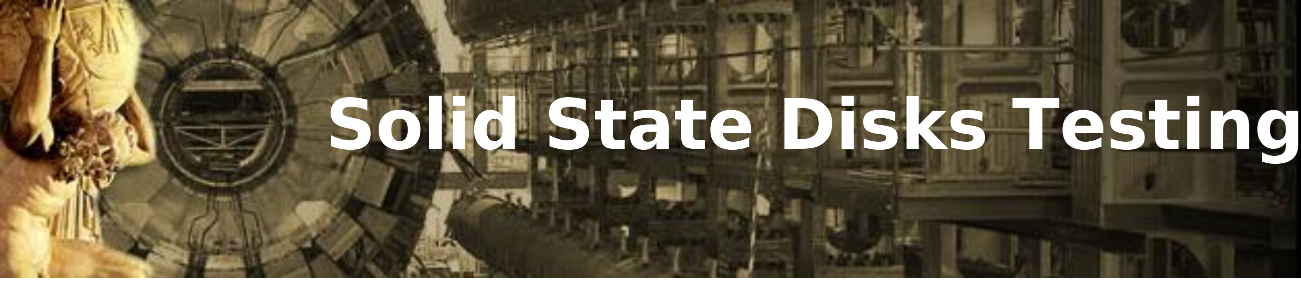


Solid State Disks Testing with PROOF

Sergey Panitkin, Robert Petkus, Ofer Rind

BNL

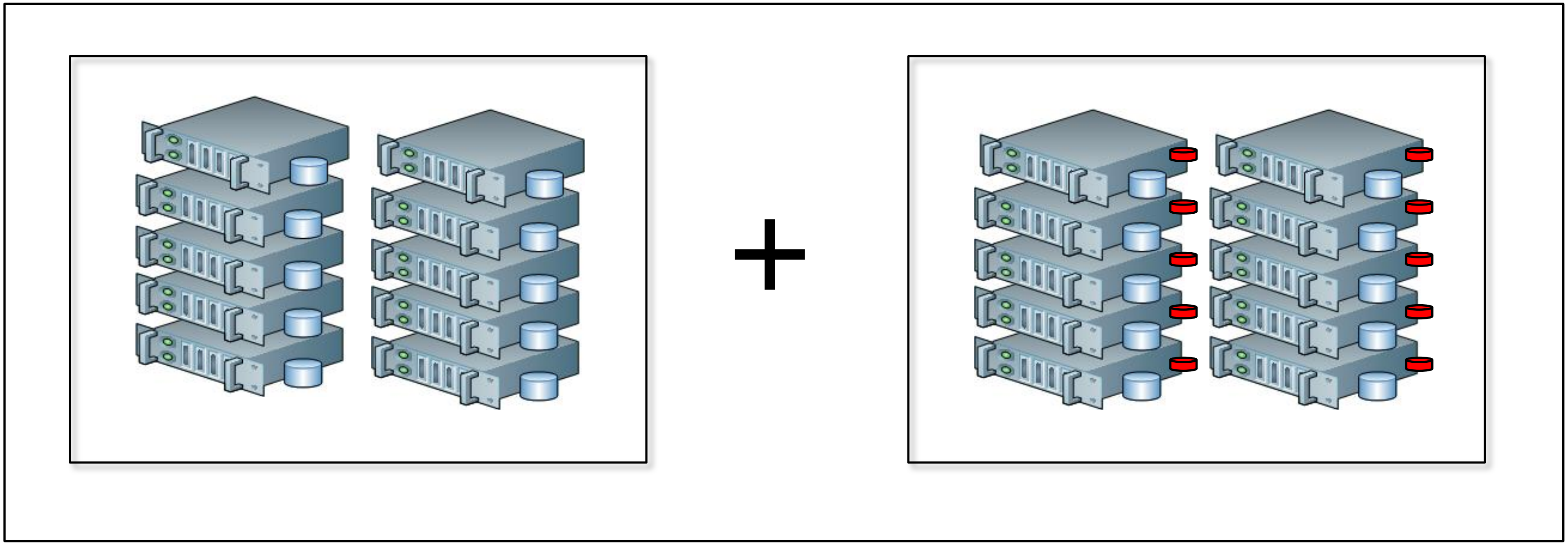
Current BNL PROOF Farm Configuration

“Old farm”-Production

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

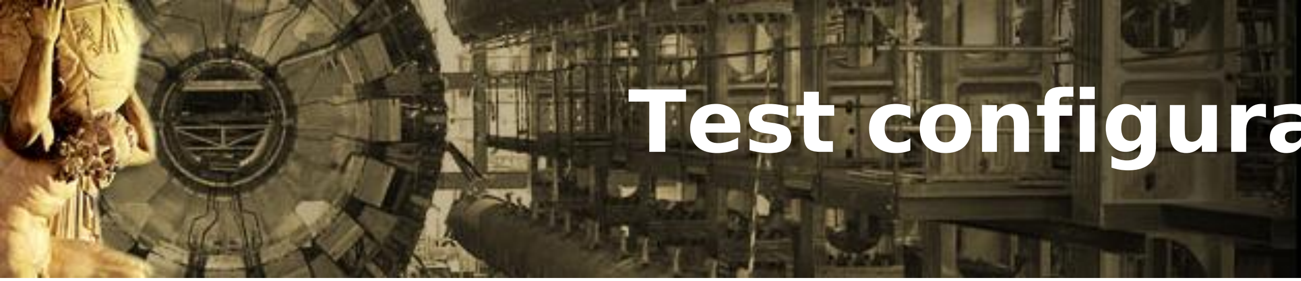
“New Farm” – test site

- 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 ~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

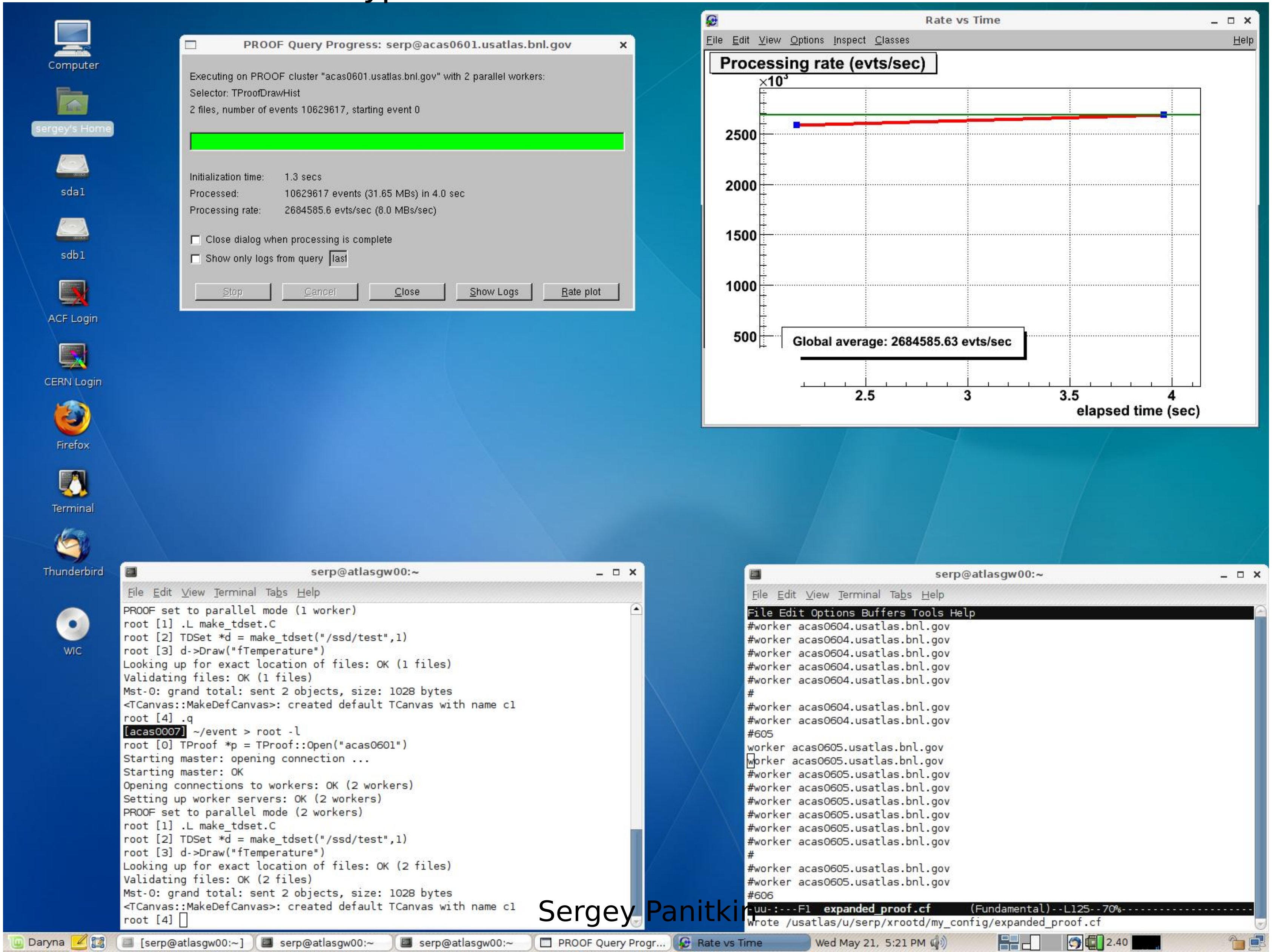


Test configuration

- 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



PROOF Query Progress: serp@acas0601.usatlas.bnl.gov

Executing on PROOF cluster "acas0601.usatlas.bnl.gov" with 2 parallel workers:
Selector: TProofDrawHist
2 files, number of events 10629617, starting event 0

Initialization time: 1.3 secs
Processed: 10629617 events (31.65 MBs) in 4.0 sec
Processing rate: 2684585.6 evts/sec (8.0 MBs/sec)

☐ Close dialog when processing is complete
☐ Show only logs from query last

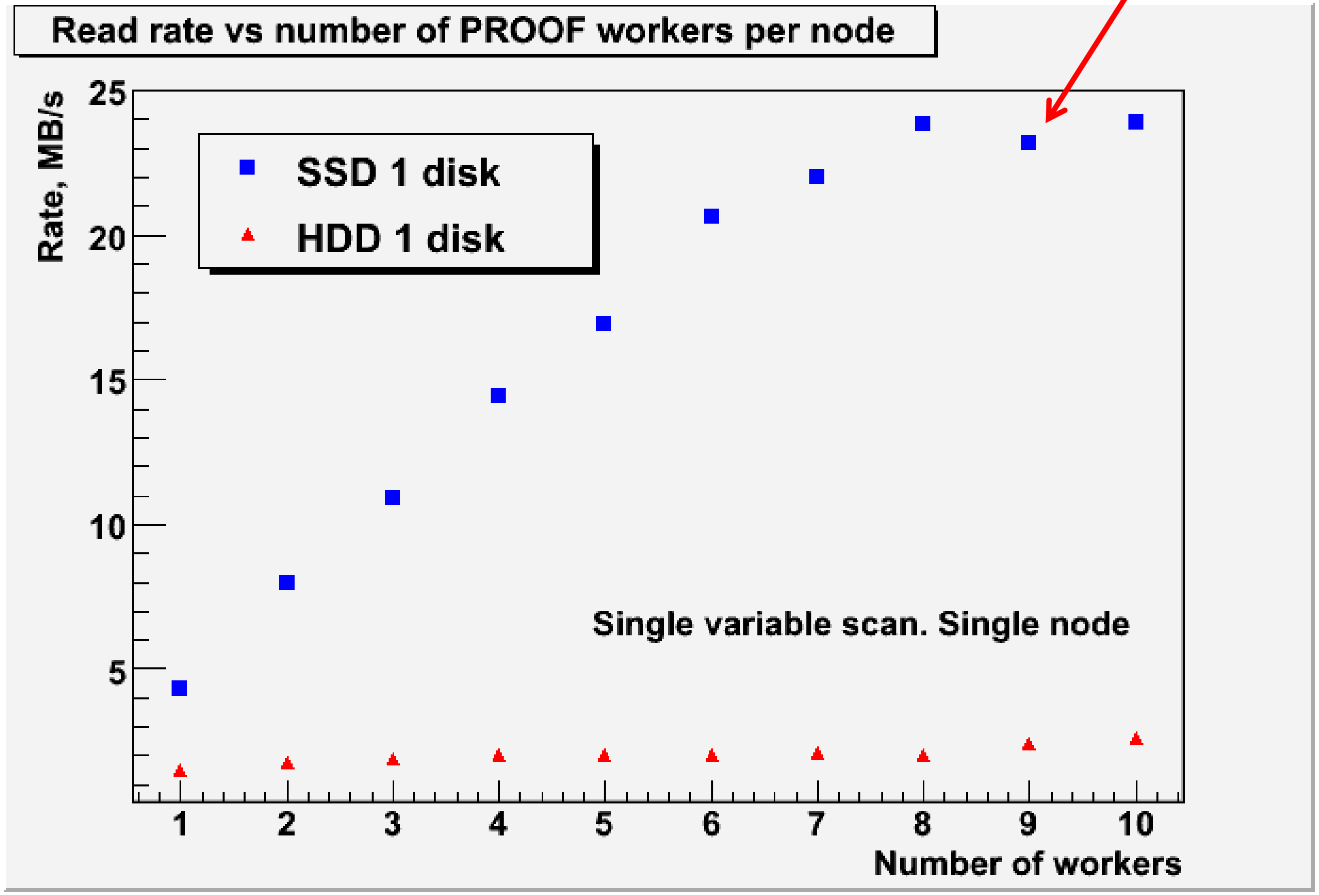
Global average: 2684585.63 evts/sec

```
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]
```

```
File Edit Options Buffers Tools 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
```

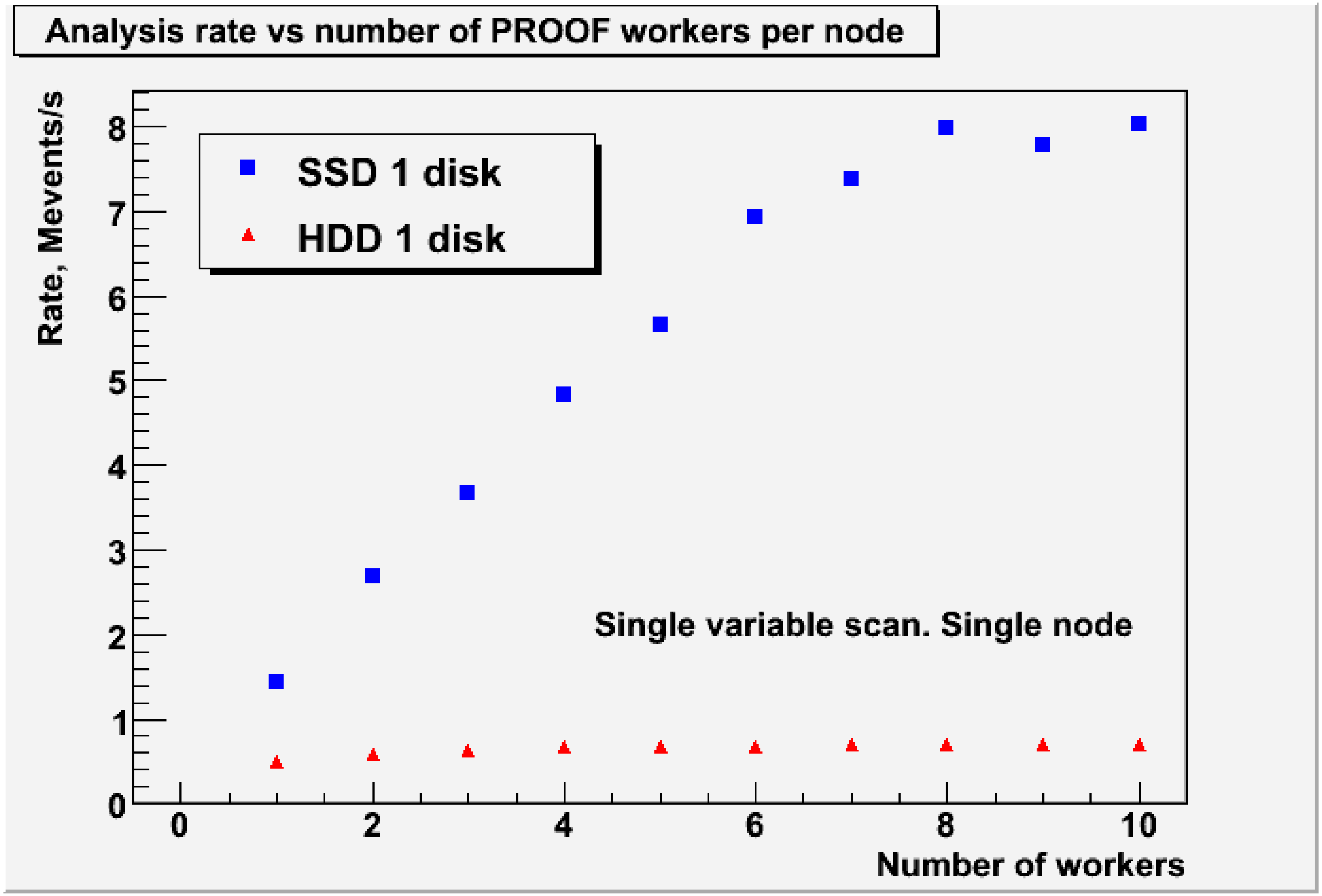
SSD vs HDD

CPU limited



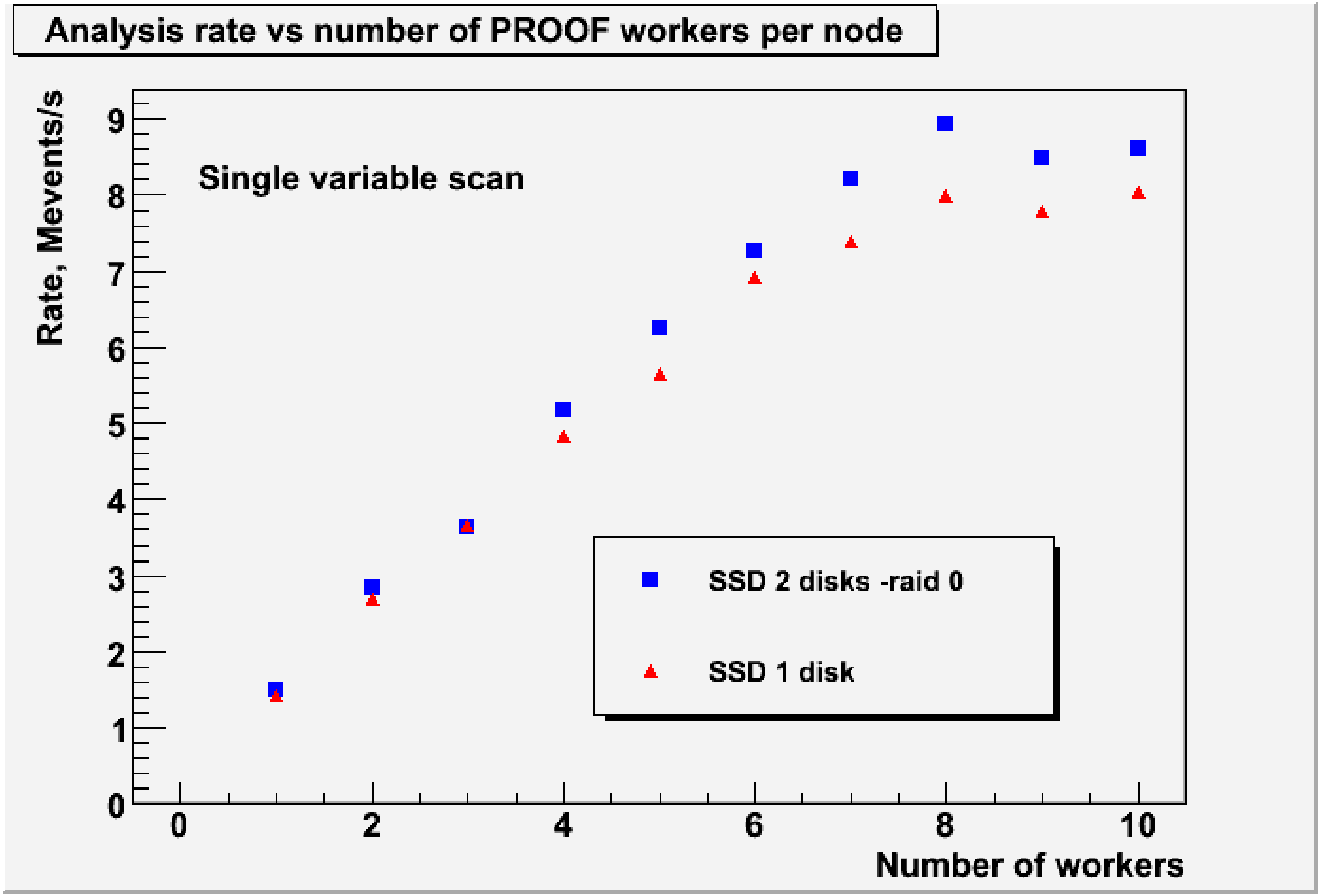
- SSD holds clear speed advantage
- ~ 10 times faster in concurrent read scenario

SSD vs HDD

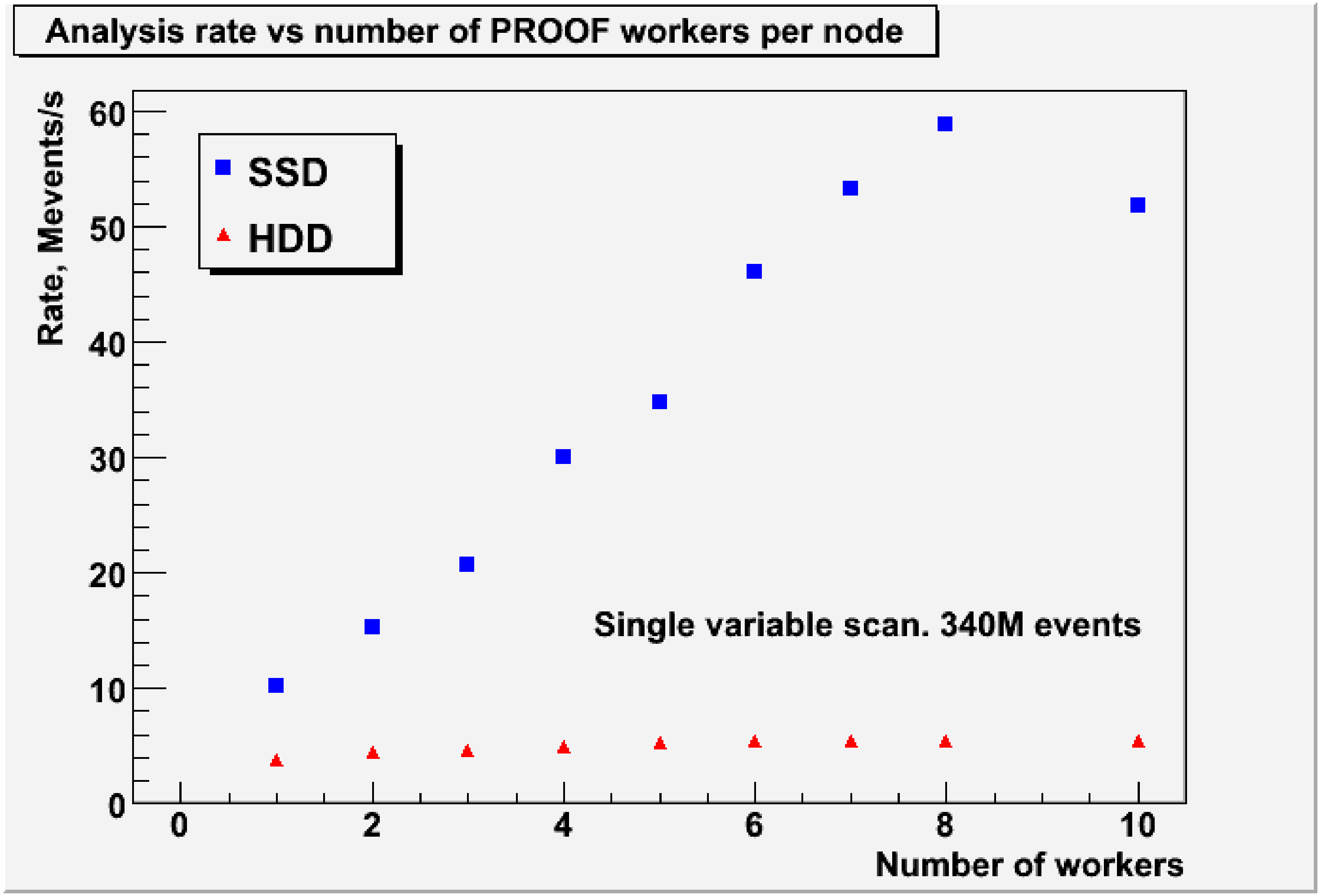


With 1 worker :   5.3M events,   15.8 MB read out of ~3 GB of data on disk
With 8 workers: 42.5M events, 126.5 MB read out of ~24 GB of data

SSD: single disk vs RAID



SSD 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