

Rob Snihur

QCD mtg., Apr. 14, 2004.

Optimizing stntuple analysis

- Problem
- Tests
- Conclusion
- Recommendation

Problem

- Stntuples

- Standard ✍ must include lots of info ✍ much disk!
- One solution: only read in necessary blocks
- Another approach: parallelize

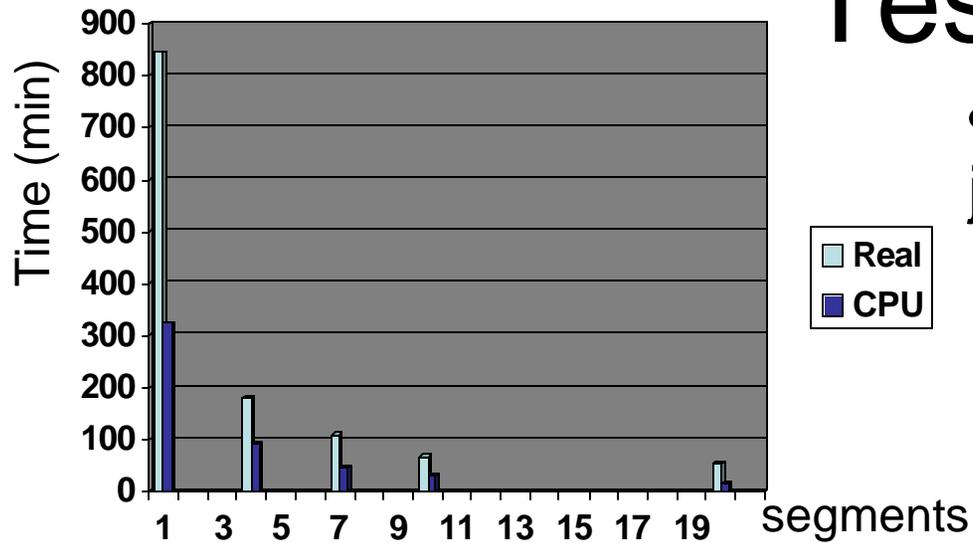
- Basics

- CAF disk breakdown of stntuple samples

• jet100	100 GB	
• jet070	100 GB	
• jet050	200 GB	(What I ran on)
• jet020	500 GB	
TOTAL	1.5 TB	

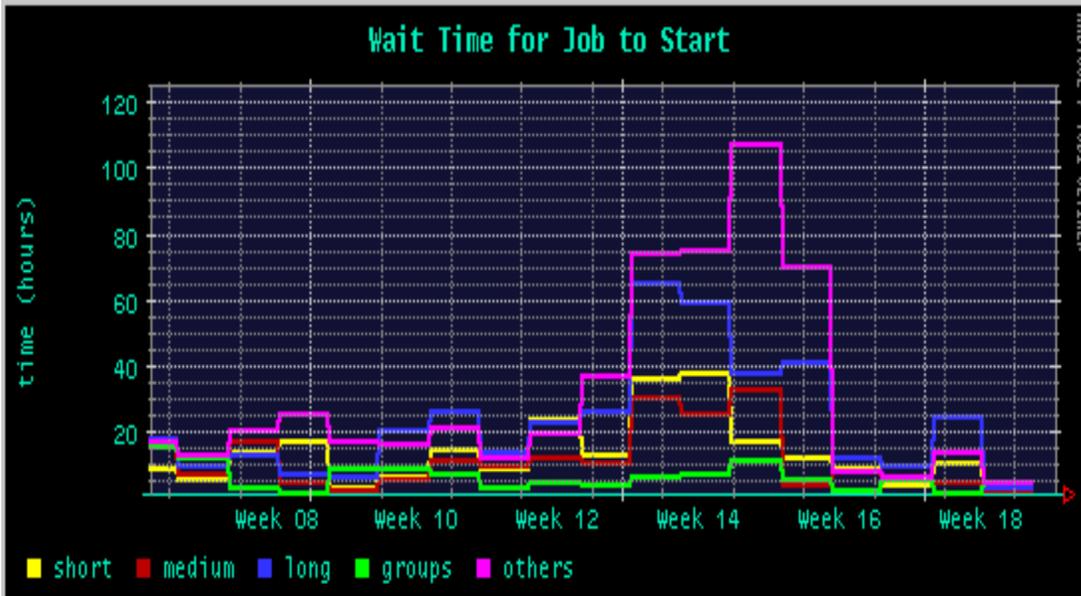
- All on a **single** disk
- Access stntuples in root via rootd
- Run root anywhere
 - Where optimally?

Tests



- CAF (2 x 1.6 GHz CPU, 3 jobs/CPU)

- Split job up into n parallel segments
- Allows use of “short” queue

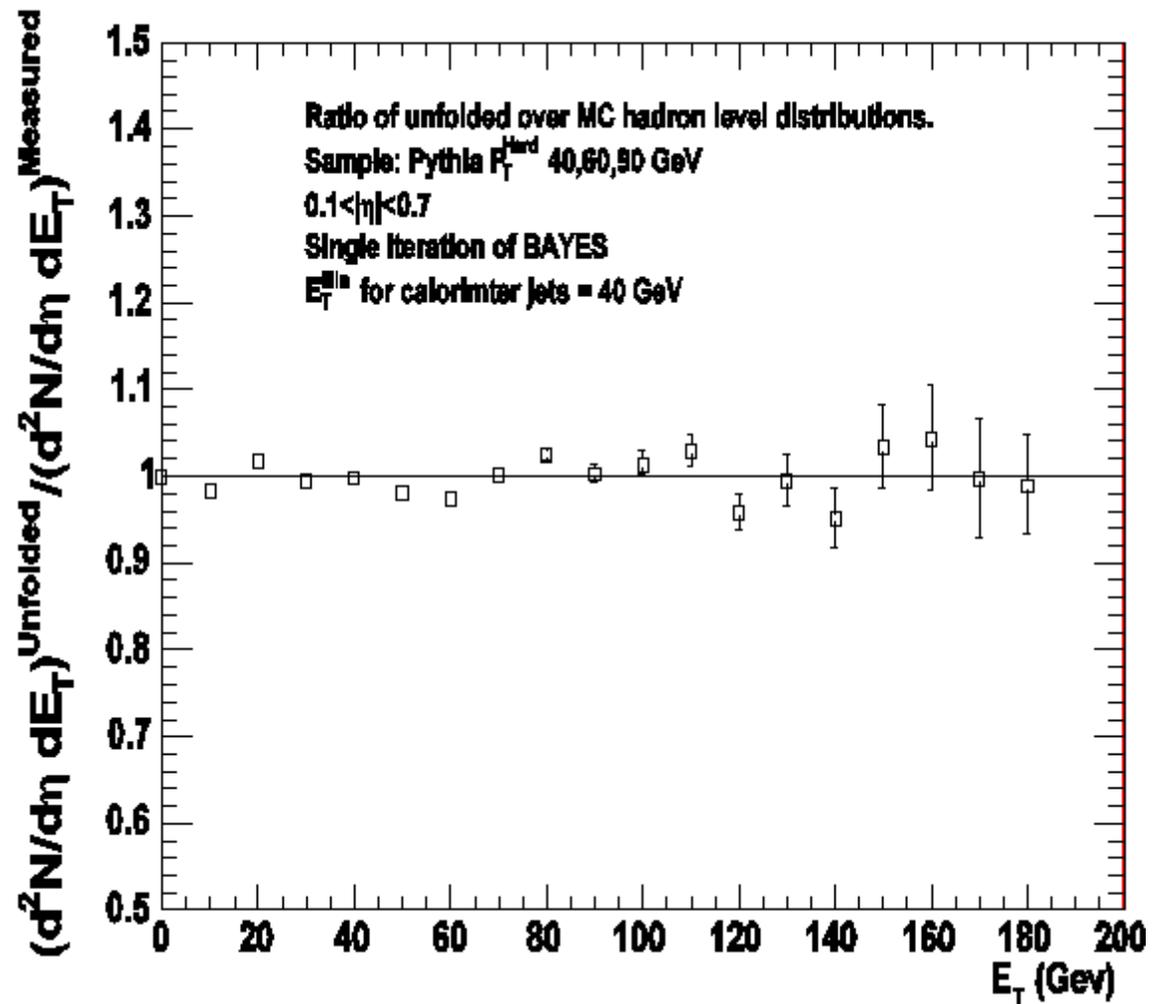


- non-CAF

- Desktops in CDF trailers
 - fcdflnx3 8 x 0.7 GHz
 - nucl05 1 x 2.8 GHz
- Remote nodes
 - Access stntuples at CDF
 - Access stntuples locally

Conclusions

- Split into many root jobs
 - I can provide general tools
- Add up histograms (& Trees) ipso facto
- CAF
 - Optimal if wait time is small
- Split stntuples on multiple disks



Summary

- Parallelizing yields radical improvements in analysis time
 - CPU matters!
 - CPU time will only increase as analyses mature
 - I/O issues
 - want CPU close to stntuples (recommendation for remote sites)
 - want to reduce contention
- Recommendation for analysis (I will provide instructions)
 - Set up analysis to run root in parallel on CAF
 - If wait time is small: HUGE time savings
 - If wait time is large: run on local desktop(s)
- Recommendation for short term (I volunteer)
 - Split stntuples on multiple disks ✍ HUGER time savings!
- Future
 - QCD group has large data samples ✍ pioneer
 - stntuples in dCache (?)
 - Stntuples on DCAFs (?)
 - PROOF (= Parallel Root Facility) ?