SSD tests at PROOF farm at BNL

Michael Ernst, Sergey Panitkin, Robert Petkus, Ofer Rind, Torre Wenaus

BNL

ATLAS

March, 24
CHEP 2009
Prague, Czech Republic

BROOKHAVEN NATIONAL LABORATORY

Outline

- Introduction
- Tests
  - Interactive analysis straw man tests
    - Single node performance
      - SSD vs HDD
      - RAID vs single disk
    - Farm performance
  - Physics analysis tests
    - SSD vs HDD
    - RAID v single disk
- Summary and Discussion

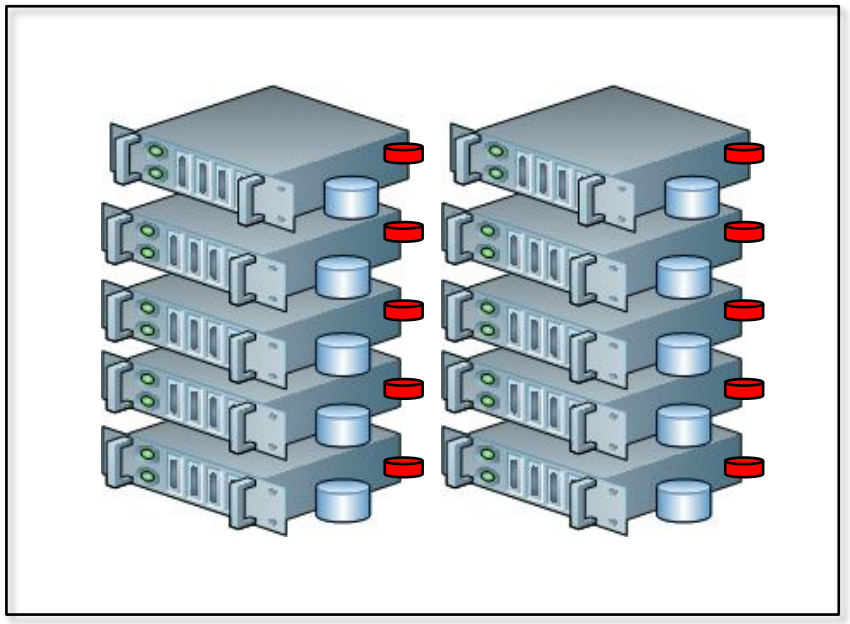
Motivation

- **P**arallel **ROO**T **F**acility, a system for the **interactive** or **batch** analysis of **very large sets** of **Root** data files on a **cluster of computers**
- Speed up the query processing by employing inherent parallelism in event data
- PROOF uses Xrootd for communication, load balancing, data discovery and file serving
- Can run on commodity hardware
- Well suited for (if not geared to) analysis farms with distributed **local** storage. Computing Element=Storage Element
  - Local data processing is encouraged – **automatic matching of code with data**
- Hence, matching between I/O demand and local disk throughput for a single node is important, especially for multi-core machines

PROOF Farm Configuration

"Test Farm at BNL

- 10 nodes - 16 GB RAM each
- 10x 2x4cores: 2.0 GHz Kentsfield CPUs
- 750 GB HDD
- 64 GB SSD space
- 1Gb network
- Scientific Linux 4.2
- Sever al versions of root
- PROOF and Xrootd installed
- Ganglia and XrdMon monitoring
- Part of Atlas T1 facility



Solid State Disks Used for Tests

- Model: Mtron MSP-SATA7035064
- Capacity 64 GB
- Average access time ~0.1 ms (typical HD ~10ms)
- Sustained read ~120MB/s
- Sustained write ~80 MB/s
- IOPS (Sequential/ Random) 81,000/18,000
- Write endurance >140 years @ 50GB write per day
- MTBF 1,000,000 hours
- 7-bit Error Correction Code

Tests Motivation

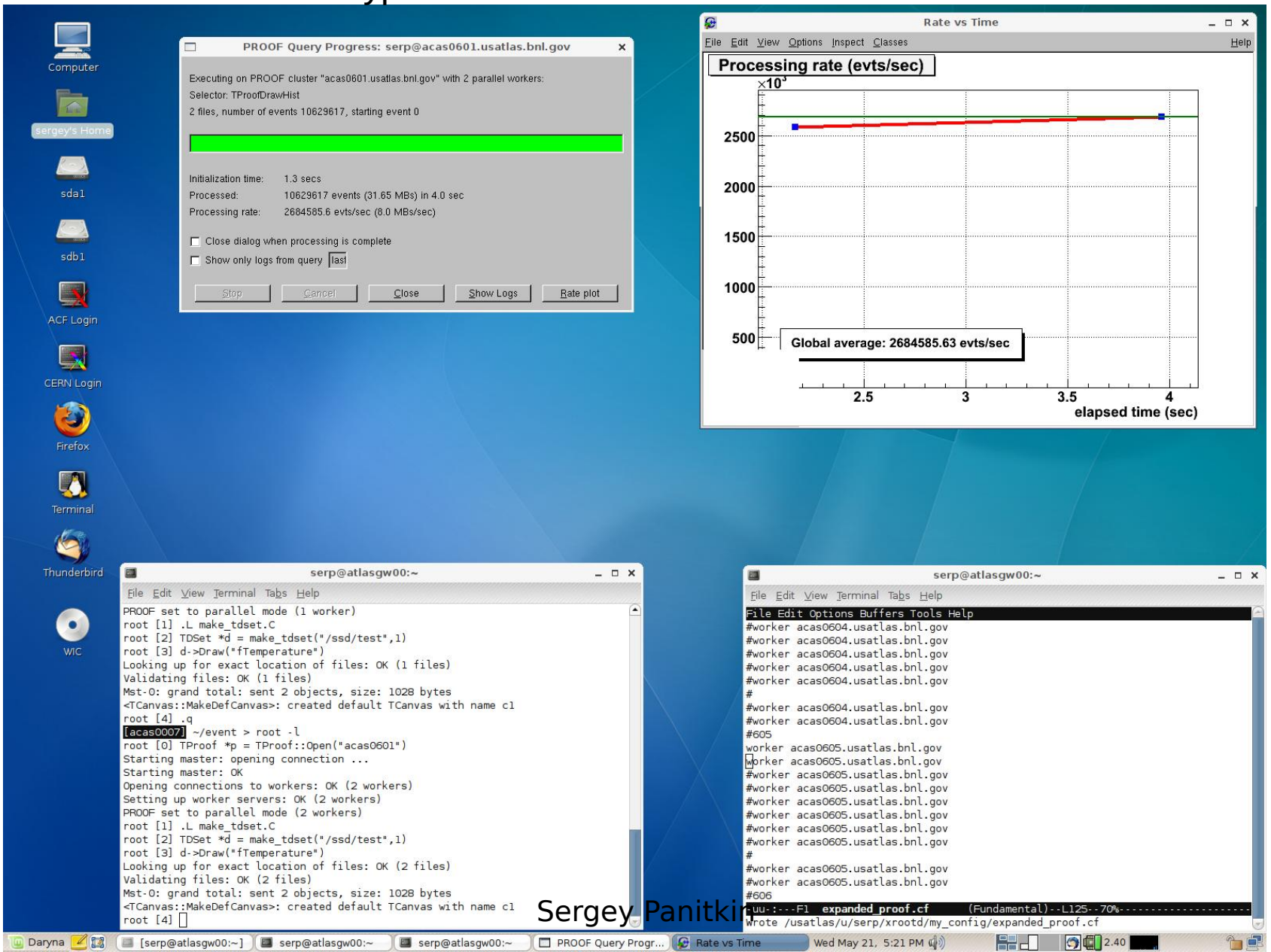
- **“Interactive analysis” test**
  - Emulates interactive, command prompt root session
  - Plot one variable, scan ~10E7 events, in ROOT tree , ala D3PD analysis
  - “PROOF Bench” suit of benchmark scripts used to generate data. Part of ROOT distribution.
  - http://root.cern.ch/twiki/bin/view/ROOT/ProofBench
  - Study scenario with sparse data access and minimal processing.
  - Data simulate HEP events in root trees ~1k per event
  - Single ~3+ GB file per PROOF worker in this tests
- **“Realistic” analysis test**
  - H->4l analysis of simulated Atlas data (by G. Carillo, U. Wisconsin Madison)
  - CPU intensive
  - Atlas D3PD data format
- **General Idea:** Look at read performance of disks in PROOF context

Additional Test Details

- 1+1 or 1+8 nodes PROOF farm configurations
- 2x4 cores, 2.0 GHz Kentsfield CPUs per node, with16 GB of RAM per node
- All default settings in software and OS
- Root 5.18.00 for “interactive analysis” test
- Root 5.20 for H->4l analysis tests
- Use PROOF provided information about analysis and read rates
- Additional hardware monitoring via Ganglia
- Single user environment. No ambient load on the farm.
- Reboot before every test to avoid memory caching effects

SSD Tests

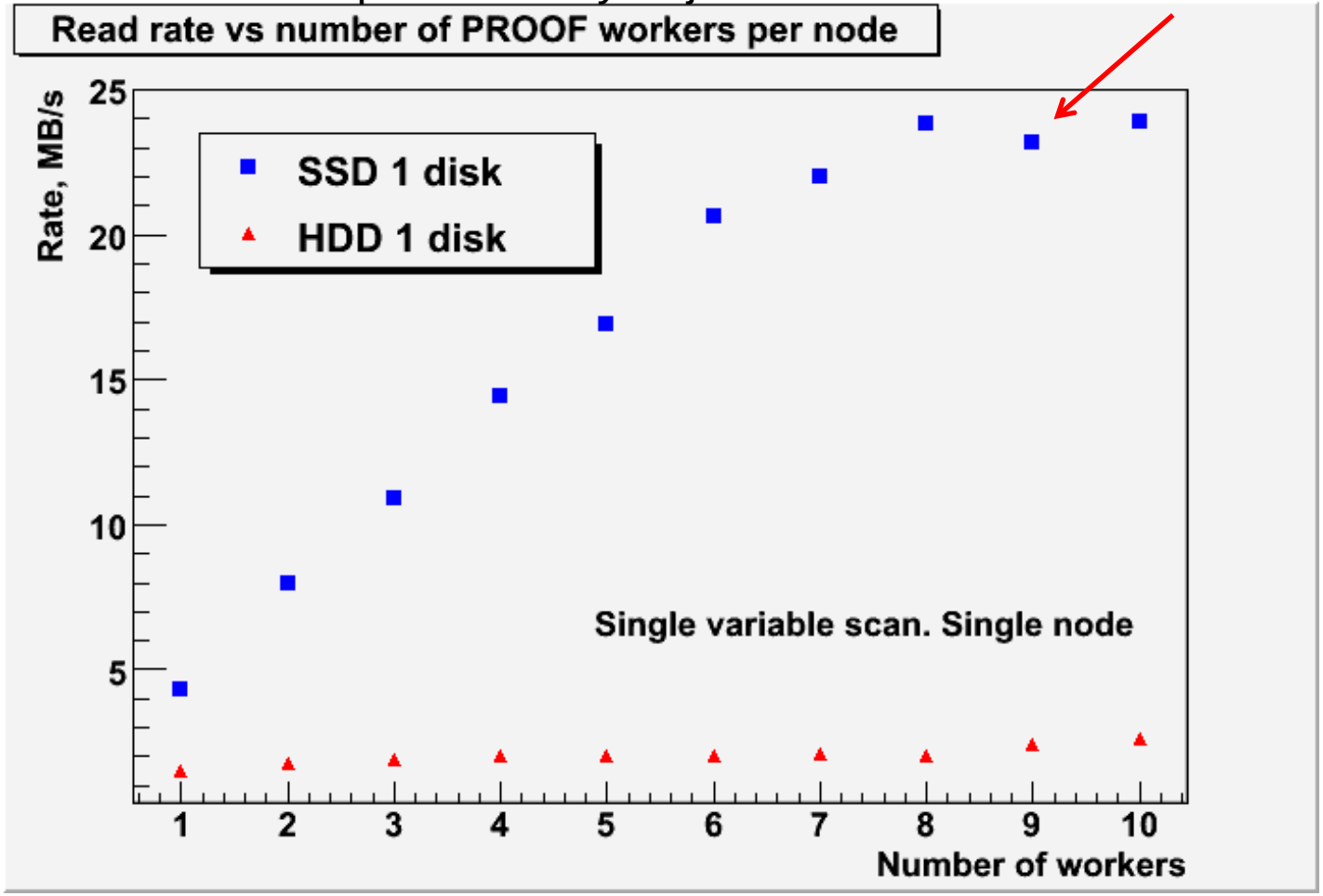
Typical test session in root



Interactive analysis. SSD vs HDD

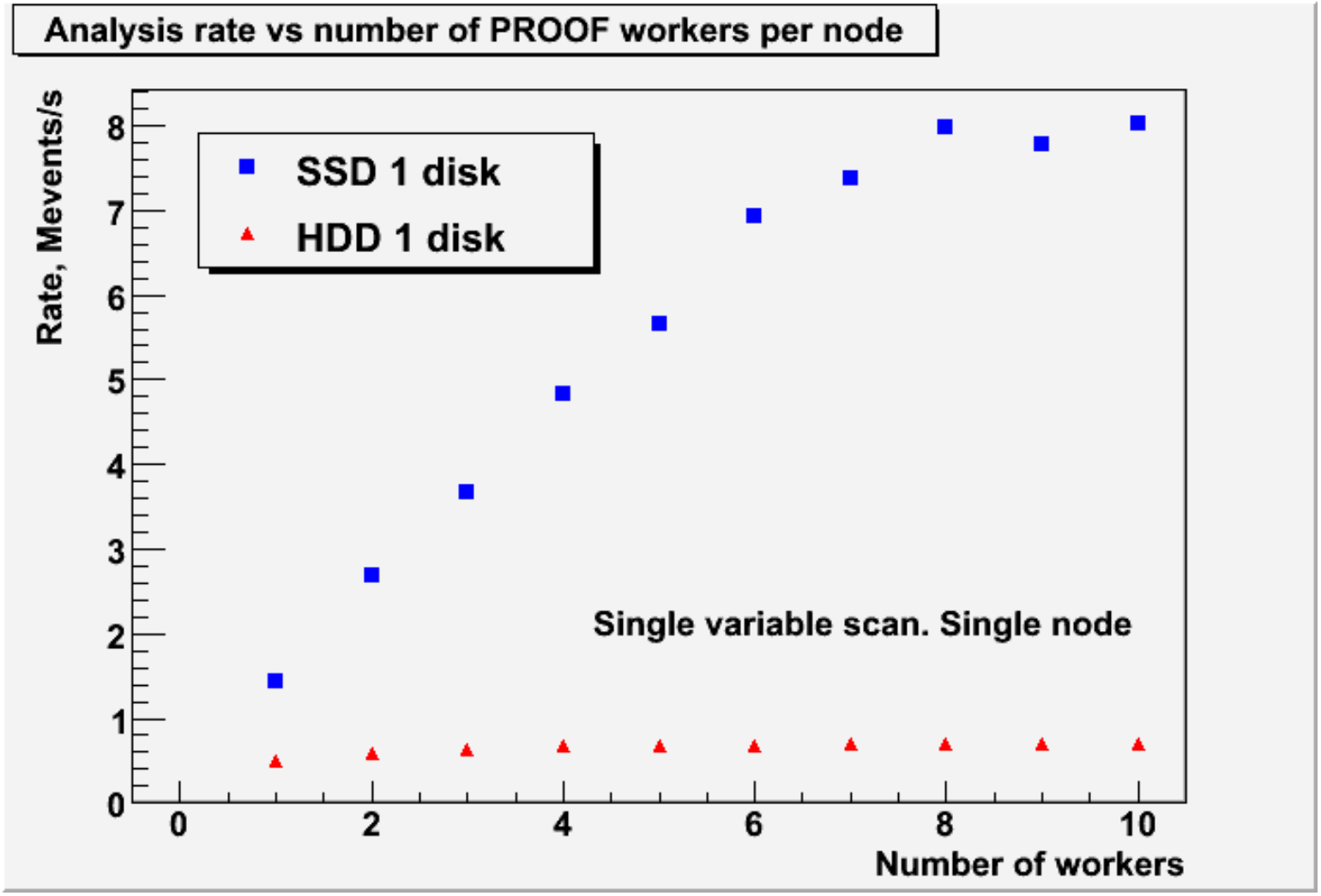
- Worker is a PROOF parallel analysis job

CPU limited



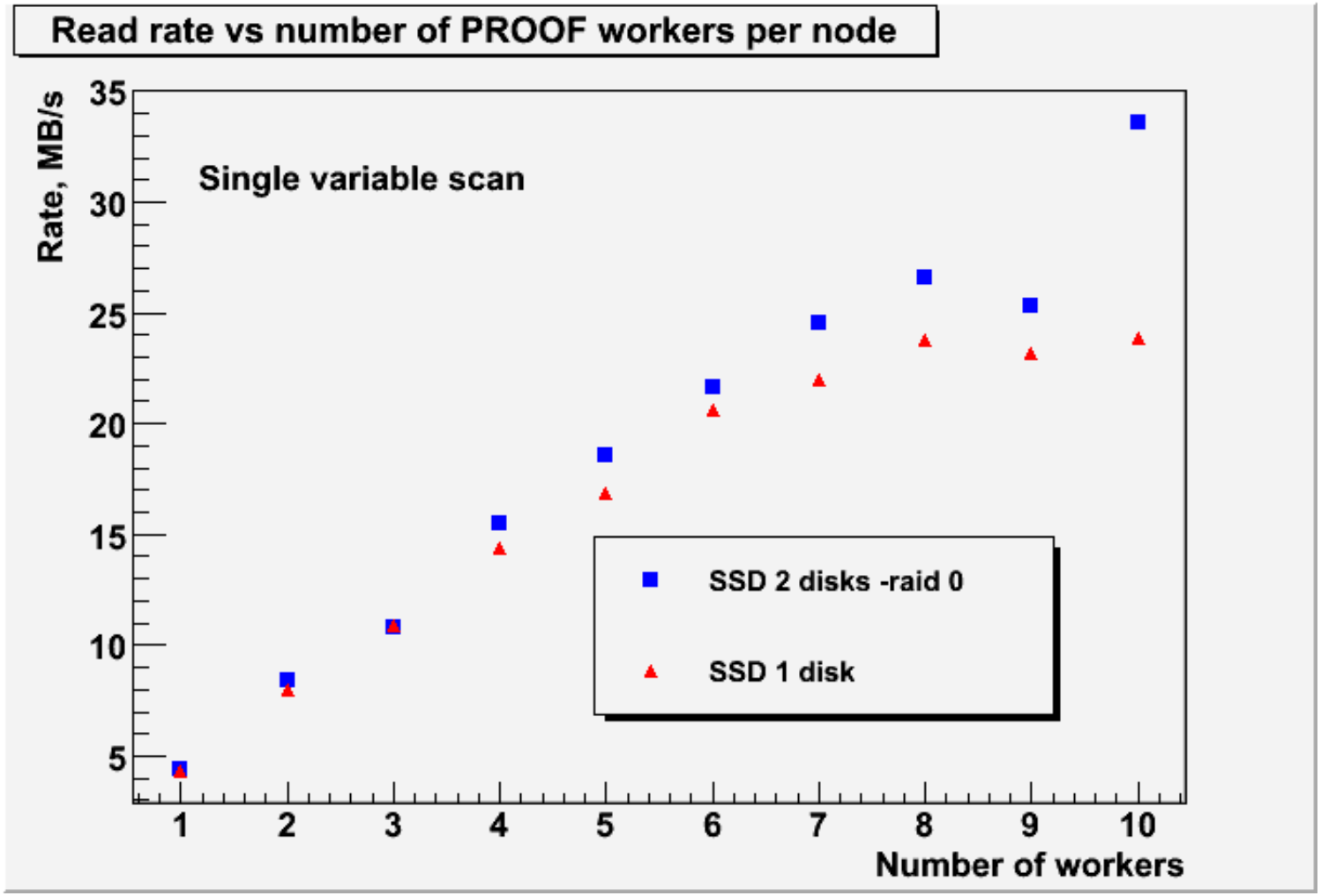
- SSD holds clear speed advantage
- ~Up to10 times faster in concurrent read scenario

SSD vs HDD



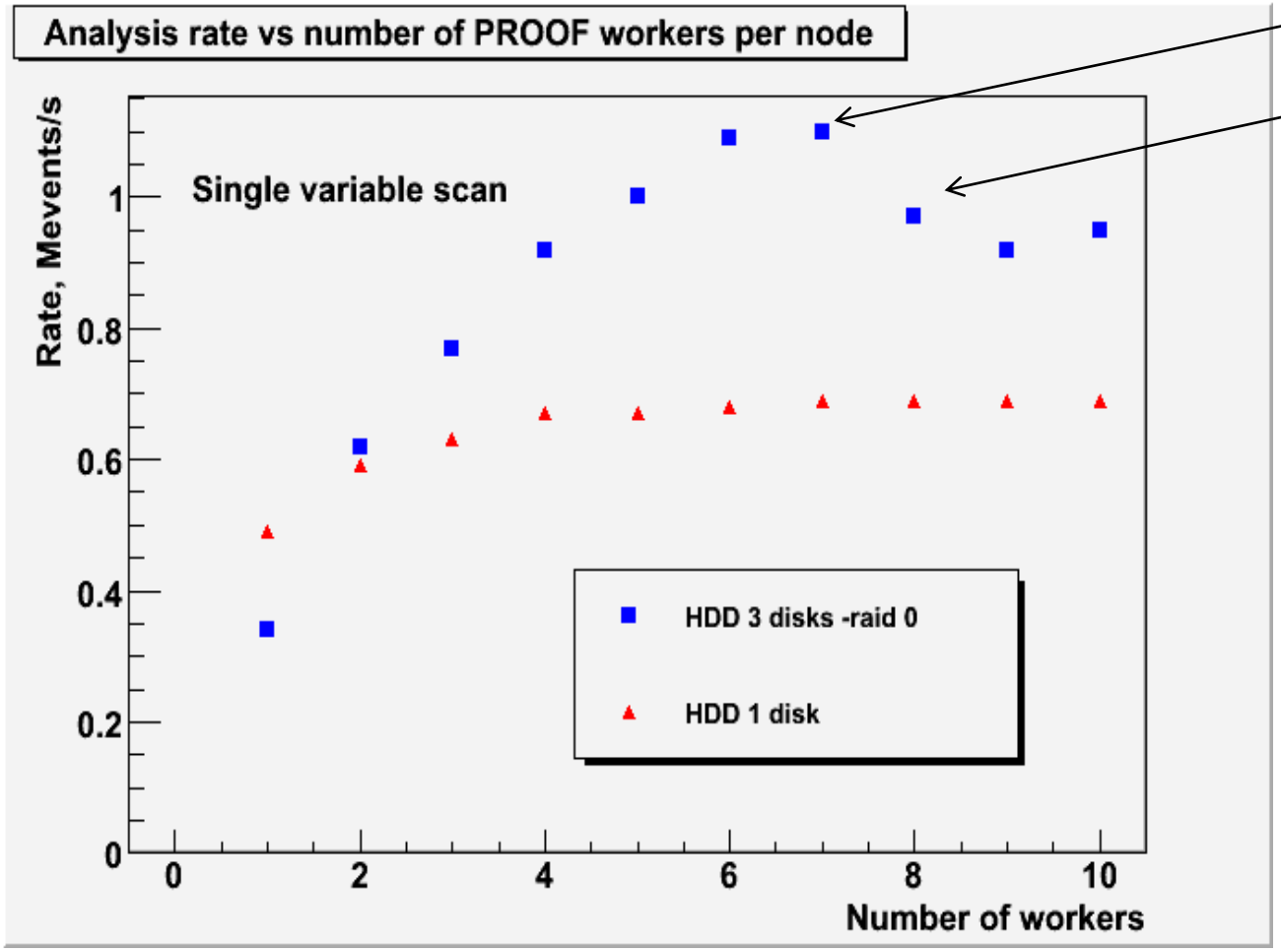
With 1 worker : 5.3M events, 15.8 MB read out of ~3 GB of data on disk
With 8 workers: 42.5M events, 126.5 MB read out of ~24 GB of data

SSD: single disk vs RAID



- SSD RAID has minimal impact until 8 simultaneously running jobs
- Behavior at 8+ workers is not explored in details yet

HDD single disk vs RAID



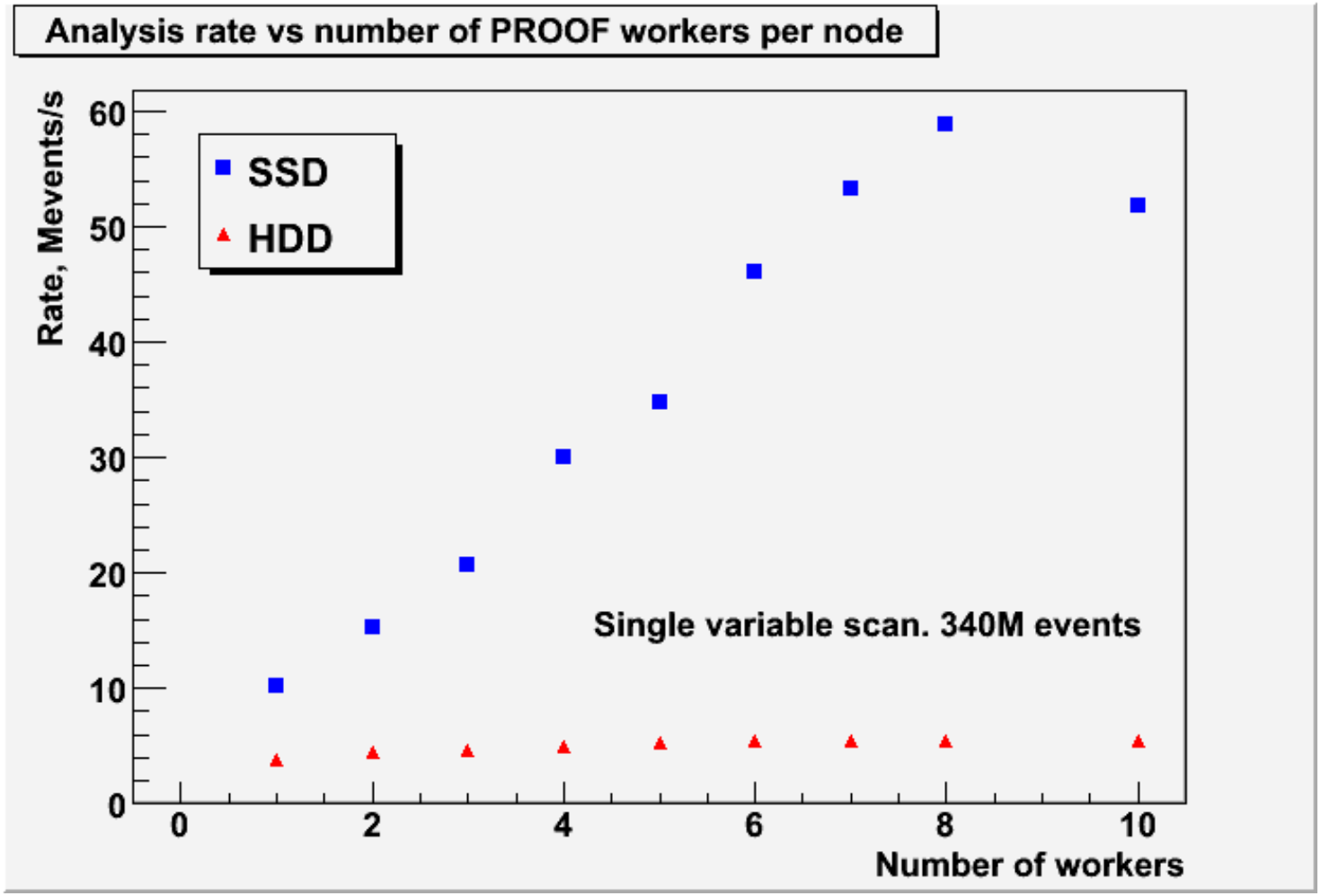
I/O limited?

I/O and CPU limited?

3x750GB disks in RAID 0 (software RAID) vs 1x500GB drive

1 disk shows rather poor scaling in this tests
3 disk raid supports 6 workers?

SSD vs HDD. 8 node farm



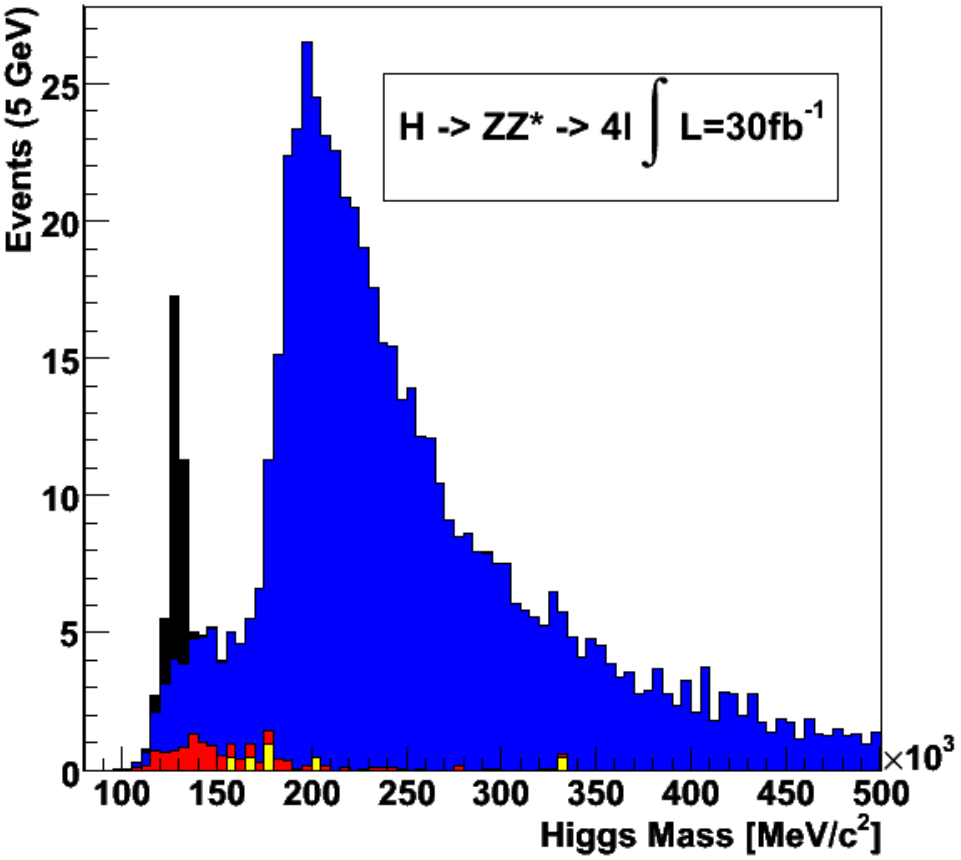
Aggregate (8 node farm) analysis rate as a function of number of workers per node

Almost linear scaling with number of nodes

Second use case details

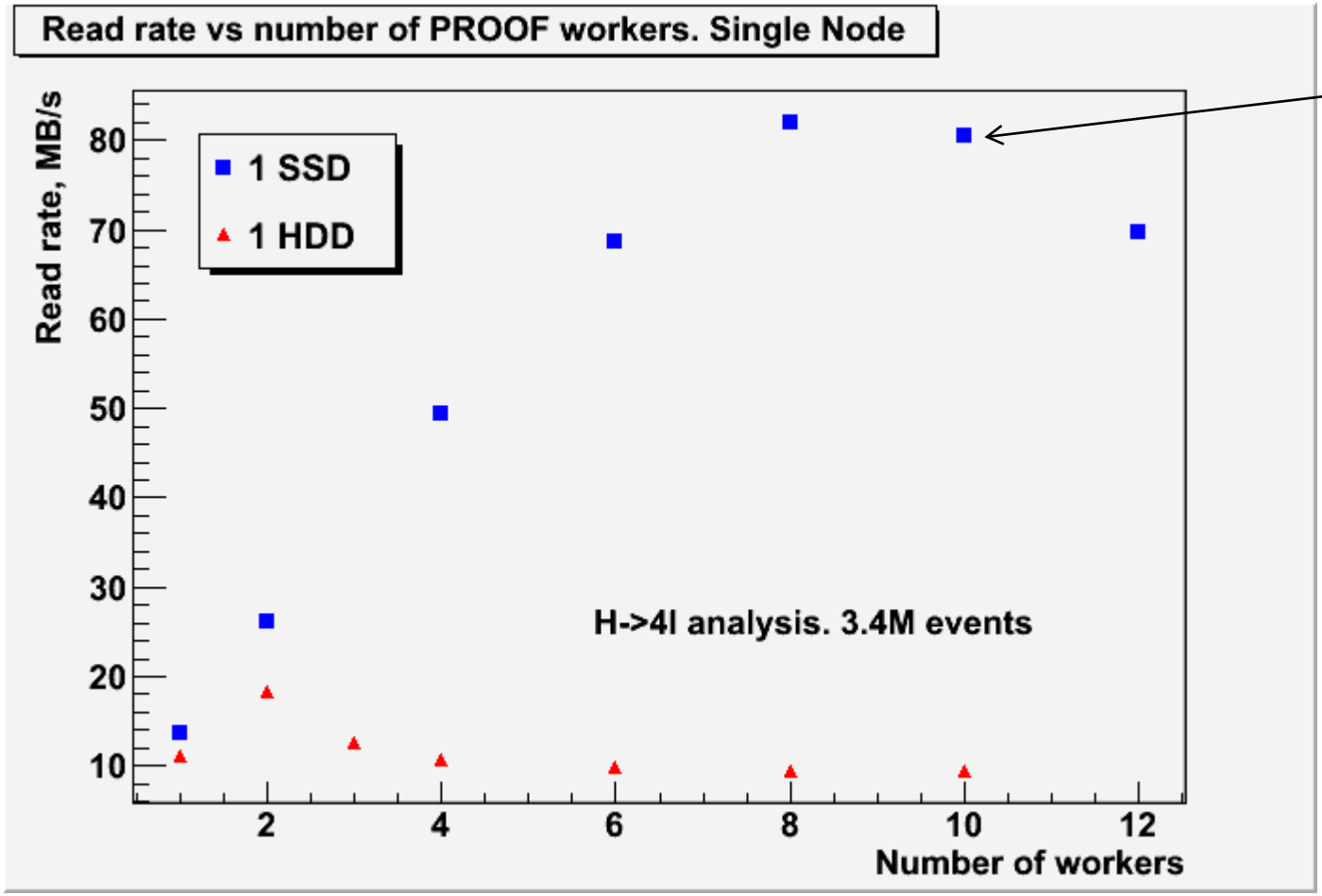
Courtesy German Carrillo, UWM

- Higgs decay into 4-lepton analysis
- 200 D3PD files, ~3.4M events
- 46.4 GB of data
- Analysis include TMinuit fits
- CPU intensive, I/O intensive



- 8 cores, 2.0 GHz Kentsfield CPUs
- 16 GB RAM
- Mtron SSD 64GB
- 750 GB SATA HDD (7200 rpm class)

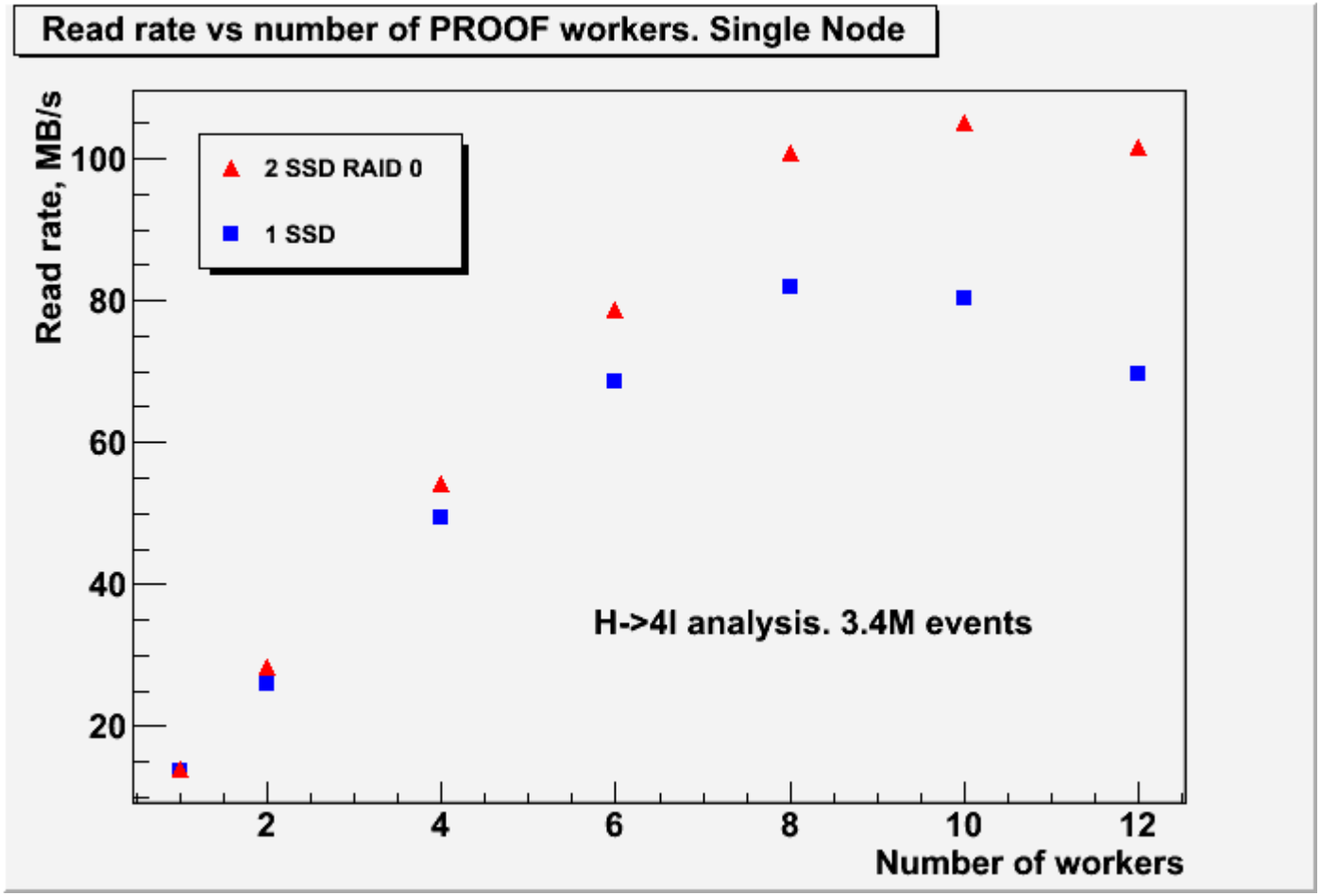
H->4l analysis. SSD vs HDD



CPU Limited

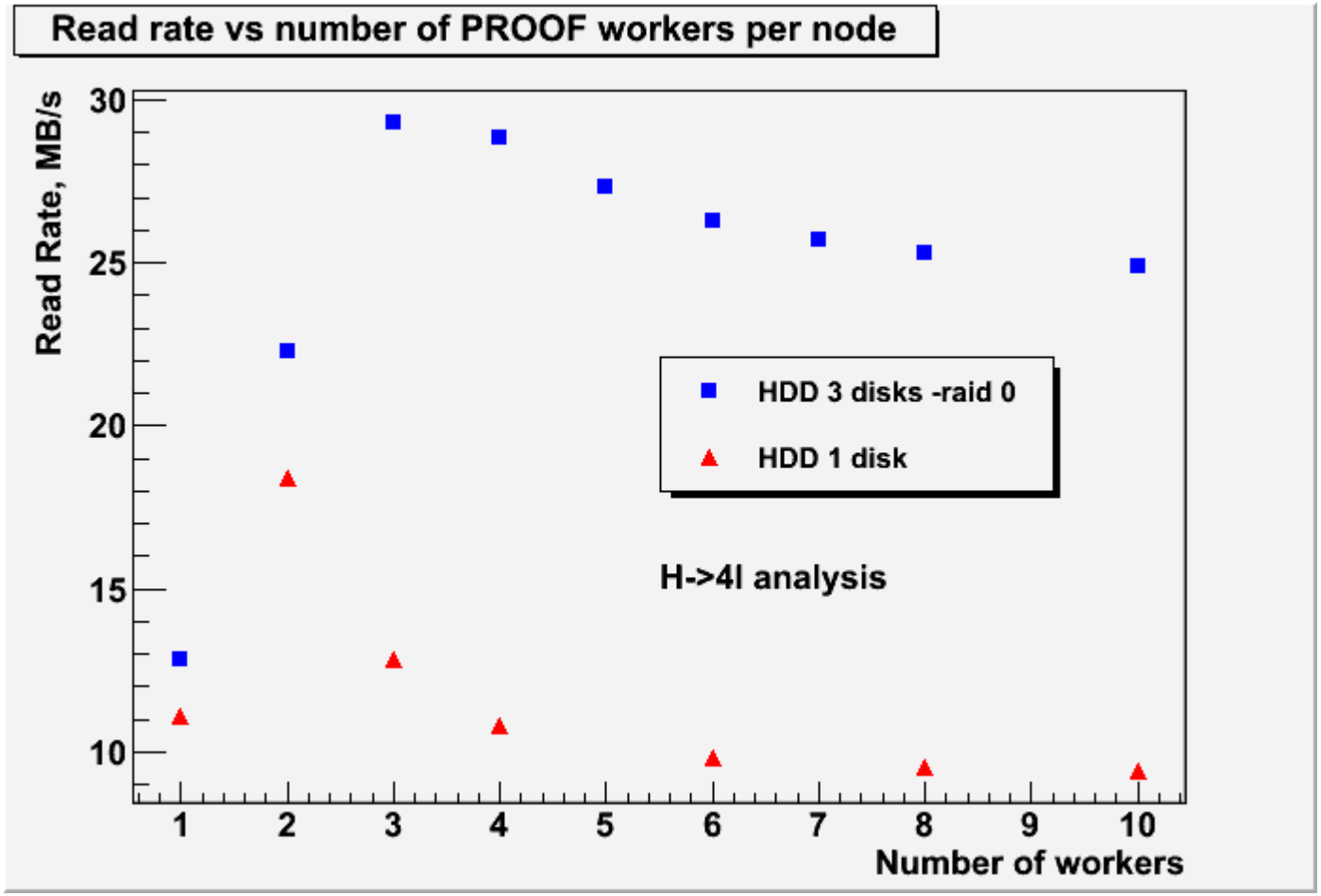
SSD is about 10 times faster at full load
Best HDD performance at 2 worker load
Single analysis job generates ~10 -14 MB/s load with given hardware

H->4l analysis. SSD RAID 0



SSD 2 disk RAID 0 shows little impact up to 4 worker load

H->4l analysis. HDD: single vs RAID



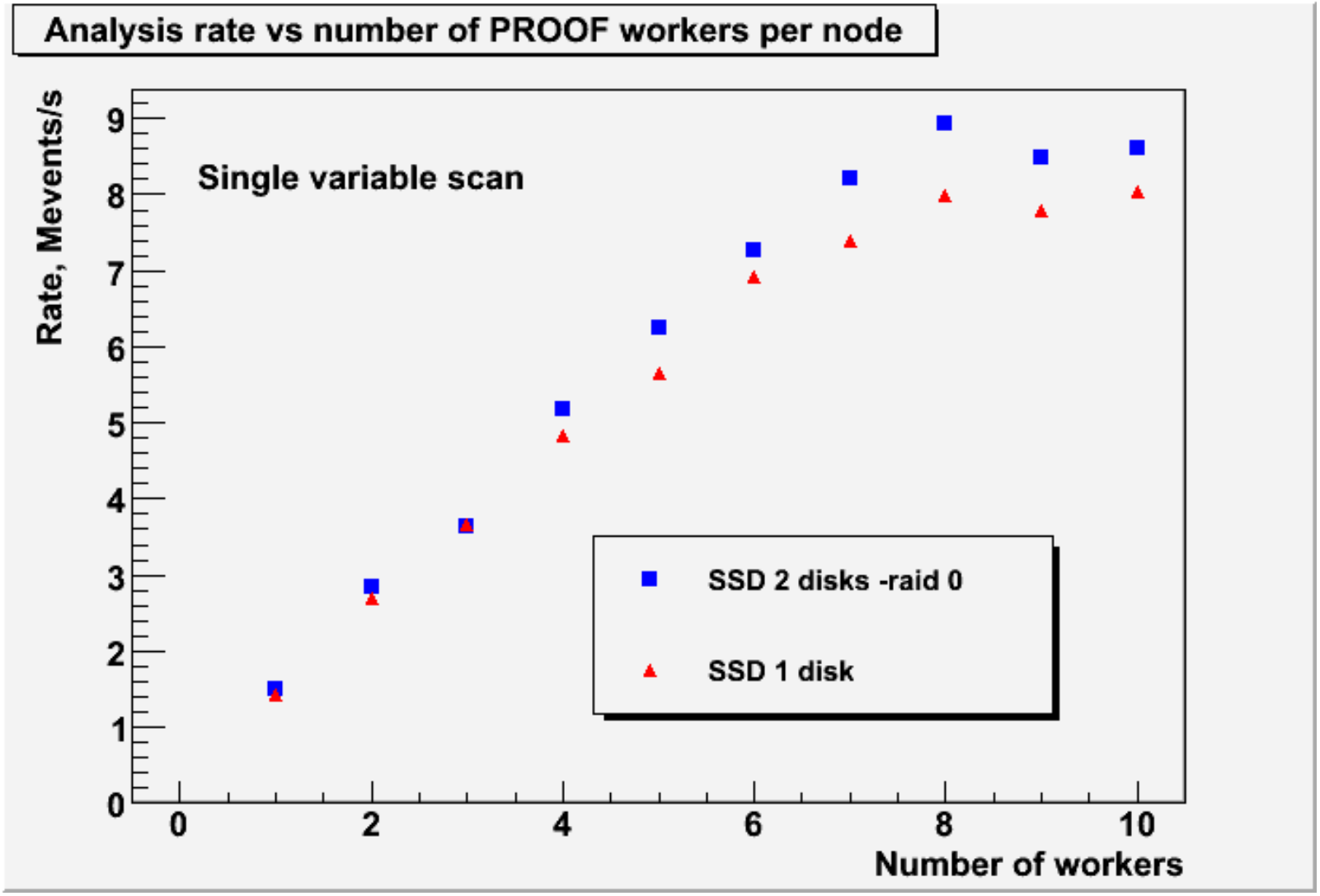
3x750 GB HDD RAID peaks at ~3 worker load
Single HDD disk peaks at 2 worker load, then performance rapidly deteriorates

Summary and Discussion

- SSD technology offer significant performance advantage in concurrent analysis environment
- We observed~x10 better read performance than HDD in our test
- The main issue, in PROOF context, is matching of local I/O demand and supply
- Some observations from our tests
  - Single analysis worker in PROOF can generate ~10-15 MB/s read load
  - One SATA HDD can sustain ~2-3 PROOF workers
  - HDD RAID array can sustain ~ 3 to 6 workers
  - One Mtron SSD can sustain ~8 workers, almost at peak performance
  - SSD RAID is nice, but not really necessary with current hardware
- Currently the main issue with SSD is size (and cost) .
- Multi tiered local disk sub-system, with automatic pre-staging of data from HDD to SSD may be a promising solution which can provide both capacity and speed. Efficient data management is needed.
- We plan to investigate this option.

The End

SSD: single disk vs RAID



H-4l analysis rate. SSD vs HDD