

# Solid State Disks Testing with PROOF

**Sergey Panitkin, Robert Petkus, Ofer Rind**

**BNL**

**ATLAS**



**BROOKHAVEN**  
NATIONAL LABORATORY

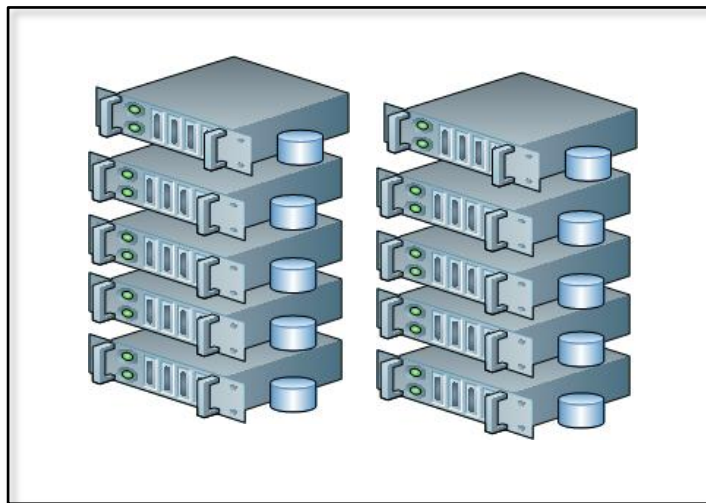
# Current BNL PROOF Farm Configuration

## “Old farm”-Production

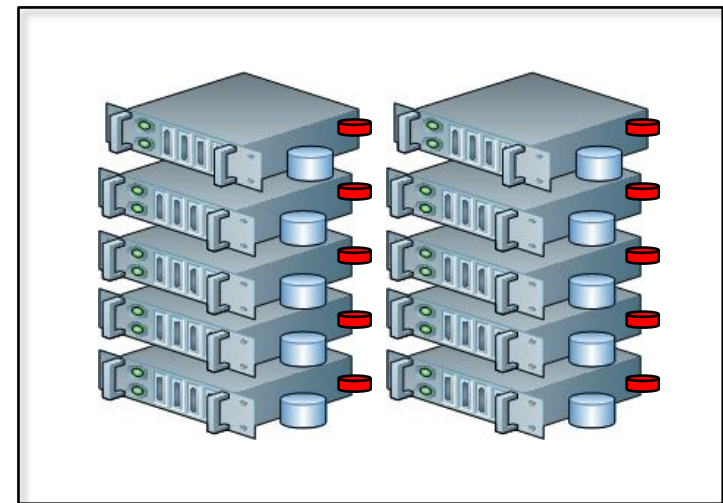
- 10 nodes – 4 GB RAM each
- 40 cores: 1.8 GHz Optrons
- 20 TB of HDD space (10x4x500 GB)

## “New Farm” – test site

- 10 nodes - 16 GB RAM each
- 80 cores: 2.0 GHz Kentsfields
- 5 TB of HDD space (10x500 GB)
- 640 GB SSD space (10x64 GB)



+



# New Solid State Disks@ BNL

- ◆ Model: Mtron MSP-SATA7035064
- ◆ Capacity 64 GB
- ◆ Average access time  $\sim 0.1$  ms (typical HD  $\sim 10$ ms)
- ◆ Sustained read  $\sim 120$ MB/s
- ◆ Sustained write  $\sim 80$  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





# Test configuration

- ◆ 1+1 or 1+8 nodes PROOF farm configurations
- ◆ 2x4 core Kentsfield CPUs per node, 16 GB RAM per node
- ◆ All default settings in software and OS
- ◆ Different configuration of SSD and HDD hardware depending on tests
- ◆ Root 5.18.00 – latest production version
- ◆ “PROOF Bench” suit of benchmark scripts to simulate analysis in root. Part of root distribution.
  - ◆ <http://root.cern.ch/twiki/bin/view/ROOT/ProofBench>
  - ◆ Data simulate HEP events ~1k per event
  - ◆ Single ~3+ GB file per PROOF worker in this tests
- ◆ Reboot before every test to avoid memory caching effects
- ◆ This set of tests emulates interactive, command prompt root session
  - ◆ Plot one variable, scan ~10E7 events, ala D3PD analysis
- ◆ Looking at read performance of I/O subsystem

# 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:

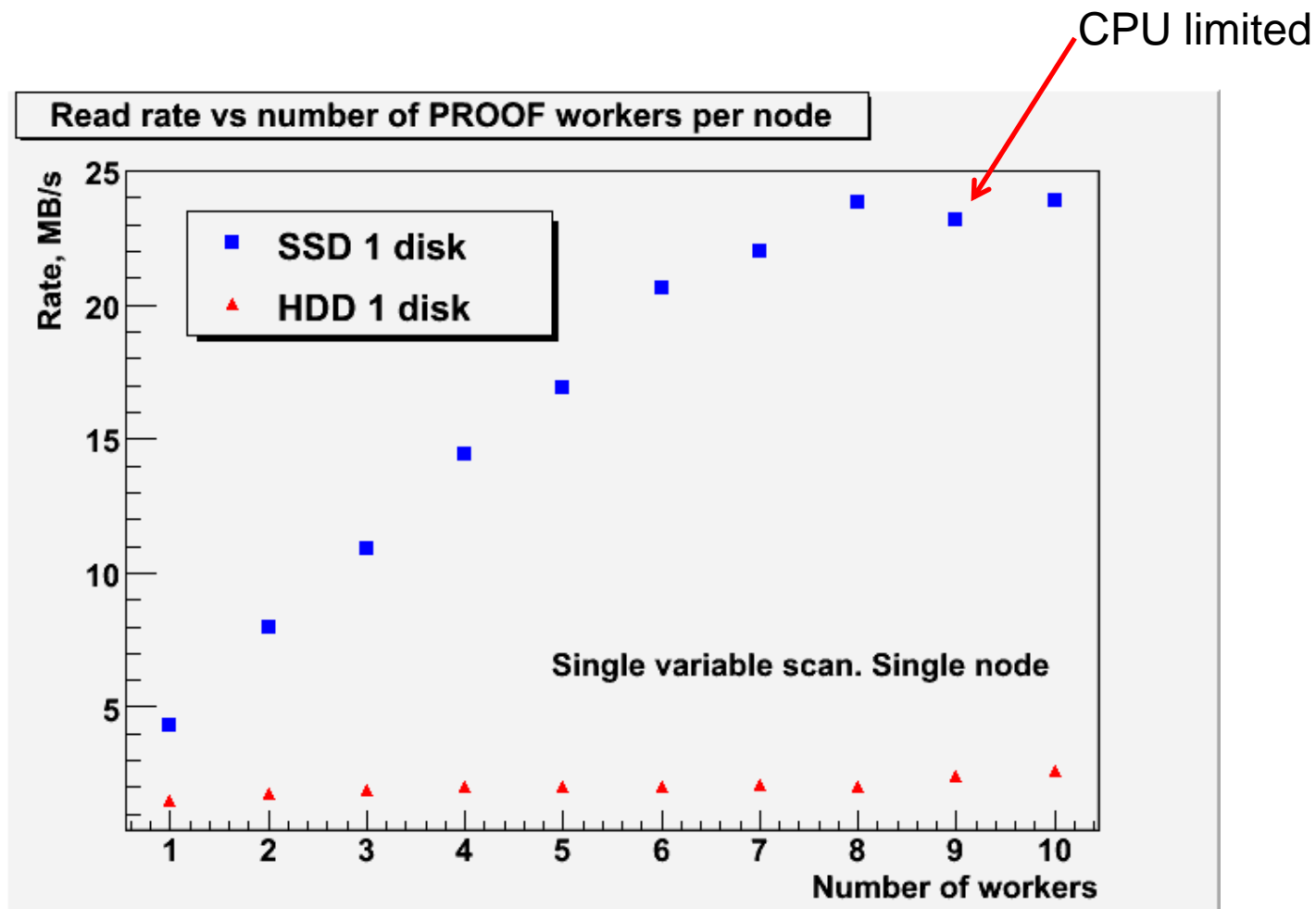
- PROOF Query Progress: serp@acas0601.usatlas.bnl.gov**: A dialog box showing the progress of a query. It indicates that 2 files with a total of 10629617 events are being processed. The processing rate is 2684585.6 evts/sec (8.0 MBs/sec). The initialization time is 1.3 seconds. The processed data is 10629617 events (31.65 MBs) in 4.0 seconds. There are checkboxes for "Close dialog when processing is complete" and "Show only logs from query [last]". Buttons for Stop, Cancel, Close, Show Logs, and Rate plot are visible at the bottom.
- Rate vs Time**: A line graph showing the processing rate in events per second (evts/sec) over time. The y-axis is labeled "Processing rate (evts/sec)" and has a multiplier of  $\times 10^6$ . The x-axis is labeled "elapsed time (sec)" and ranges from 2.5 to 4.0. The graph shows a nearly horizontal line at approximately 2684.58563 million evts/sec. A text box at the bottom of the graph states "Global average: 2684585.63 evts/sec".
- Terminal (serp@atlasgw00:~)**: A terminal window showing the execution of the PROOF test. The commands and output are as follows:

```
serp@atlasgw00:~  
File Edit View Terminal Tabs Help  
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:~)**: A terminal window showing the output of the PROOF test. The output is as follows:

```
serp@atlasgw00:~  
File Edit View Terminal Tabs Help  
#worker acas0604.usatlas.bnl.gov  
#worker acas0604.usatlas.bnl.gov  
#worker acas0604.usatlas.bnl.gov  
#worker acas0604.usatlas.bnl.gov  
#worker acas0604.usatlas.bnl.gov  
#  
#worker acas0604.usatlas.bnl.gov  
#worker acas0604.usatlas.bnl.gov  
#605  
worker acas0605.usatlas.bnl.gov  
worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#worker acas0605.usatlas.bnl.gov  
#606  
uu:---F1 expanded_proof.cf (Fundamental)---L125--70%-----  
Wrote /usatlas/u/serp/xrootd/my_config/expanded_proof.cf
```

The desktop taskbar at the bottom shows the system date and time as "Wed May 21, 5:21 PM". The system tray includes icons for Daryna, [serp@atlasgw00:~], serp@atlasgw00:~, serp@atlasgw00:~, PROOF Query Progr..., Rate vs Time, and the system clock showing 2.40.

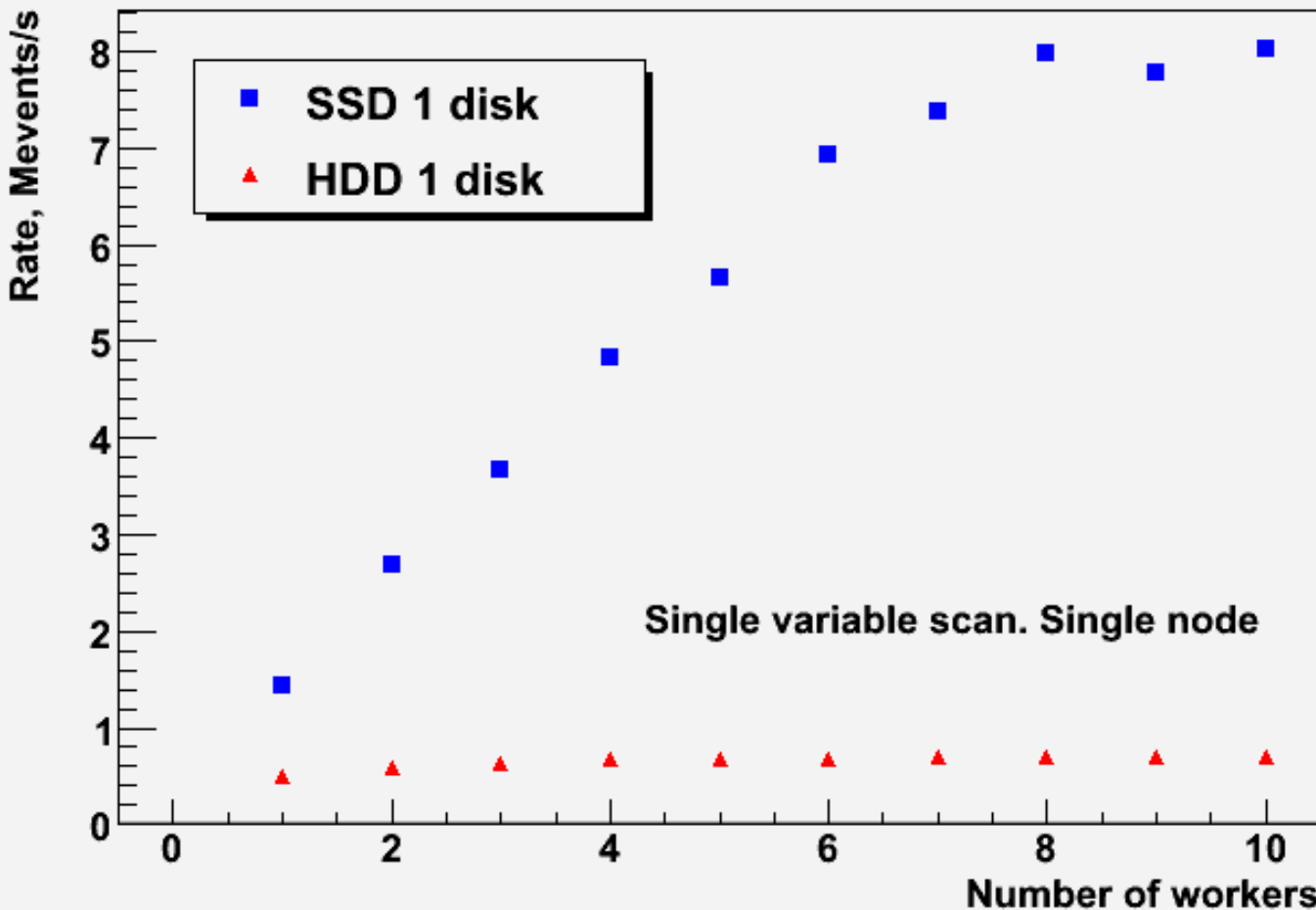
# SSD vs HDD



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

# SSD vs HDD

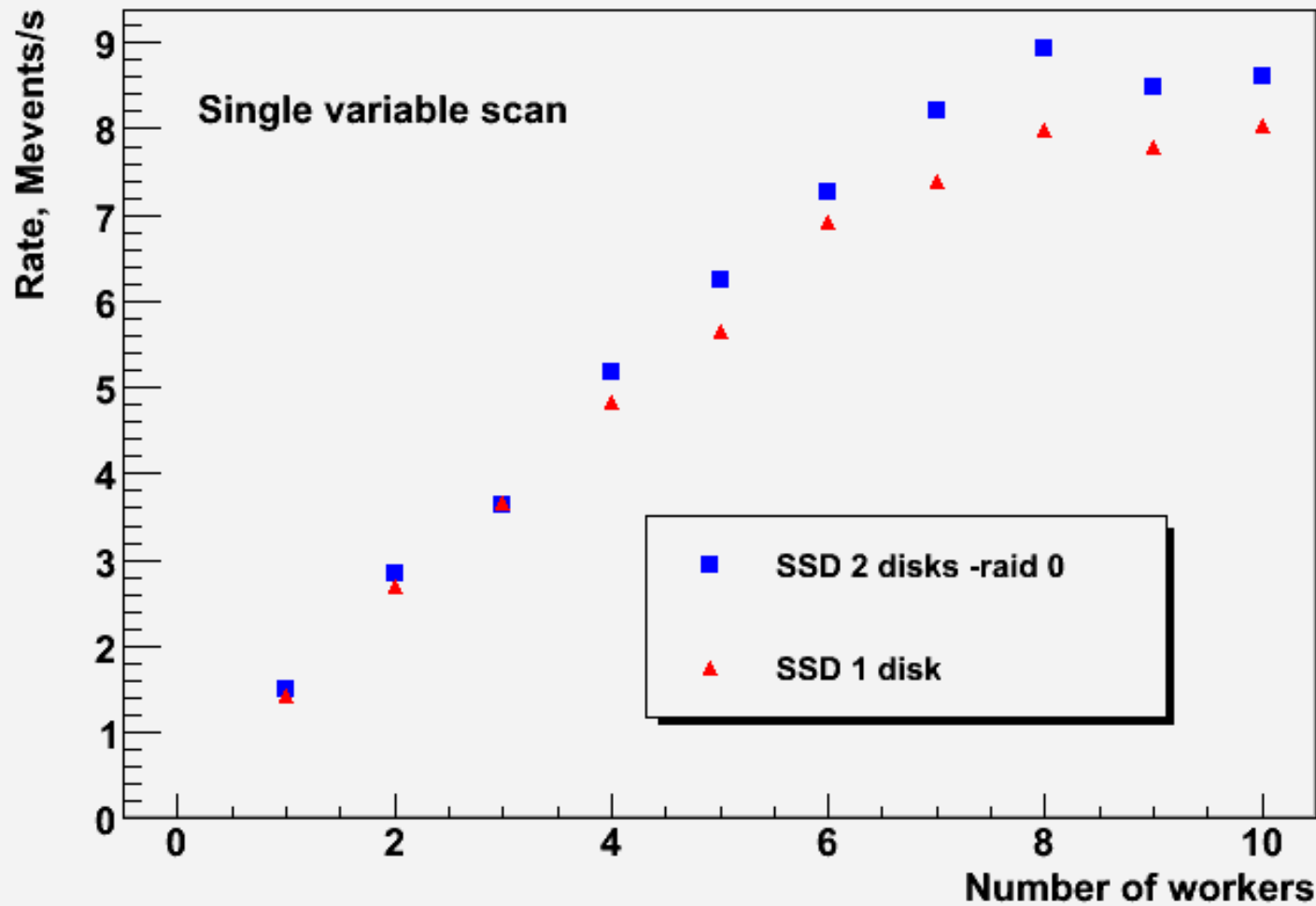
Analysis rate vs number of PROOF workers per node



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

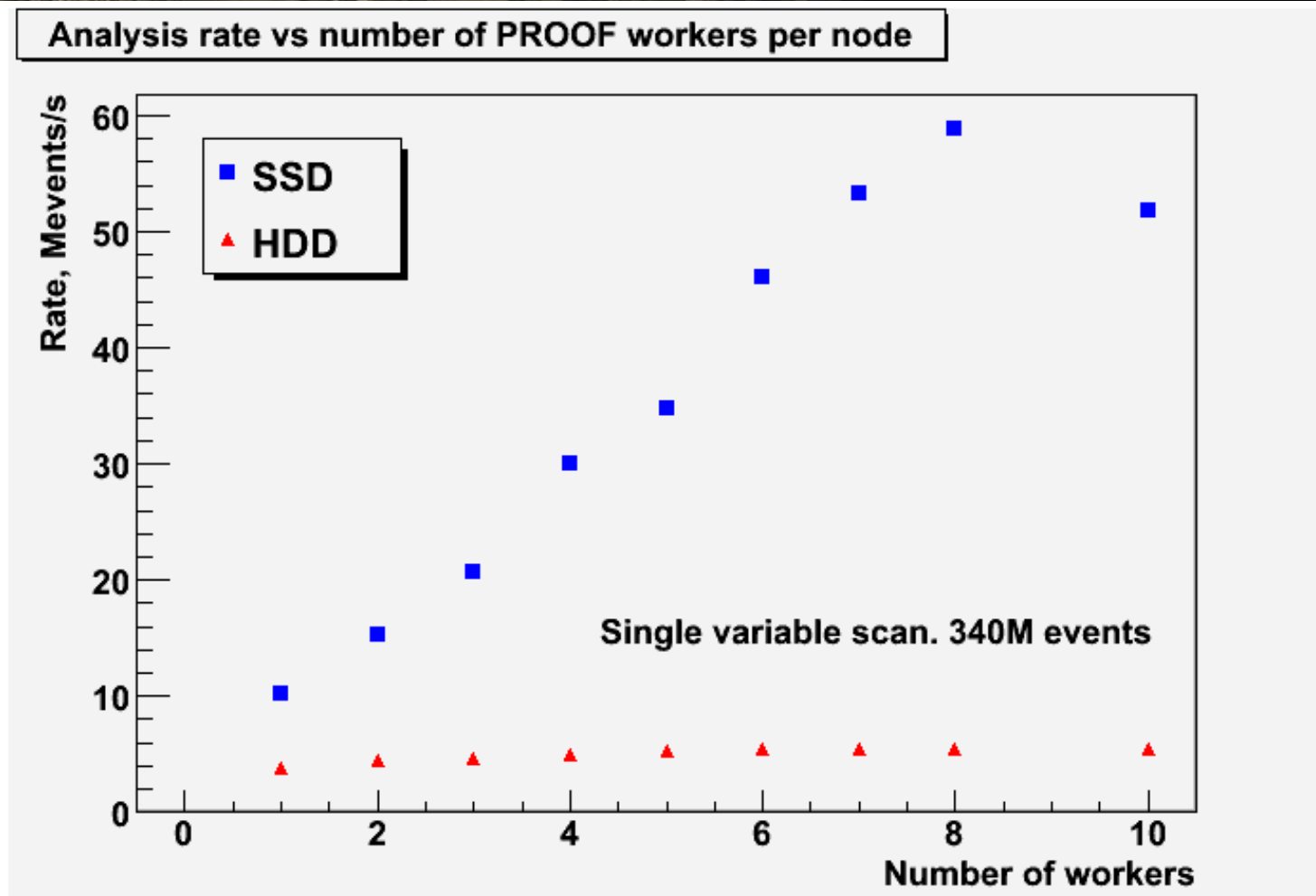
# SSD: single disk vs RAID

Analysis rate vs number of PROOF workers per node





# 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



# Summary

- ◆ SSD offer significant performance advantage in concurrent analysis environment
- ◆ ~x10 better read performance than HD in our test
- ◆ More results will be shown tomorrow
- ◆ More tests are planned
  - ◆ ARA on AODs and DPDs
  - ◆ Different hardware configurations