SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Do SAS® High-Performance Statistical Procedures Really Perform "Highly"? A Comparison of HP and Legacy Procedures

Diep T. Nguyen
Patricia Rodríguez de Gil
Anh P. Kellermann
Yan Wang
Jessica Montgomery
Sean Joo
Jeffrey D. Kromrey

UNIVERSITY OF SOUTH FLORIDA

#SASGF

Do SAS® High-Performance Statistical Procedures Really Perform “Highly”?
A Comparison of HP and Legacy Procedures

Introduction

HP PROCEDURES:

Respond to the growth of big data, and computer capabilities

1. HPGENSELECT®
2. HPREG®
3. HPCORR®
4. HPNLMOD®
5. HPSUMMARY®
6. HPLOGISTIC®

PURPOSE OF THE STUDY:

Describe differences between key HP procedures and their legacy counterparts in terms of capacity and performance. Main focus is on differences in Real Time and CPU time required for execution

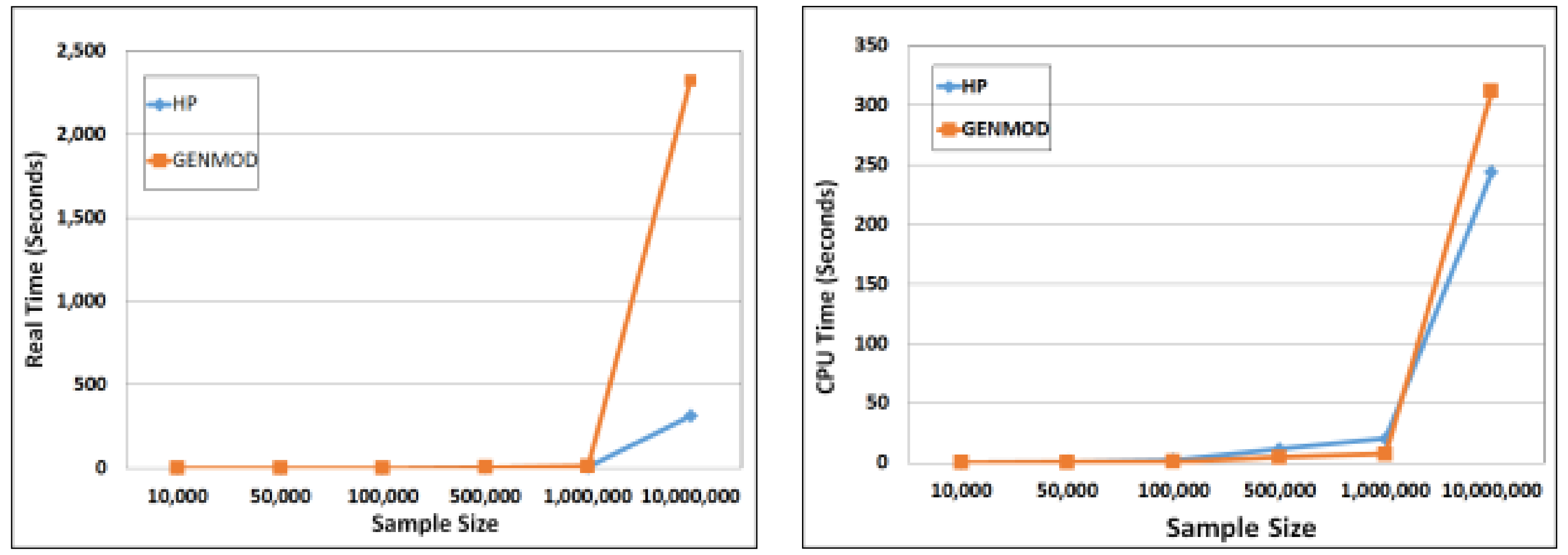
Method

SIMULATION FACTORS:

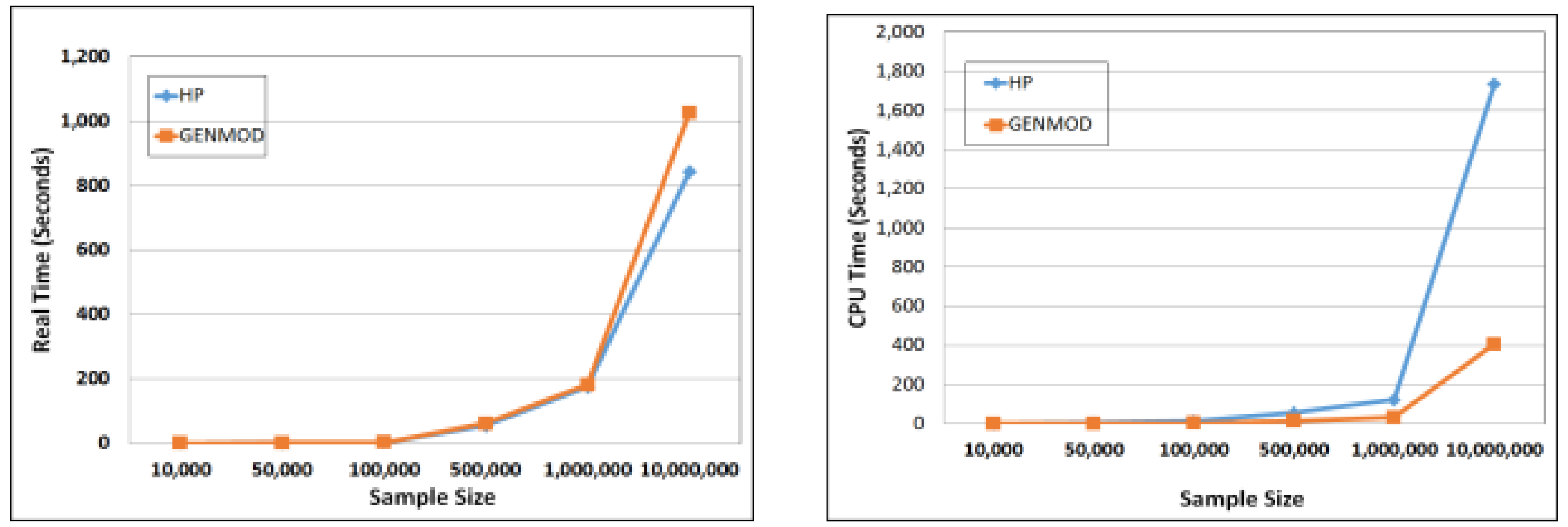
- **Sample size (*N*):** 10,000; 50,000; 100,000; 500,000; 1,000,000; 10,000,000
- **Number of variables (*k*):** 50, 100, 500, 1000
- **All procedures were run in desktop computers and high performing cluster computers**
- **For each procedure, each condition was run10 and averaged its CPU Time and Real Time**

Results

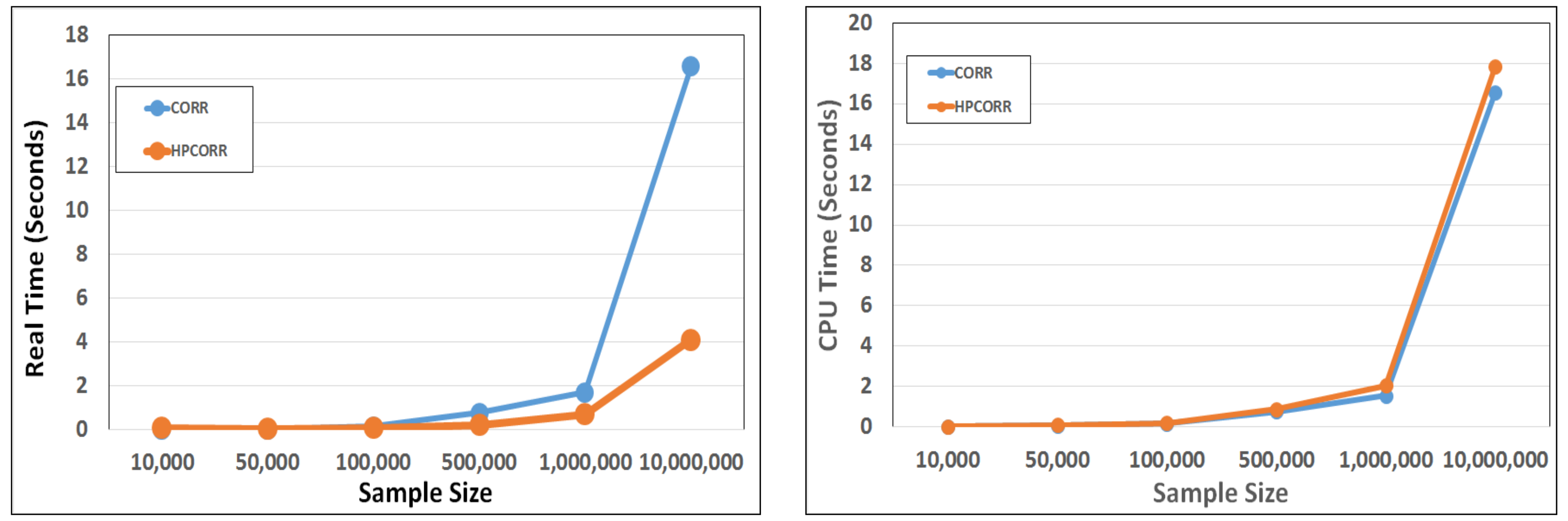
HPGENSELECT and GENMOD (Poisson; k = 50)



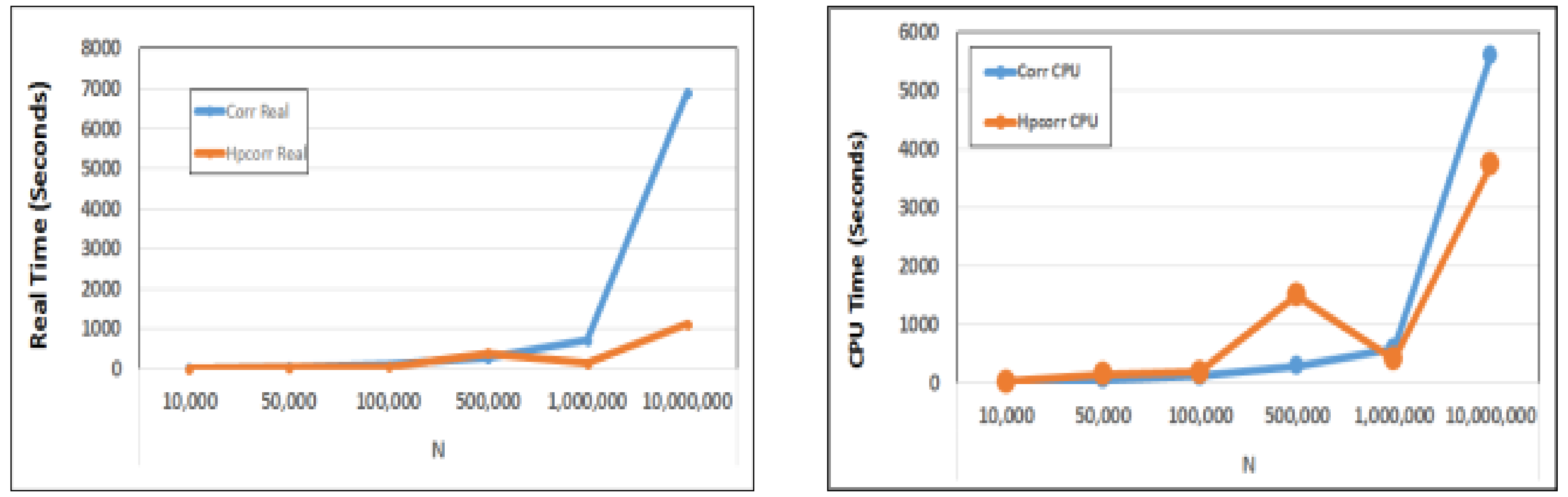
HPGENSELECT and GENMOD (Poisson; k = 100)



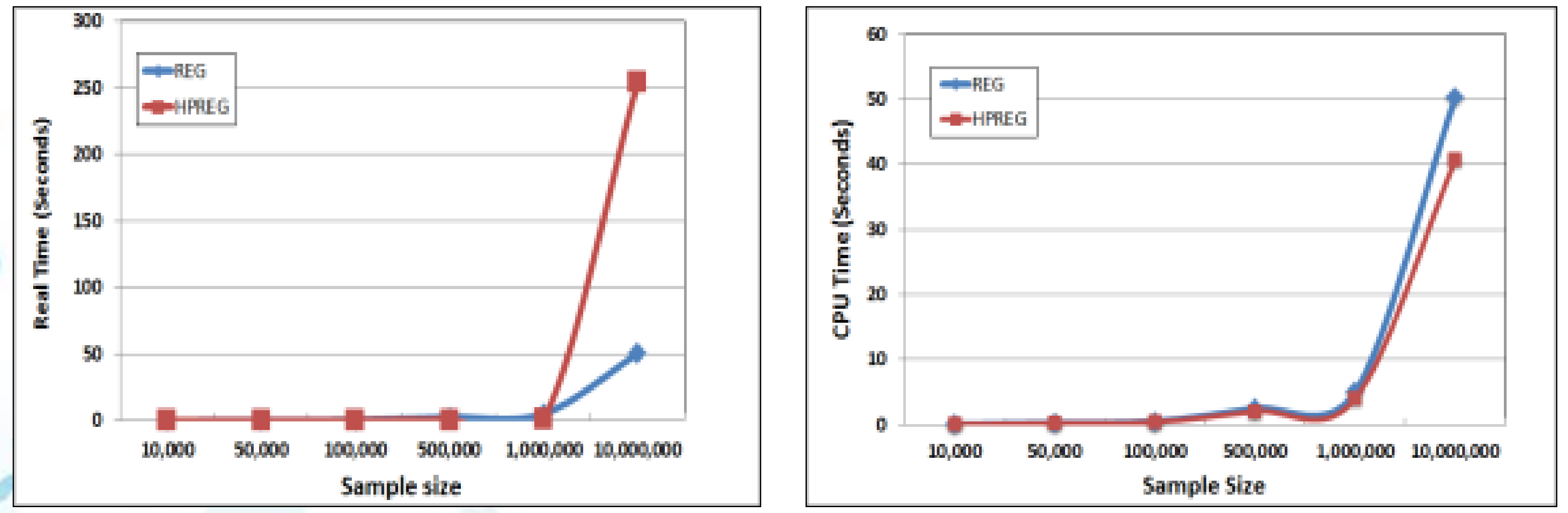
HPCORR and CORR (Cluster; k = 50)



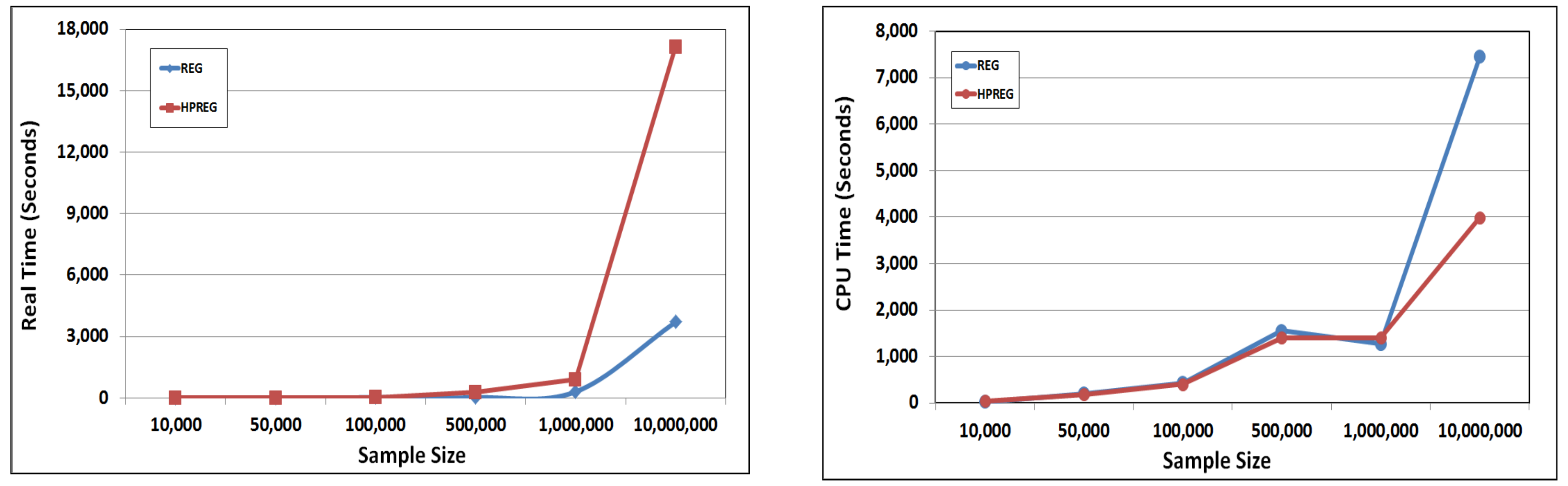
HPCORR and CORR (Cluster; k = 1000)



HPREG and REG (Cluster; k = 50)



HPREG and REG (Cluster; k = 1000)



Conclusions

- There was nearly no divergence between CPU Time and Real Time across procedural pairs for the smallest sample sizes in the study (i.e. 10,000, 50,000, and 100,000)
- Legacy procedures performed well even when sample sizes were pretty large and had a large number of variables (i.e. 1 million observations and 100 variables)
- However, as sample size exceeded a certain threshold, greater improvements associated with some HP procedures (HPGENSELECT, HPCORR, HPREG) were observed, suggesting that the efficiency built into this new code provided greater enhancements as the sample size went up
- Patterns for Real Time and CPU Time in cluster computers were similar to what were found in the personal computers

#SASGF

Do SAS® High-Performance Statistical Procedures Really Perform "Highly"? A Comparison of HP and Legacy Procedures

Introduction

HP PROCEDURES:

Respond to the growth of big data, and computer capabilities

1. HPGENSELECT ®
2. HPREG ®
3. HPCORR ®
4. HPNLMOD ®
5. HPSUMMARY ®
6. HPLOGISTIC ®

PURPOSE OF THE STUDY:

Describe differences between key HP procedures and their legacy counterparts in terms of capacity and performance. Main focus is on differences in Real Time and CPU time required for execution

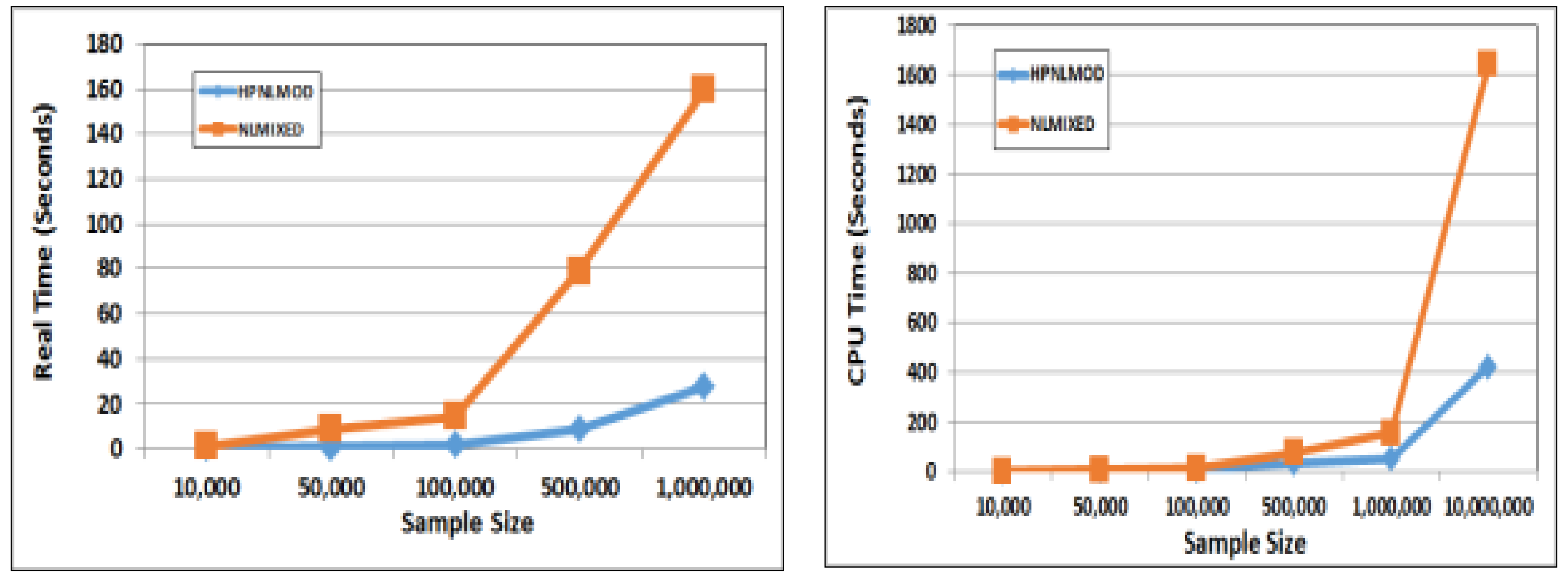
Method

SIMULATION FACTORS:

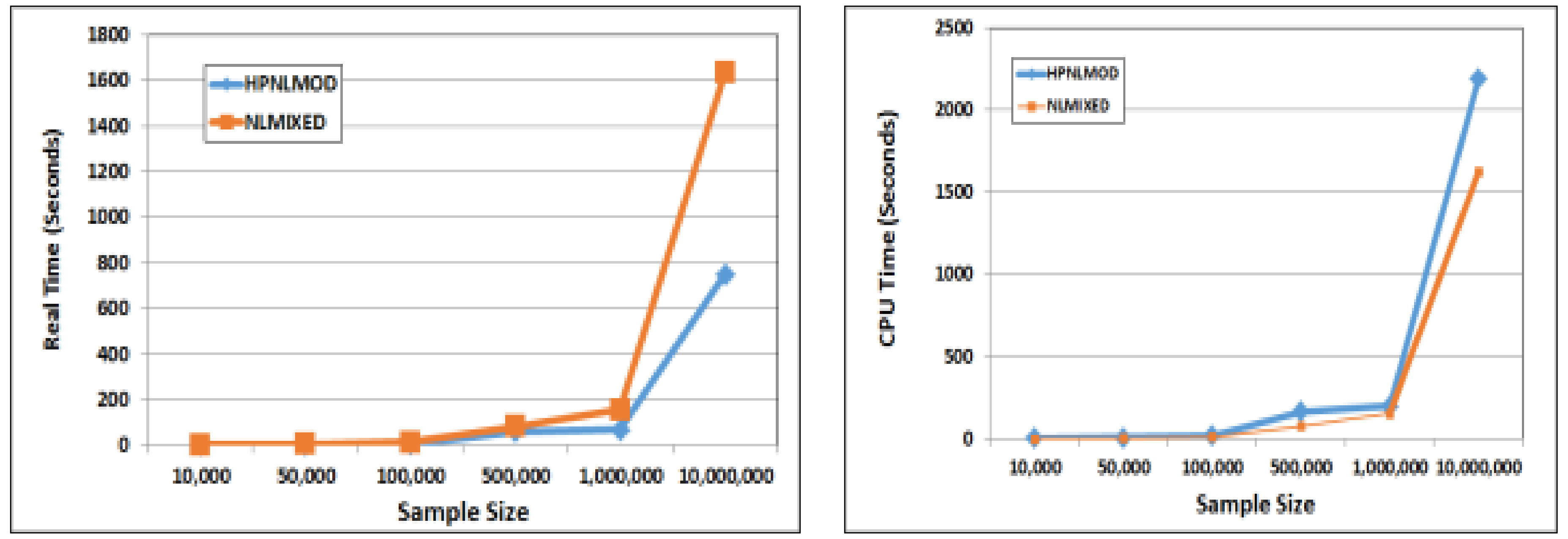
- **Sample size (*N*):** 10,000; 50,000; 100,000; 500,000; 1,000,000; 10,000,000
- **Number of variables (*k*):** 50, 100, 500, 1000
- **All procedures were run in desktop computers and high performing cluster computers**
- **For each procedure, each condition was run10 and averaged its CPU Time and Real Time**

Results Continued

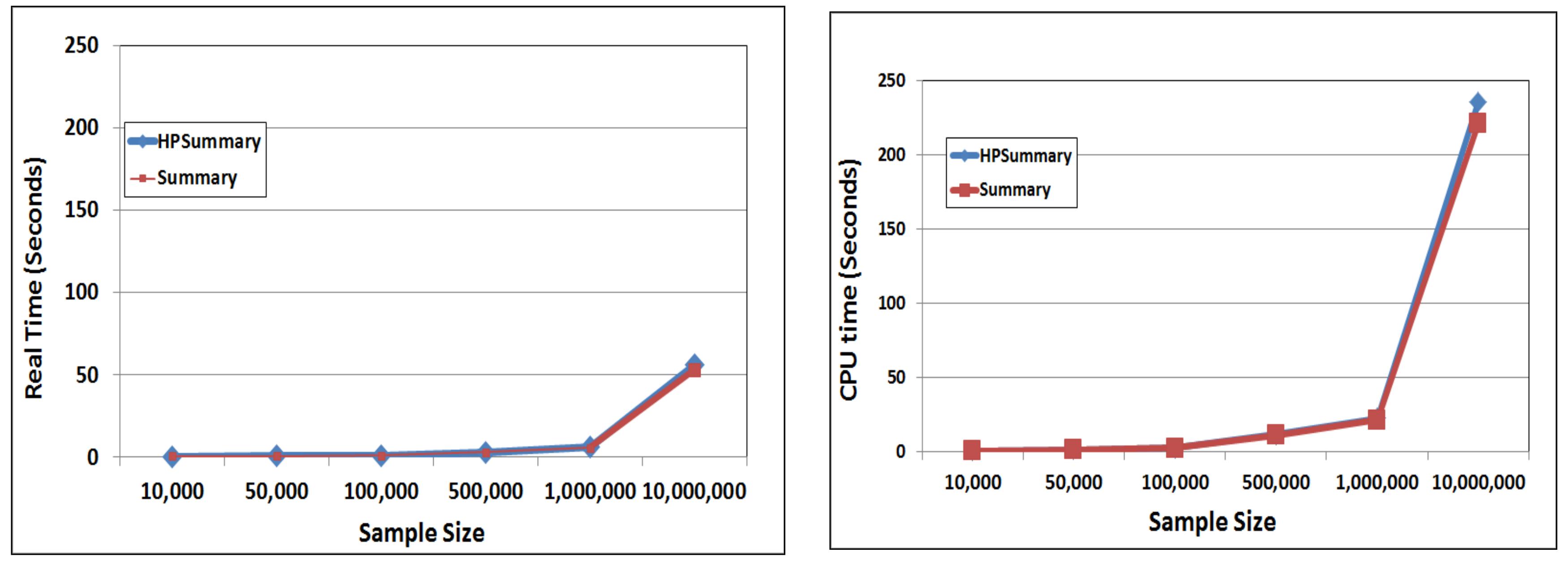
HPNLMOD and NLMIXED (Cluster; *k* = 10)



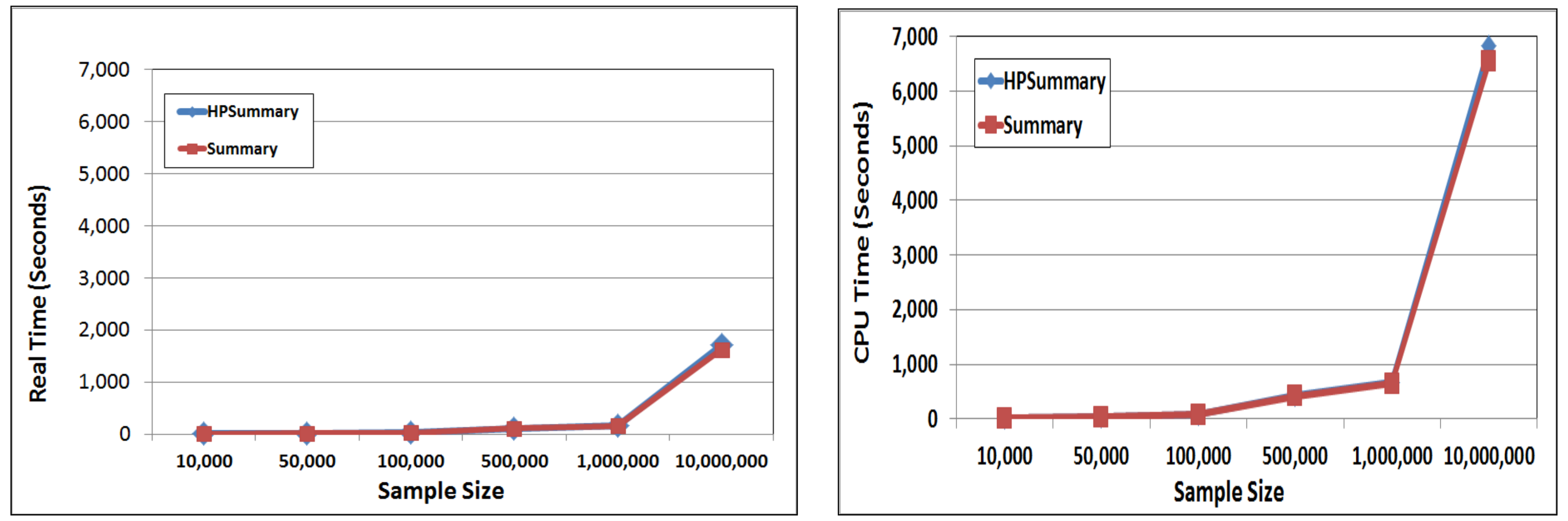
HPNLMOD and NLMIXED (Cluster; *k* = 20)



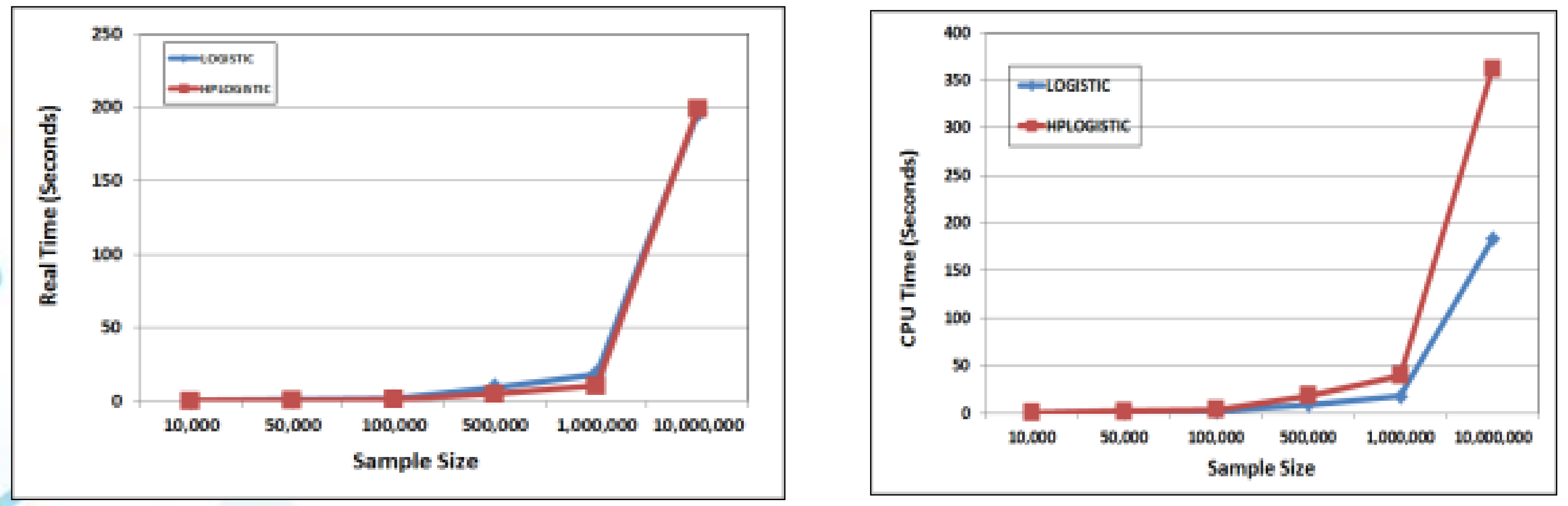
HPSUMMARY and SUMMARY (Cluster; k = 50)



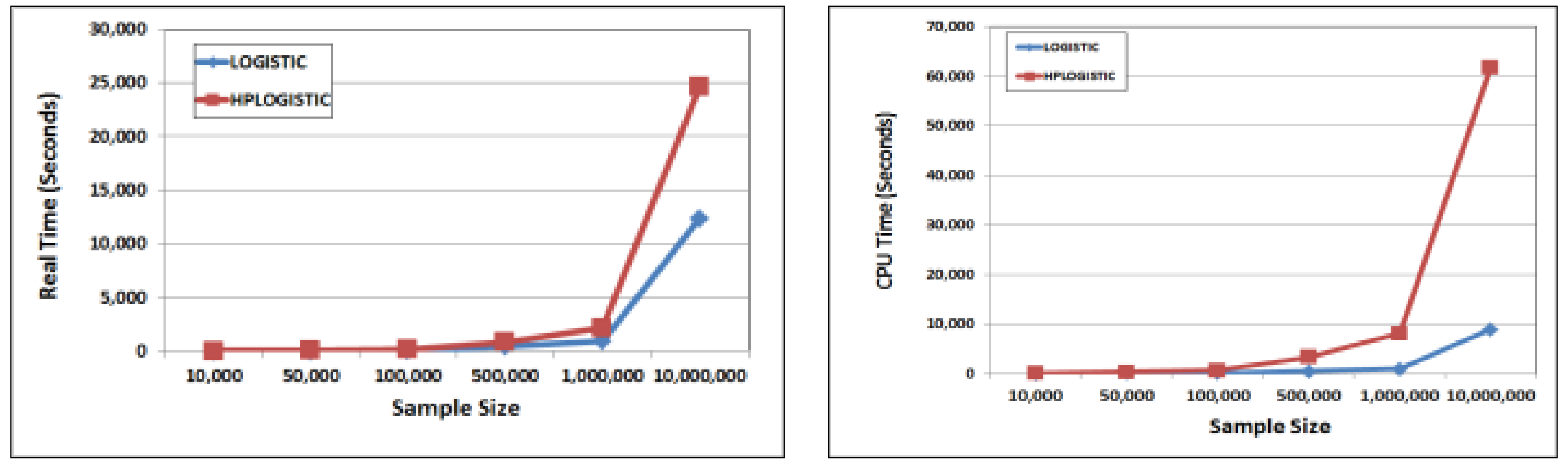
HPSUMMARY and SUMMARY (Cluster; k = 1000)



HPLOGISTIC and LOGISTIC (PC, k = 100)



HPLOGISTIC and LOGISTIC (PC, k = 1000)



Conclusions Continued

- For HPNLMOD & NLMIXED: the legacy procedures could not run when K was larger than 20 while the HP procedures could still perform well in both PC and cluster.
- For the HPSUMMARY &SUMMARY: the pair could run well in cluster but the SUMMARY procedure was shut down when k=500 in combination with N>=100,000 or k=1000 with N>=50,000.
- The pair of HPGENSELECT and GENMOD with Poisson distribution couldn't run when k>100 because the computer couldn't compute the positive definite covariance matrix.

#SASGF

Contact the Author!

Diep T. Nguyen
Doctoral Candidate
diepnguyen@usf.edu