



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

- ◆ 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



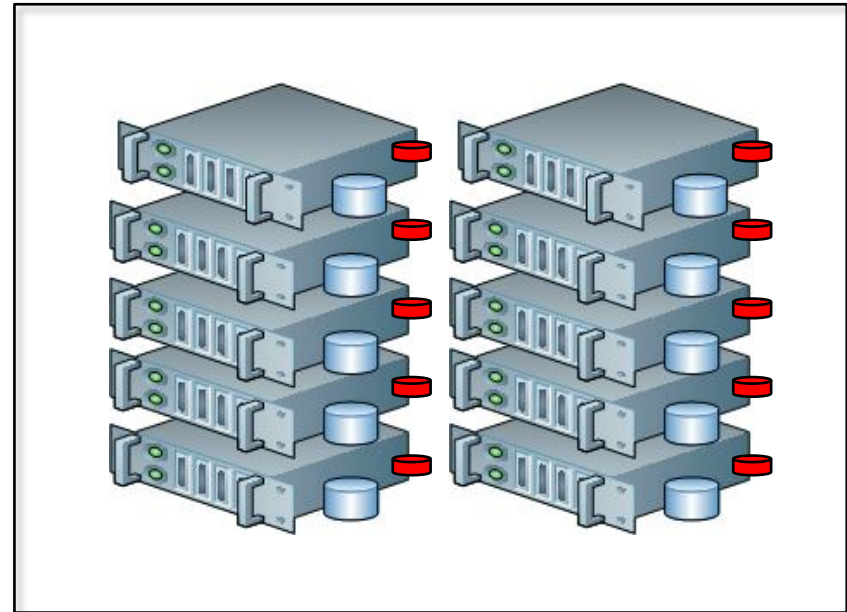
Motivation

- ◆ **Parallel ROOT Facility**, 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 ~ 10 ms)
- ◆ Sustained read ~ 120 MB/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



◆ “Interactive analysis” test

- ◆ Emulates interactive, command prompt root session
- ◆ Plot one variable, scan $\sim 10E7$ 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 $\sim 1k$ per event
- ◆ Single $\sim 3+$ 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, with 16 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->4I 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

The screenshot displays a Linux desktop environment with a blue background. On the left side, there is a vertical dock containing icons for Computer, sda1, sdb1, ACF Login, CERN Login, Firefox, Terminal, Thunderbird, and WIC. The main workspace contains three windows:

- PROOF Query Progress: serp@acas0601.usatlas.bnl.gov**: A dialog box showing the progress of a query. It indicates that 10629617 events (31.65 MBs) were processed in 4.0 seconds, resulting in a processing rate of 2684585.6 evts/sec (8.0 MBs/sec). A green progress bar is visible. Buttons for Stop, Cancel, Close, Show Logs, and Rate plot are at the bottom.
- Rate vs Time**: A line graph showing the processing rate in events per second (evts/sec) over an elapsed time of 2.5 to 4.0 seconds. The y-axis is scaled by $\times 10^6$. The rate is constant at approximately 2684.58563. A text box at the bottom of the graph states: "Global average: 2684585.63 evts/sec".
- Terminal (serp@atlasgw00:~)**: Shows the execution of the PROOF test script. The output includes: "PROOF set to parallel mode (1 worker)", "root [1] .L make_tdset.C", "root [2] TDSet *d = make_tdset('/ssd/test',1)", "root [3] d->Draw('fTemperature*)", "Looking up for exact location of files: OK (1 files)", "Validating files: OK (1 files)", "Mst-0: grand total: sent 2 objects, size: 1028 bytes", "<TCanvas::MakeDefCanvas>: created default TCanvas with name c1", "root [4] .q", "[acas0007] ~/event > root -l", "root [0] TProof *p = TProof::Open('acas0601*)", "Starting master: opening connection ...", "Starting master: OK", "Opening connections to workers: OK (2 workers)", "Setting up worker servers: OK (2 workers)", "PROOF set to parallel mode (2 workers)", "root [1] .L make_tdset.C", "root [2] TDSet *d = make_tdset('/ssd/test',1)", "root [3] d->Draw('fTemperature*)", "Looking up for exact location of files: OK (2 files)", "Validating files: OK (2 files)", "Mst-0: grand total: sent 2 objects, size: 1028 bytes", "<TCanvas::MakeDefCanvas>: created default TCanvas with name c1", "root [4] [".
- Terminal (serp@atlasgw00:~)**: Shows the output of the expanded proof configuration file. It lists worker nodes: "#worker acas0604.usatlas.bnl.gov" and "#worker acas0605.usatlas.bnl.gov".

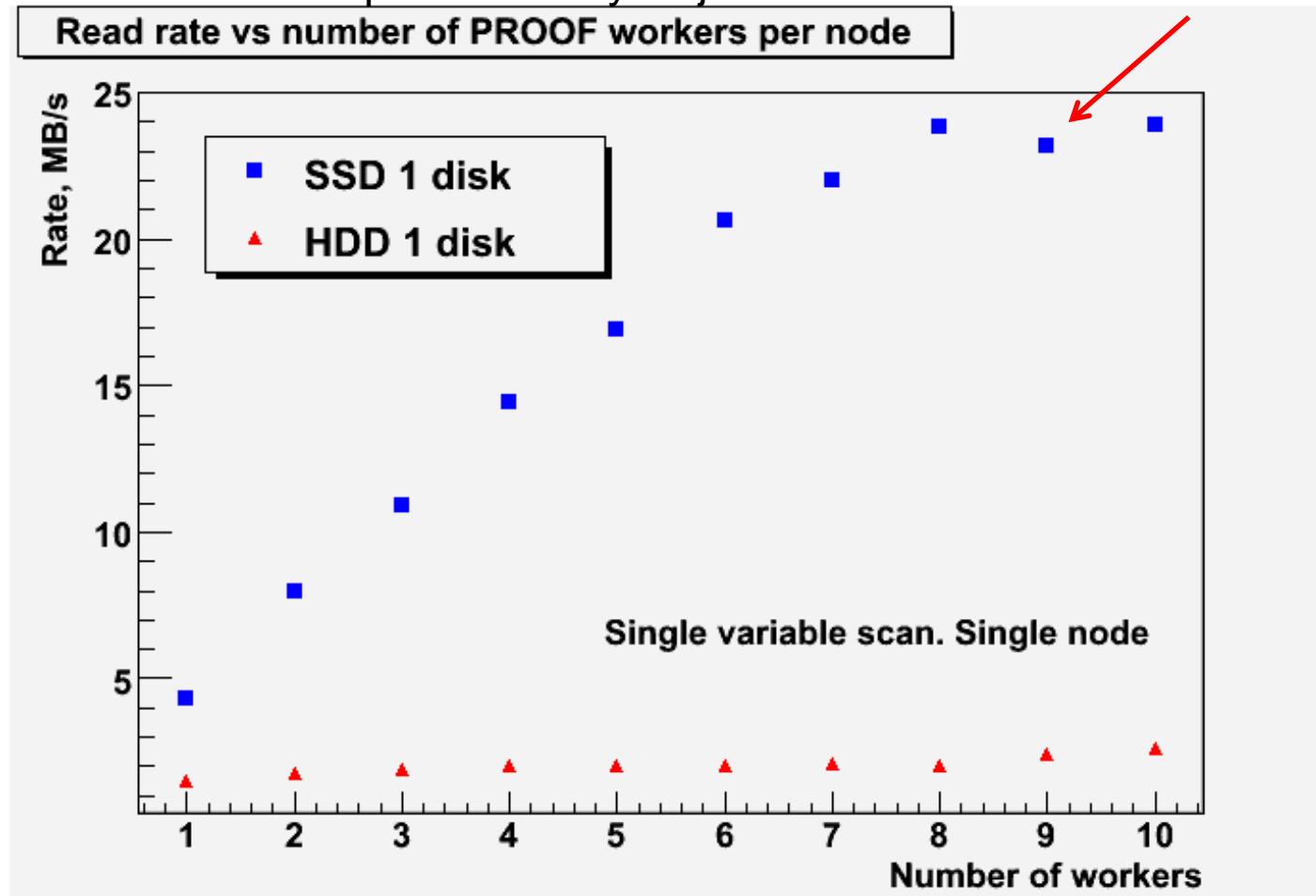
The system tray at the bottom shows the date and time as "Wed May 21, 5:21 PM".

Sergey Panitkin

Interactive analysis. SSD vs HDD

- Worker is a PROOF parallel analysis job

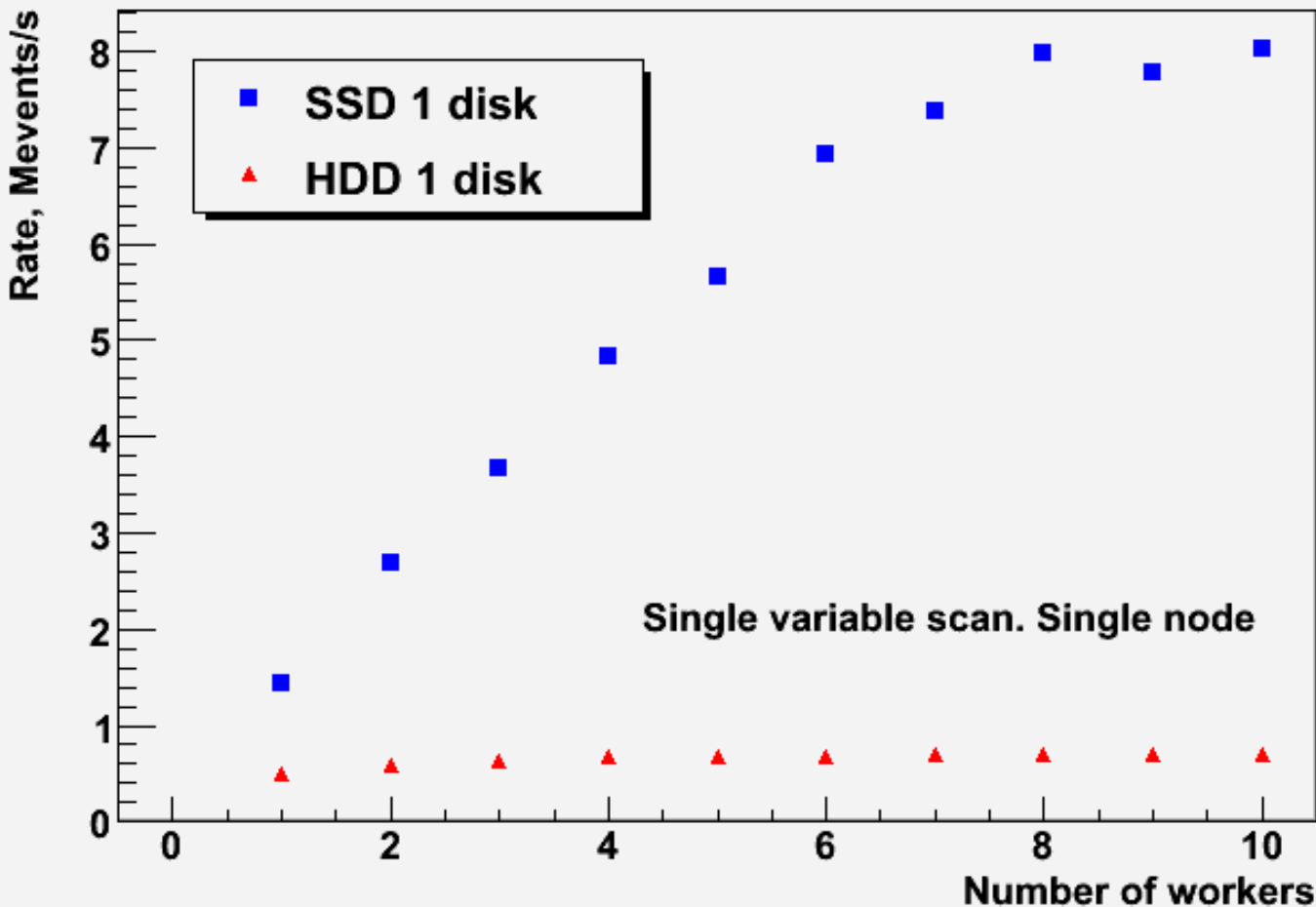
CPU limited



- SSD holds clear speed advantage
- ~Up to 10 times faster in concurrent read scenario

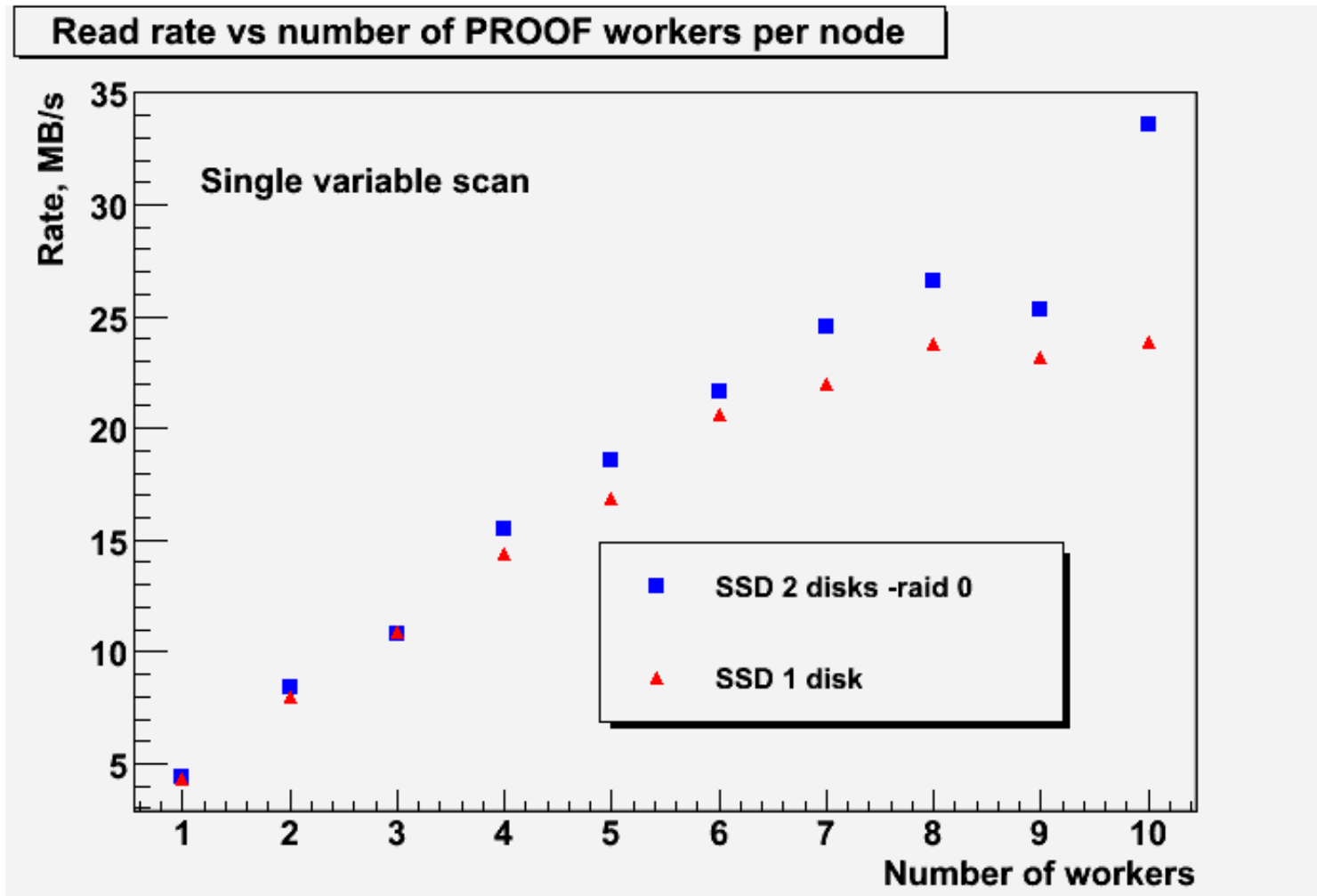
SSD vs HDD

Analysis rate vs number of PROOF workers per node



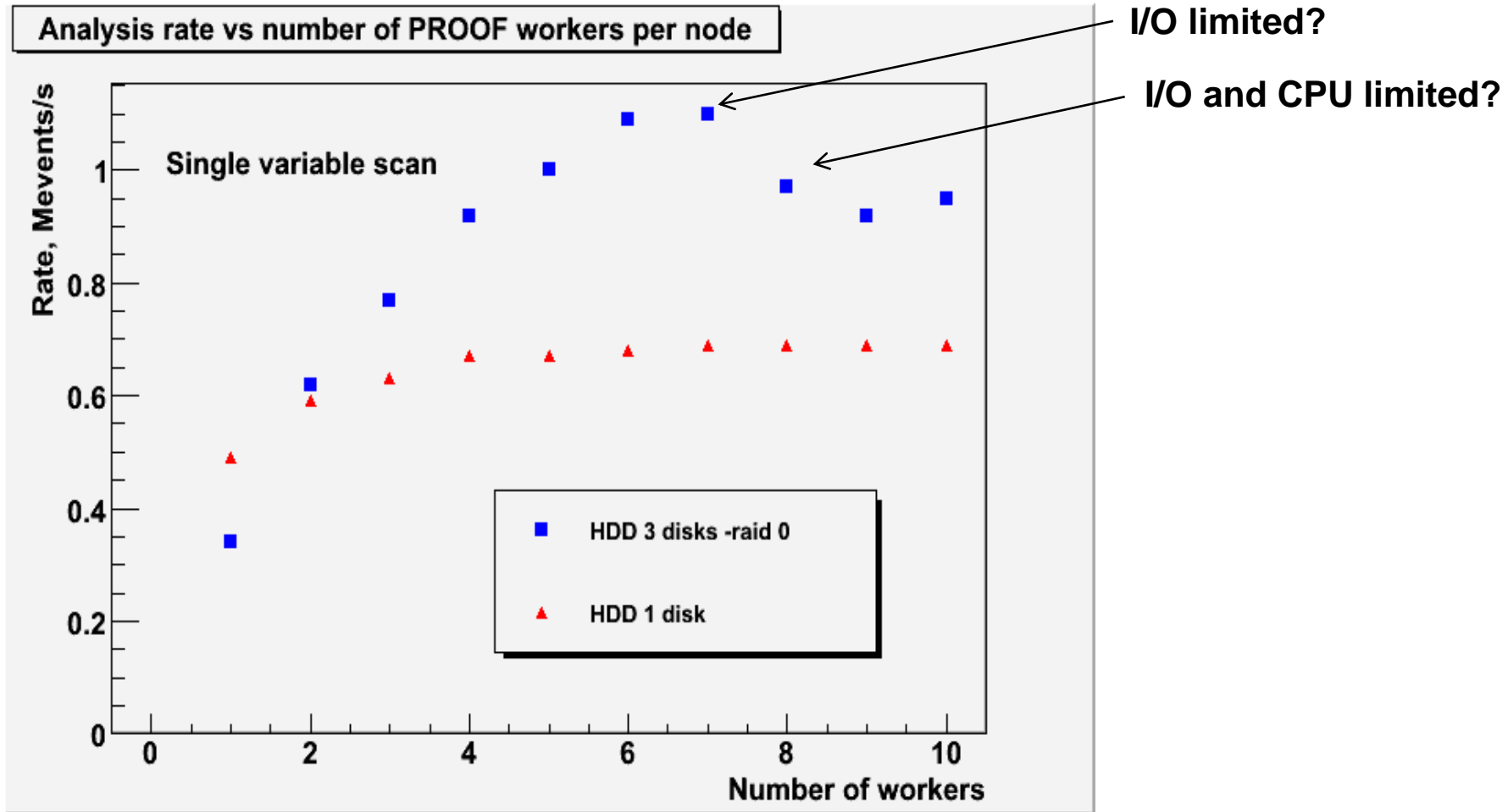
With 1 worker : 5.3M events, 15.8 MB read out of ~3 GB of data on disk
With 8 workers: 42.5M events, 126.5 MB read out of ~24 GB of data

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

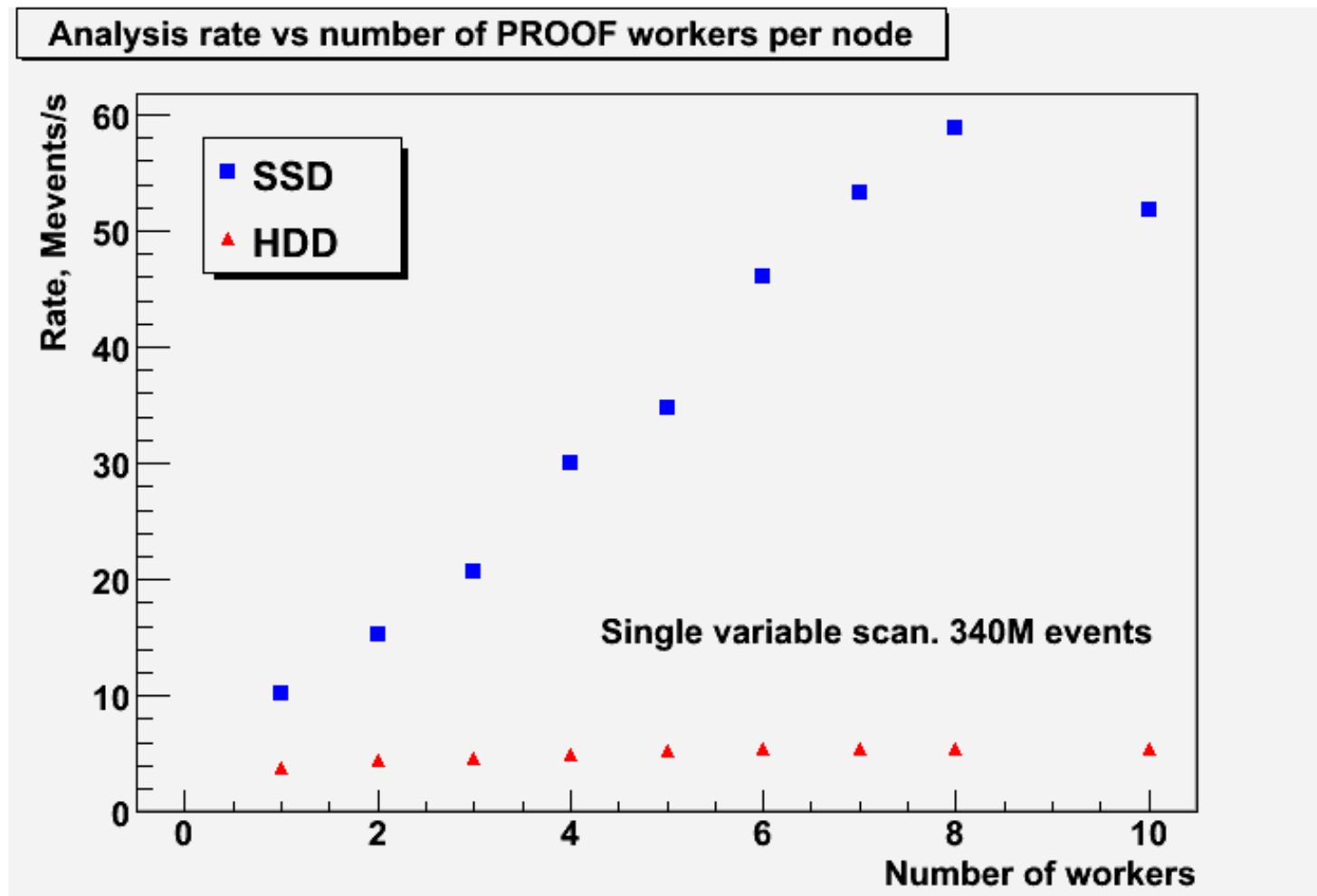
HDD single disk vs RAID



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

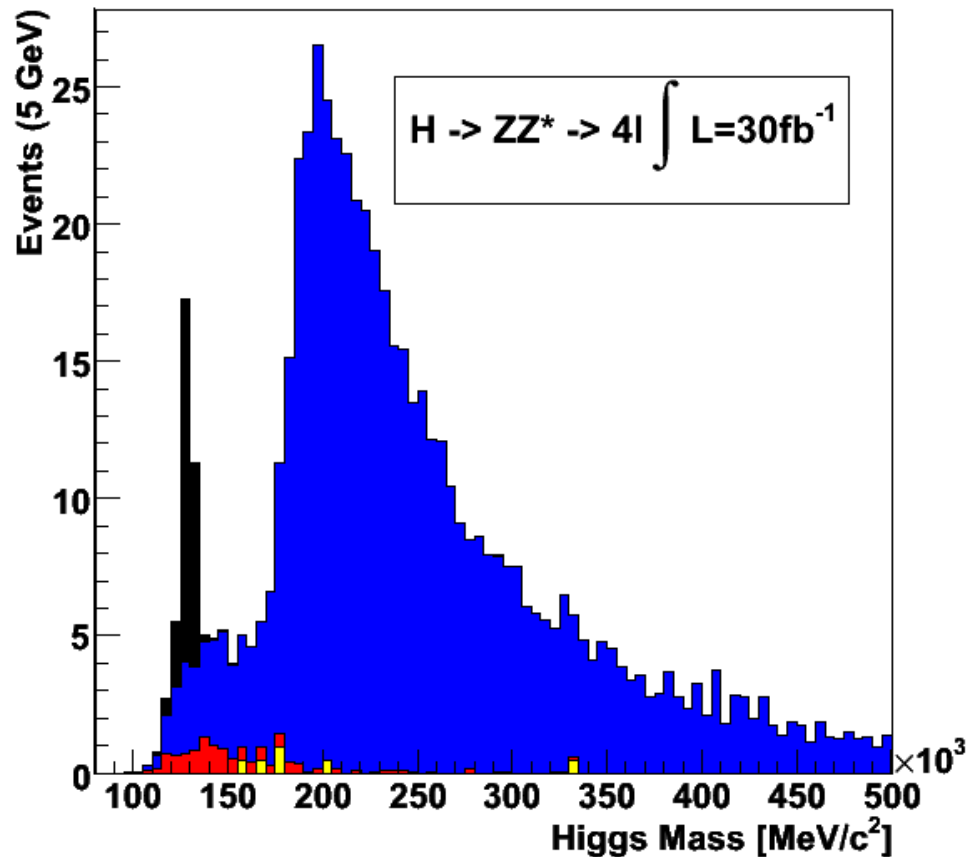
Almost linear scaling with number of nodes

Second use case details

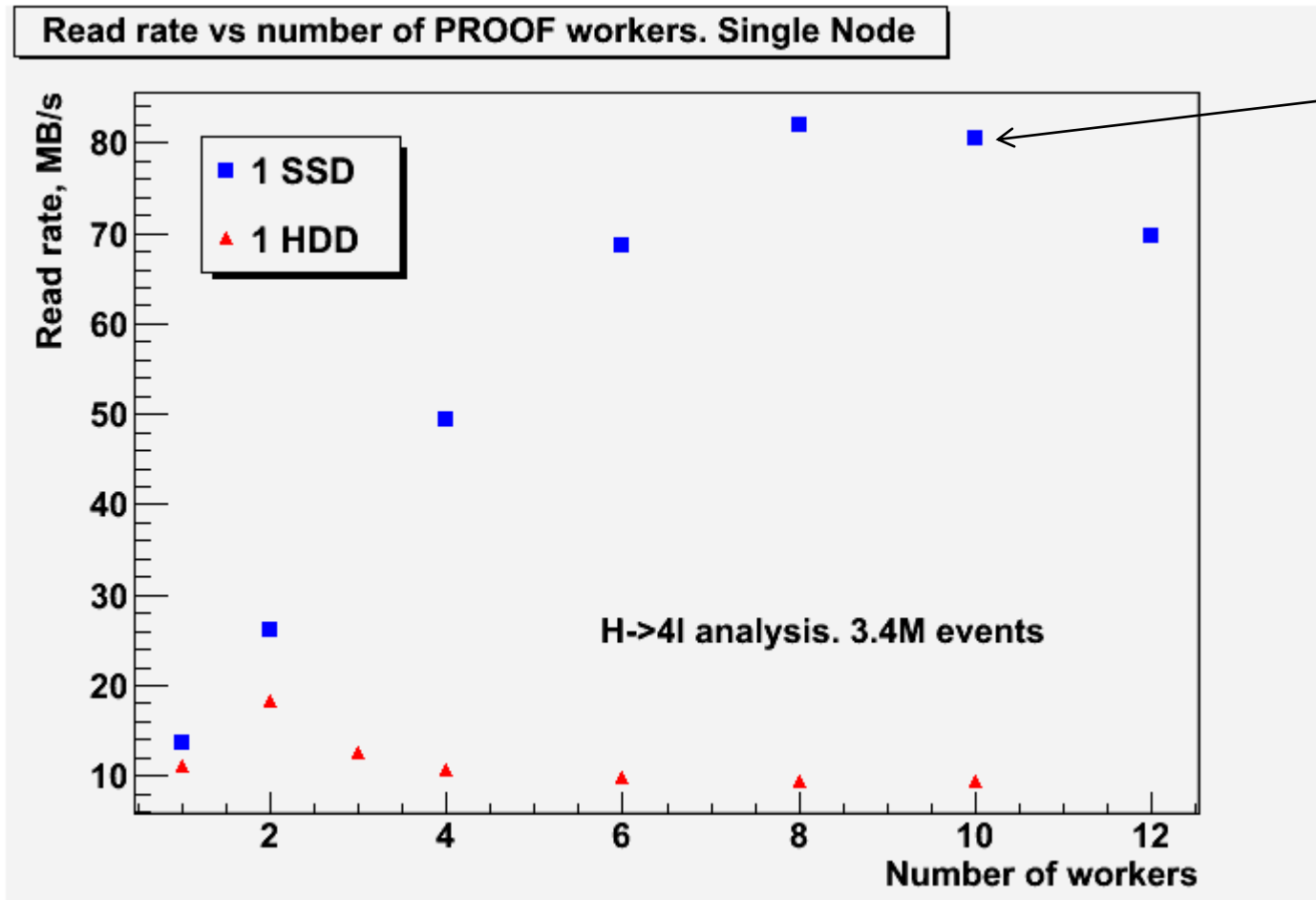
- Higgs decay into 4-lepton analysis
- 200 D3PD files, $\sim 3.4\text{M}$ events
- 46.4 GB of data
- Analysis include TMinuit fits
- CPU intensive, I/O intensive

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

Courtesy German Carrillo, UWM

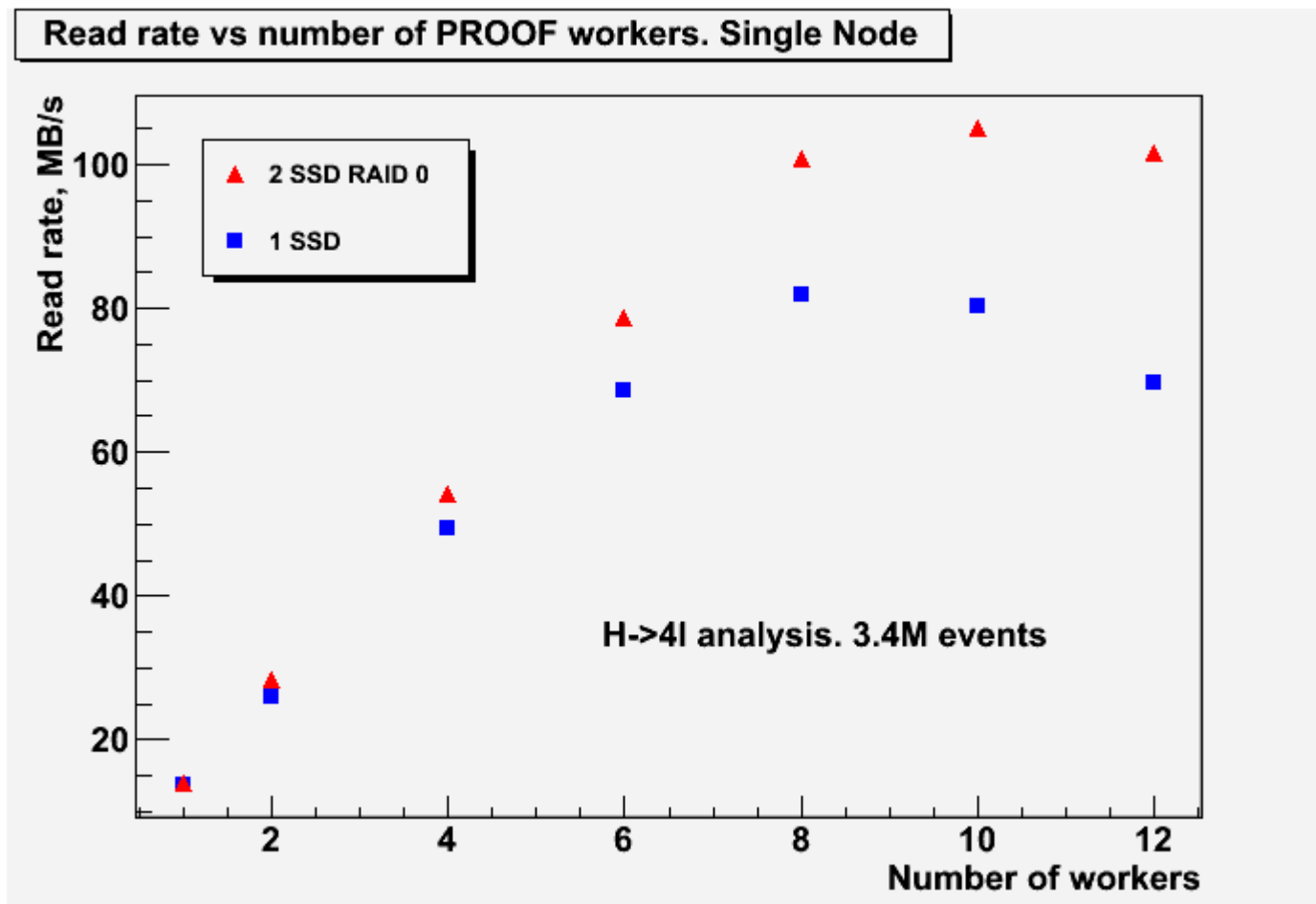


H->4l analysis. SSD vs HDD



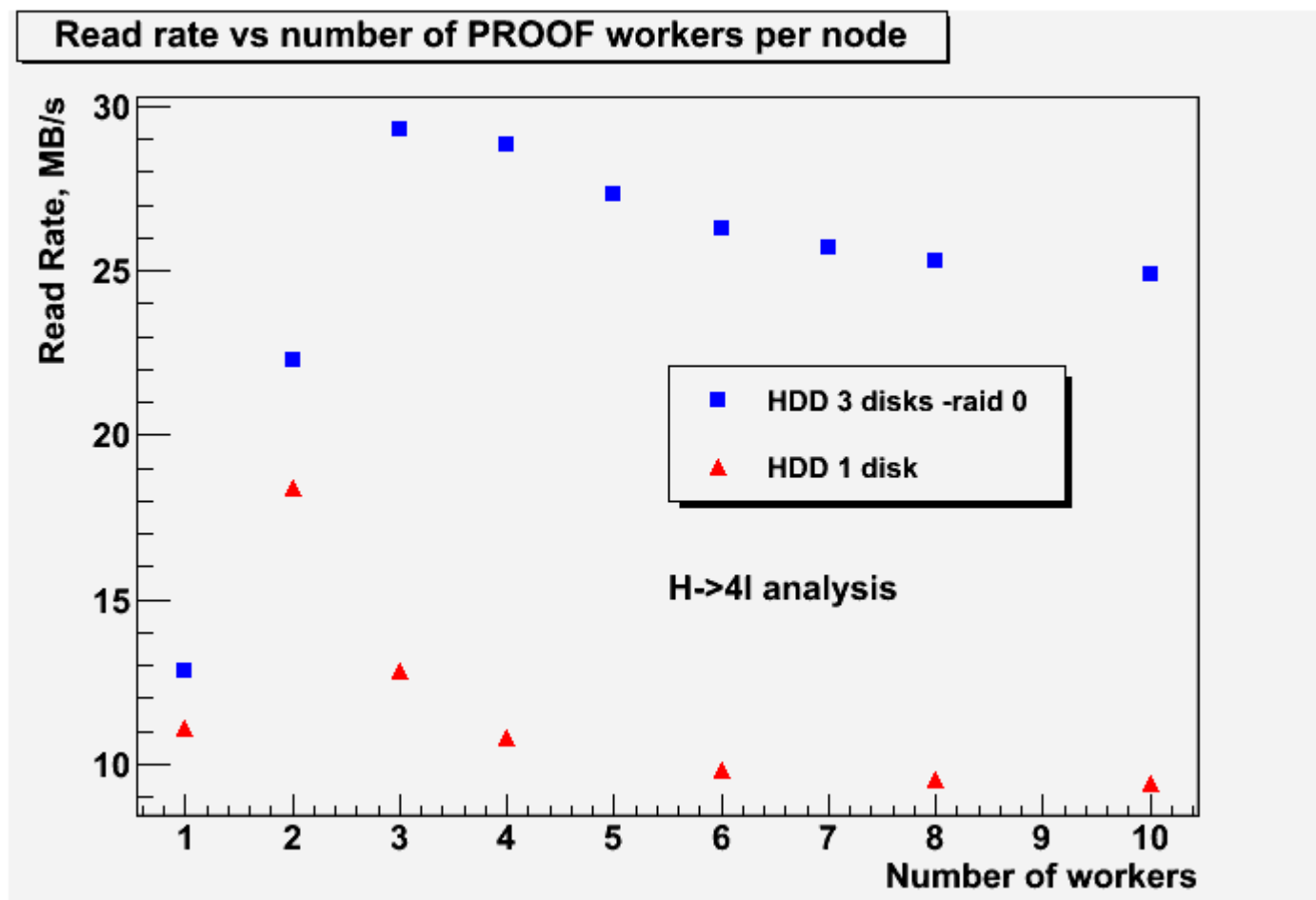
SSD is about 10 times faster at full load
Best HDD performance at 2 worker load
Single analysis job generates ~10 -14 MB/s load with given hardware

H->4l analysis. SSD RAID 0



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

H->4I 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

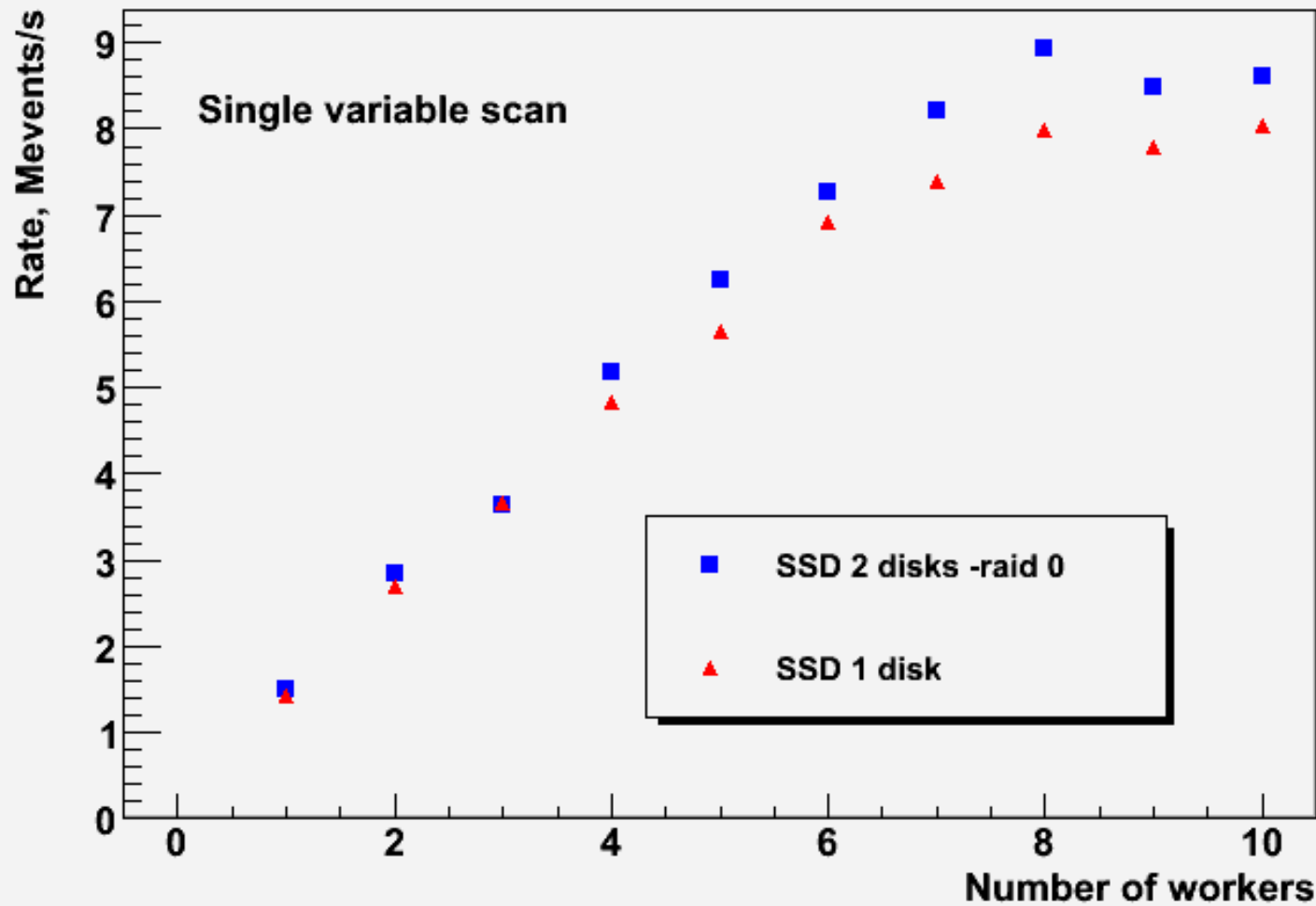
- ◆ SSD technology offer significant performance advantage in concurrent analysis environment
- ◆ We observed ~x10 better read performance than HDD in our test
- ◆ The main issue, in PROOF context, is matching of local I/O demand and supply
- ◆ Some observations from our tests
 - ◆ Single analysis worker 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

Analysis rate vs number of PROOF workers per node



H-4I analysis rate. SSD vs HDD

Analysis rate vs number of PROOF workers. Single Node

