

Parallelization in Action with SAS Analytic Procedures

Robert Cohen
Senior Research Statistician
Linear Models R&D

Your Rise and Shine Menu

- Parallelization adds value to the IVC

Marketing: I should have slept in

- Multithreading to provide parallel execution

Boring: I should have left when I had the chance

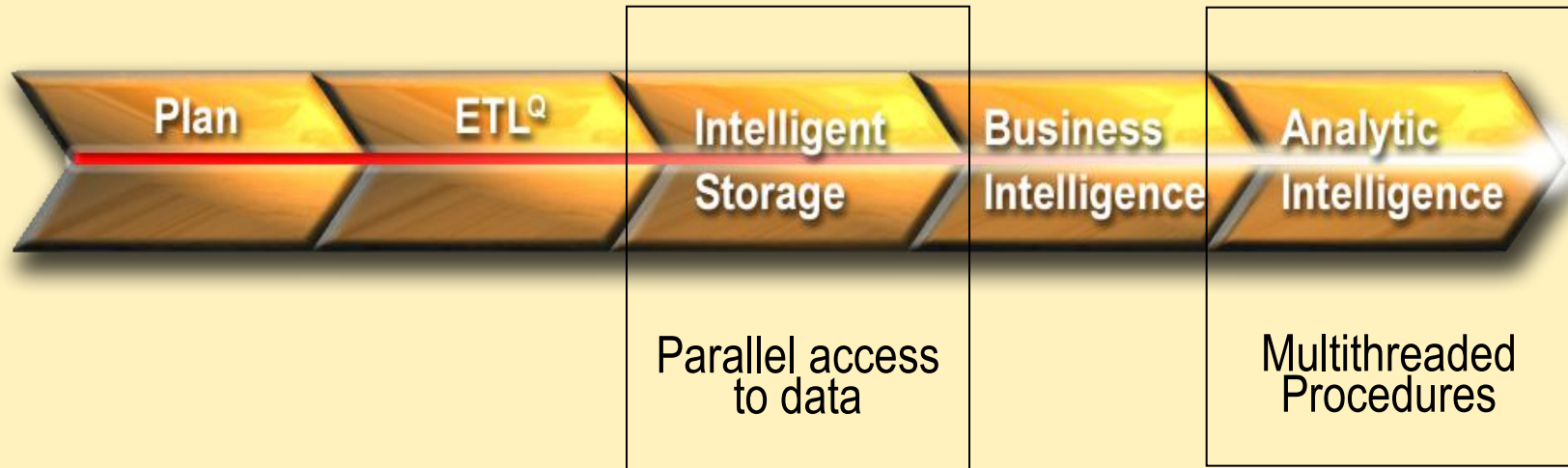
- How do you measure scalability

Insulting: This guy thinks I'm a 10 year old

- Selected demonstrations

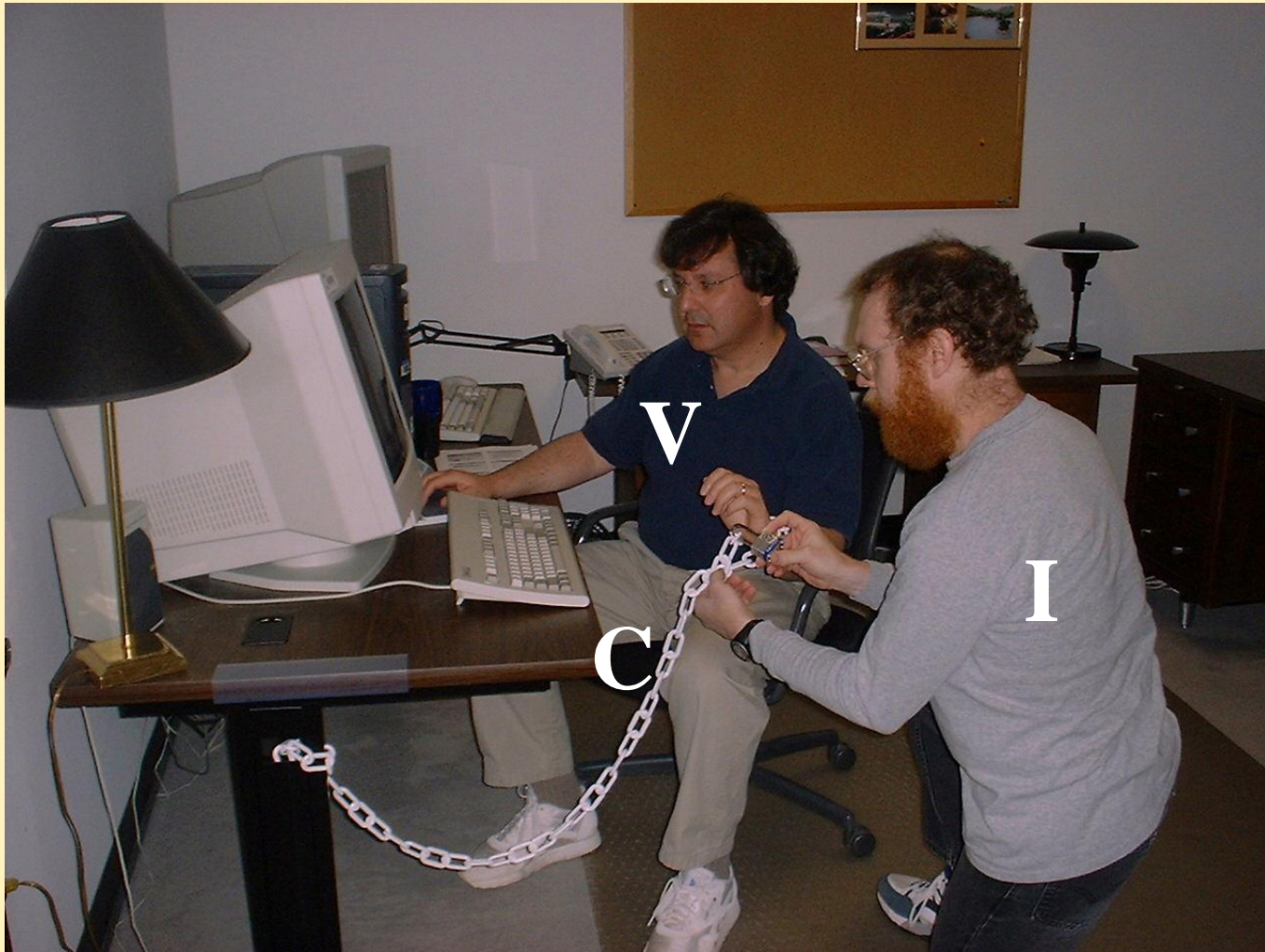
Deceiving: The truth, but not the whole truth

IVC: Parallelization Adds Value



- Complete today's analyses faster
- Analyze tomorrow's problems within today's time constraints

The IVC in Action

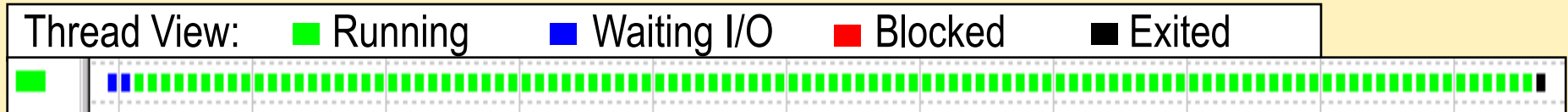


Changes You Have to Make in Your Legacy Code

TINSTAAFL

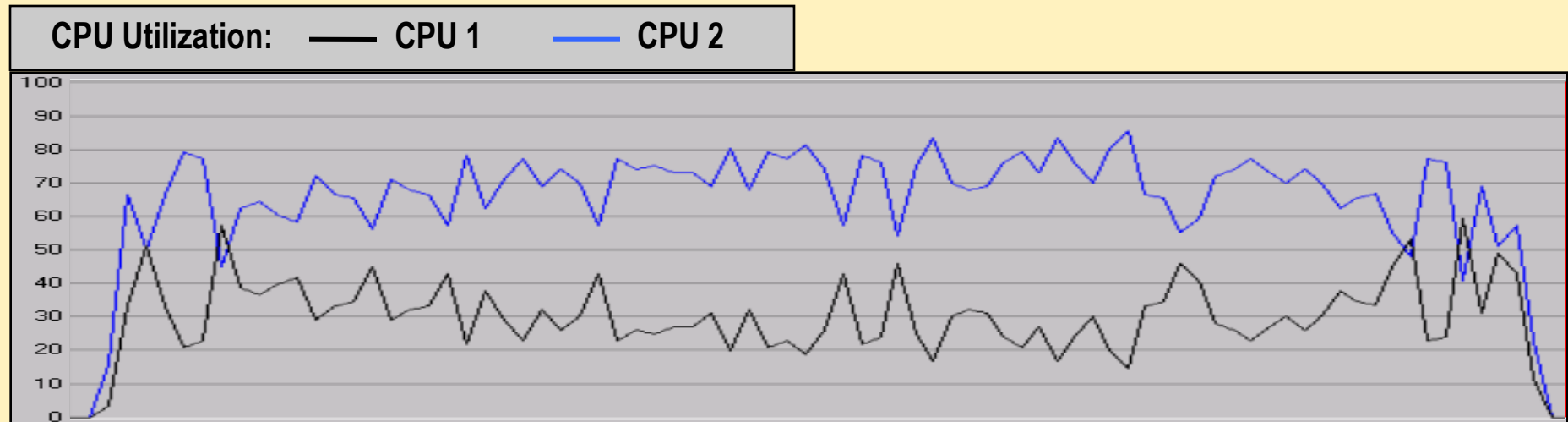
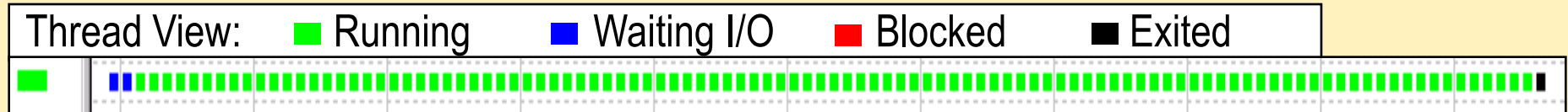


Unthreaded GLM: 2 CPU Box



- GLM runs in a single thread
- GLM never blocks this thread
- GLM work is NOT done in parallel

Unthreaded GLM: 2 CPU Box



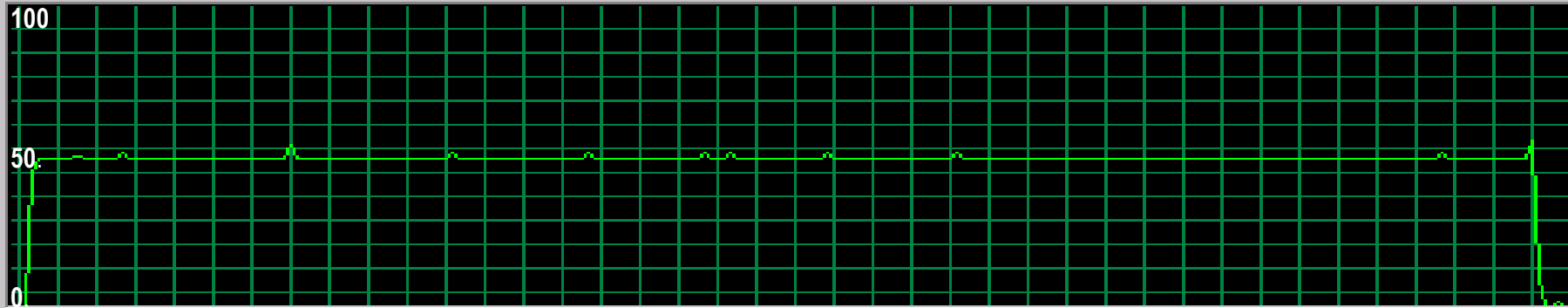
Unthreaded GLM: 2 CPU Box

Thread View: ■ Running ■ Waiting I/O ■ Blocked ■ Exited

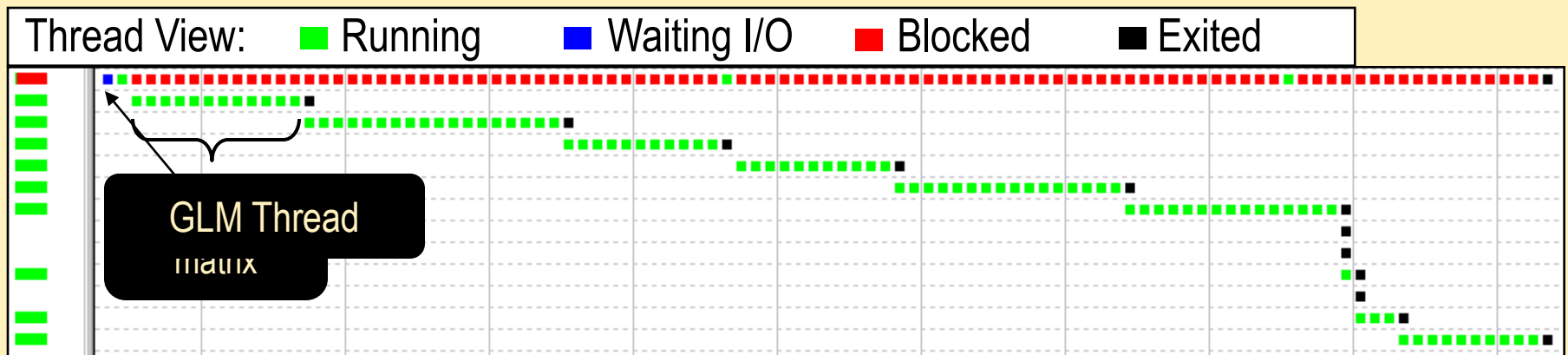


Combined CPU Utilization

CPU Usage History

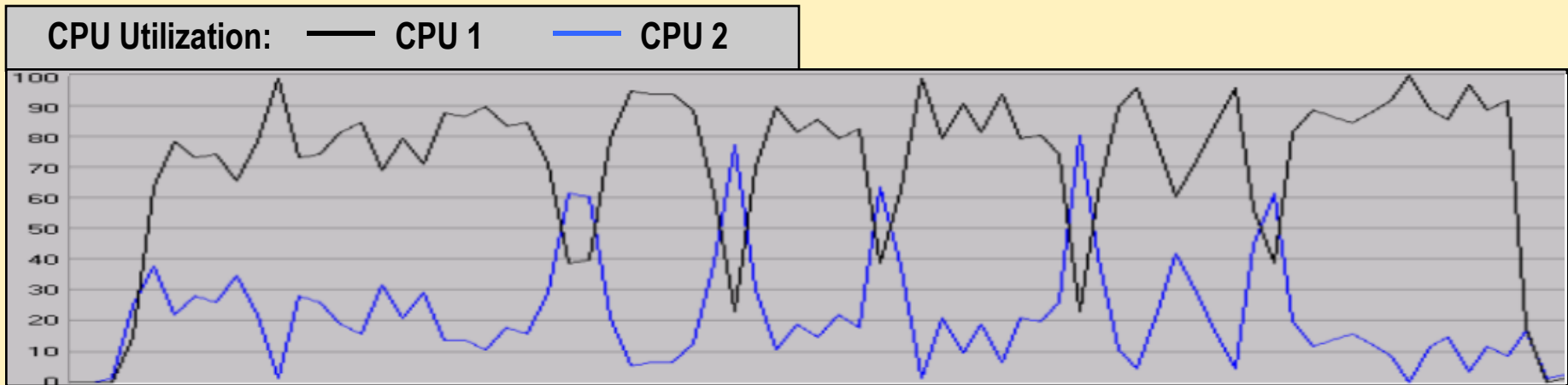
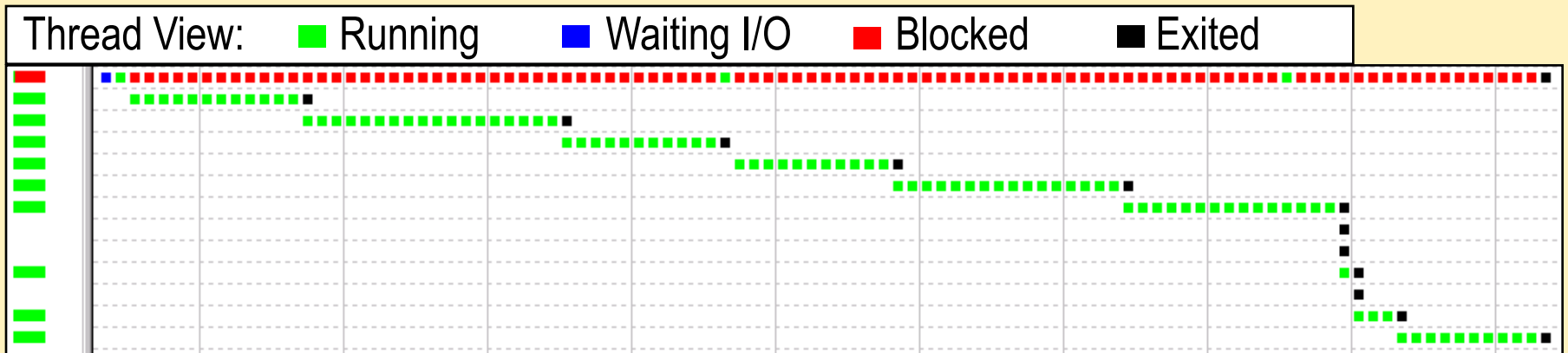


Multithreaded GLM: 1 Active Thread 2 CPU Box

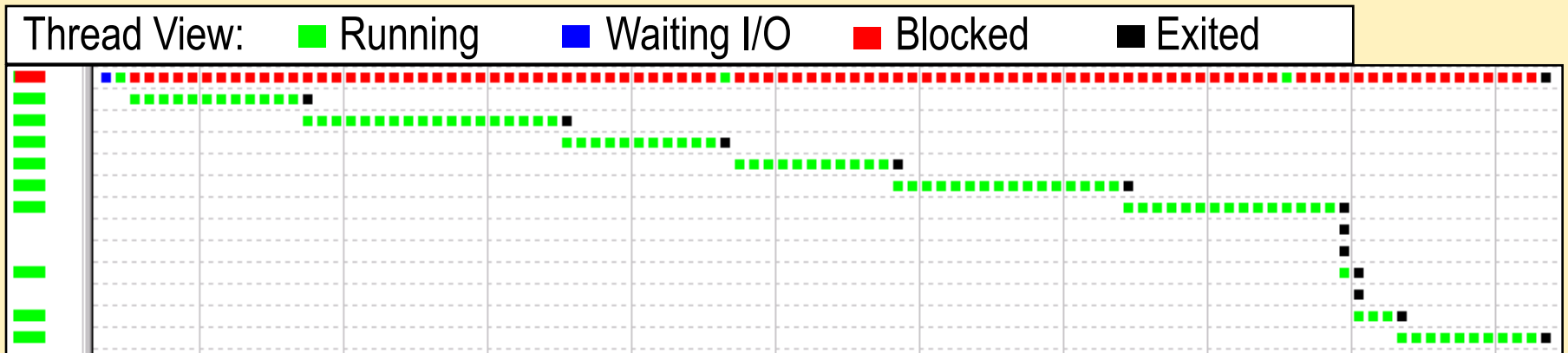


- Worker threads used for specific tasks
- GLM thread blocks while a worker thread is active
- GLM does not execute in parallel

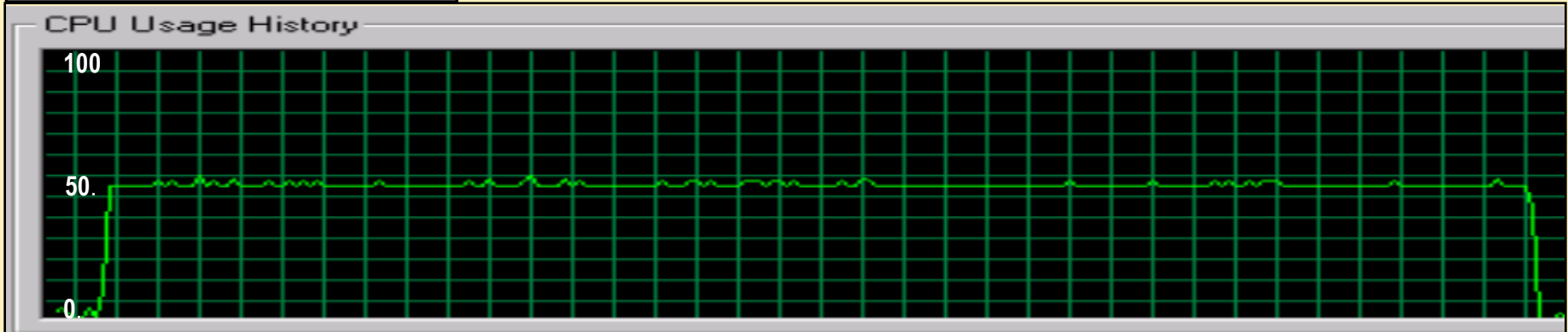
Multithreaded GLM: 1 Active Thread 2 CPU Box



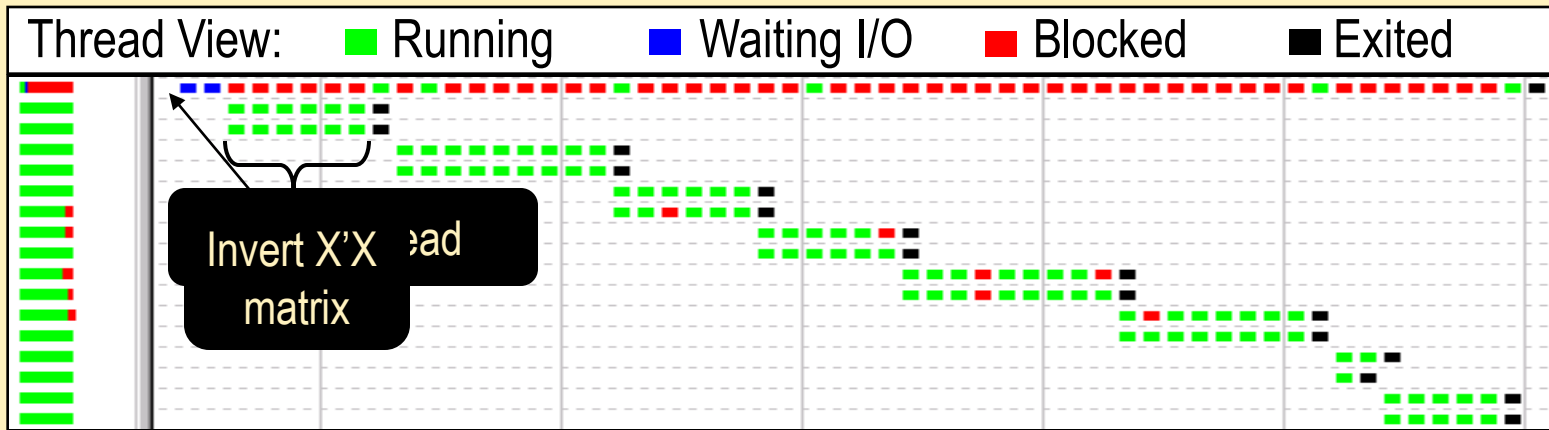
Multithreaded GLM: 1 Active Thread 2 CPU Box



Combined CPU Utilization

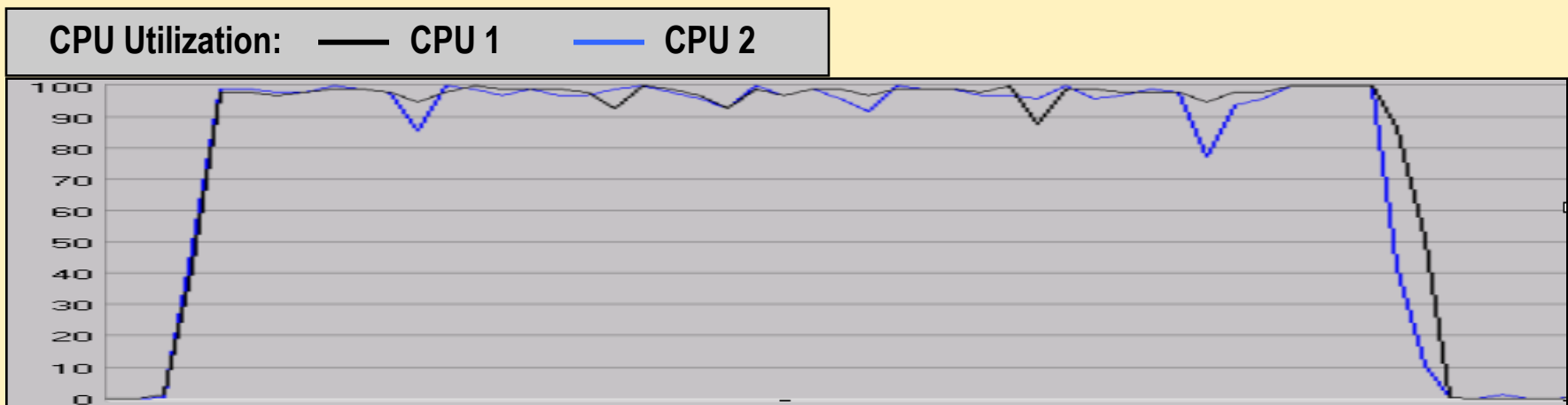
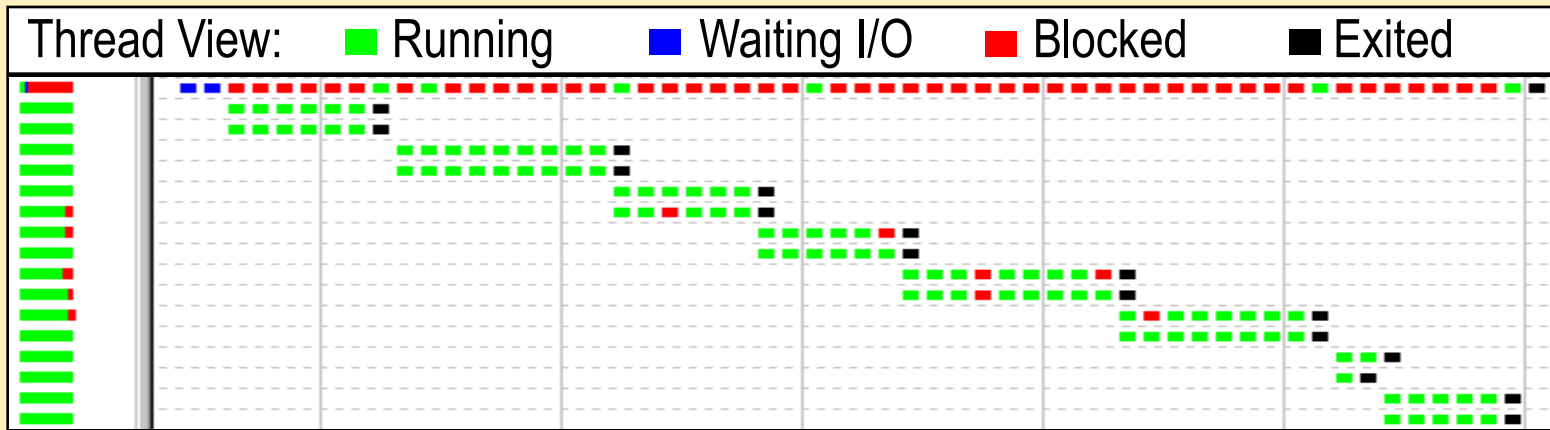


Multithreaded GLM: 2 Active Threads 2 CPU Box

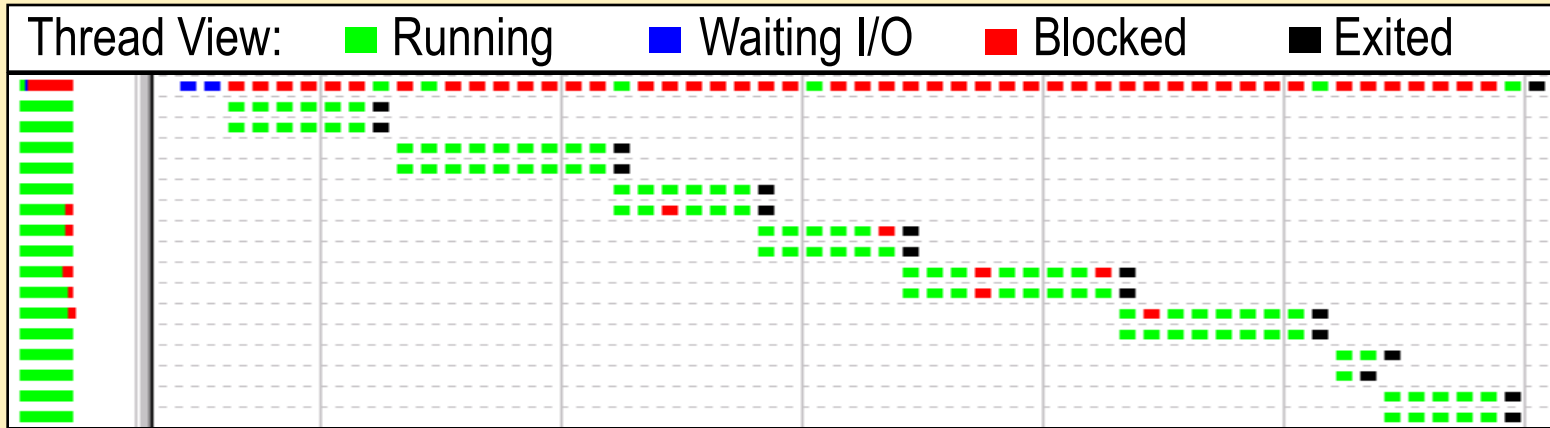


- GLM thread spawns off worker threads
- Two independent worker threads per task
- Work is done in parallel

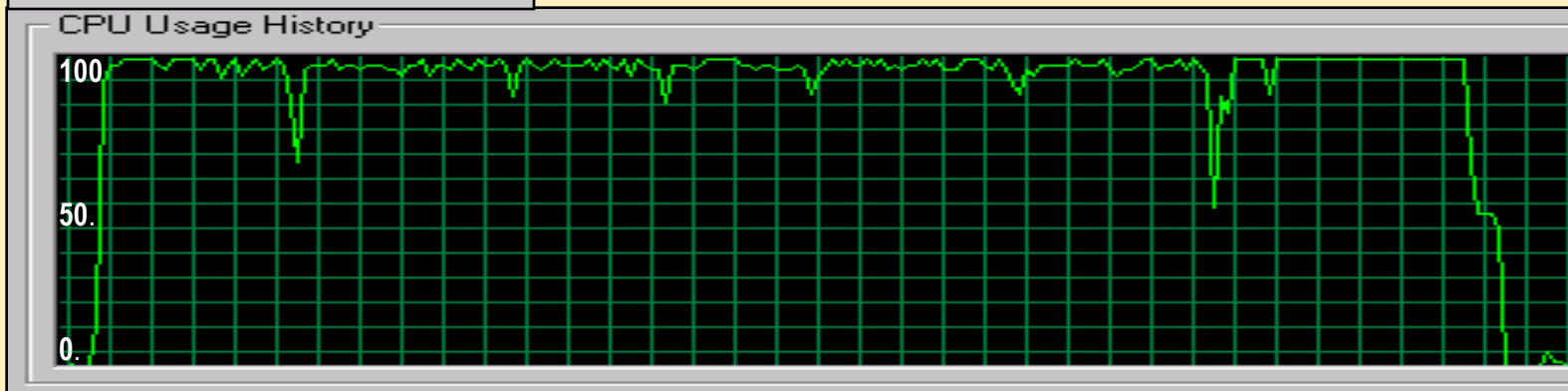
Multithreaded GLM: 2 Active Threads 2 CPU Box



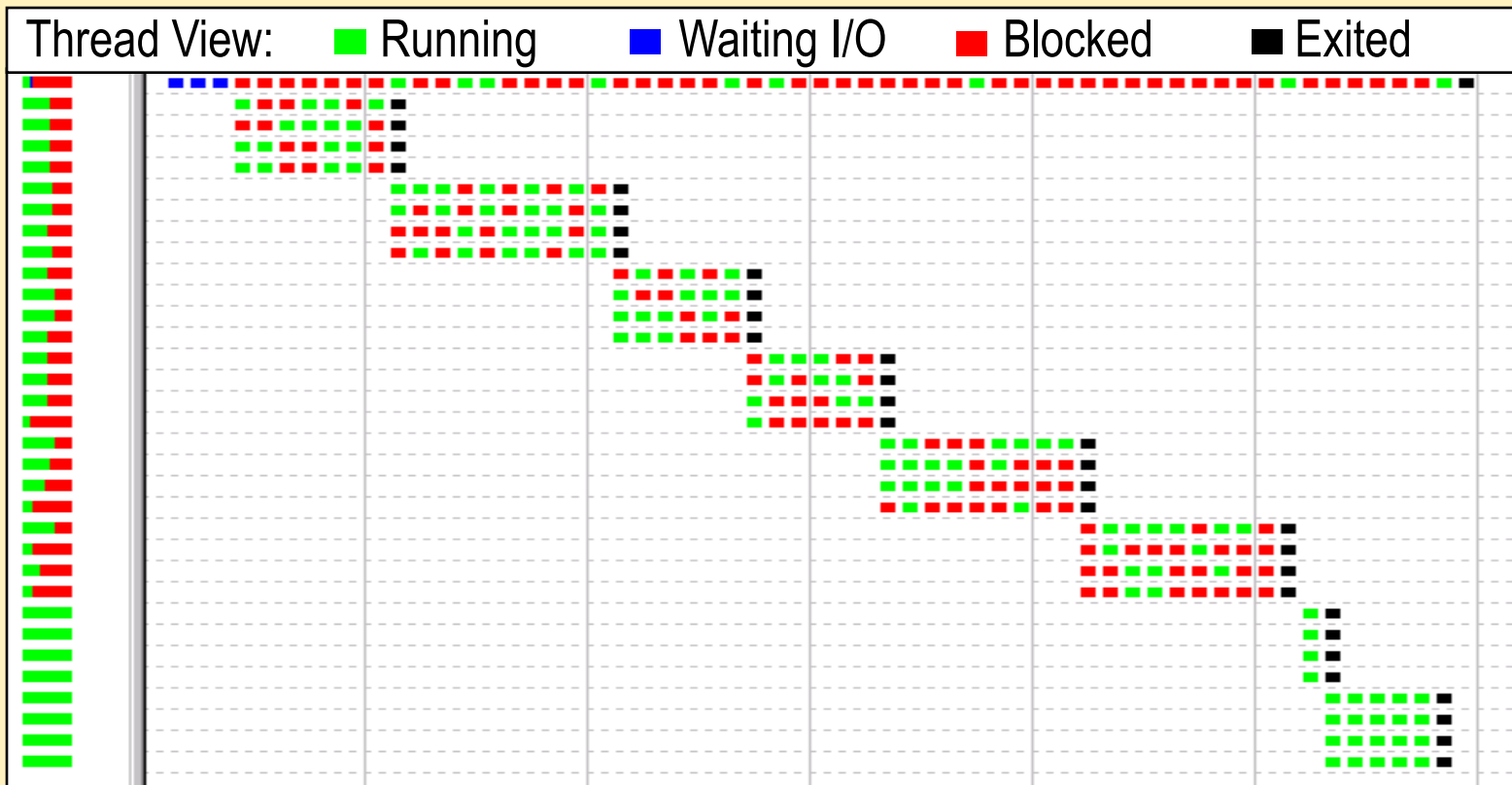
Multithreaded GLM: 2 Active Threads 2 CPU Box



Combined CPU Utilization

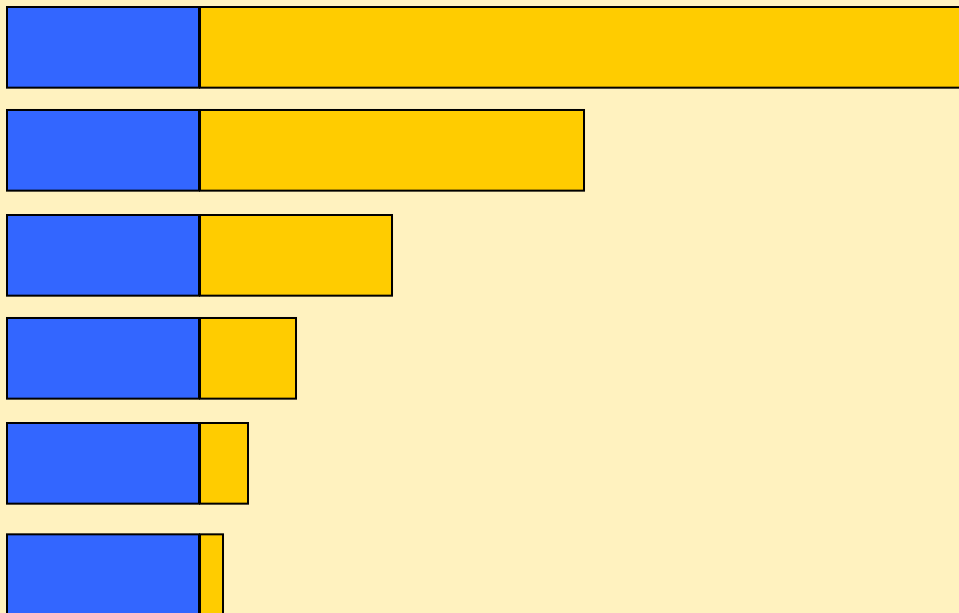


Multithreaded GLM: 4 Active Threads 2 CPU Box



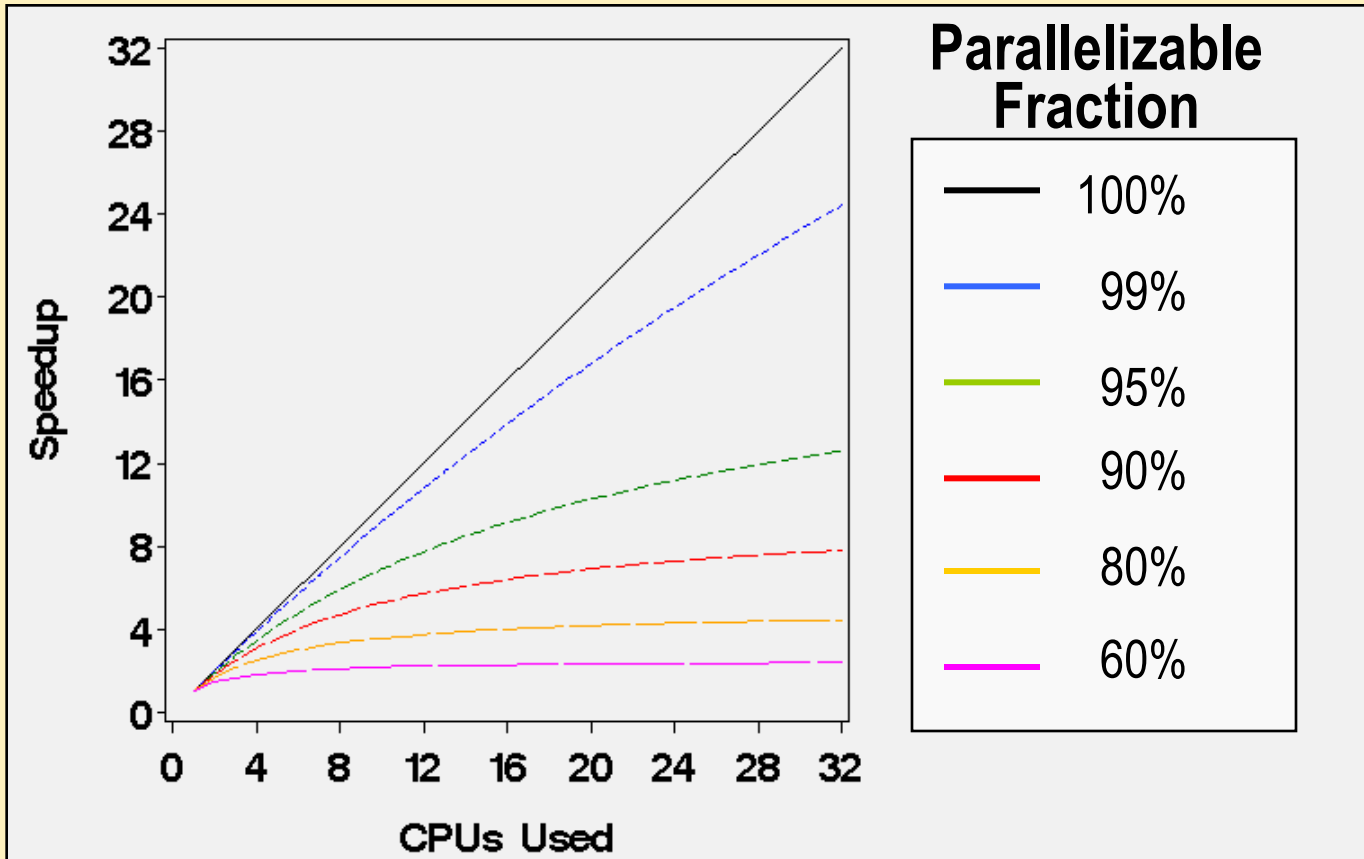
Amdahl's Law

Not Scalable Scalable PF = 80%



CPUs	Speedup
1	1.00
2	1.67
4	2.50
8	3.33
16	4.00
32	4.44

Amdahl's Law



Scalability in PROC REG: Wide Data and Scalar I/O

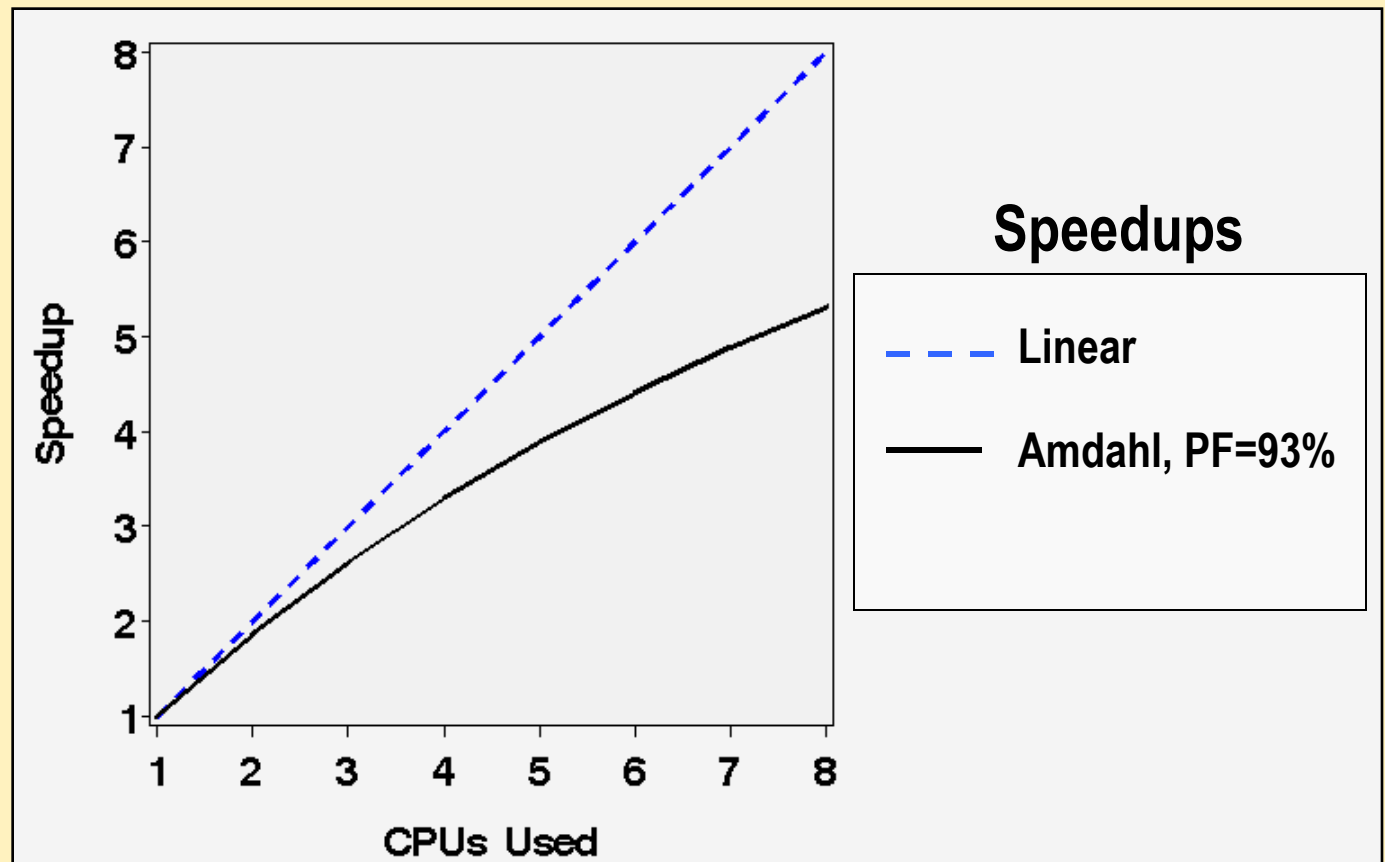
Test Details

50,000 observations

500 predictors

Stepwise Selection

Scalar I/O



Scalability in PROC REG: Wide Data and Scalar I/O

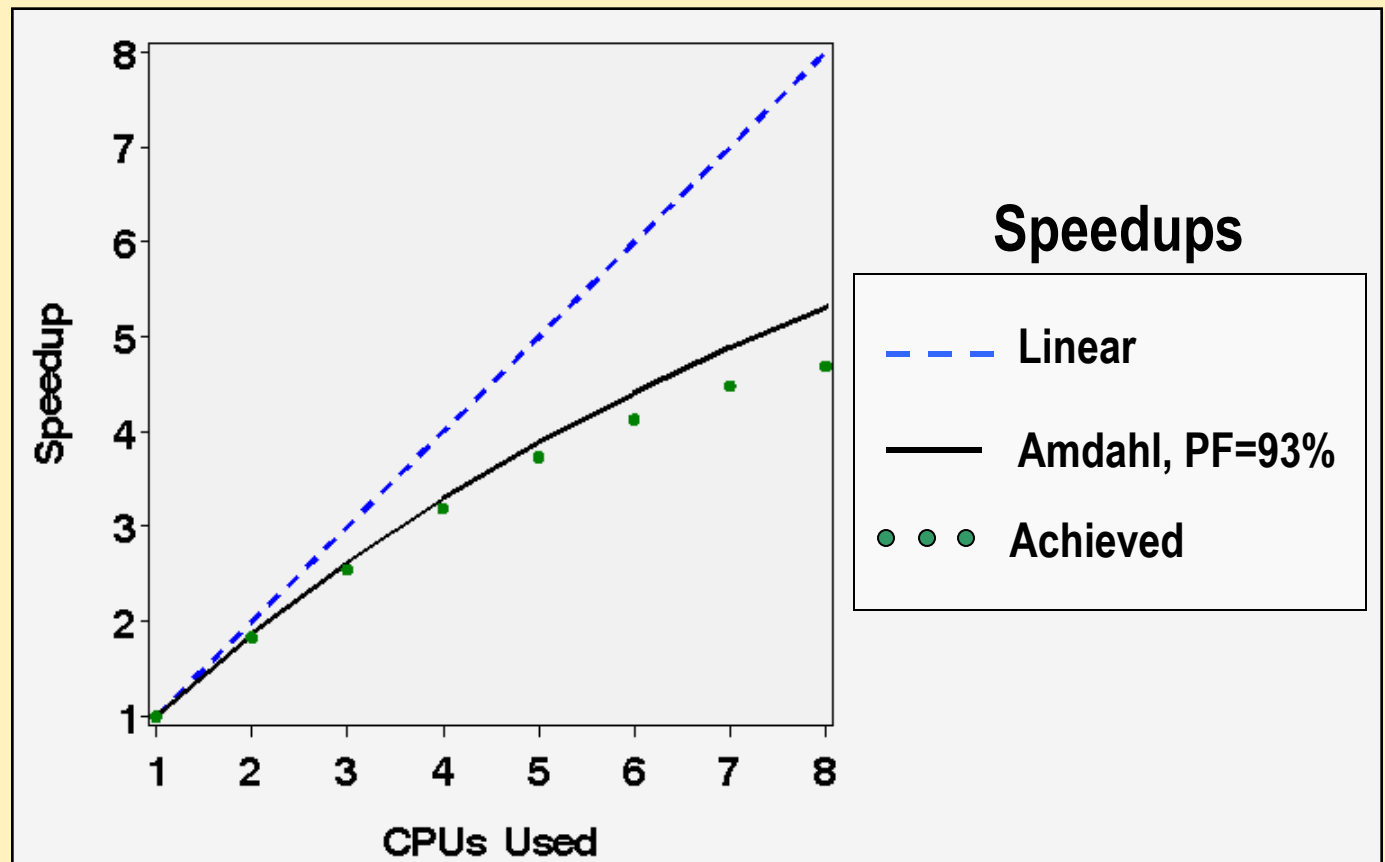
Test Details

50,000 observations

500 predictors

Stepwise Selection

Scalar I/O



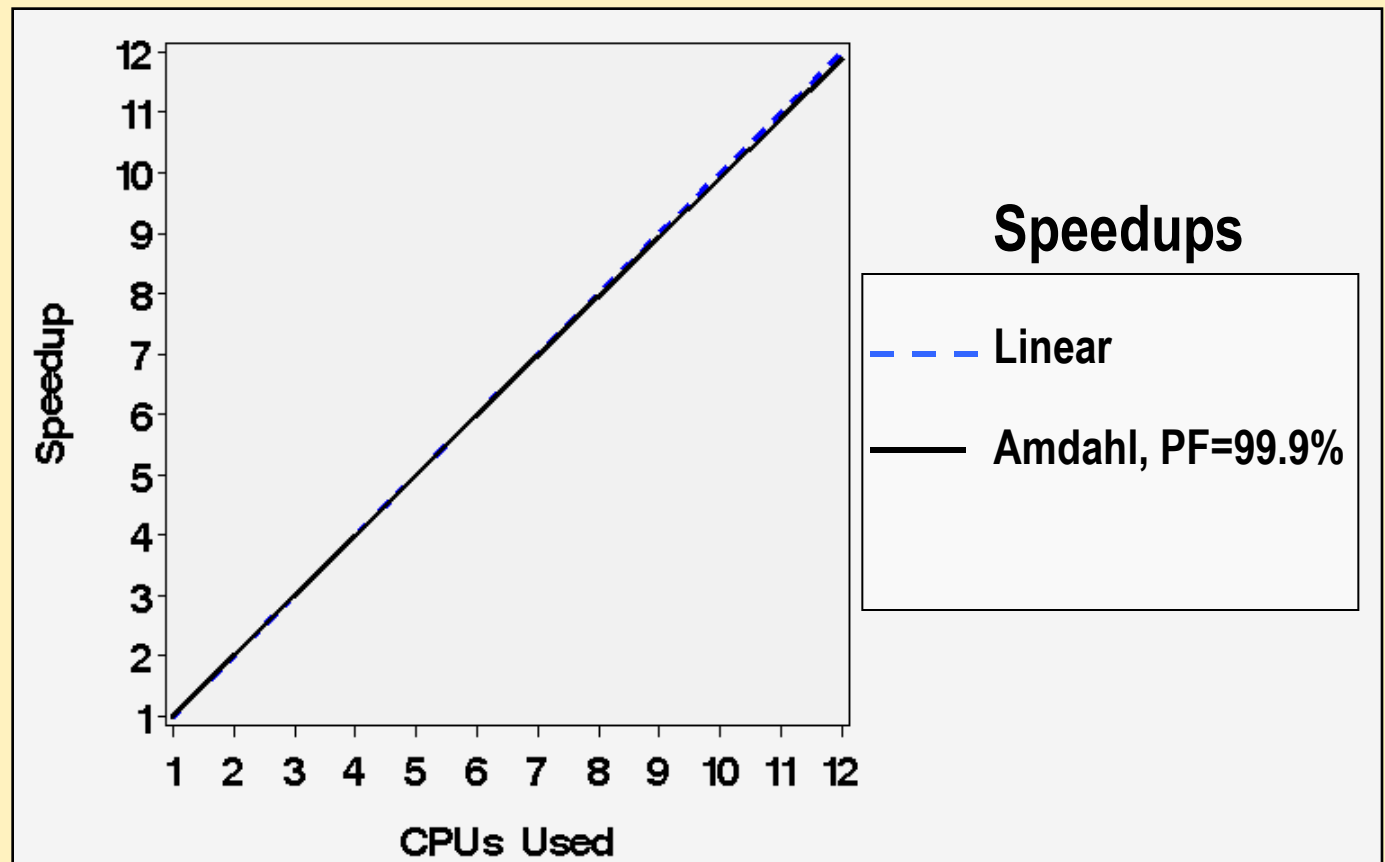
Scalability in PROC REG: Narrow Data, Parallel I/O

Test Details

4 million observations

20 predictors

Parallel I/O



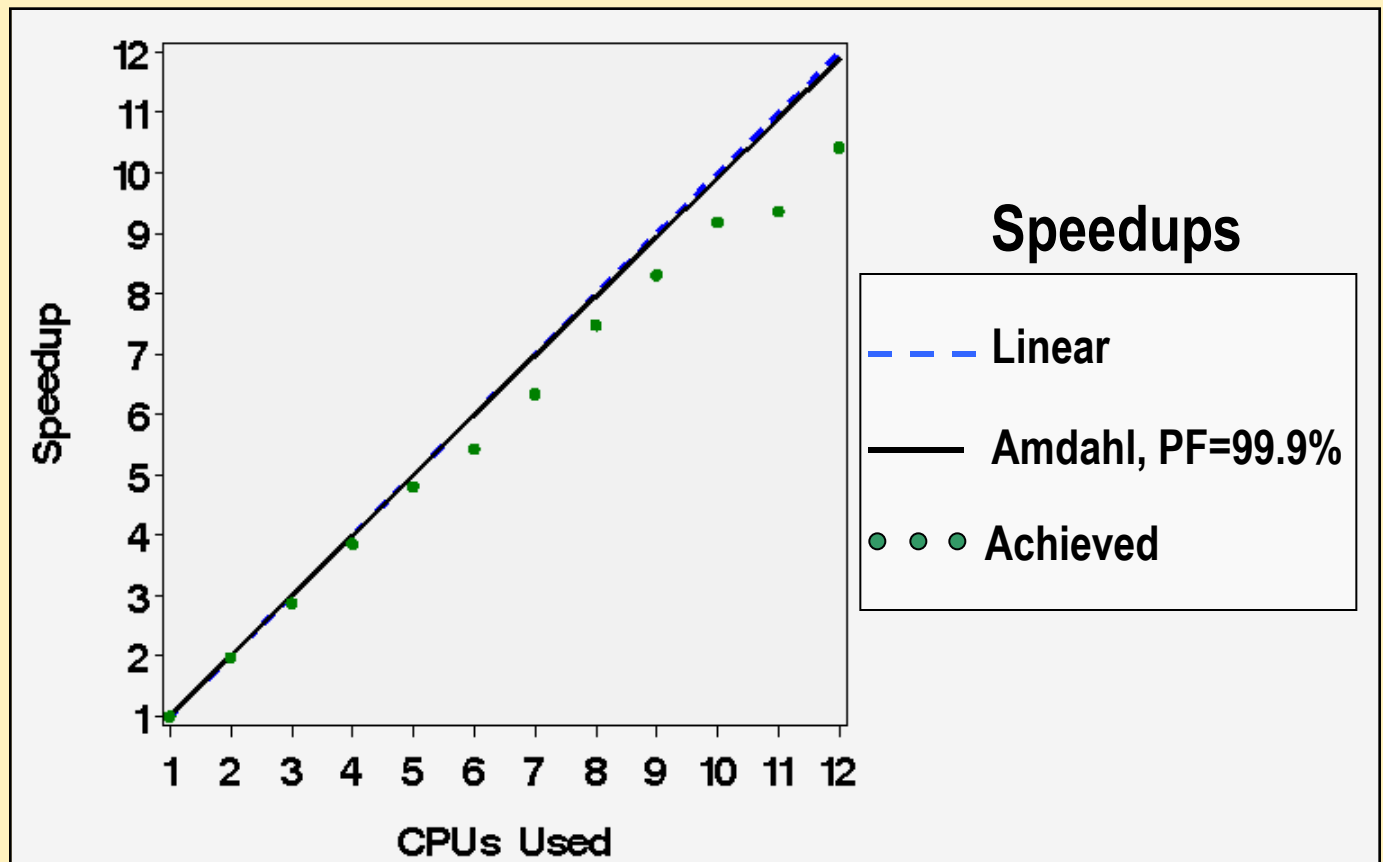
Scalability in PROC REG: Narrow Data, Parallel I/O

Test Details

4 million observations

20 predictors

Parallel I/O



Scalability in PROC DMREG

Test Details

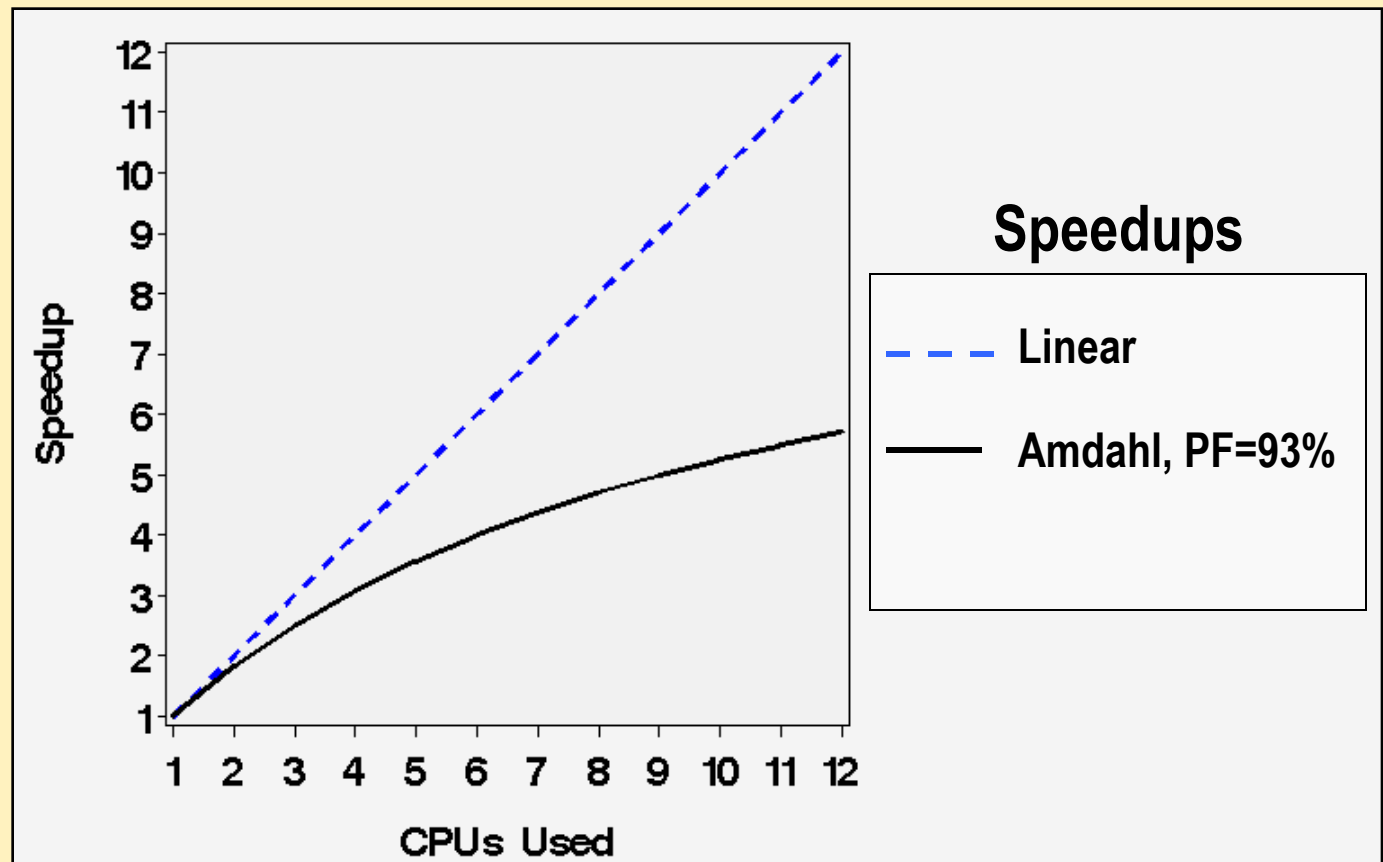
500,000 observations

Predictors:

50 continuous
15 classification

Logistic model

Parallel I/O



Scalability in PROC DMREG

Test Details

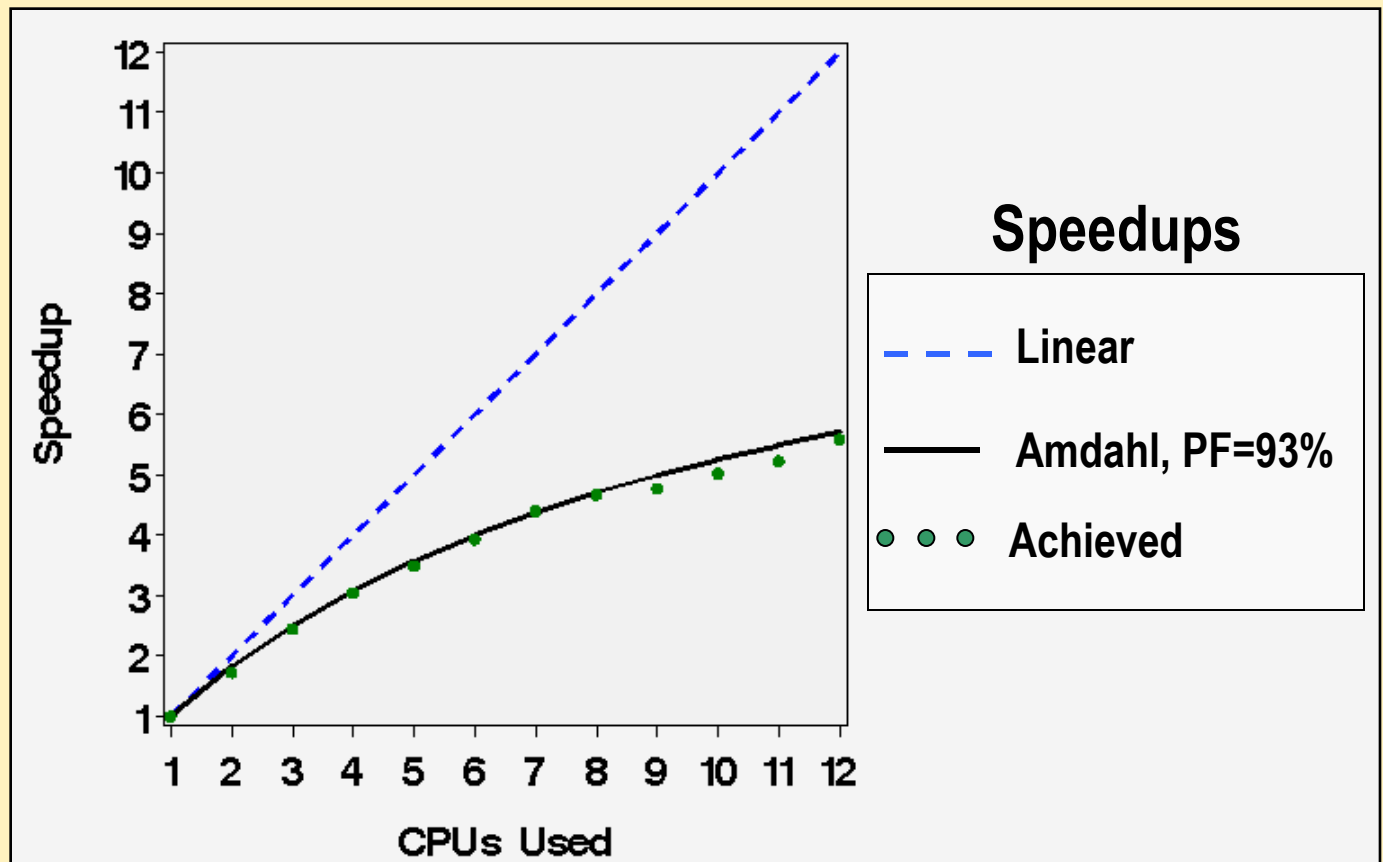
500,000 observations

Predictors:

50 continuous
15 classification

Logistic model

Parallel I/O



Baseline Speedup and Scalability in PROC DMREG

Test Details

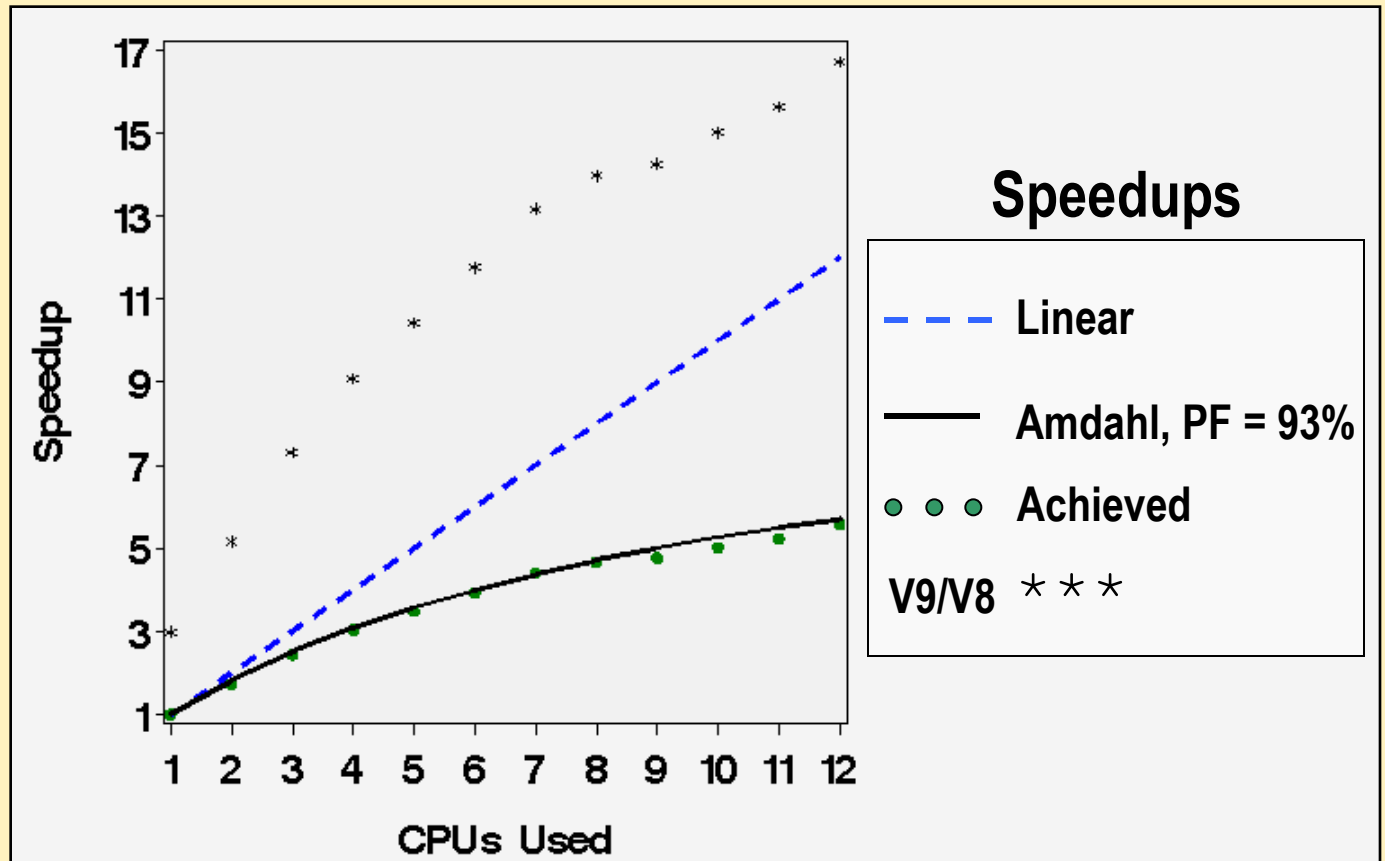
500,000 observations

Predictors:

50 continuous
15 classification

Logistic model

Parallel I/O



Speedups

- -- Linear
- — Amdahl, PF = 93%
- ● ● Achieved
- V9/V8 * * ***

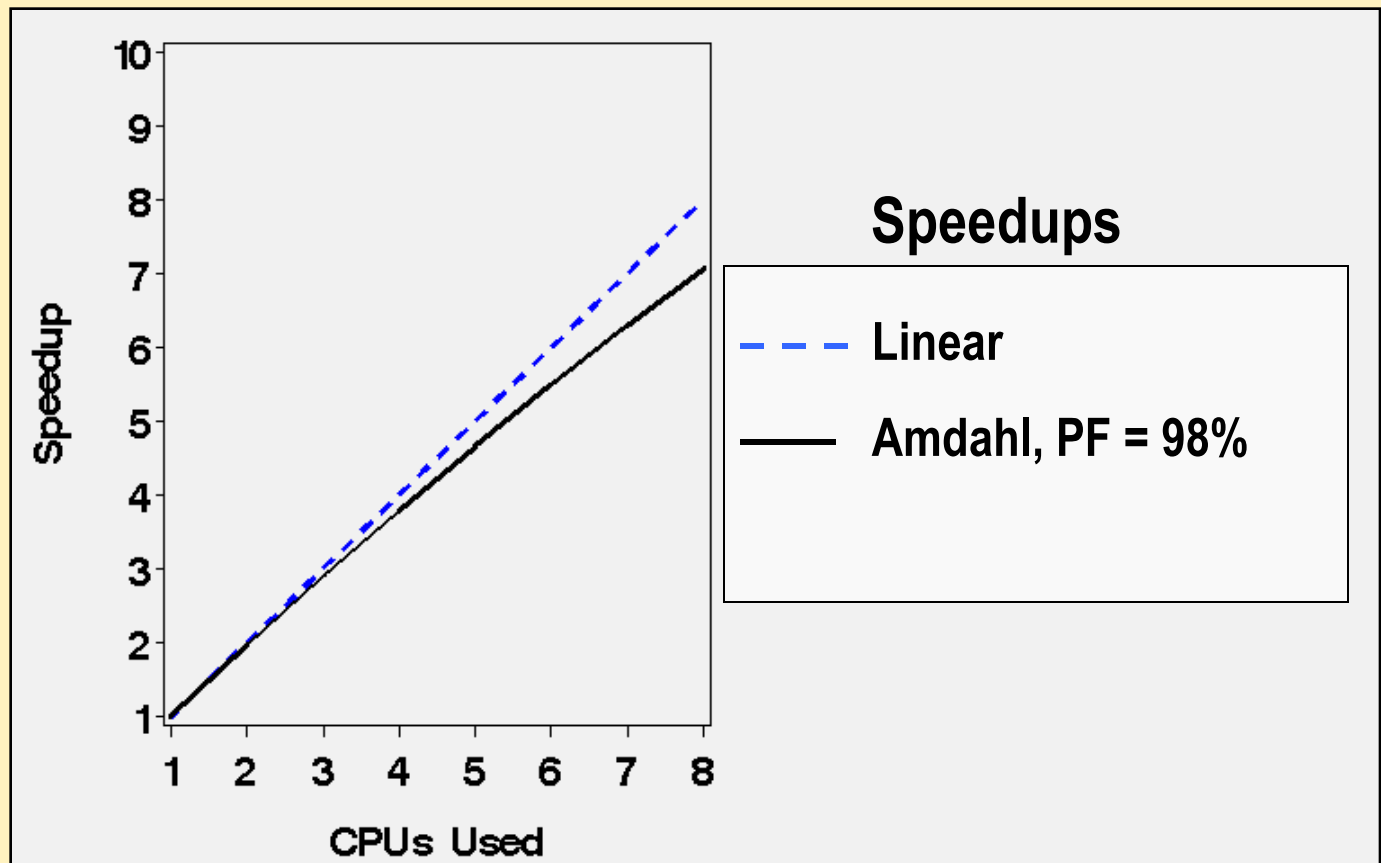
Scalability in PROC GLM

Test Details

6000 observations

4 classification
variables

2000 parameters



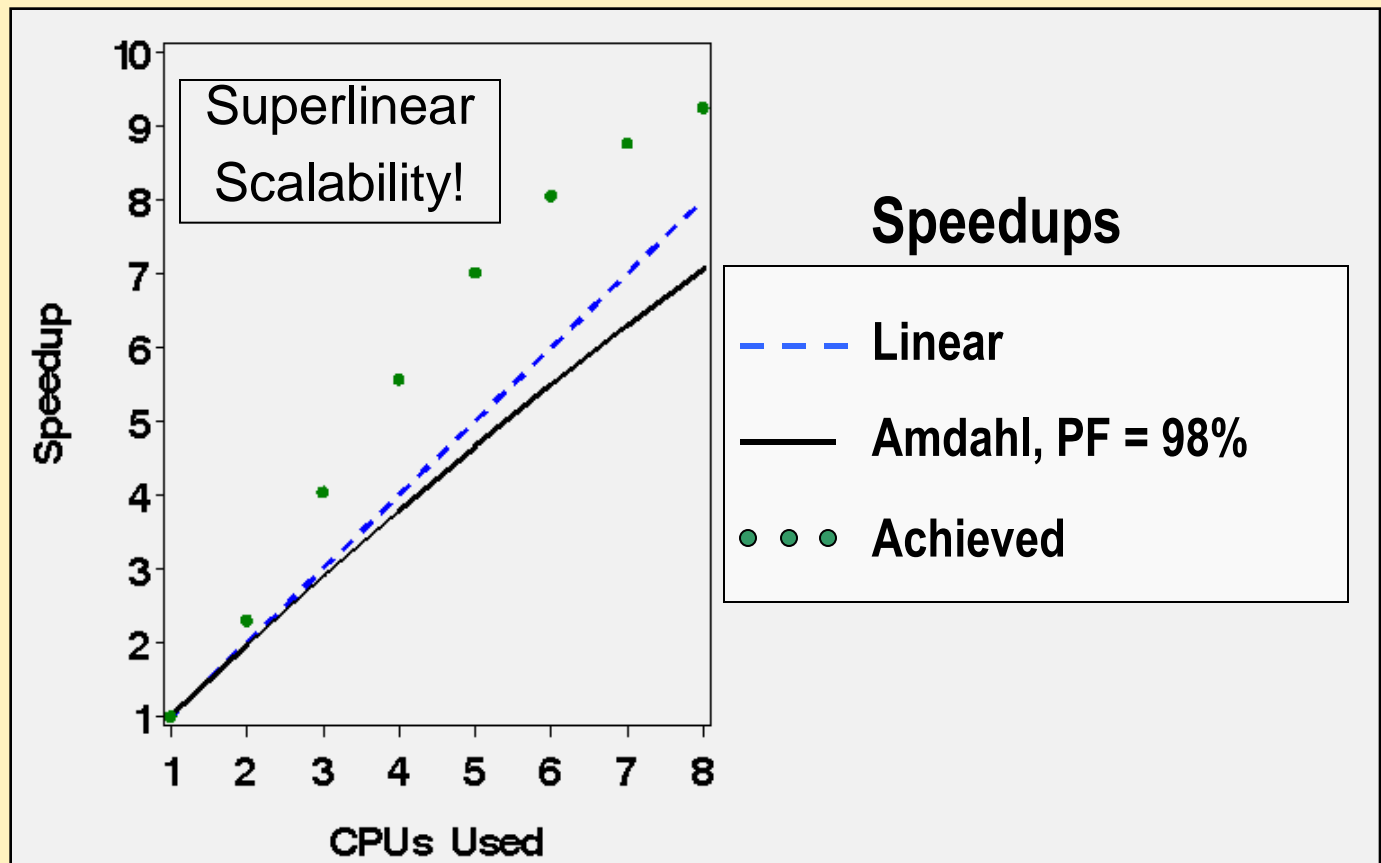
Scalability in PROC GLM

Test Details

6000 observations

4 classification
variables

2000 parameters



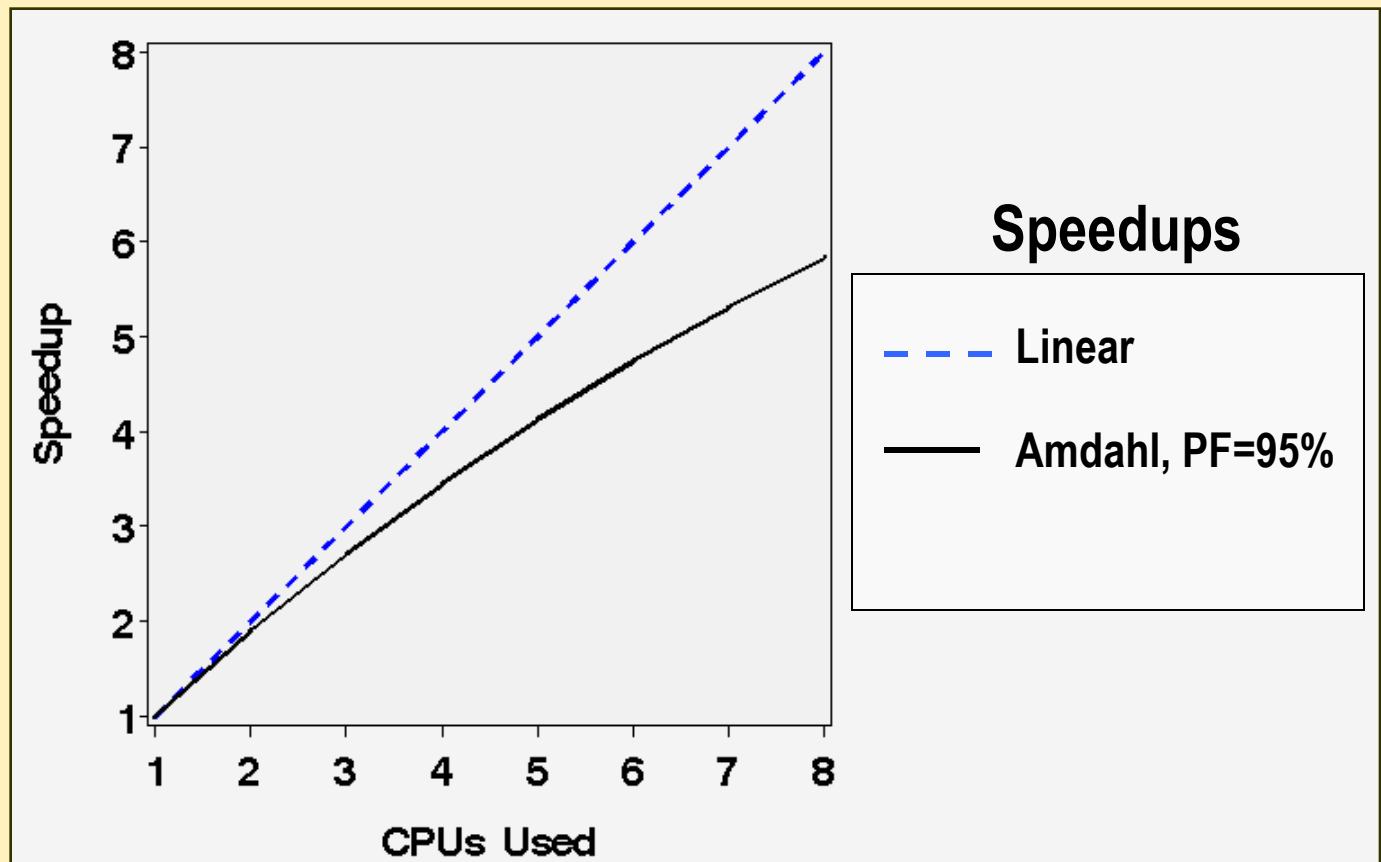
Scalability in PROC LOESS

Test Details

4000 observations

18 models evaluated

Confidence limits for
selected model



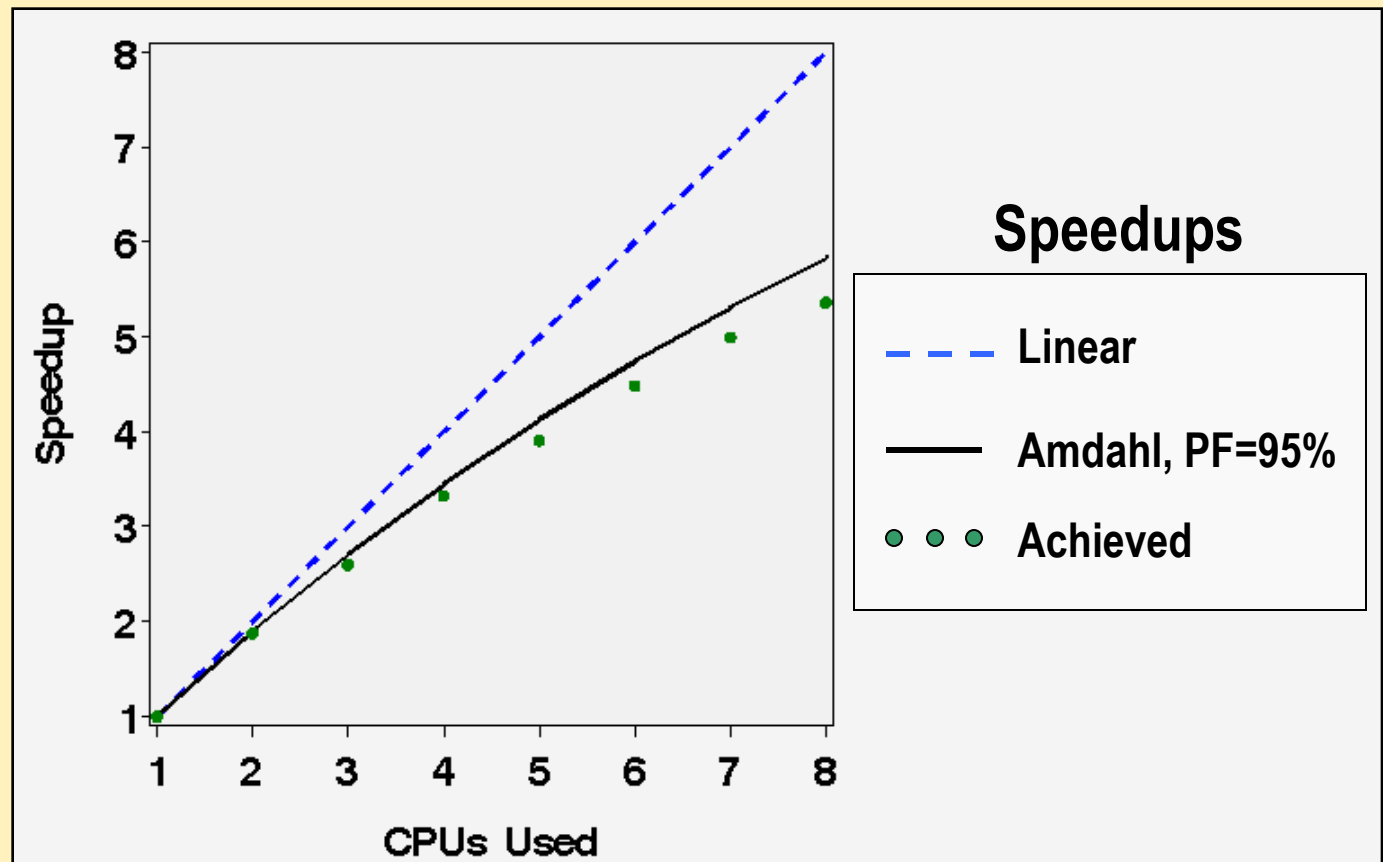
Scalability in PROC LOESS

Test Details

4000 observations

18 models evaluated

Confidence limits for
selected model



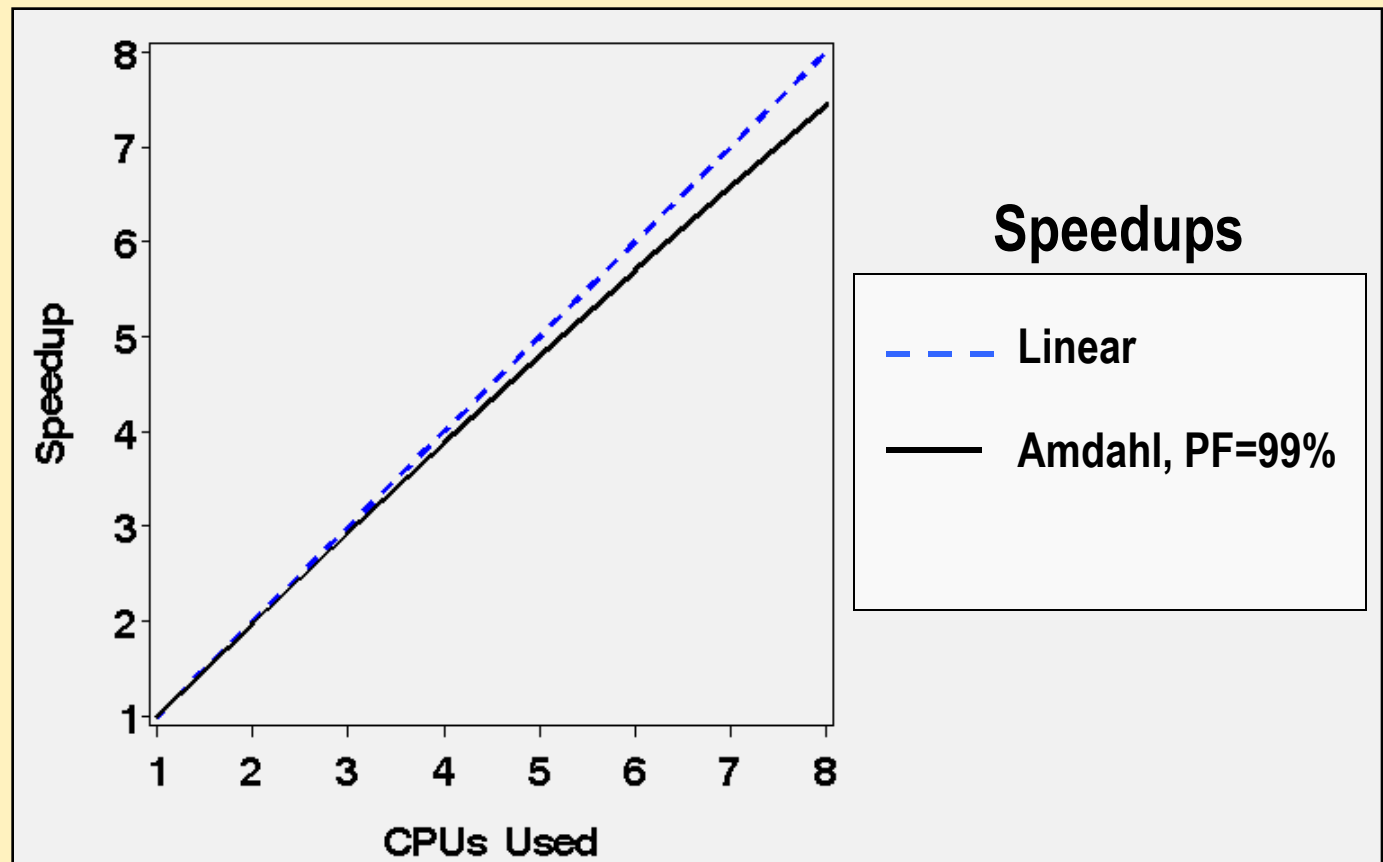
Scalability in PROC LOESS

Test Details

4000 observations

1 model specified

Confidence limits for
specified model



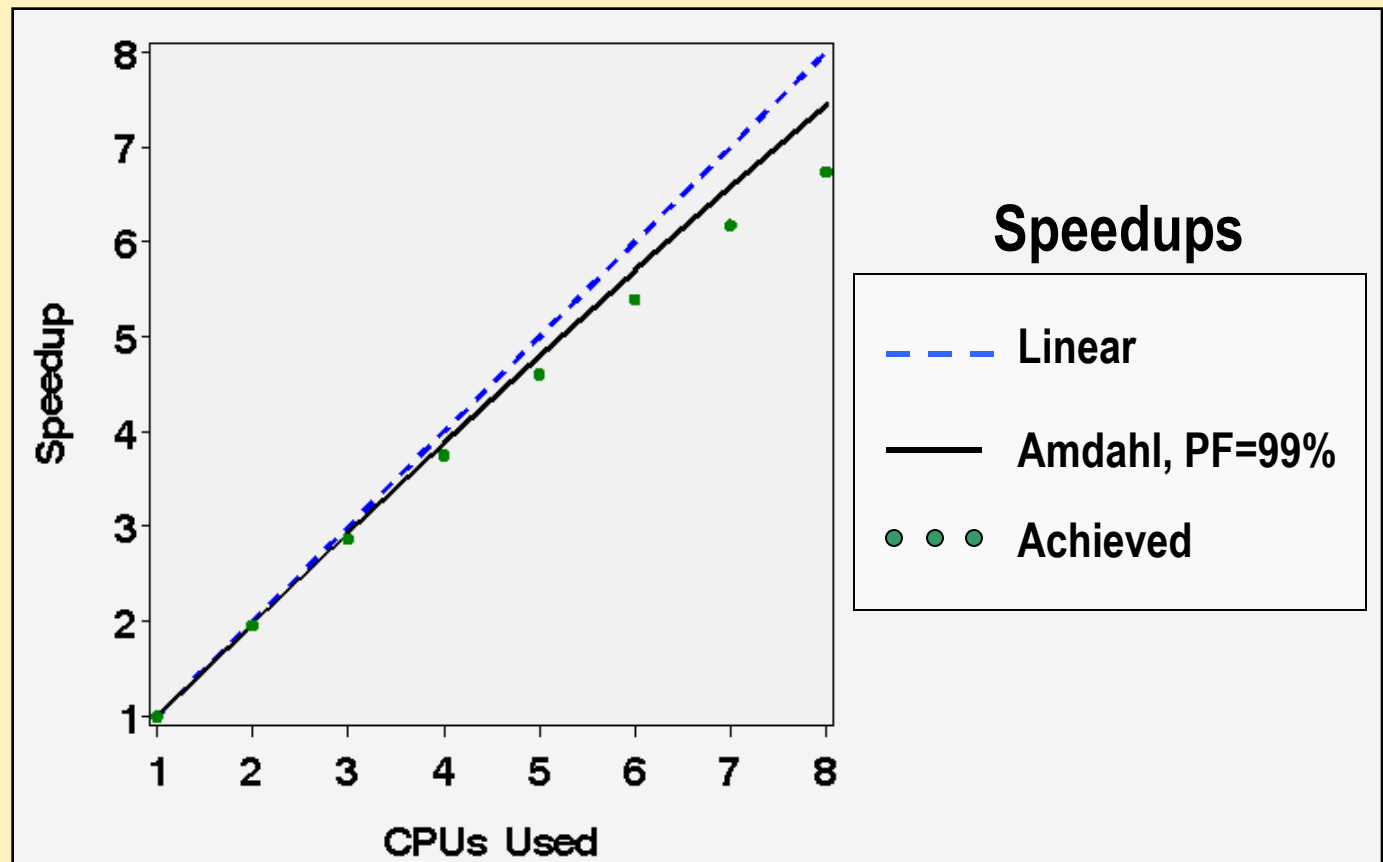
Scalability in PROC LOESS

Test Details

4000 observations

1 model specified

Confidence limits for
specified model



Partially Multithreaded Procedures

■ Base SAS

- PROC SORT
- PROC SUMMARY
- SQL (Group by, Order by)

■ Enterprise Miner

- PROC DMDB
- PROC DMREG
- PROC DMINE

■ SAS/STAT

- PROC GLM
- PROC LOESS
- PROC REG
- PROC ROBUSTREG

NOTE: Not all usages of these procedures are scalable.

Your mileage may vary!

Reading Between the Lines

- Parallelization adds value to the IVC

Analyze bigger volumes of data

- Multithreading to provide parallel execution

Not as boring as I feared

- How do you measure scalability

Predicting scalability is a subtle task

- Selected demonstrations

Some of my jobs will run faster in SAS 9

Questions

and hopefully answers