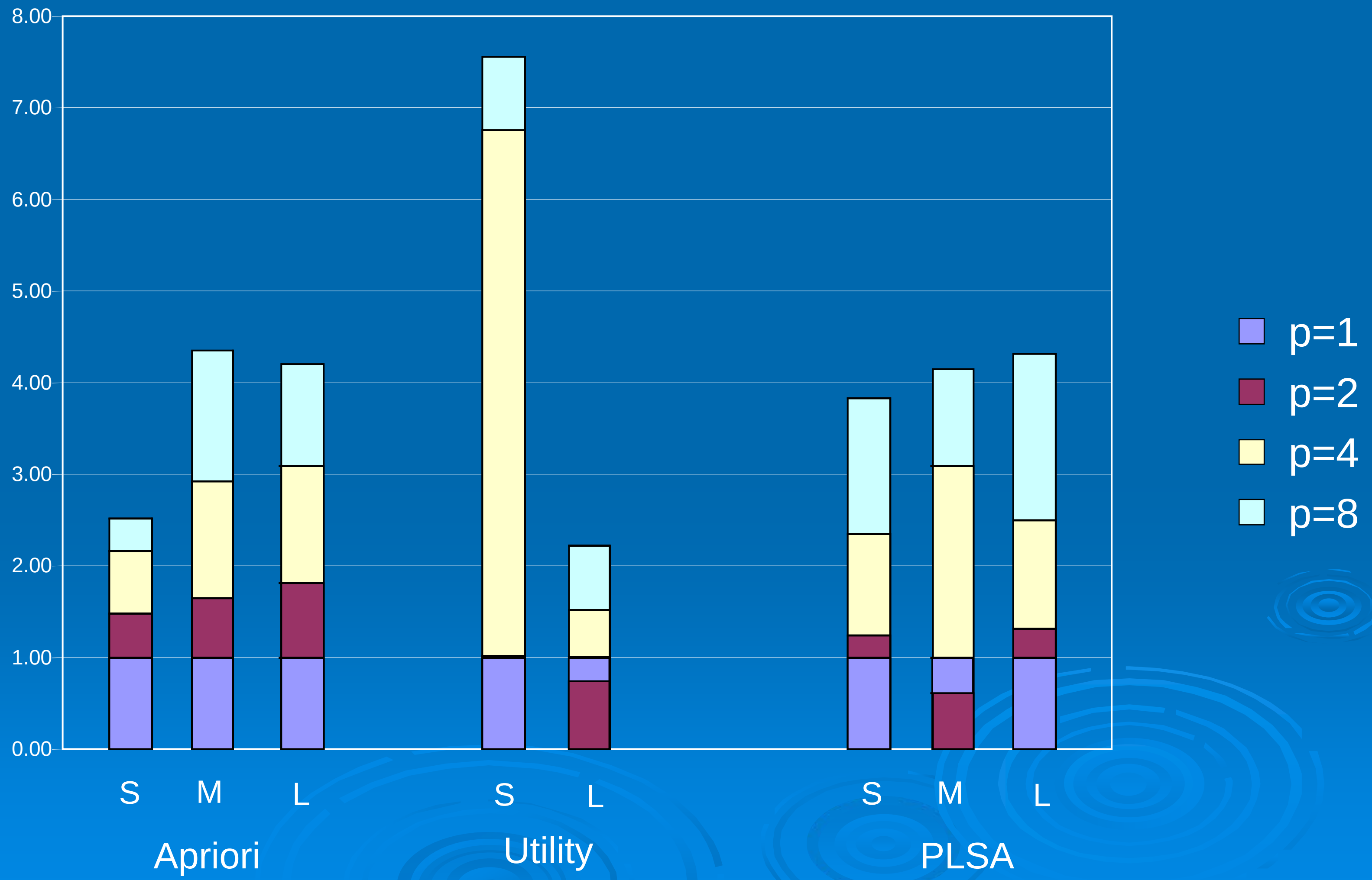


Execution Speedups
Classification

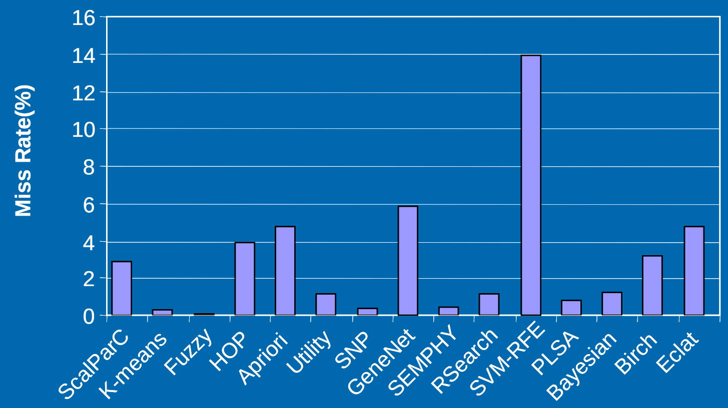


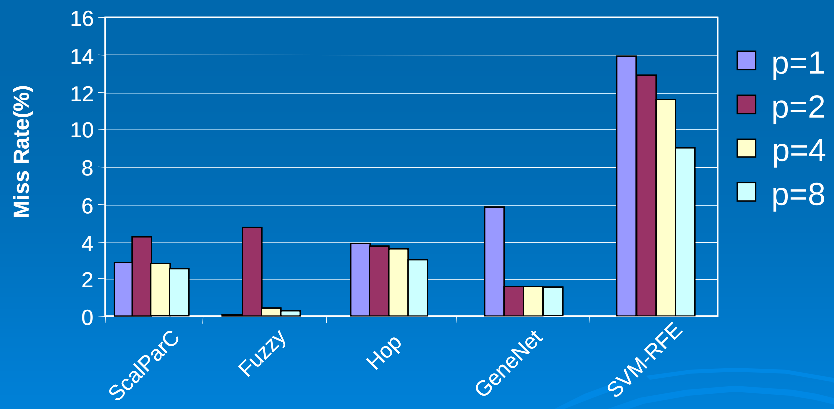
Execution Speedups

ARM & Optimization



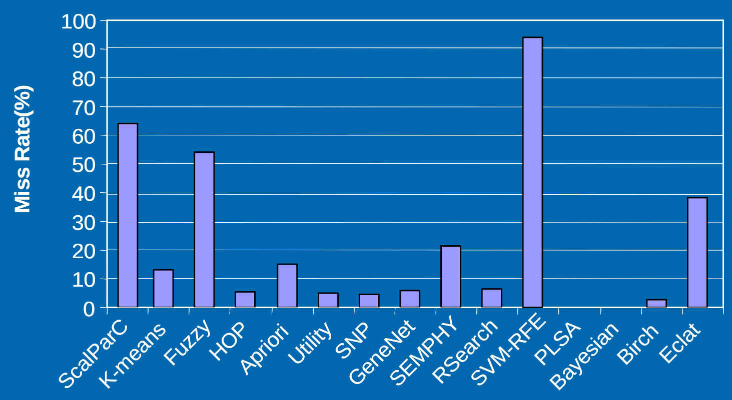
Memory Hierarchy Behavior
L1 Data Miss Rates

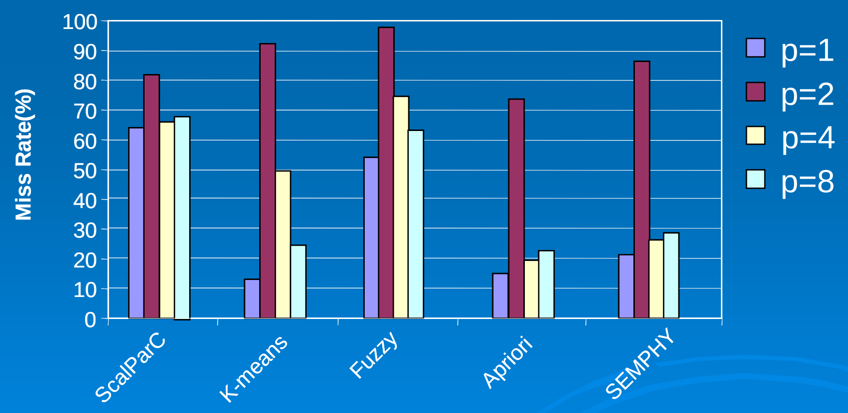




- L1 data miss rates are usually small, two categories:
  - Very small (less than 1.5%)
  - Larger (2-14%)
- L1 data miss rates are higher in 2-processor cases
- L1 instruction cache misses are very low (on average 0.11%)
  - Kernels dominate execution

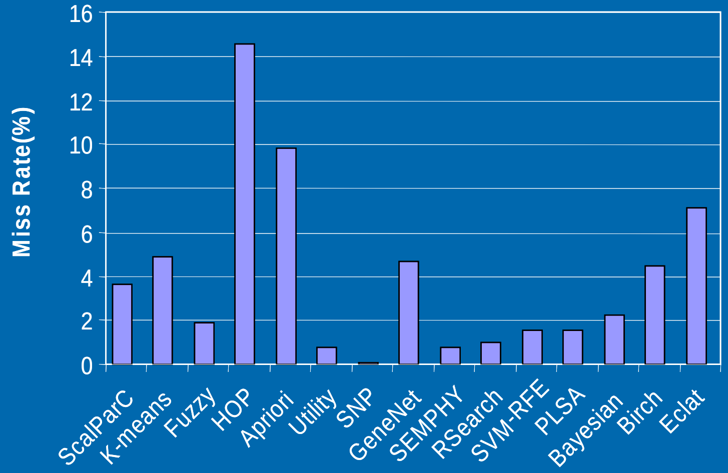
Memory Hierarchy Behavior
L2 Cache Miss Rates

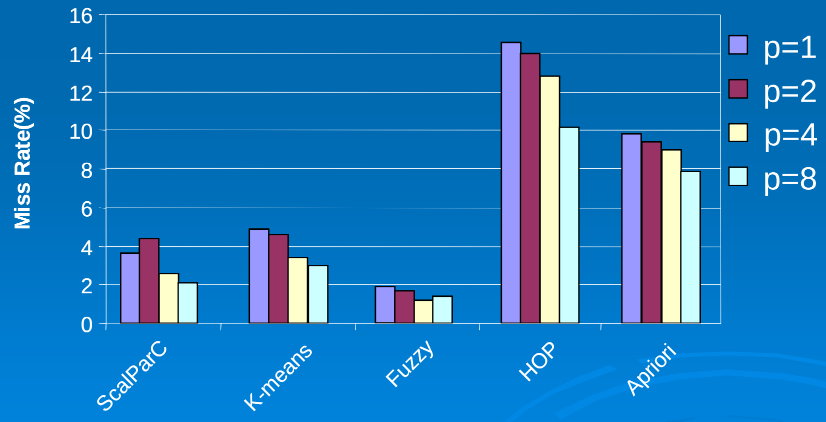




- L2 miss rates can be quite high
  - 94.2% for SVM-RFE
- Streaming nature
- Low L1 miss rates
- SVM-RFE has the worst L2 miss rate
  - 8.44% of all data references require off-chip memory access

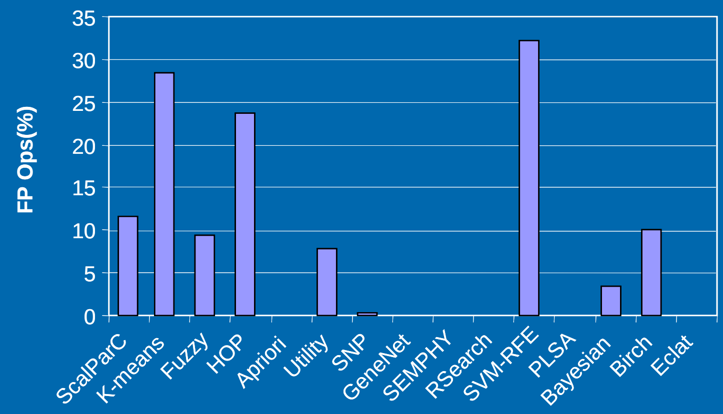
Instruction Efficiency
Branch Misprediction Rate

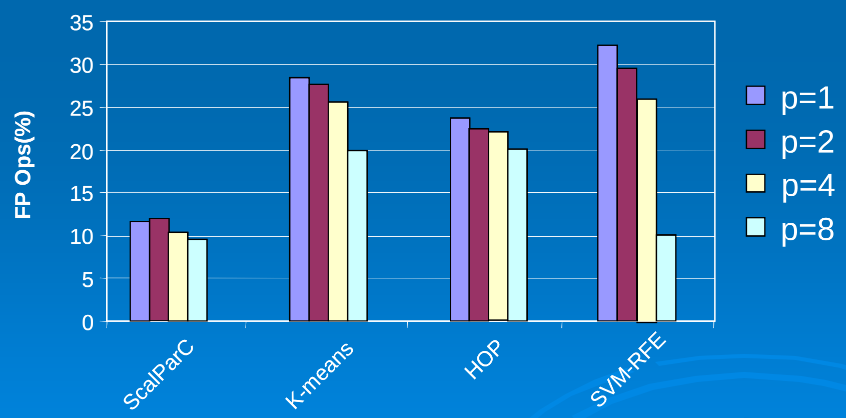




- Branch prediction performs well for most of the applications
- In most applications, the branch misprediction rate decreases with the increasing number of processors.

Instruction Efficiency
Fraction of Floating Point Operations

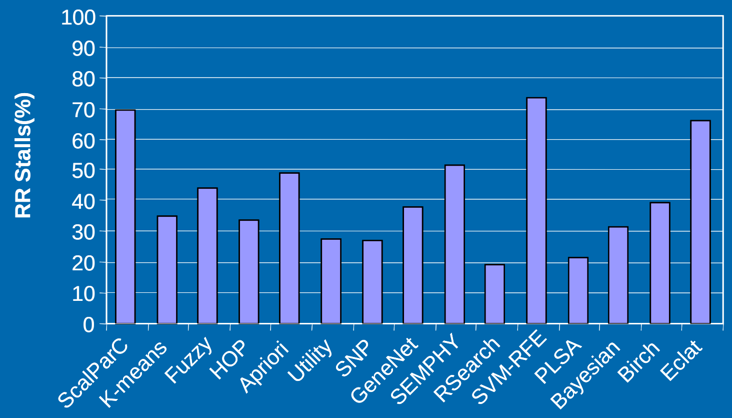


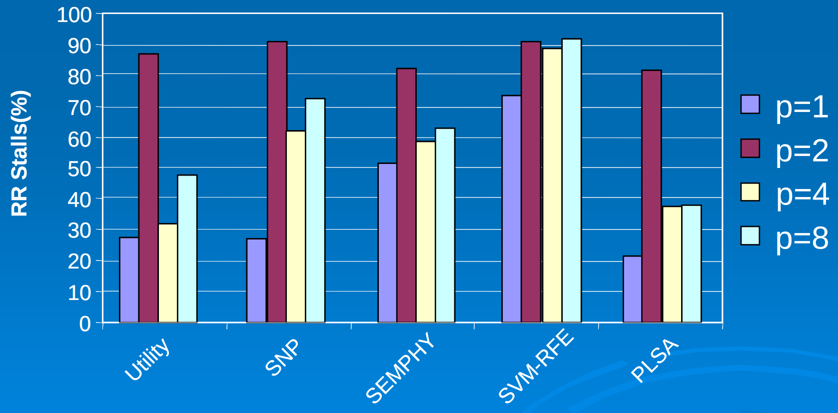


- Several benchmarks are floating point operation intensive.
- As the number of processors increase, the percentage of floating point operations decrease

Instruction Efficiency

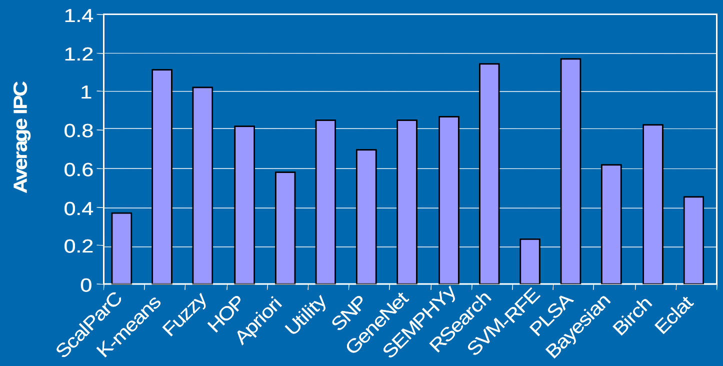
Resource Related Stalls

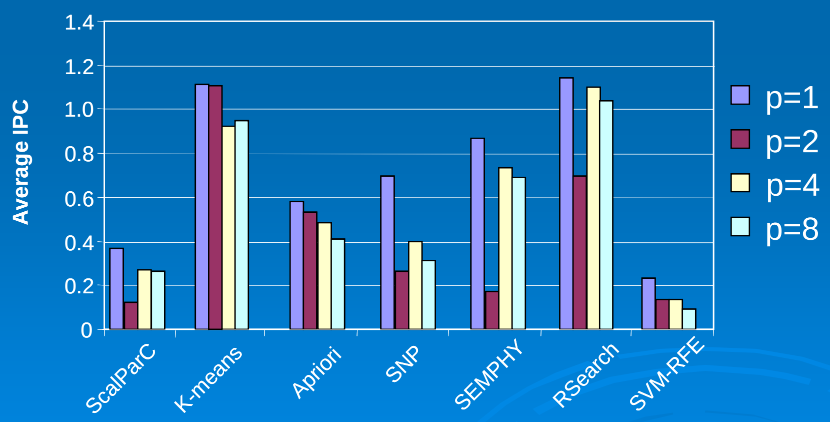




- SVM-RFE has the highest stall rates
  - Memory accesses
- As the number of floating point operations increases, the processor is able to utilize its resources better
- As the number of processors increases, the resource related stalls increase
  - Synchronization

Instruction Efficiency
Instructions per Cycle





- Some applications suffer from low IPCs
  - Combination of high L2 miss rates, cost of synchronization
- Parallel versions have lower IPC
  - Proc = 2, generally the worst IPC

Related Work

- Chen et al. at Intel recently analyzed the performance scalability of bioinformatics workloads
  - Their workloads are incorporated into MineBench
  - Results are compared where applicable
- Srinivasan et al. explore cache misses and algorithmic optimizations for SVM-RFE
- Sanchez et al. perform architectural analysis on biological sequence alignment
- A few other recently-developed bioinformatics benchmark suites:
  - BioInfoMark – University of Florida
  - BioBench – University of Maryland
  - BioPerf – Georgia Tech, University of New Mexico
  - All contain several applications in common (*Blast*, *FASTA*, *Clustalw*, *Hmmer*, etc.)

Conclusions

- Data mining applications are essential given the rate of data growth
- Current systems design approach may not be sufficient
  - Need for data mining specific optimizations
- MineBench – a new benchmark suite that encompasses many algorithms found in data mining
- Initial findings:
  - Data mining applications are unique in terms of performance characteristics
  - There exists much room for optimization with regards to data mining workloads

Thank You

Email: boz283@eecs.northwestern.edu
Web: http://www.ece.northwestern.edu/~boz283

Project Web Page
http://cucis.ece.northwestern.edu/projects/DMS

MineBench can be downloaded at
http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html