How to build inexpensive, low-maintenance, scalable supercomputer cluster?

B. Hari Haran

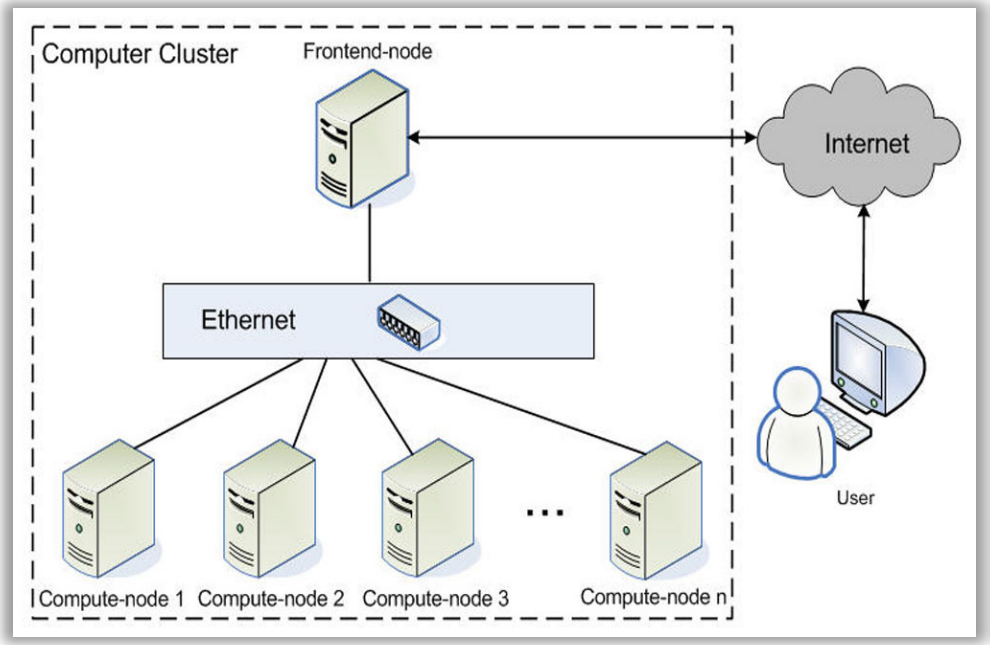
Cosmic Ray Laboratory, Ooty

Outline

- Computer cluster
- GRAPES-3 experiment
- Storage and Computing Demands
- Cluster design, assembly, and installation
- Challenges
- Monitoring
- Maintenance & security
- Applications

Computer cluster

- A computer cluster is a set of computers connected that work together and that can be viewed as a single system.



#1

Sunway TaihuLight, China
93.0 PFLOPS

#2

MilkyWay-2, China
33.9 PFLOPS

#3

Piz Daint, Switzerland
19.6 PFLOPS

Importance

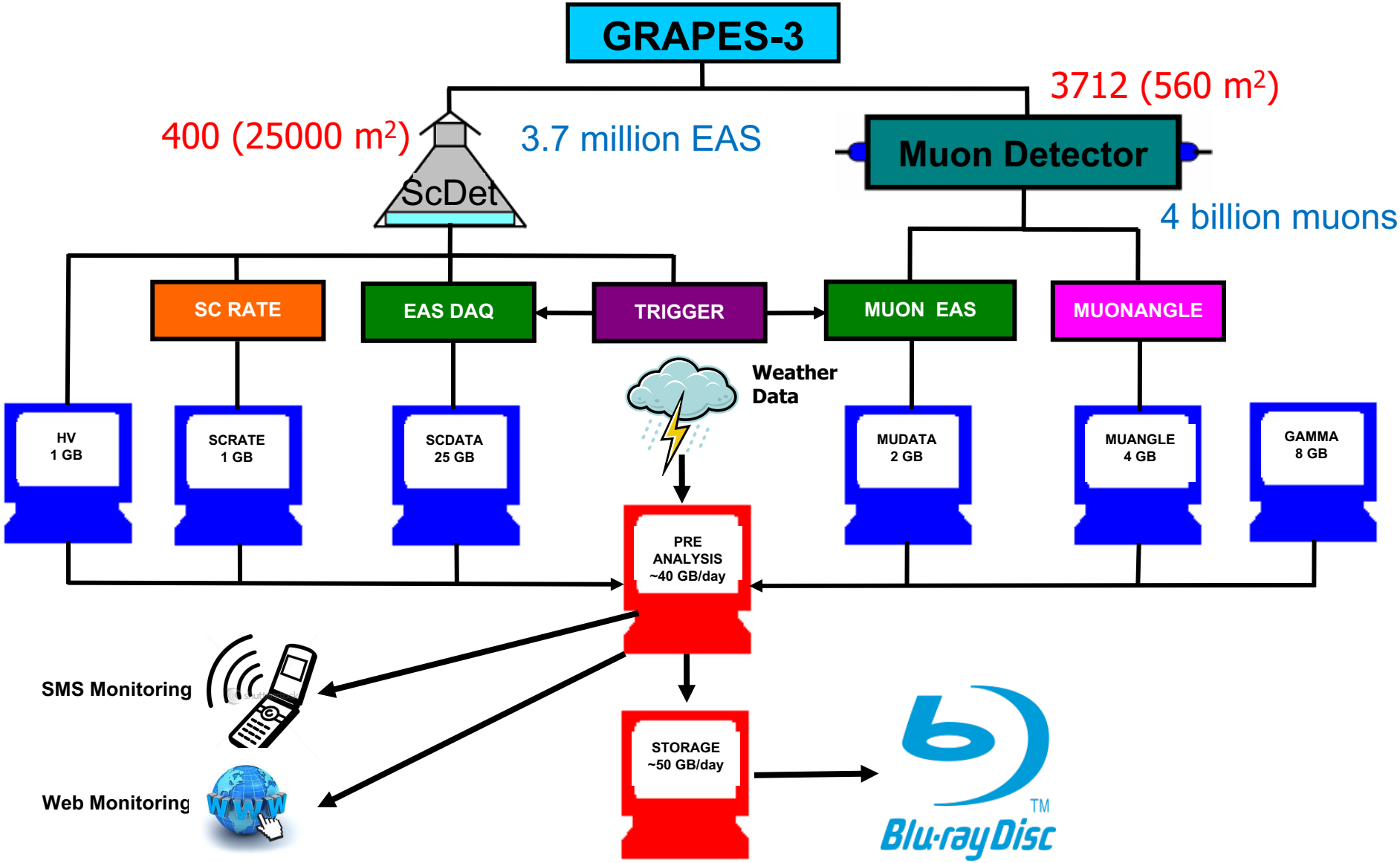
- Large scale computing
- Large data storage (PetaByte scale)
- Simulations
  - Weather forecast (Earthquake, storm, flood)
  - Industry (Aviation, automobile)
  - Health (Genetic)

GRAPES-3 experiment

Gamma Ray Astronomy at PeV EnergieS-3

Ooty, India, altitude 2200 m above msl

GRAPES-3 DAQ


GRAPES-3
400 (25000 m²)
ScDet
3.7 million EAS
3712 (560 m²)
Muon Detector
4 billion muons
SC RATE
EAS DAQ
TRIGGER
MUON EAS
MUONANGLE
Weather Data
HV 1 GB
SCRATE 1 GB
SCDATA 25 GB
MUDATA 2 GB
MUANGLE 4 GB
GAMMA 8 GB
PRE ANALYSIS ~40 GB/day
SMS Monitoring
Web Monitoring
STORAGE ~50 GB/day
Blu-ray Disc

Storage and computing demands

- Storing and managing 20 years of data
- Fast access
- Physics analysis (frequent)
- Simulations (frequent)
- Storage and management of simulated data
- Simulations are time consuming

Simulations are time consuming

- 6 m. X 60 s. X 16 mod. X 3 kHz = 17.3 X 10^6 muons

	Experiment	Simulation
Time	6 m.	24 h
Size	Few MB	Few hundred GB

- High statistics essential for precision study

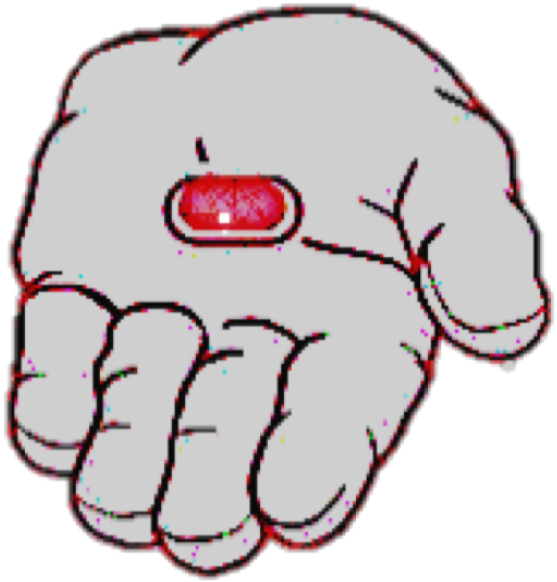
Various simulations

- CORSIKA (Sensitive to I/P)
  - Muon angular distribution (10 GeV – 10 TeV)
    - Geomagnetic storm
    - Thunderstorm
  - Extensive Air shower (10 TeV – 10 PeV)
  - Multi-TeV gamma rays
- Detector simulation Geant4
- In-house simulation programs
  - G3Sim (P.K. Mohanty et al., *Review of Scientific Instruments 83 (2012) 043301*)
  - GRAPES-3 EAS

Storage and computing demands

- Storage target
  - Experiment: 200 TB includes
    - Compressed binary
    - ROOT files
    - Processed
  - Simulation: 100 TB
- Computing target
  - Minimum 1000 job, 1 GB RAM/job
  - Analysis of one year data in one day

Available solutions



COMERCIAL
- Expensive
- Fixed configuration
- AMC

CUSTOM
- Inexpensive (Hardware & Software)
- Customized configuration
- In-house maintenance, no AMC

Blue pill or Red pill ?

Early computer clusters

- 1st Generation (2006)
  - Nodes: 8
  - Jobs: 8 X 2 = 16
  - RAM: 8 X 2 GB = 16 GB
  - Storage: 12 TB
- 2nd Generation (2008)
  - Nodes: 34
  - Jobs: 34 X 8 = 272
  - RAM: 34 X 8 GB = 272 GB
  - Storage: 150 TB

3rd generation

- Prototype-1 (2012)
  - Intel Xeon E5645 @ 2.40 GHz
  - Jobs: 2 X 12 = 24
  - RAM: 24 GB
  - Storage: 40 TB
- Prototype-2 (2013)
  - Intel Xeon E5-2650 @ 2.00 GHz
  - Jobs: 2 X 16 = 32
  - RAM: 32 GB
  - Storage: 60 TB

Cluster configuration

- Nodes: 40
- Jobs: 40 X 32 = 1280
- RAM: 40 X 32 = 1280 GB
- Storage: 780 TB

Hardware configuration

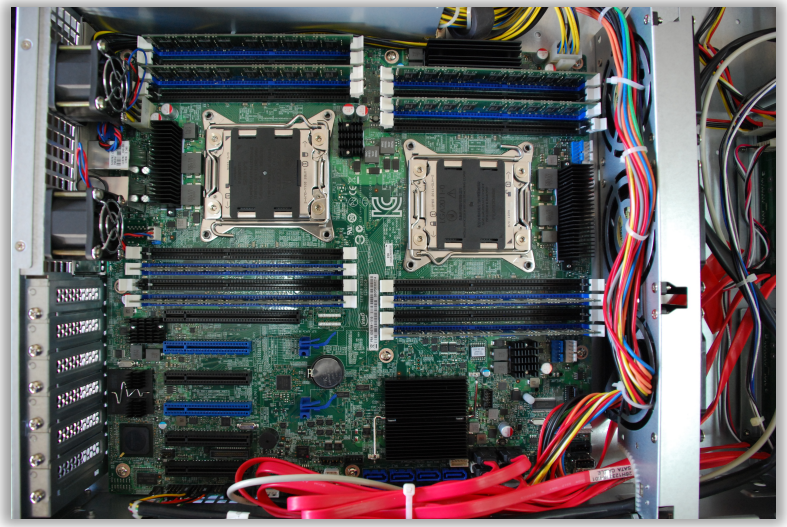
- CPU: Intel Xeon E5-2650 @ 2.00 GHz
- MB: Intel server board DBS2600CP2
- RAM: Kingston 4 GB DDR3 (KVR1333D3E9S/4)
- HDD: Seagate Constellaton 3TB (ST33000650NS/ST3000NM00333)
- Chassis
  - Primesource chassis RM-210 with 500W SMPS
  - Primesource chassis RM4201 with 2X1200W SMPS
- RAID controller: Intel RAID RS2WG160

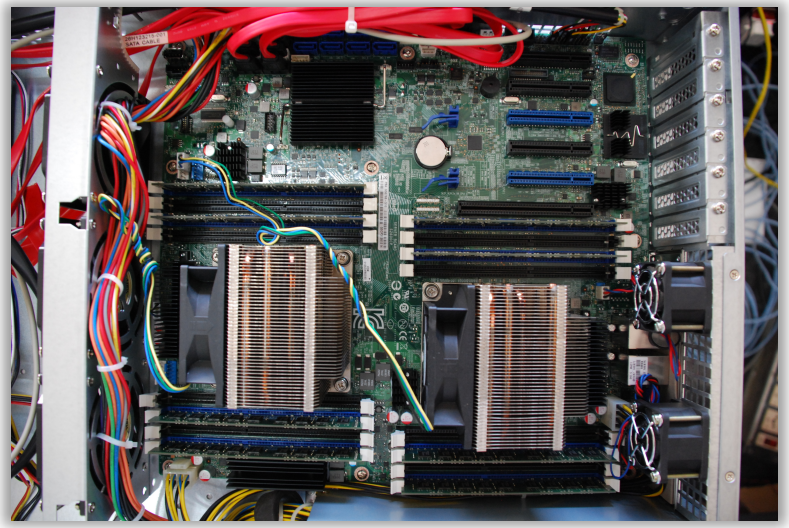
Software configuration

- OS: Rocks 6.1, CentOS based
- Sun Grid Engine
- Glusterfs
- In-house monitoring tools
- Ganglia
- CORISKA, ROOT, Geant4, G3ANALYSIS, etc..

Please Pay
$0

Cluster assembly





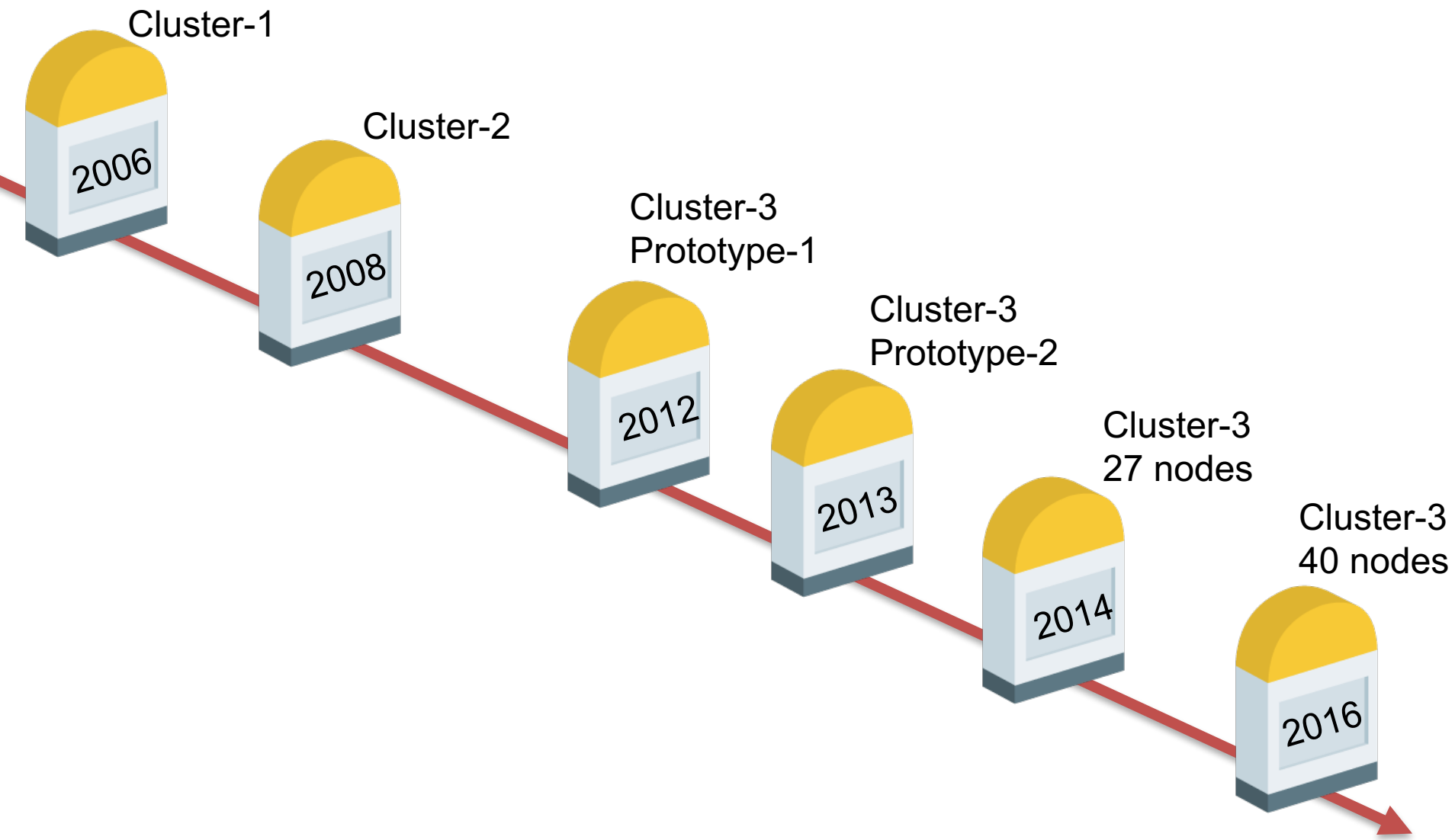
Assembled by team of 6 people

Cluster assembly

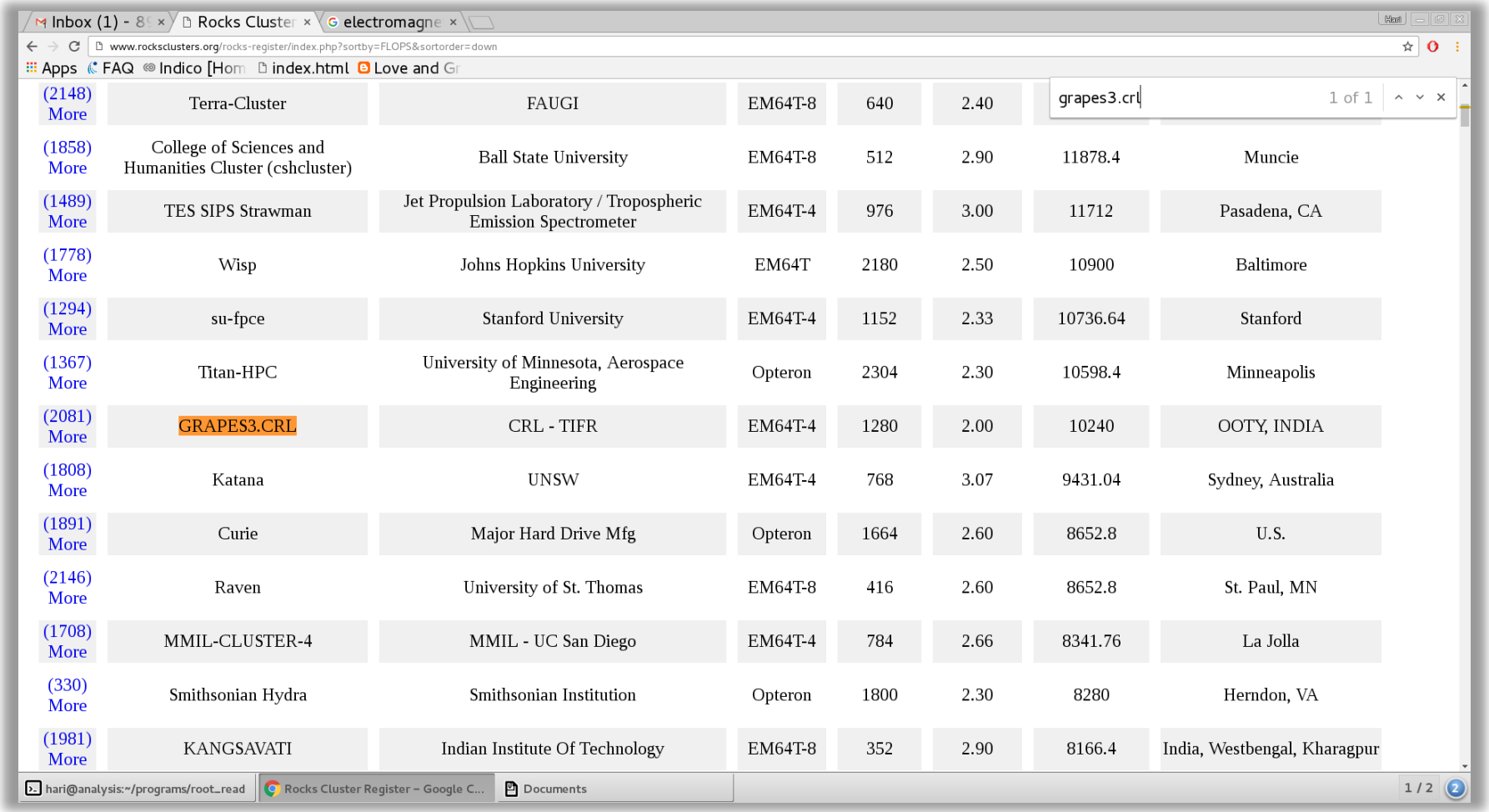
Installed at
17” 42U racks

Powered by
2 X 40 kVA UPS

Timeline


Cluster-1
2006
Cluster-2
2008
Cluster-3
Prototype-1
2012
Cluster-3
Prototype-2
2013
Cluster-3
27 nodes
2014
Cluster-3
40 nodes
2016

Cluster ranking



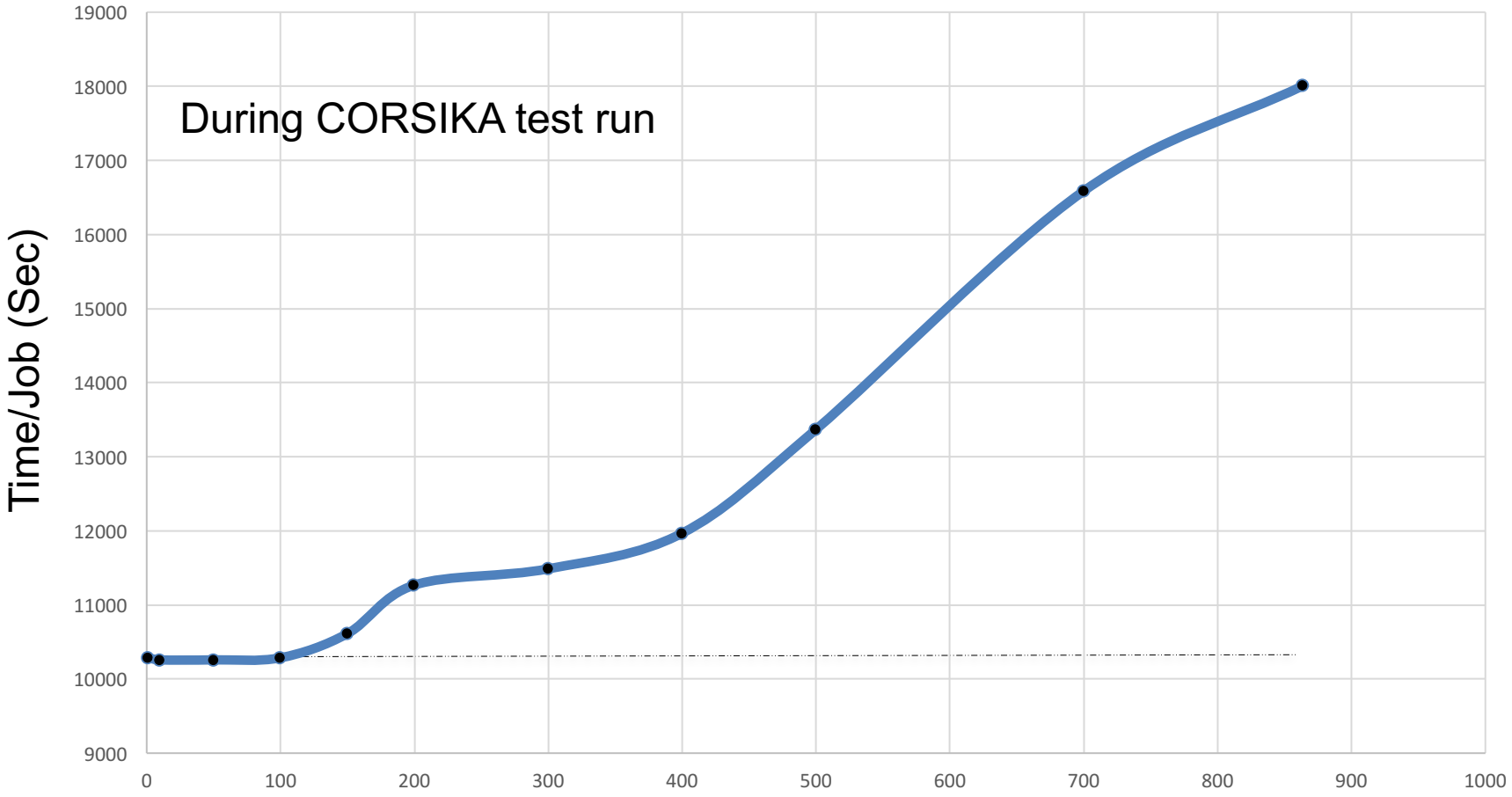
(2148) More	Terra-Cluster	FAUGI	EM64T-8	640	2.40		
(1858) More	College of Sciences and Humanities Cluster (cshcluster)	Ball State University	EM64T-8	512	2.90	11878.4	Muncie
(1489) More	TES SIPS Strawman	Jet Propulsion Laboratory / Tropospheric Emission Spectrometer	EM64T-4	976	3.00	11712	Pasadena, CA
(1778) More	Wisp	Johns Hopkins University	EM64T	2180	2.50	10900	Baltimore
(1294) More	su-fpce	Stanford University	EM64T-4	1152	2.33	10736.64	Stanford
(1367) More	Titan-HPC	University of Minnesota, Aerospace Engineering	Opteron	2304	2.30	10598.4	Minneapolis
(2081) More	GRAPES3.CRL	CRL - TIFR	EM64T-4	1280	2.00	10240	OOTY, INDIA
(1808) More	Katana	UNSW	EM64T-4	768	3.07	9431.04	Sydney, Australia
(1891) More	Curie	Major Hard Drive Mfg	Opteron	1664	2.60	8652.8	U.S.
(2146) More	Raven	University of St. Thomas	EM64T-8	416	2.60	8652.8	St. Paul, MN
(1708) More	MMIL-CLUSTER-4	MMIL - UC San Diego	EM64T-4	784	2.66	8341.76	La Jolla
(330) More	Smithsonian Hydra	Smithsonian Institution	Opteron	1800	2.30	8280	Herndon, VA
(1981) More	KANGSAVATI	Indian Institute Of Technology	EM64T-8	352	2.90	8166.4	India, Westbengal, Kharagpur

36th position in rocks cluster ranking
http://www.rocksclusters.org/rocks-register

Innovative cooling

- No air conditioning
- Ambient air through ducts
- Dust filter
- <25℃ throughout year
- 3 kW to remove 24 kW heat

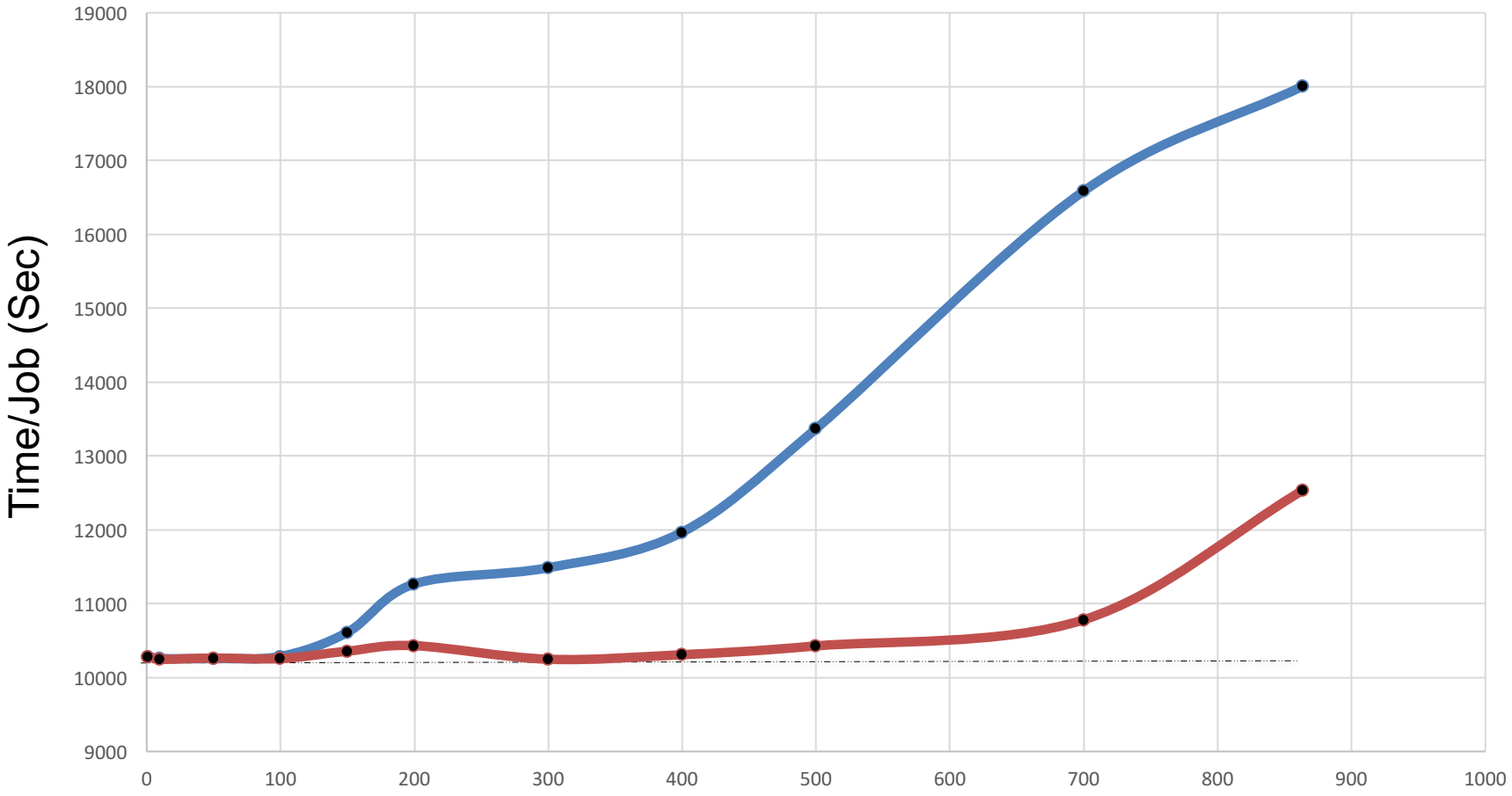
Challenges - Network saturation



With 48 X 1 Gbps LAN

No. of Jobs submitted

Challenges - Network saturation



With 48 X 1 Gbps LAN

No. of Jobs submitted

With Dell power connect 5548
2 X 10 Gbps opt., 48 X 1 Gbps

Challenges - Distributed storage

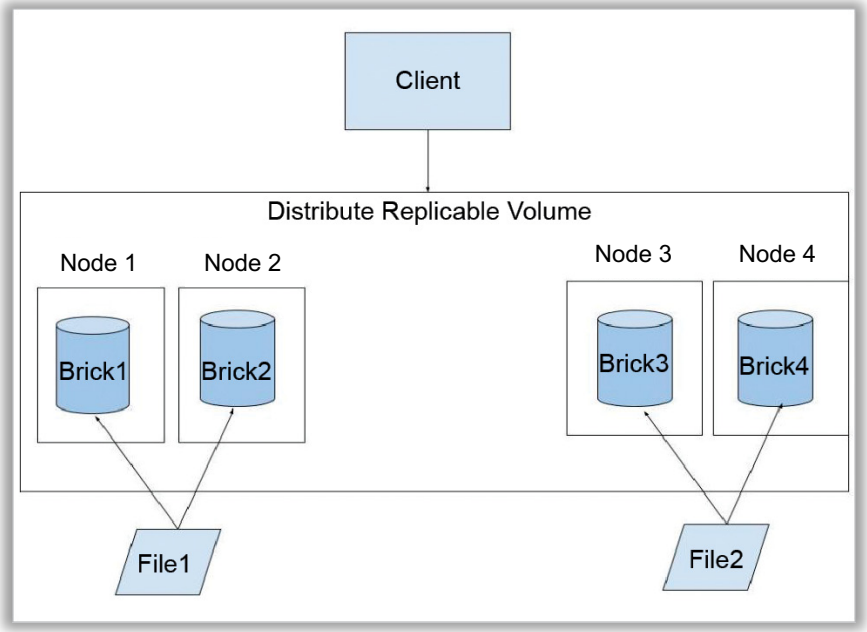
- A storage volume grown without affecting existing data
- Distributed over network
- Speed
- Redundancy
- Scalable
- Solution for increasing storage demand

Challenges - Distributed storage

- Open-source
  - Glusterfs, Ceph, DRBD, BeeGFS, HDFS
- Users of glusterfs are
  - Pandora (music service)
  - box.net (file sharing service)
  - NTT (Nippon Telegraph and Telephone)
  - Some universities

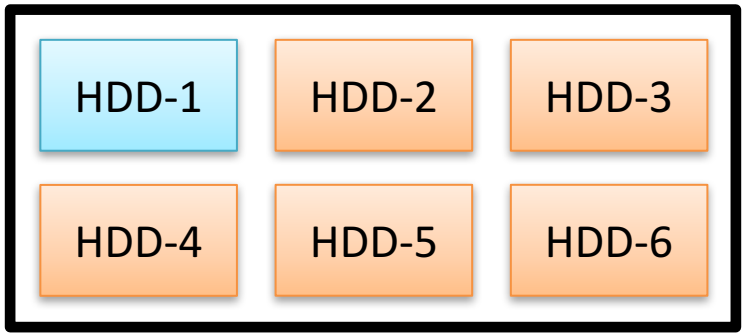
Various options

- Distributed
- Replicated
- Striped



Challenges - Distributed storage

In one cluster node



3 TB HDDs

HDD-1 has a minimal linux (16 GB)

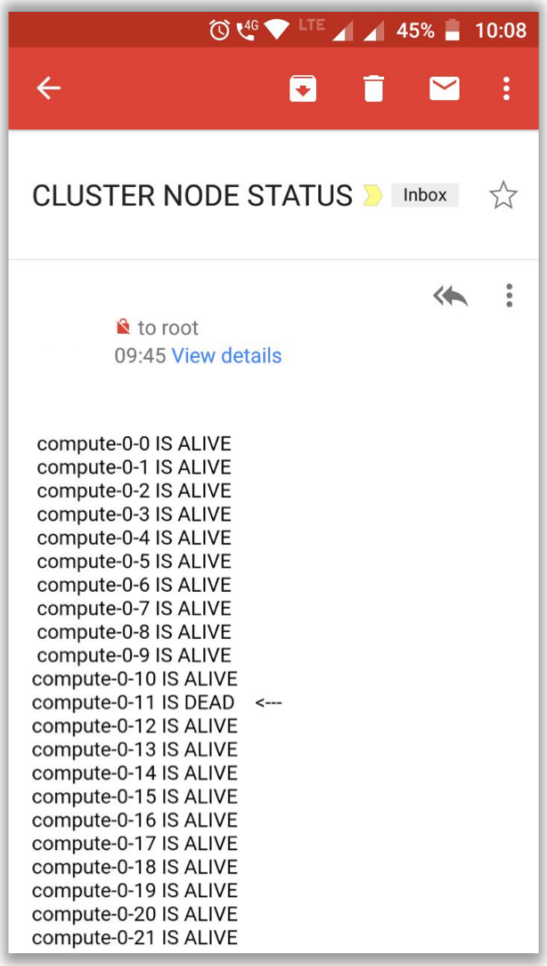
So, 40(20) X 3 TB = 120(60) TB

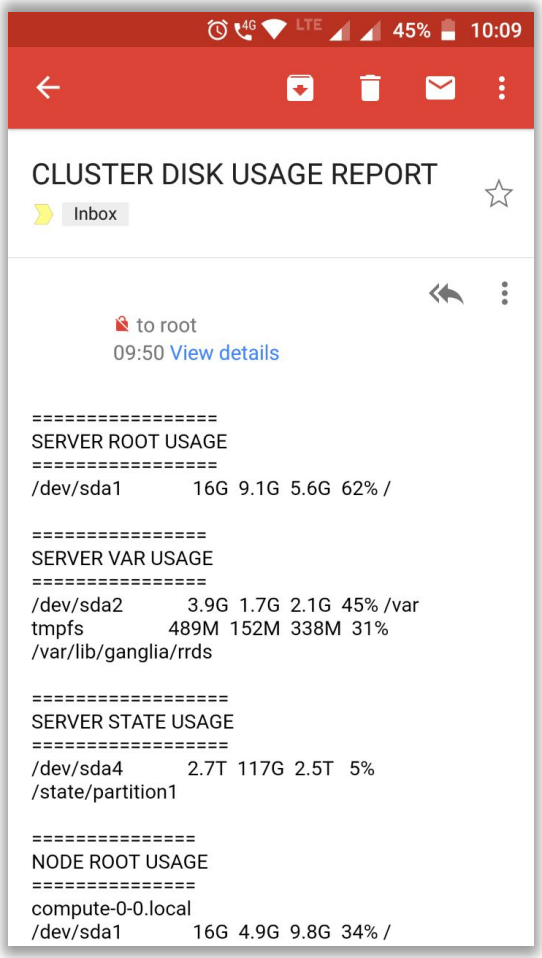
Remaining 5 X 3 TB = 15 TB
Soft-RAID5 4 X 3 TB = 12 TB

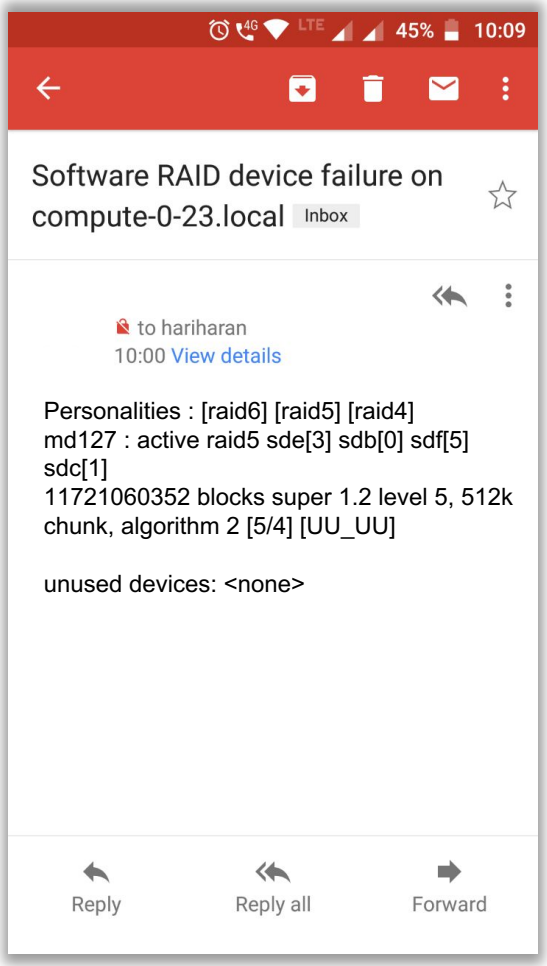
So, 40(20) X 12 TB = 480(240) TB

Monitoring - In-house monitoring

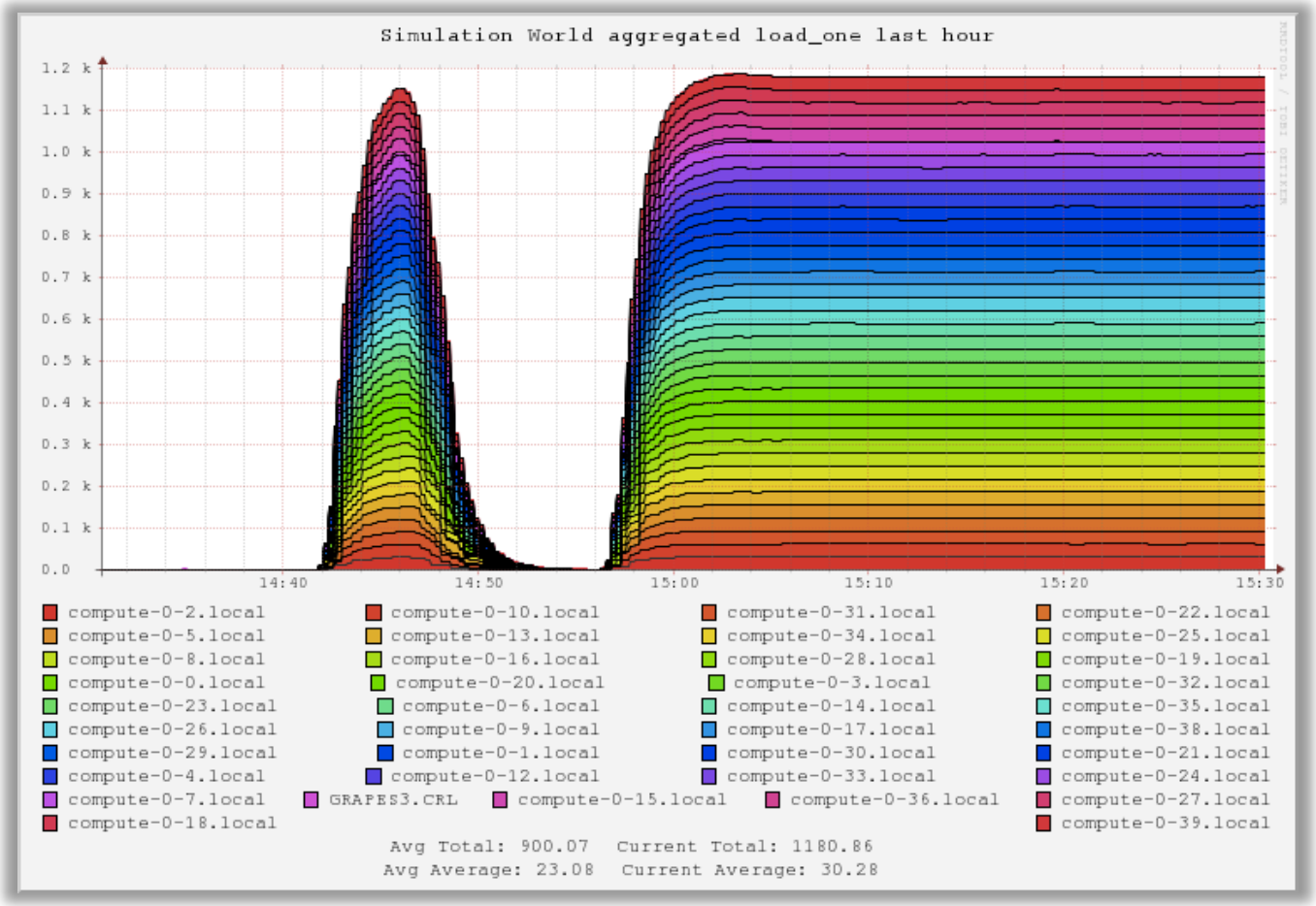
Using only shell commands !!!







Monitoring - Ganglia



At full load

Standard tool to monitor metrics

Maintenance & security

- Maintained by 2 people
- No AMC
- In 2016, operated with ~5% downtime
- In 2017, operated with <1 % downtime

- Registered users from registered IP
- Limited number of logins

Applications - Published

- 2 million jobs executed
- Transient Weakening of Earth's Magnetic Shield Probed by a Cosmic Ray Burst
  (P.K. Mohanty et al., *Phys. Rev. Lett. 117, 171101 (2016)*)
- Was the cosmic ray burst detected by the GRAPES-3 muon telescope on 22 June 2015 caused by a transient weakening of the geomagnetic field or by an interplanetary anisotropy?
  (*P.K. Mohanty et al., Physical Review D 97, 082001 (2018)*)
- More than 50 trials, each 5 hours, few TB
- Hundreds of events in 20 years of data

Applications - Ongoing

- Atmospheric electric field simulations
  - 7.4 X 10^{10} EAS
  - 2 months, 40 TB
- Monte-Carlo to estimate chance coincidence
  - 1 trillion events
  - 1 day
    - 4 years on single CPU
    - 1 year on a quad core machine

Bonus

- Another success story of **SearchEnabler**

  https://thetechpanda.com/2012/07/07/searchenabler-hardware-details-explained-how-they-built-their-own-data-center/

Cost-Benefit Analysis Of Having Our Own Infrastructure

We have to do a lot of web crawling and data processing to provide metrics and analytics to our customers. We need servers and web crawlers that run 24 x 7.

- **Cost Factor** – We explored the cloud services like Amazon EC2 and Microsoft Azure and almost all of them charge based upon the compute cycles. Our web crawlers will be running every second which eats up huge amounts of compute cycles resulting in higher costs.The cost of third party infrastructure increases linearly as we scale higher but it nearly stabilizes if we can build and maintain our own data center.
- **Building Capability** – As we keep on working with our own set of infrastructure, we can come to know and tackle all the possible problems. Typically, it is very hard to shift your setup from a third party infrastructure to your private one. It will also be easier for us to scale when the need to expand our infrastructure arises.

How we built our own data center?

1. Hardware

We designed our data center with the goal of maximum availability using redundancy in just about every thing. So that, in case if some thing goes down, availability should not be an issue.

- **Servers Built Using Commodity Hardware** – All our servers use desktop based components such as Intel Core i3 processors, 16 GB of RAM and 3 Tera Bytes of storage space in each server. We have used multiple hard disk drives, Ethernet cards, Routers and Switches in our hardware setup for maximum availability.

Thanks

Backup Slides

Top 10

Secure | https://en.wikipedia.org/wiki/TOP500#Top_10_ranking

Top 10 ranking [edit]

Top 10 positions of the 50th TOP500 in November 2017[15]

Rank	Rmax Rpeak (PFLOPS)	Name	Model	Processor	Interconnect	Vendor	Site country, year	Operating system
1	93.015 125.436	Sunway TaihuLight	Sunway MPP	SW26010	Sunway[16]	NRCPC	National Supercomputing Center in Wuxi China, 2016[16]	Linux (Raise)
2	33.863 54.902	Tianhe-2	TH-IVB-FEP	Xeon E5–2692, Xeon Phi 31S1P	TH Express-2	NUDT	National Supercomputing Center in Guangzhou China, 2013	Linux (Kylin)
3	19.590 25.326	Piz Daint	Cray XC50	Xeon E5-2690v3, Tesla P100	Aries	Cray	Swiss National Supercomputing Centre Switzerland, 2016	Linux (CLE)
4	19.136 28.192	Gyoukou	ZettaScaler-2.2 HPC system	Xeon D-1571, PEZY-SC2	Infiniband EDR	ExaScaler	Japan Agency for Marine-Earth Science and Technology Japan, 2017	Linux (CentOS)
5	17.590 27.113	Titan	Cray XK7	Opteron 6274, Tesla K20X	Gemini	Cray	Oak Ridge National Laboratory United States, 2012	Linux (CLE, SLES based)
6	17.173 20.133	Sequoia	Blue Gene/Q	A2	Custom	IBM	Lawrence Livermore National Laboratory United States, 2013	Linux (RHEL and CNK)
7	14.137 43.902	Trinity	Cray XC40	Xeon E5–2698v3, Xeon Phi	Aries	Cray	Los Alamos National Laboratory United States, 2015	Linux (CLE)
8	14.015 27.881	Cori	Cray XC40	Xeon Phi 7250	Aries	Cray	National Energy Research Scientific Computing Center United States, 2016	Linux (CLE)
9	13.555 24.914	Oakforest-PACS	Fujitsu	Xeon Phi 7250	Intel Omni-Path	Fujitsu	Kashiwa, Joint Center for Advanced High Performance Computing Japan, 2016	Linux
10	10.510 11.280	K computer	Fujitsu	SPARC64 VIIIfx	Tofu	Fujitsu	Riken, Advanced Institute for Computational Science (AICS) Japan, 2011	Linux

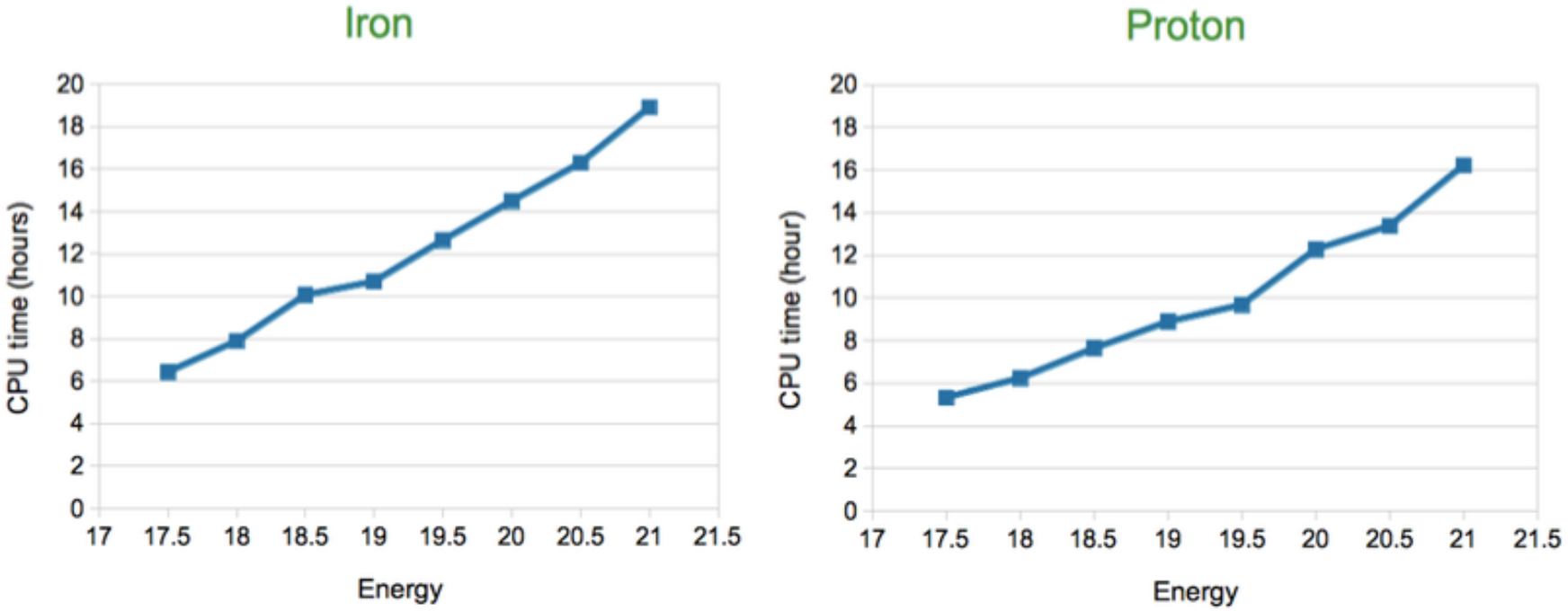
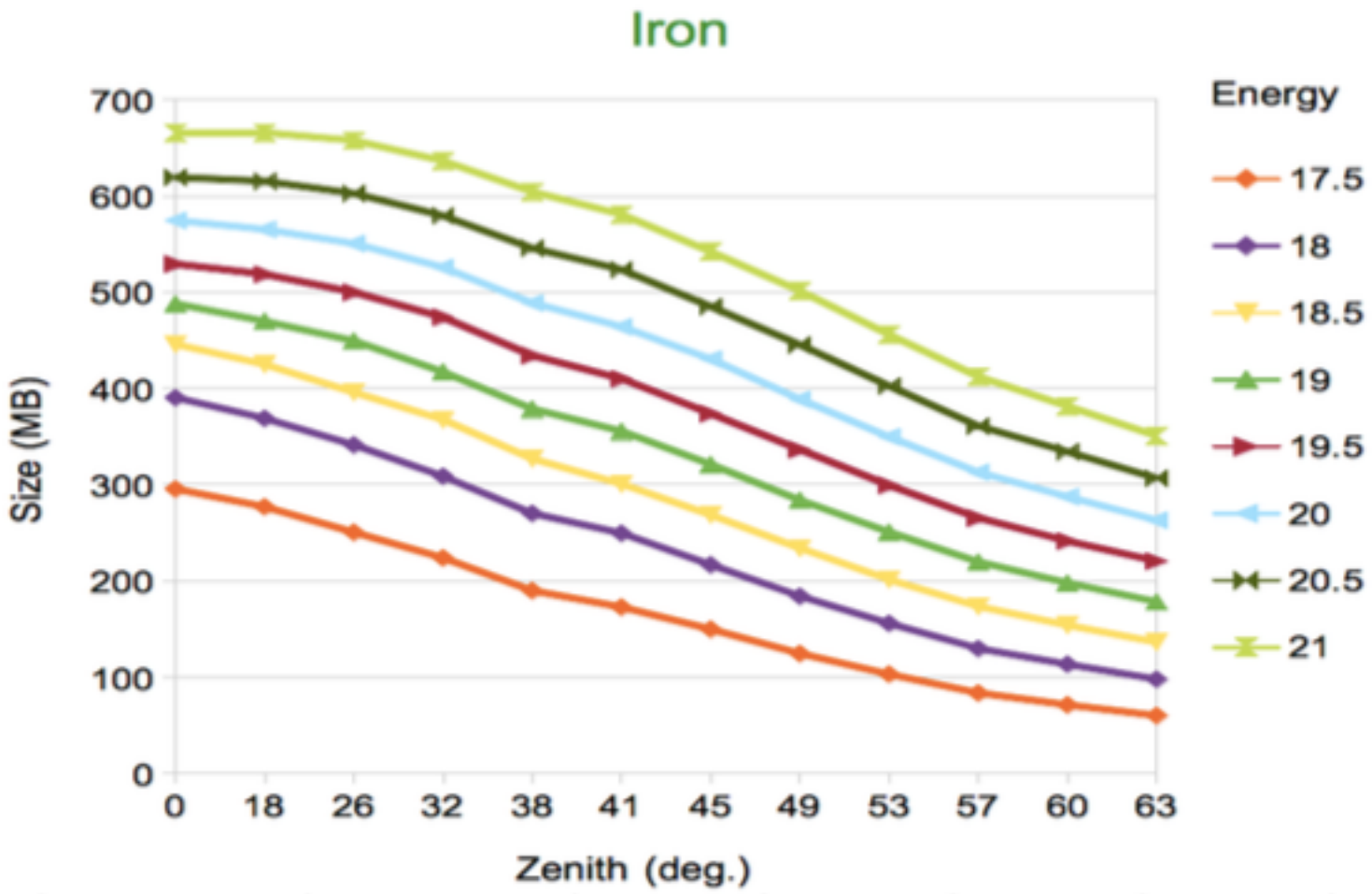


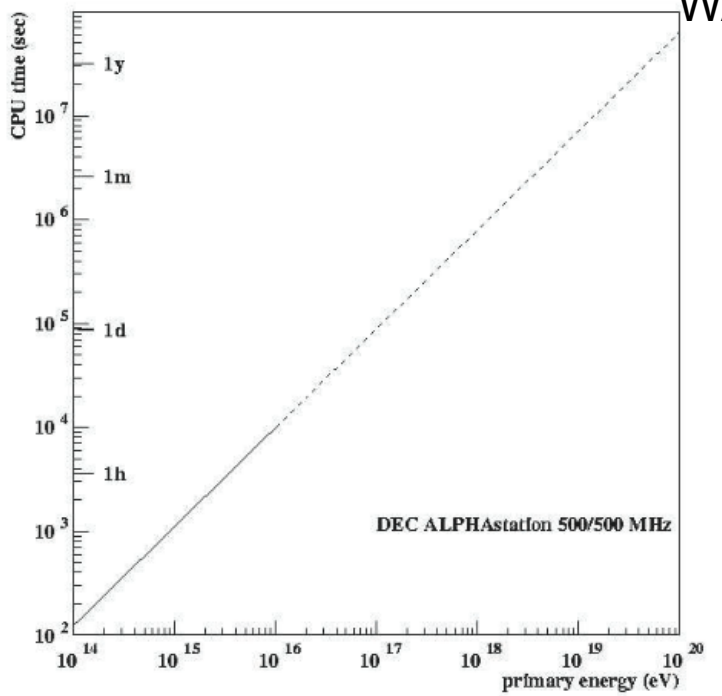
Figure 2. Left: CORSIKA average execution time as a function of the logarithm of the energy for iron primary cosmic rays. Right: same plot for proton cosmic rays

doi:10.1088/1742-6596/368/1/012015



doi:10.1088/1742-6596/368/1/012015

michael.wommer@kit.edu
WAPP Darjeeling, 2009-12-16



- storage amount and cpu time increase linearly with primary energy
- one single $10^{20} eV$ shower needs $\approx$1 year CPU time, several TB storage amount