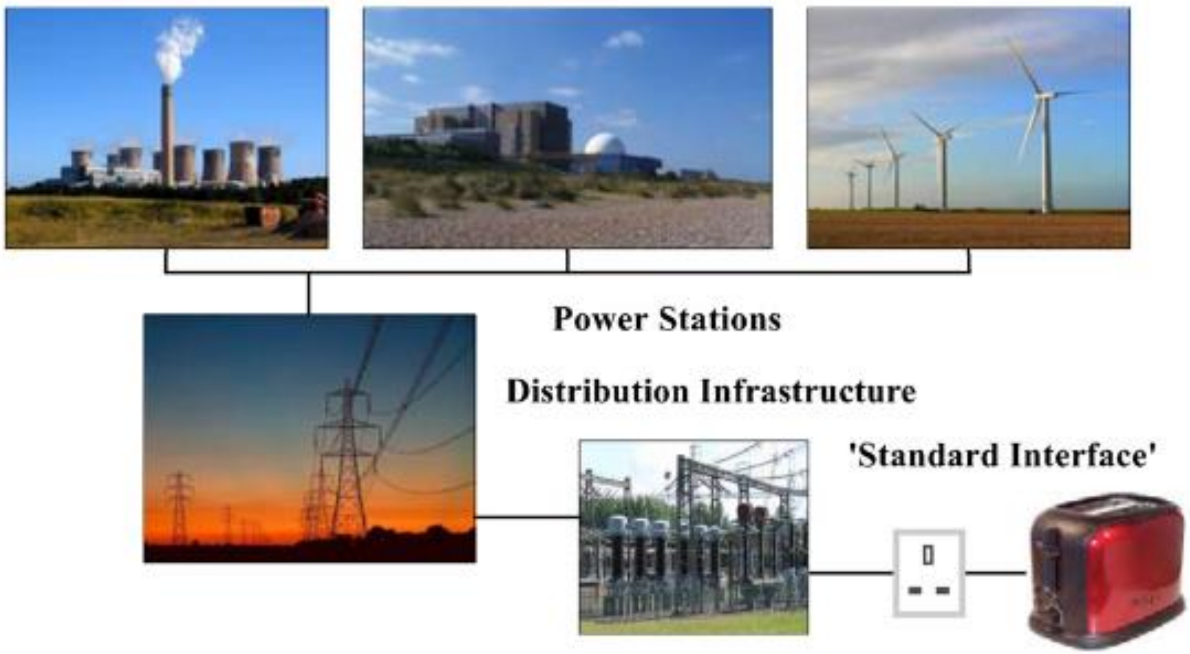
Tata Institute of Fundamental Research
टाटा मूलभूत अनुसंधान संस्थान

Grid computing at LHC and CMS Tier-II centre at TIFR



Brij Kishor Jashal
Email – brij.jashal@tifr.res.in

Outline:

- Grid computing
- Architecture overview
- Grid middleware
- Grid networking
- CMS data model
- T2_IN_TIFR

  --Resources

  --Site performance and status

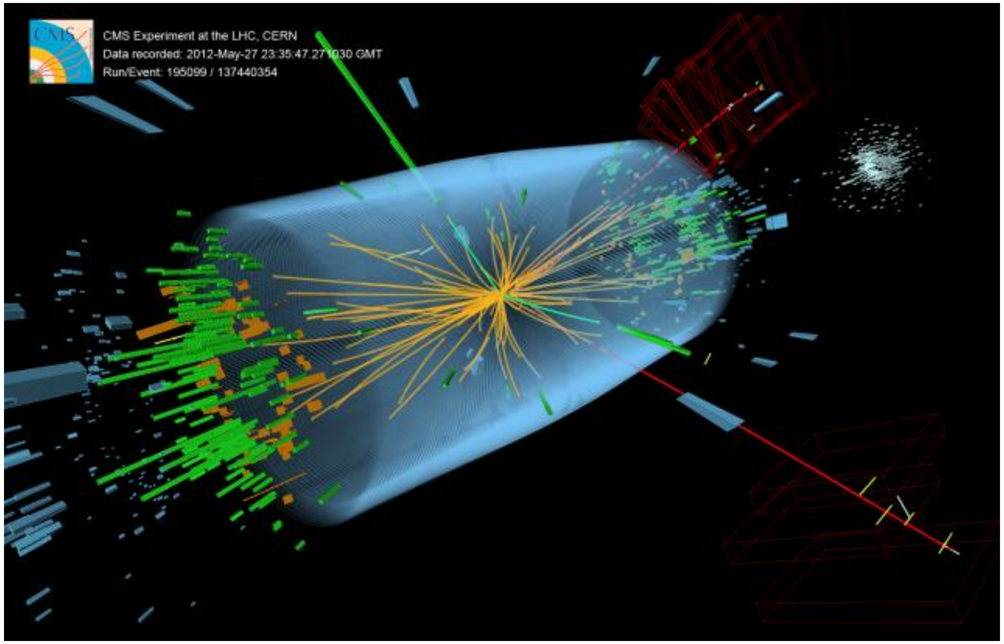
  --Recent upgrades

  --Future

- Preparation for Run2

Scale of LHC computing

- Higgs event in CMS: 2012
- Nobel prize in Physics 2013
- Made possible by grid computing



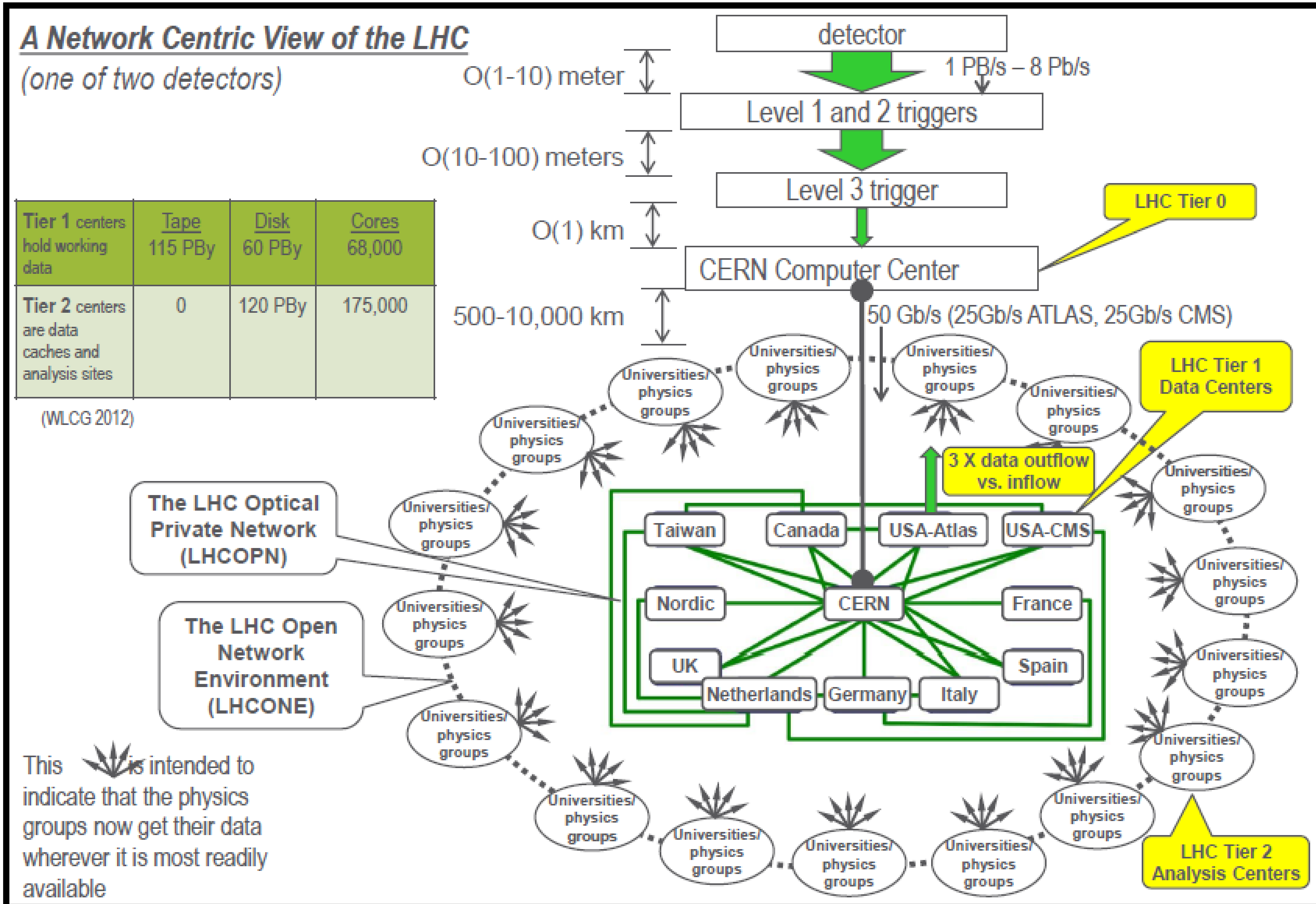
1 Higgs event out of 10^{12} proton – proton collisions

- CMS designed to observe a billion (1×10^9) collisions/sec.
- Data rate out of the detector of more than 1,000,000 Gigabytes/sec (1 PBy/s)
- Compression techniques reduce the output data rate to about 25Gb/s that must be transported, managed and analyzed to extract the science.
- 50 Gb/s for 7x24 is distributed to physics groups around the world
- Around the world 6000 people from 50 countries

Scale of LHC computing

CERN – TIFR Latency – 160 ms at present

Scale of LHC computing



	Tape	Disk	Cores
Tier 1 centers hold working data	115 PBy	60 PBy	68,000
Tier 2 centers are data caches and analysis sites	0	120 PBy	175,000

(WLCG 2012)

Computing characteristics at LHC

- Large numbers of independent events (millions/sec) – "Job granularity"
- Large data sets – mostly read-only
- Modest I/O rates – few MB/sec per processor
- Modest floating point requirement – HEP-SPEC06 performance. (which matches with batch jobs ~10%)

Computation and storage needs can not be met at single site.

Therefore

- Scaling up is complex once you exceed the capabilities of single geographical installation

High Performance computing

- HPC systems tend to focus on tightly coupled parallel jobs, and as such they must execute within a particular site with low-latency interconnects
- Granularity largely defined by the algorithm
- Hard to schedule different workloads
- Reliability and speed is very important
- Achieved by super computers

High Throughput computing

- HTC systems are independent, sequential jobs that can be individually scheduled on many different computing resources across multiple administrative boundaries
- Granularity can be selected to fit the environment
- Mixing workload is easy
- Sustained throughput is the key goal
- **Achieved by Grid computing technology**

The Promise of Grid Technology (for the user)

- Submit your computing task
  - and the Grid ….
    - Finds convenient place for the Jobs/calculation to run
    - Optimizes use of the widely dispersed resources
    - Organizes efficient access to your data
      - Data placement, migration, replication, caching
    - Deals with authentication and security
    - Interfaces to the local site resources
    - Runs your jobs
    - Monitors progress
    - Recovers from problems

    ✓ …… and ……………Tells you when your work is complete.

Grid Computing

"Coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organization""

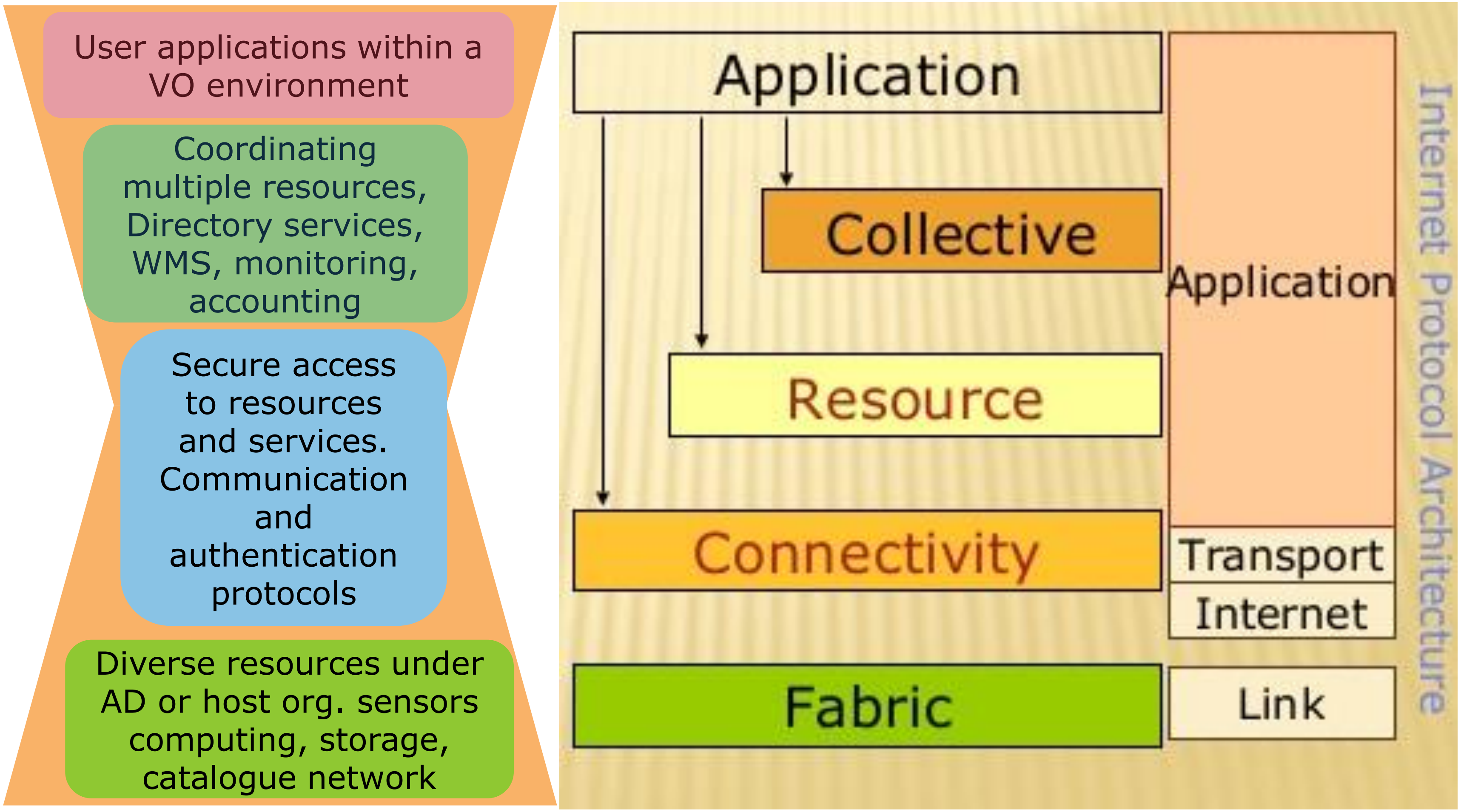
- *Coordinates resources that are not subject to centralized control …*
- *…. Using standard, open, general-purpose protocols and interfaces but still "standard" (* allows dynamic resource sharing *)*
- *…… to deliver **nontrivial** qualities of services*

Why ?

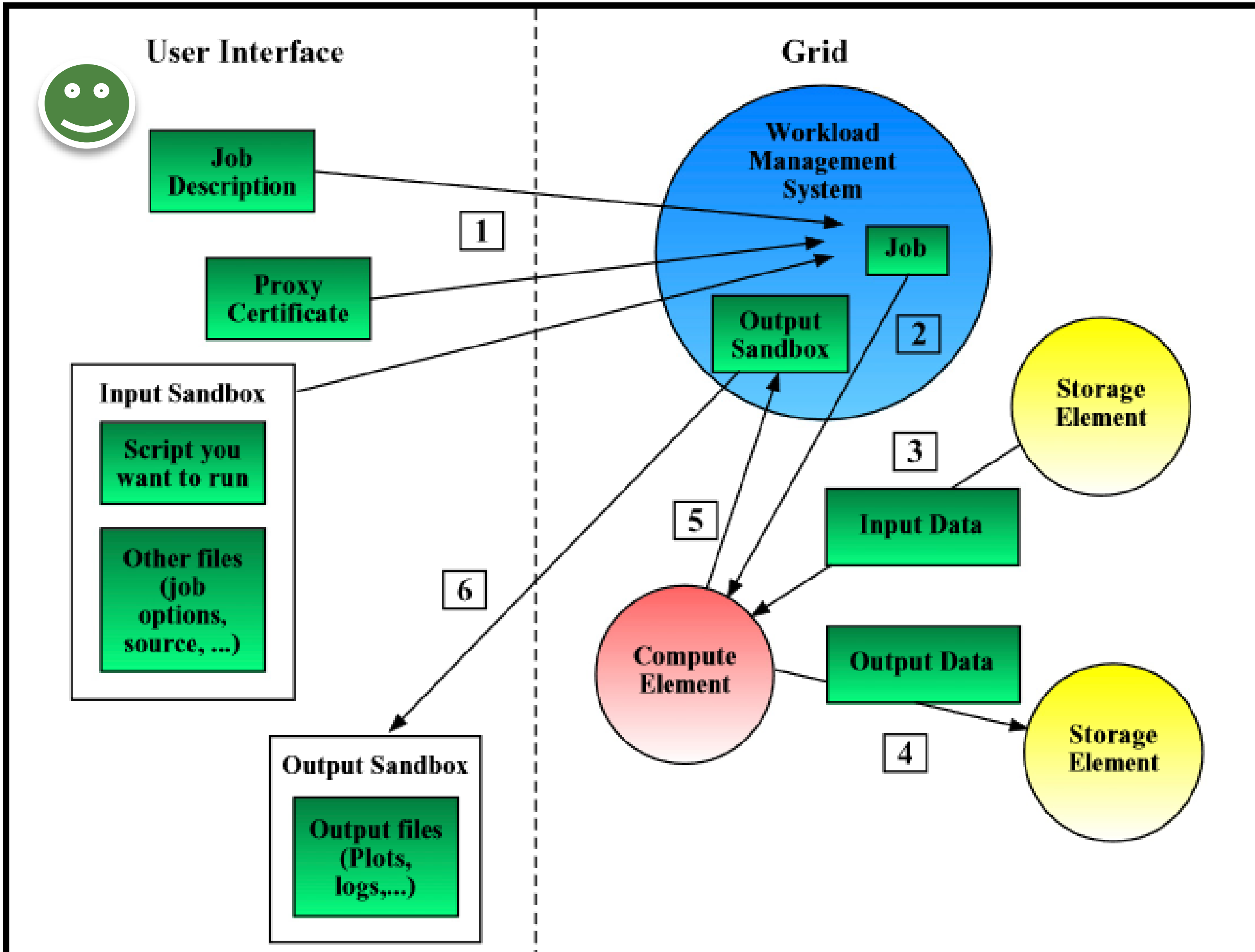
*… **So that the utility of the combined system is significantly greater than that of the sum of its parts***

Source: Ian Foster
http://dlib.cs.odu.edu/WhatIsTheGrid.pdf

Grid architecture overview



Simple job workflow

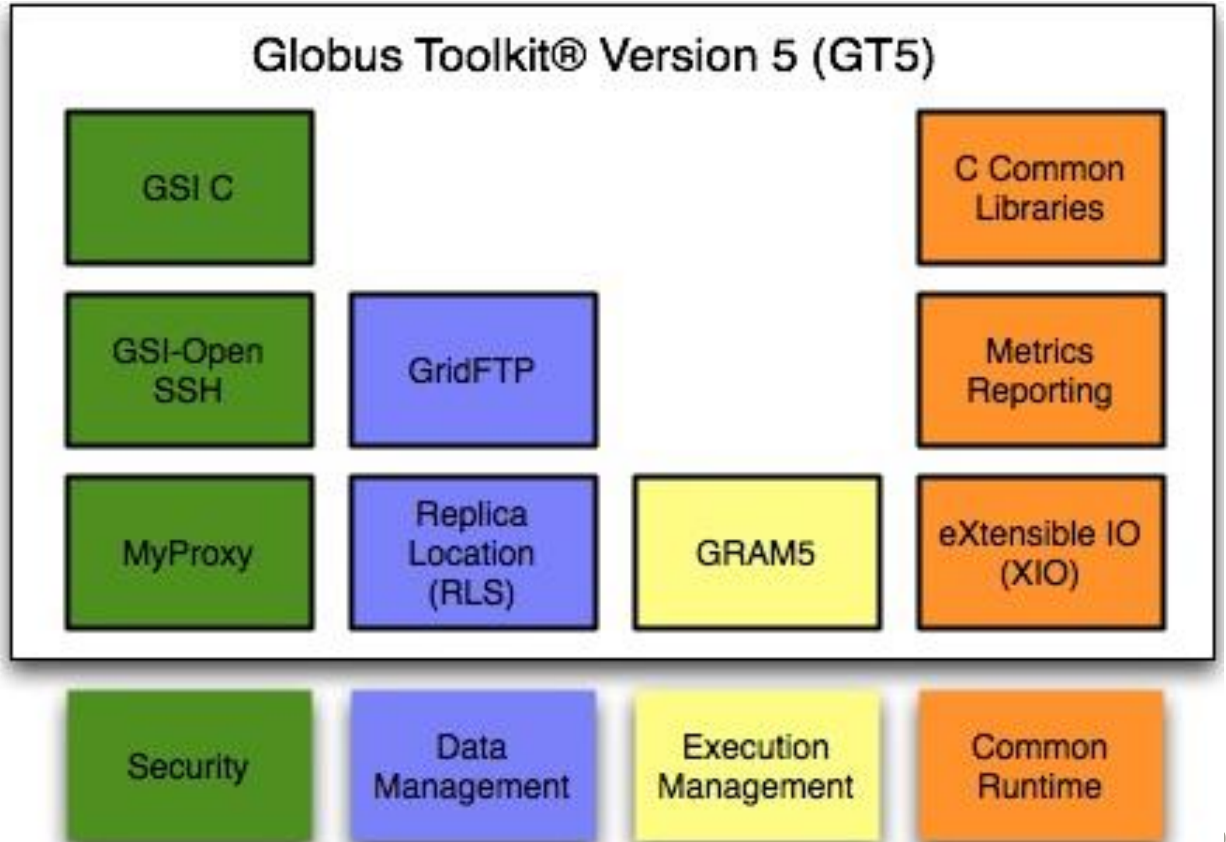


Grid Middleware

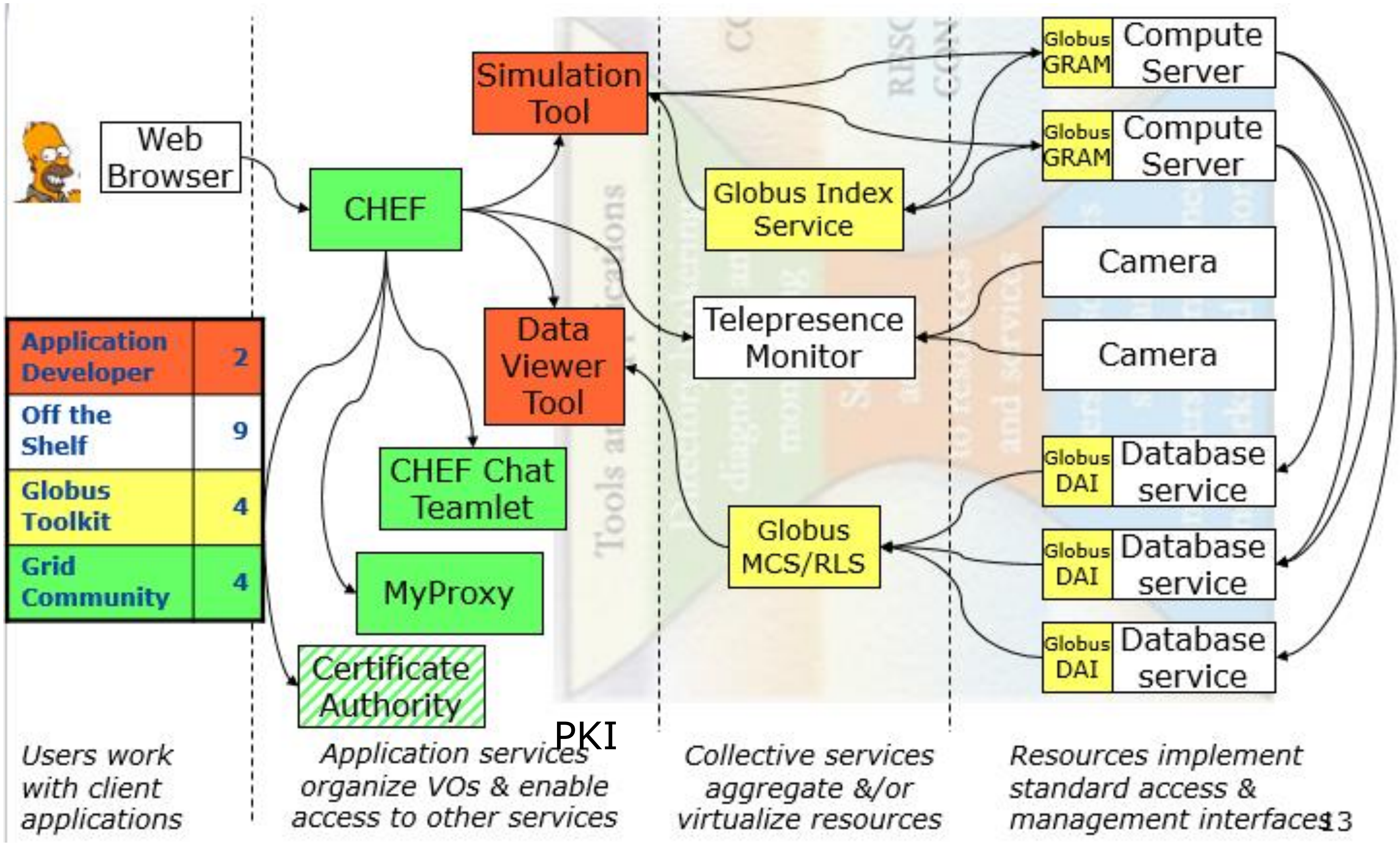
Software infrastructure between OS kernel and user application is considered middleware

"Globus" , The first middleware project in mid 90s started by Ian foster and Karl Keselman

Majority of Grid systems in the world have built upon "Globus toolkit"



Build your own grid



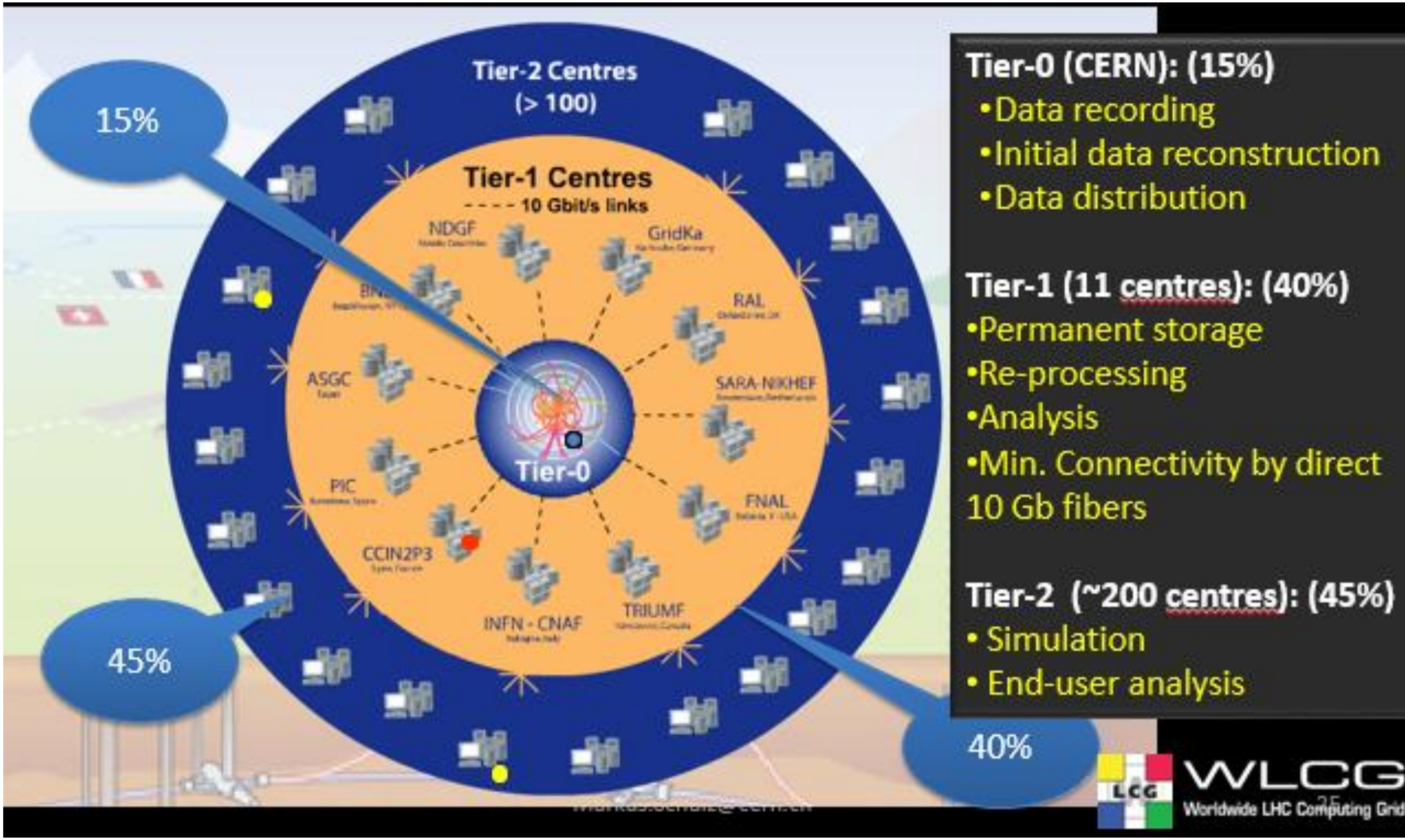
Worldwide LHC Computing Grid (WLCG)

Distributed Computing Infrastructure for LHC experiments

- Connected by High speed wide area networks
- Linking more than 300 computer centers
- Providing > 340,000 cores
- To more than 2000 (active) users
- Archiving 15PB per year

- 1 - T0 @ CERN (For all the LHC experts) (15% of total resources)
- 10 - T1s worldwide
- T2s - TIFR as National Facility CMS T2
- Many… T3s where physicists actually work

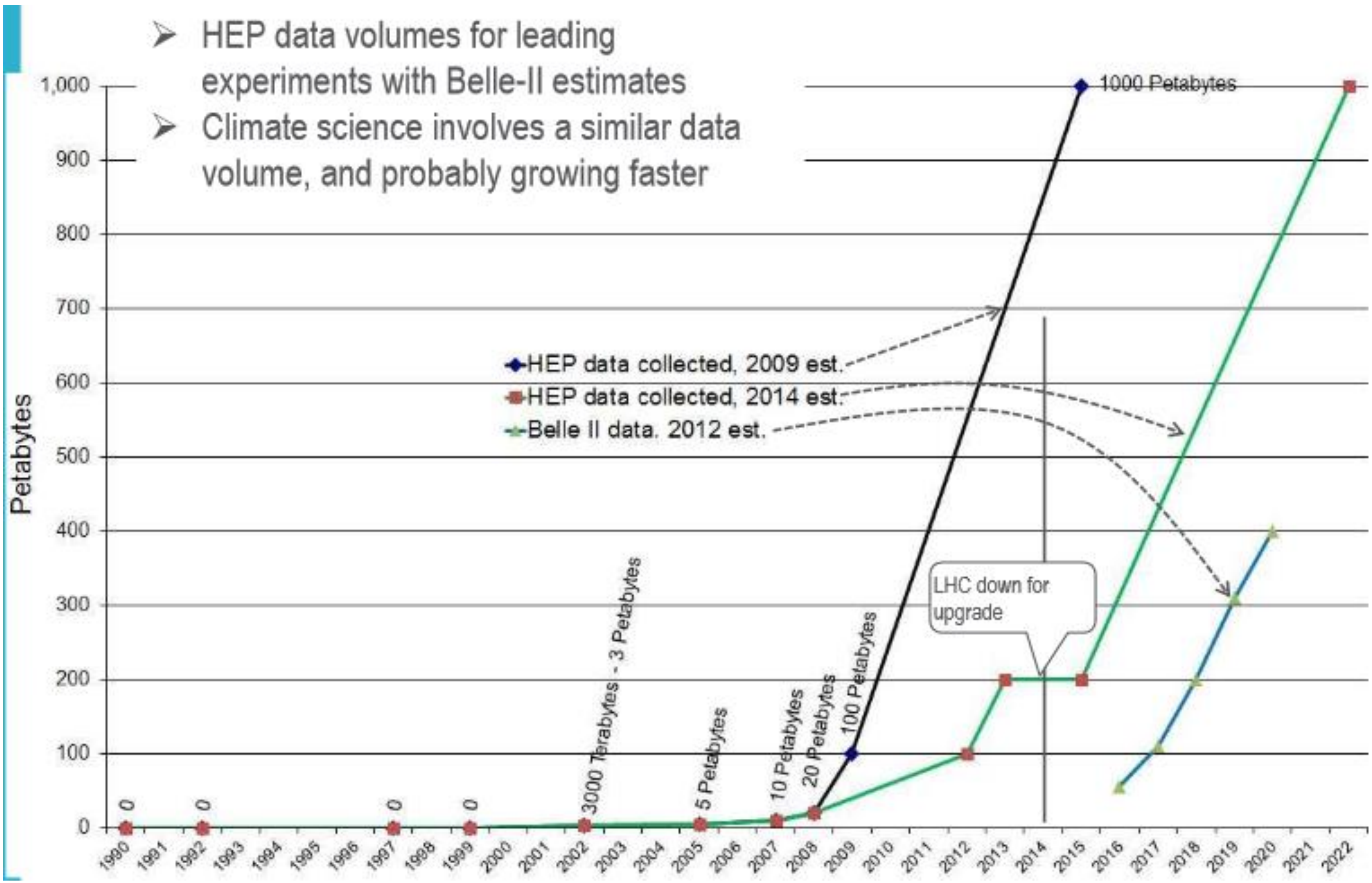
WLCG architecture



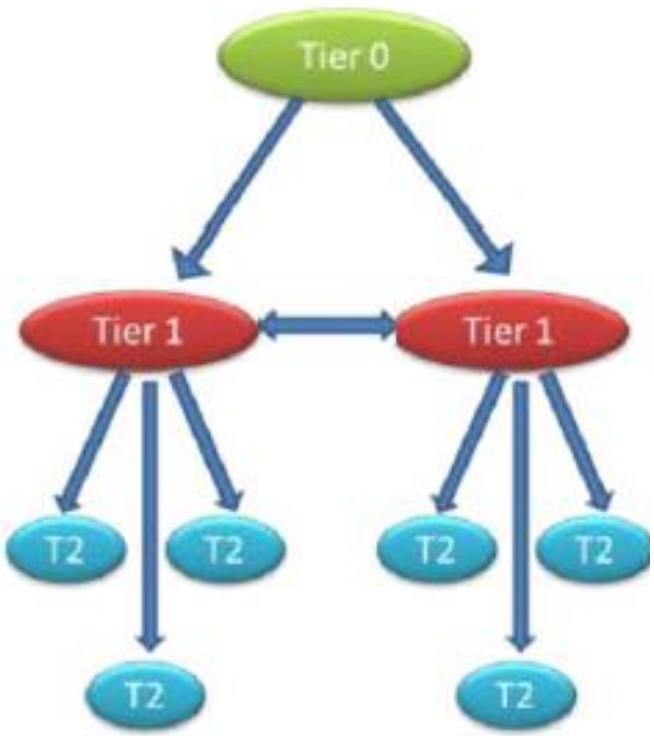
Data growth prediction

LHC data volume is predicted to grow 10 fold over the next 10 years

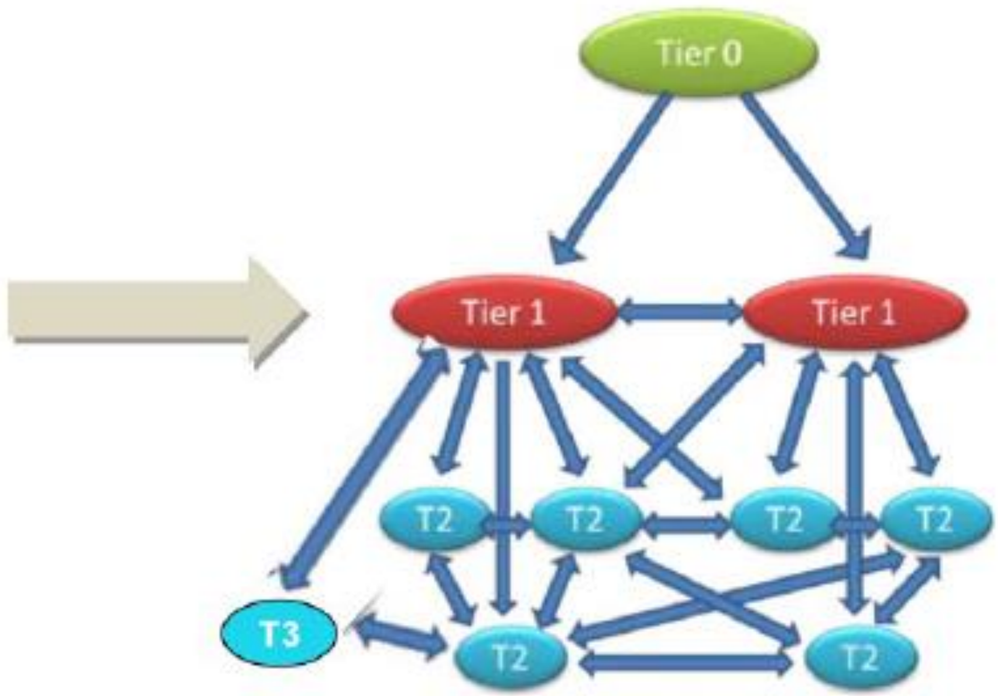
- HEP data volumes for leading experiments with Belle-II estimates
- Climate science involves a similar data volume, and probably growing faster



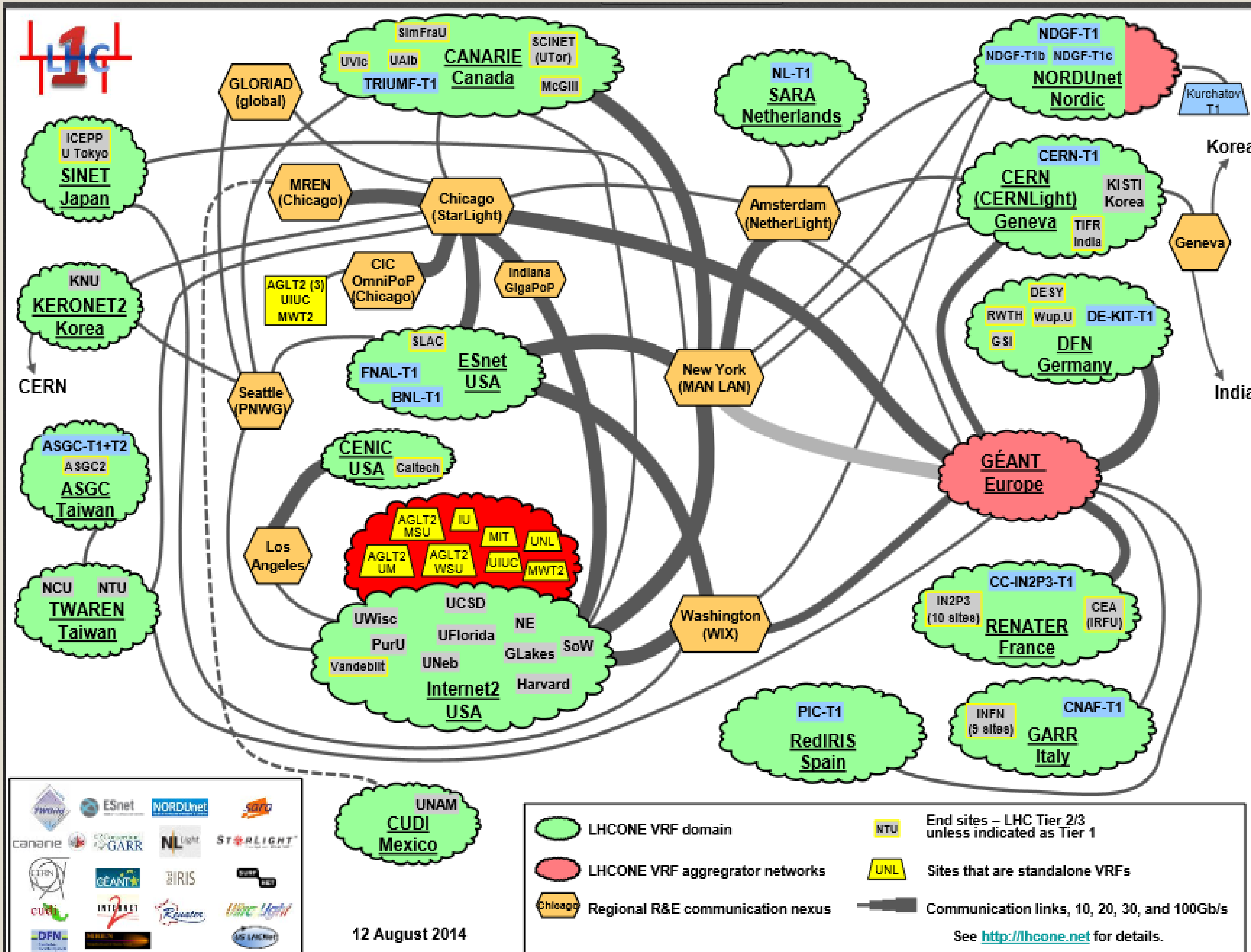
Evolution of data distribution model



Original MONARCH model



Model evolution



LHC 1
CANARIE
Canada
SimFraU
UVic
UAlb
TRIUMF-T1
SCINET (UTor)
McGill
GLORIAD (global)
NL-T1
SARA
Netherlands
NDGF-T1
NDGF-T1b
NDGF-T1c
NORDUnet
Nordic
Kurchatov T1
Korea
ICEPP U Tokyo
SINET
Japan
MREN (Chicago)
Chicago (StarLight)
Amsterdam (NetherLight)
CERN-T1
CERN (CERNLight) Geneva
KISTI Korea
TIFR India
Geneva
KNU
KERONET2
Korea
CERN
AGLT2 (3) UIUC MWT2
CIC OmniPoP (Chicago)
Indiana GigaPoP
DESY
RWTH
Wup.U
DE-KIT-T1
GSI
DFN
Germany
India
SLAC
FNAL-T1
BNL-T1
ESnet
USA
Seattle (PNWG)
New York (MAN LAN)
ASGC-T1+T2
ASGC2
ASGC
Taiwan
CENIC
USA
Caltech
GÉANT
Europe
Los Angeles
AGLT2 MSU
IU
MIT
UNL
AGLT2 UM
AGLT2 WSU
UIUC
MWT2
NCU
NTU
TWAREN
Taiwan
UCSD
UWisc
NE
UFlorida
PurU
SoW
GLakes
Vandeblit
UNeb
Harvard
Internet2
USA
Washington (WIX)
CC-IN2P3-T1
IN2P3 (10 sites)
CEA (IRFU)
RENATER
France
PIC-T1
RedIRIS
Spain
CNAF-T1
INFN (9 sites)
GARR
Italy
UNAM
CUDI
Mexico
12 August 2014
LHCONE VRF domain
LHCONE VRF aggregator networks
Regional R&E communication nexus
NTU
End sites – LHC Tier 2/3 unless indicated as Tier 1
UNL
Sites that are standalone VRFs
Communication links, 10, 20, 30, and 100Gb/s
See http://lhcone.net for details.

What has made all that possible ?

(**Optical network technology**)

Dense wave division multiplexing(DWDM) 100Gb/s per wave (optical channel)

- Transport using dual polarization–quadrature phase shift keying (DP-QPSK) technology with coherent detection
  - two independent optical signals, same frequency
  - two polarization
- Together DP and QPSK reduce required rate by factor of 4
  - Allows 100G payload(plus overhead) to fit into the spectrum

Over simplification of the optical technology involved

Data transport

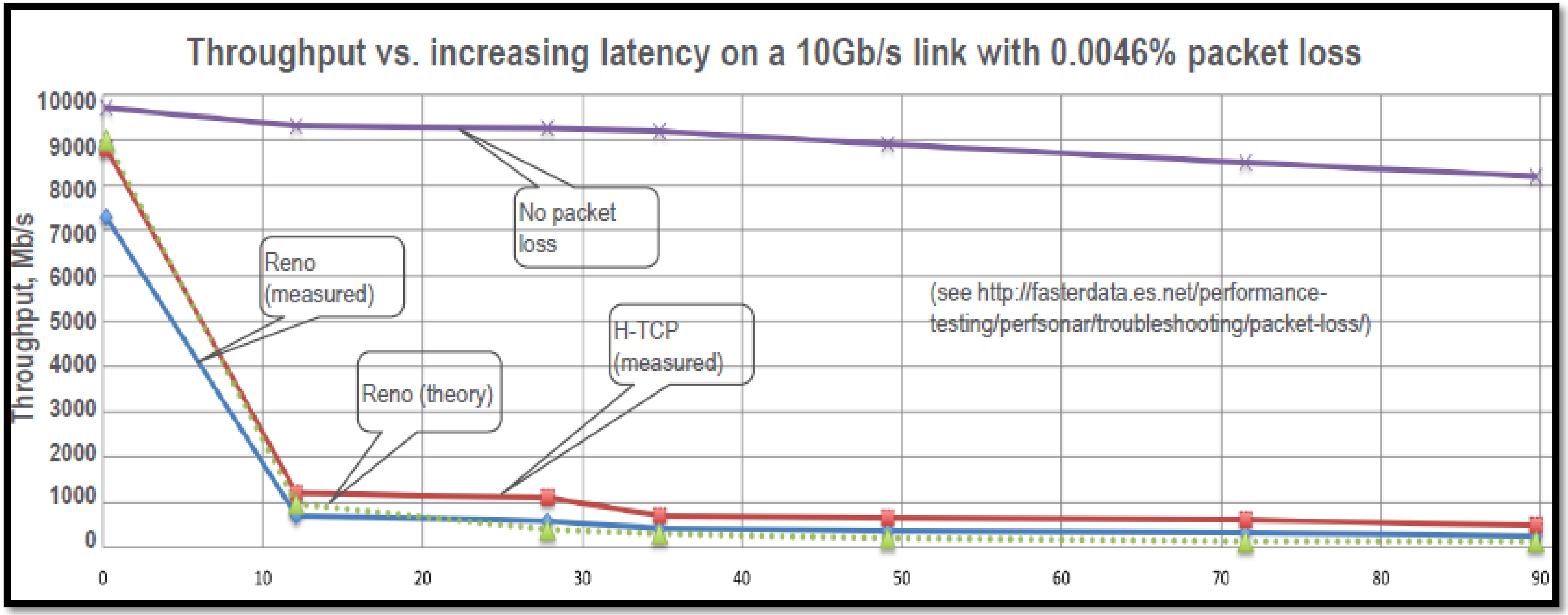
TCP remains the workhorse of the internet, including for data-intensive science

- Very sensitive to packet loss (due to bit errors)
- A single bit error can cause the loss of 1-9 kBy packets (depending on the MTU size) significantly reducing throughput

- **Reason ?**
  - Congestion avoidance algorithms added to TCP
  - Packet loss is seen by TCP's congestion control algorithms as evidence of congestion
  - Network link errors also cause of packet loss

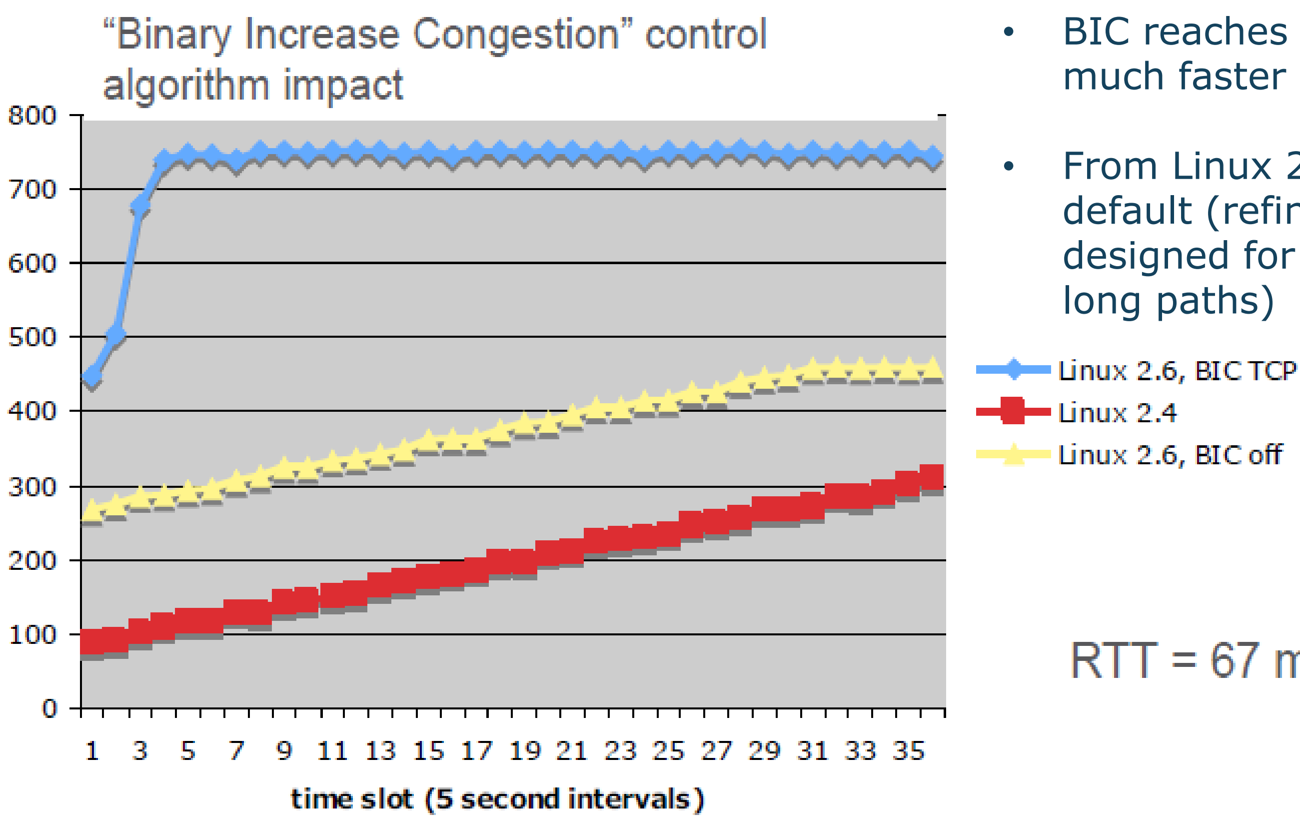
Impact of packet loss on TCP

- On 10 Gb/s LAN path the impact of packet loss is minimum
- On a 10Gb/s WAN path the impact of even very loq packet loss rate is enormous (~80X throughput reduction from TIFR to FNAL where latency is about 270ms)



Modern TCP stack

- Modern TCP stack – Kernel implementation of TCP protocol
- Important to reduce sensitivity to packet loss while still providing congestion avoidance



- BIC reaches Max throughput much faster then older algos.
- From Linux 2.6.19 CUBIC is default (refined version of BIC designed for high bandwidth, long paths)

RTT = 67 ms

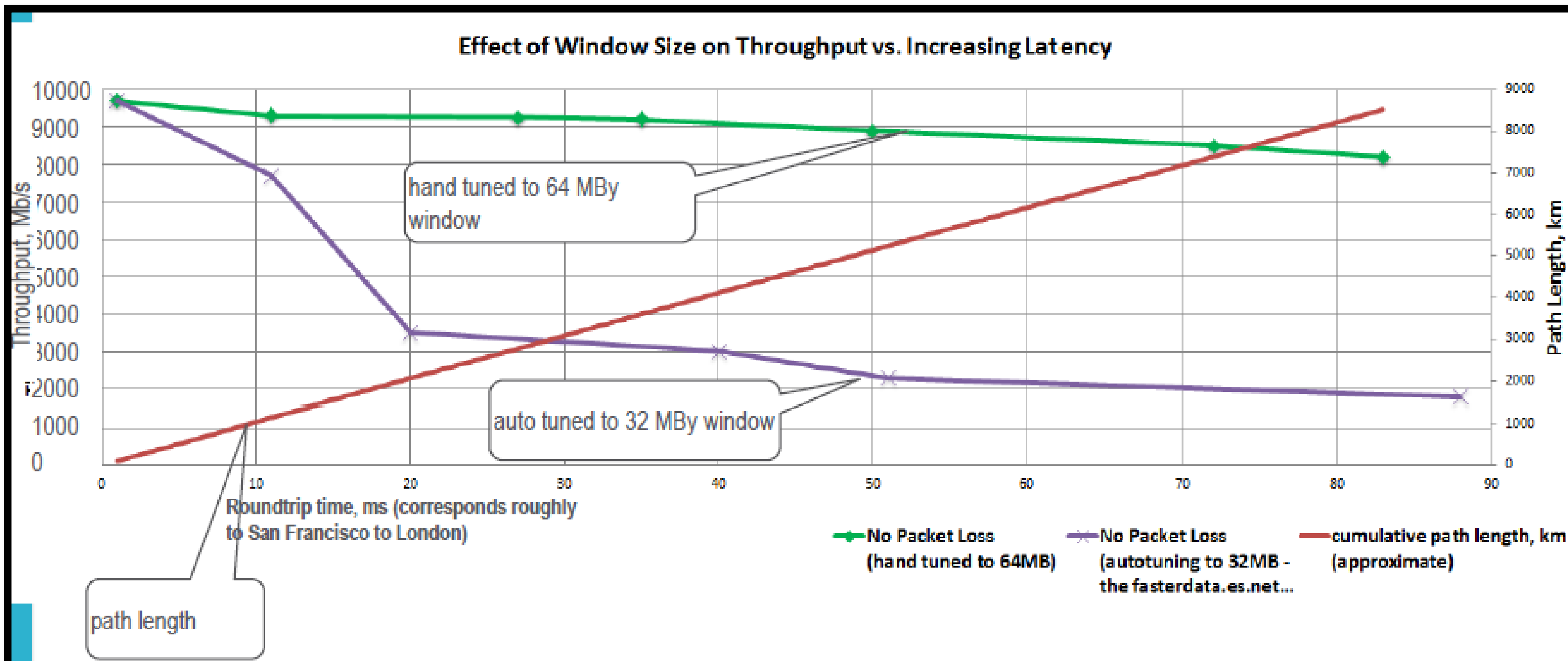
http://www.slac.stanford.edu/~ytl/thesis.pdf

System optimization for high speed transfers.

Efficiency of data movement also depends upon

Host tuning:

- Critical to use optimal windowing buffer size
- Default TCP buffer too small for todays high speed networks (64KB)
- Auto-tuning of parameters not adequate



Network Monitoring and testing

The only way to keep multi-domain, international scale networks error-free is to test and monitor continuously end-to-end to detect soft errors and facilitate their isolation and correction

PerfSONAR:

- Community efforts
- Standardize measurement
- Bundled package

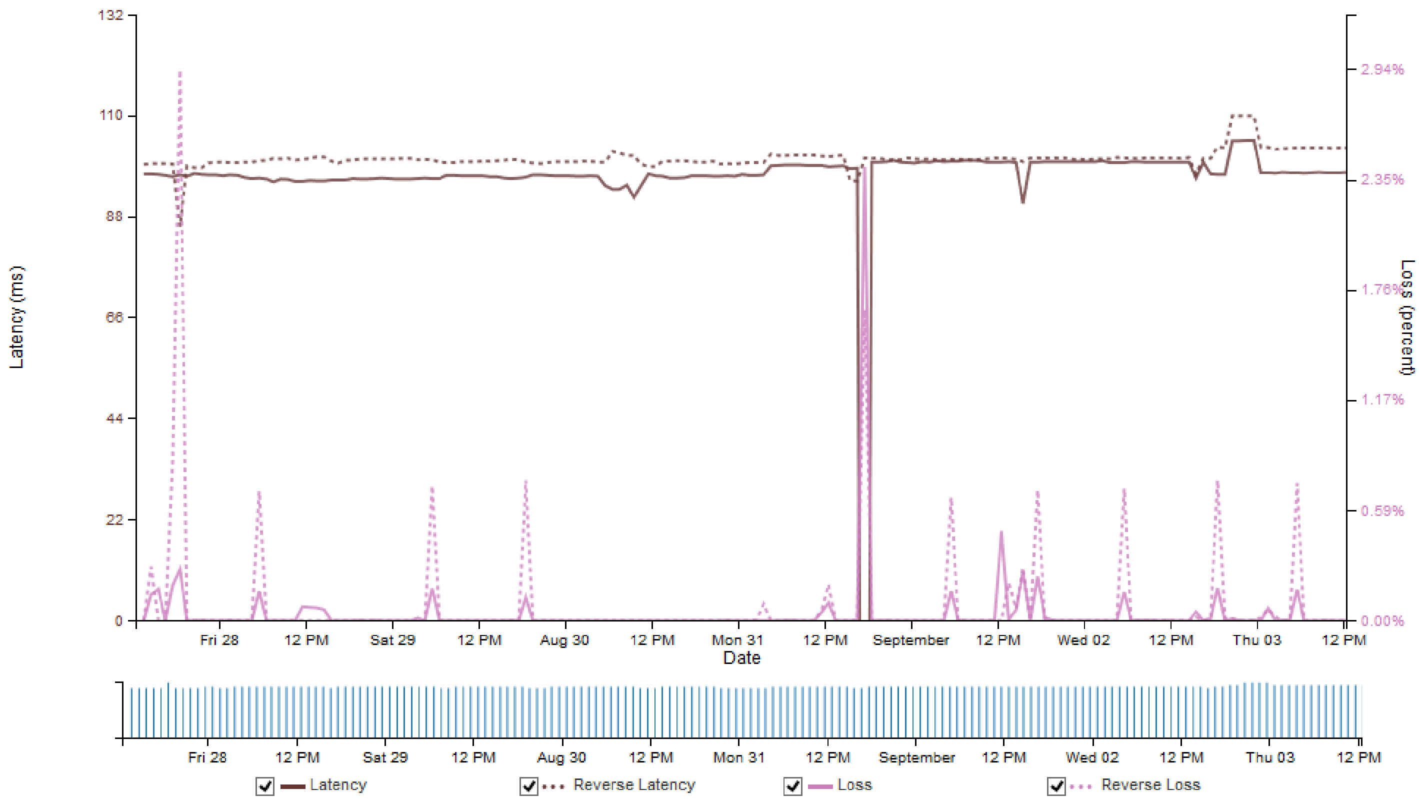
Composite view of health of LHC peers from PerfSONAR at TIFR

PerfSONAR provides a standardize way

- Test, Measure, Export
- Catalogue
- Access performance data from many different networks domain.

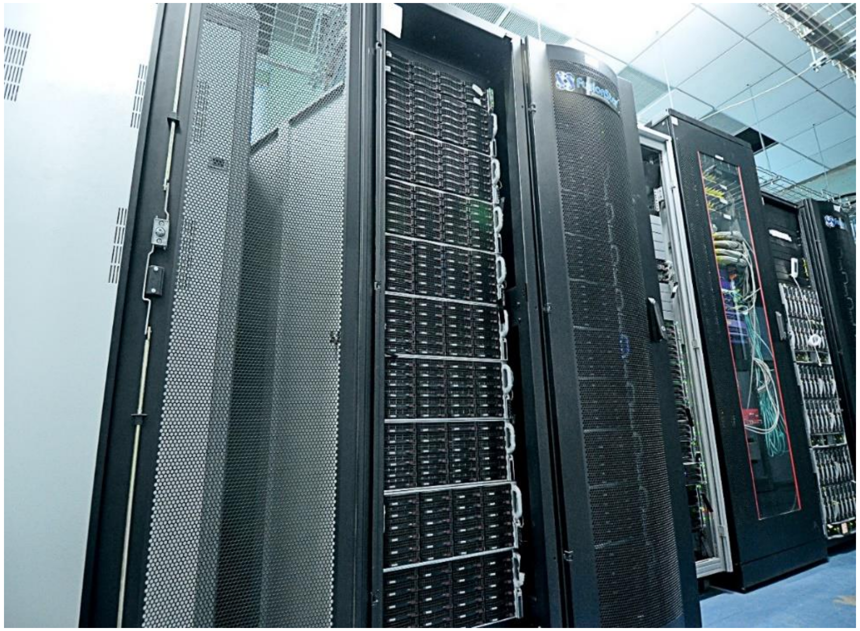
Deployed extensively throughout WLCG

- More then 1100 perfSONAR boxes deployed around the world



Latency statistics from PerfSONAR at TIFR

Grid computing facility at TIFR for CMS T2_IN_TIFR



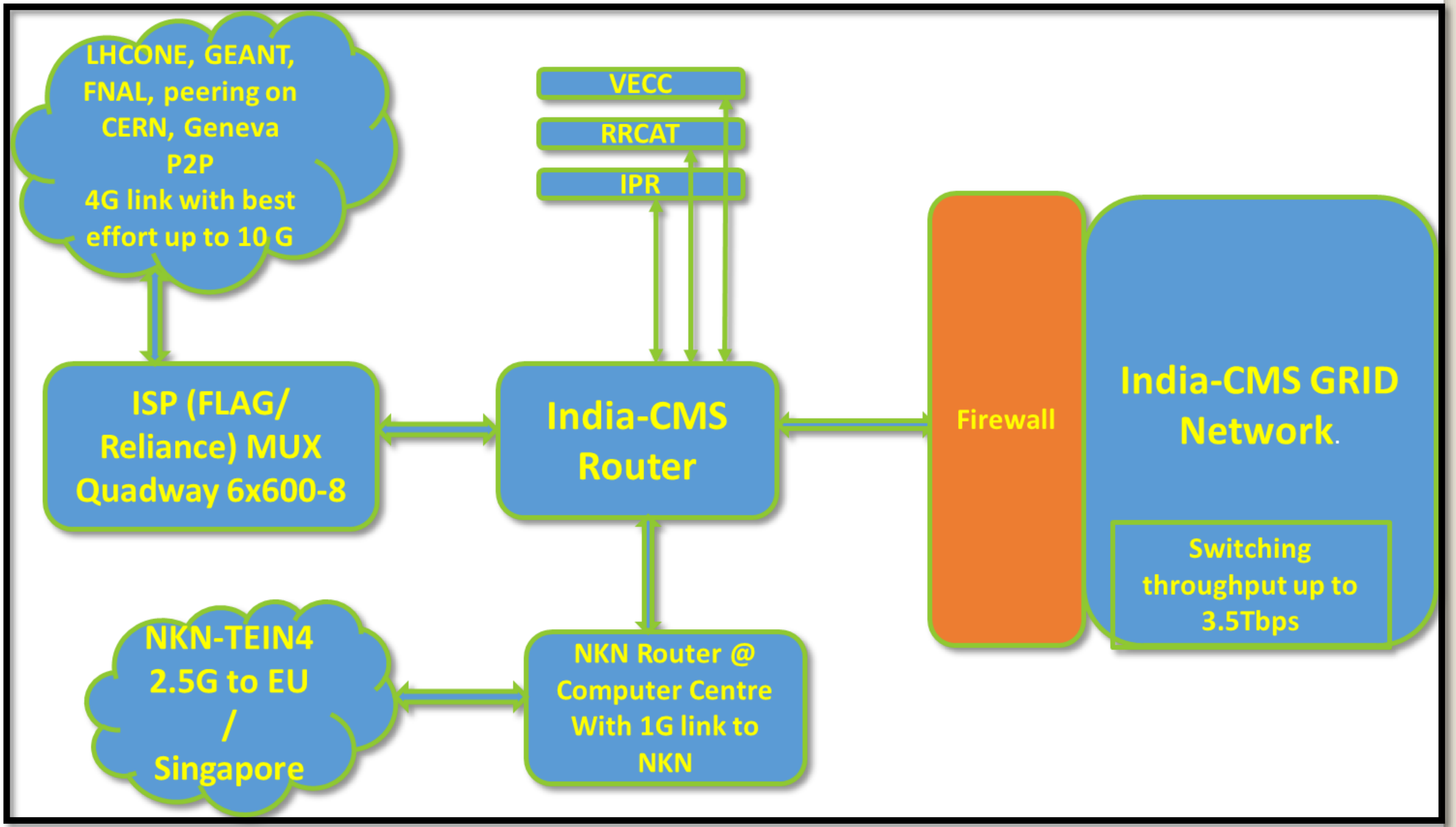
Computing

- Total no of physical cores 1024, Total average of runs executed on a machine (Special Performance Evaluation for HEP code) i.e HEP-SPEC06 is 7218.12

Storage

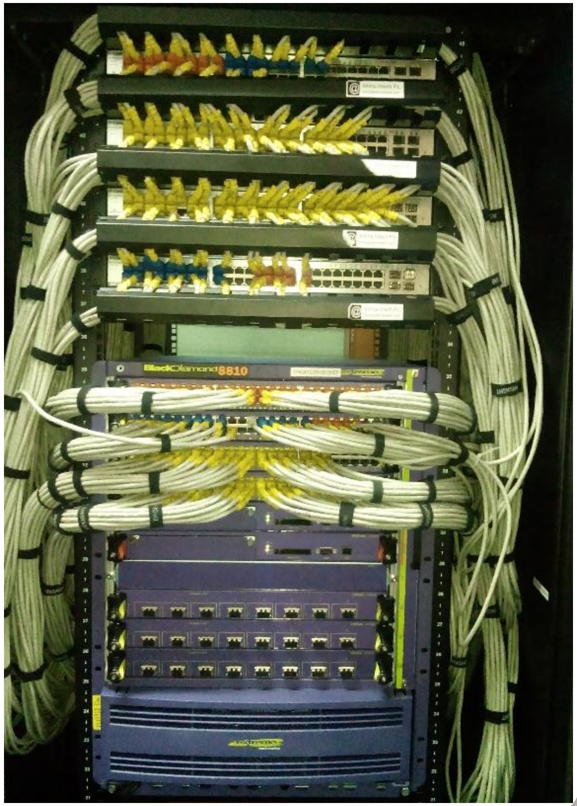
- Total Storage capacity of 28 DPM Disk Nodes is aggregated to more than 1PB (1020 TB)

Network at T2_IN_TIFR



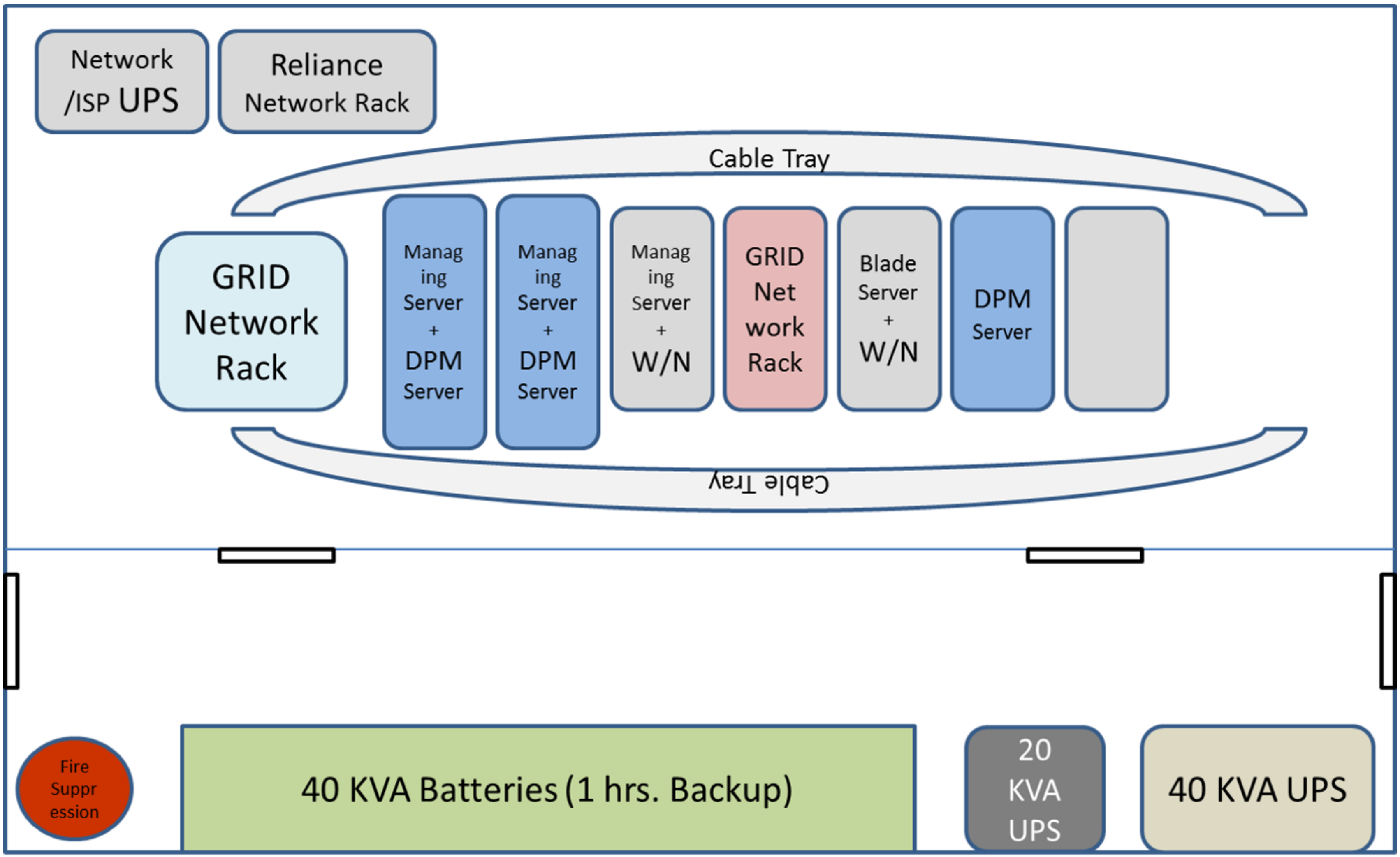
Network at T2_IN_TIFR ….. cont

- Dedicated P2P link to LHCONE, 4 G guaranteed with best effort up to 10 G.
- Planned 10G dedicated to CERN in the same budget
  - TIFR core network capable of switching throughput of upto 3.5 Tbps
  - 10G backbone between Router and Core switch
  - Backbone router - core firewall link 20G (10G+10G)
  - Storage servers moved to 10G

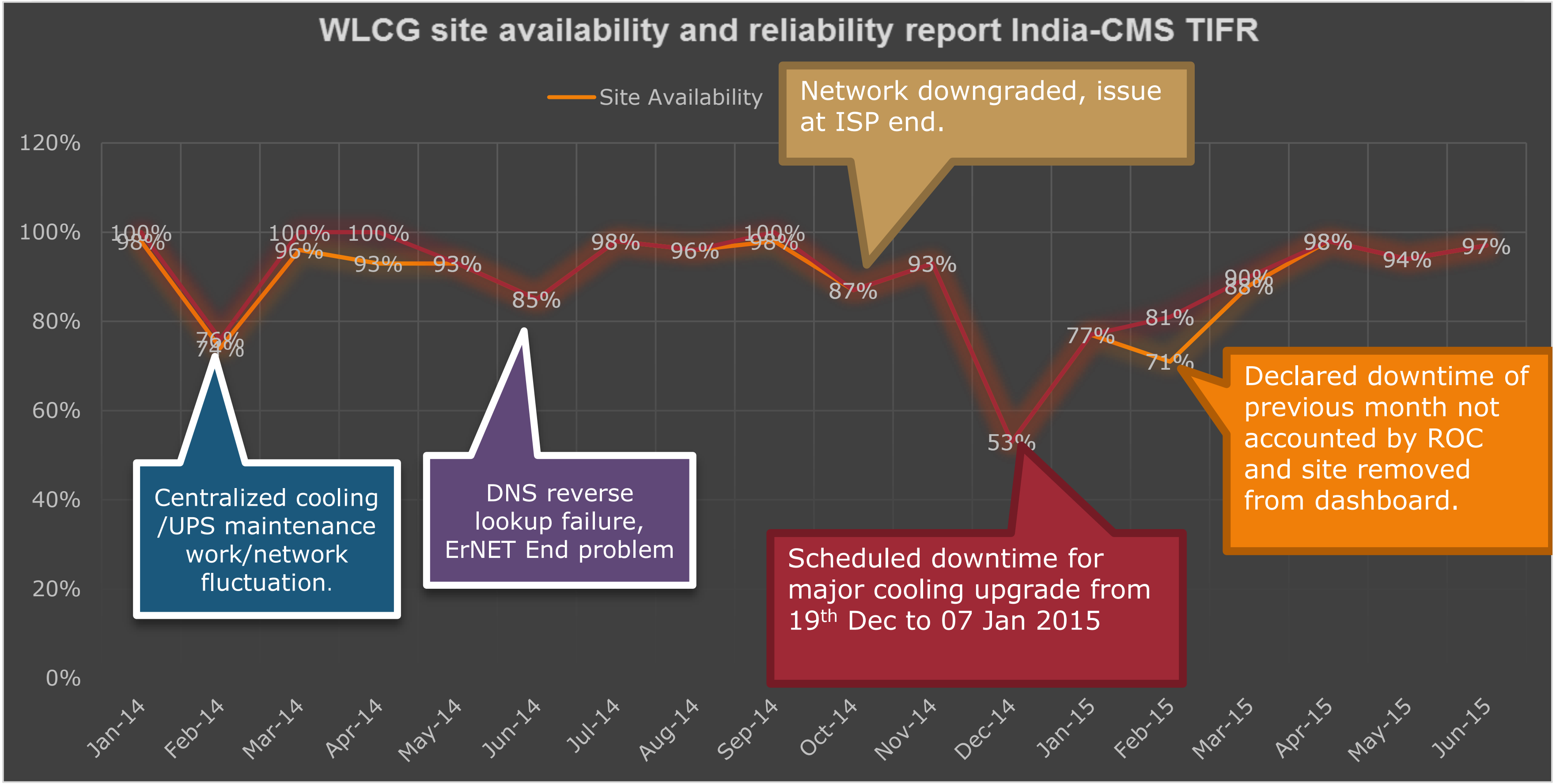


Cooling Infrastructure

- Front row cooling to improve cooling efficiency
- Capacity increased from 6K CFM to 10 K CFM

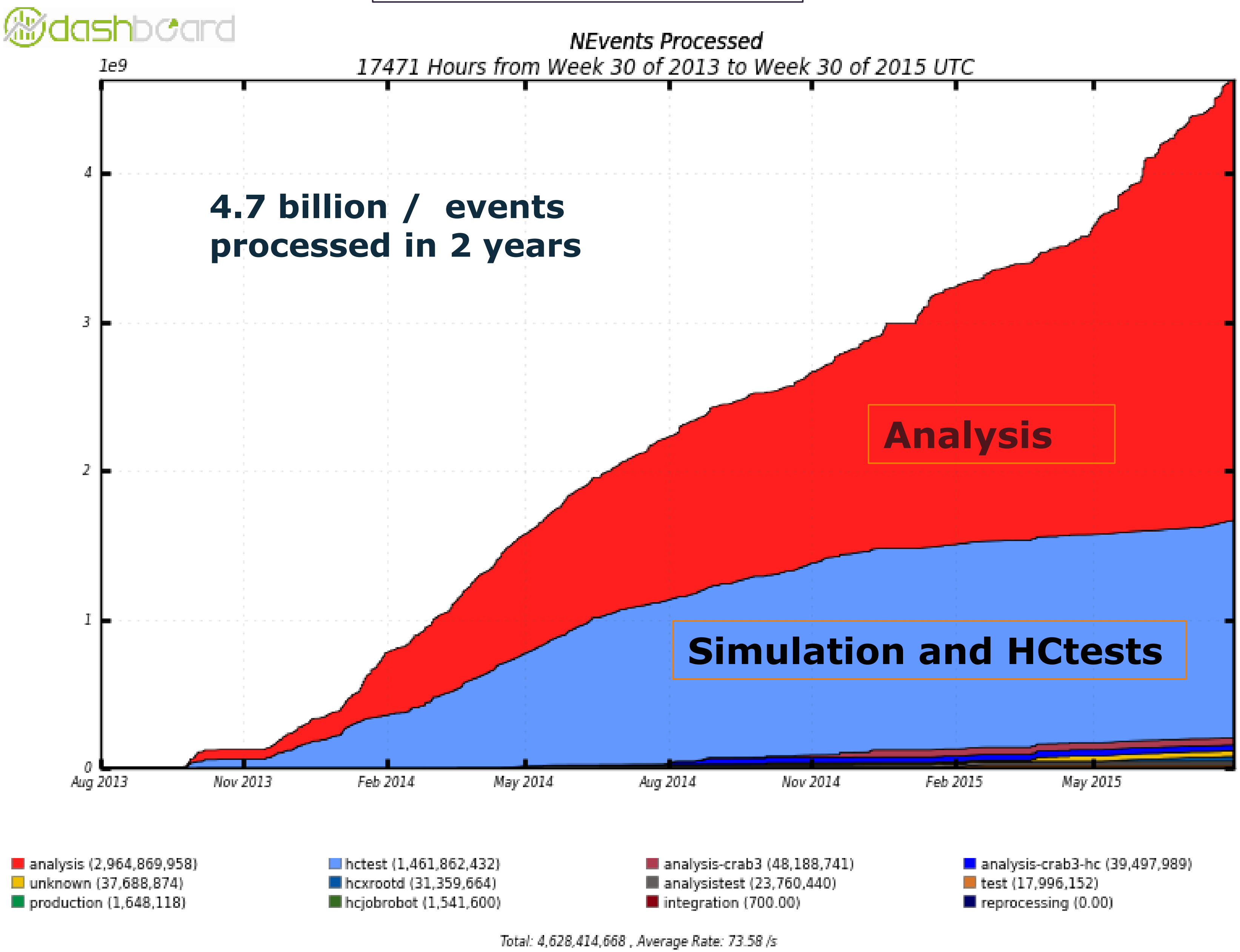


Availability and Reliability

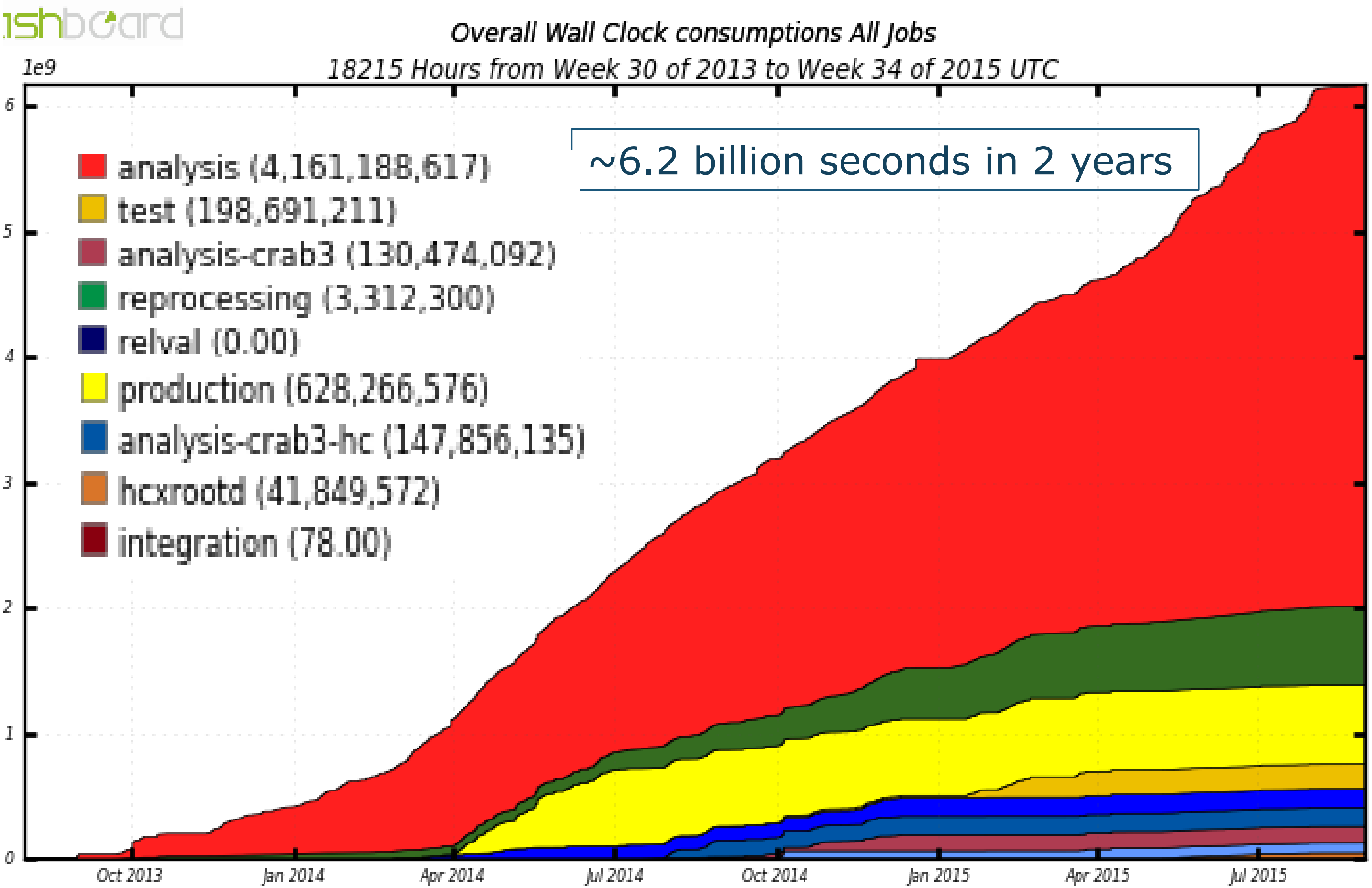


- A/R calculation based on a very elaborate CMS monitoring framework
- T2 monitored by three monitoring infrastructures – EGI, ROC TW (T1) and CMS dashboard.

Usage of T2_IN_TIFR



Usage of T2_IN_TIFR



Data transfers to T2_IN_TIFR

Downloads – 928 TB

	TOTAL	Austria	Belgium	Brazil	China	Estonia	Finland	France	Germany	Hungary	India	Italy	Netherlands
TOTAL	928 TB 1042749	3 TB 5761	12 TB 20904	1 TB 3119	45 GB 1775	372 GB 4711	303 GB 3270	43 TB 48978	103 TB 136168	1 TB 2258	3 TB 12256	223 TB 119977	113 GB 385
India	928 TB 1042749	3 TB 5761	12 TB 20904	1 TB 3119	45 GB 1775	372 GB 4711	303 GB 3270	43 TB 48978	103 TB 136168	1 TB 2258	3 TB 12256	223 TB 119977	113 GB 385

Pakistan	Portugal	Puerto-Rico	Russia	Russian-Federation	South-Korea	Spain	Switzerland	Taiwan	Thailand	Turkey	UK	USA	Ukraine	n/a
46 GB 18	7 TB 11412	3 GB 12	38 TB 23900	5 TB 11107	2 TB 3063	52 TB 76241	27 TB 191781	2 TB 940	7 GB 114	278 kB 4	86 TB 81078	303 TB 208491	61 GB 185	15 TB 74841
46 GB 18	7 TB 11412	3 GB 12	38 TB 23900	5 TB 11107	2 TB 3063	52 TB 76241	27 TB 191781	2 TB 940	7 GB 114	278 kB 4	86 TB 81078	303 TB 208491	61 GB 185	15 TB 74841

2013-08-01 00:00 to 2015-08-01 00:00 UTC

Data transfers from T2_IN_TIFR

Uploads – 763 TB

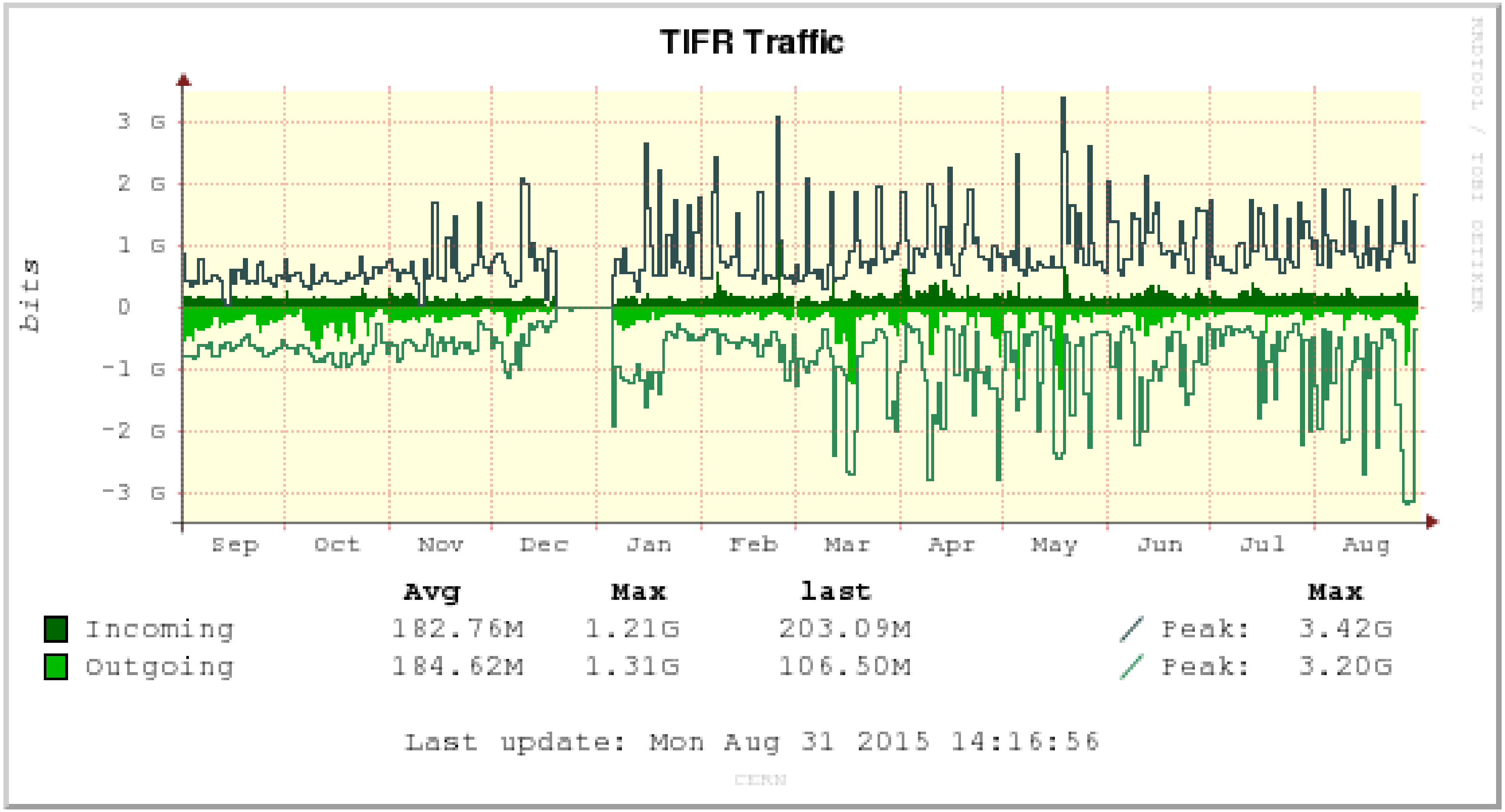
rations

sses

	TOTAL	India
TOTAL	763 TB 538578 / 394593 143985	763 TB 538578 / 394593 143985
Austria	7 TB 2428 / 1901 527	7 TB 2428 / 1901 527
Belgium	5 TB 4457 / 2034 2423	5 TB 4457 / 2034 2423
Brazil	1 TB 1691 / 563 1128	1 TB 1691 / 563 1128
China	386 GB 728 / 703 25	386 GB 728 / 703 25
Estonia	755 GB 2505 / 2205 300	755 GB 2505 / 2205 300
Finland	2 GB 139 / 112 27	2 GB 139 / 112 27
France	18 TB 11379 / 7971 3408	18 TB 11379 / 7971 3408
Germany	28 TB 19128 / 15217 3911	28 TB 19128 / 15217 3911
Greece	1 MB 56 / 56 0	1 MB 56 / 56 0
Hungary	42 GB 237 / 235 2	42 GB 237 / 235 2
India	3 TB 12256 / 11712 544	3 TB 12256 / 11712 544
Italy	32 TB 56693 / 36752 19941	32 TB 56693 / 36752 19941
Pakistan	1 TB 592 / 530 62	1 TB 592 / 530 62
Russia	1 TB 527 / 483 44	1 TB 527 / 483 44
Russian-Federation	860 GB 499 / 393 106	860 GB 499 / 393 106
South-Korea	11 GB 414 / 410 4	11 GB 414 / 410 4
Spain	286 TB 136296 / 111484 24812	286 TB 136296 / 111484 24812
Switzerland	20 TB 59811 / 36312 23499	20 TB 59811 / 36312 23499
Taiwan	13 TB 7079 / 4768 2311	13 TB 7079 / 4768 2311
UK	307 TB 155172 / 115197 39975	307 TB 155172 / 115197 39975
USA	36 TB 48516 / 29612 18904	36 TB 48516 / 29612 18904
n/a	2 TB 17975 / 15943 2032	2 TB 17975 / 15943 2032

2013-08-01 00:00 to 2015-08-01 00:00 UTC

T2_IN_TIFR traffic to LHCONE



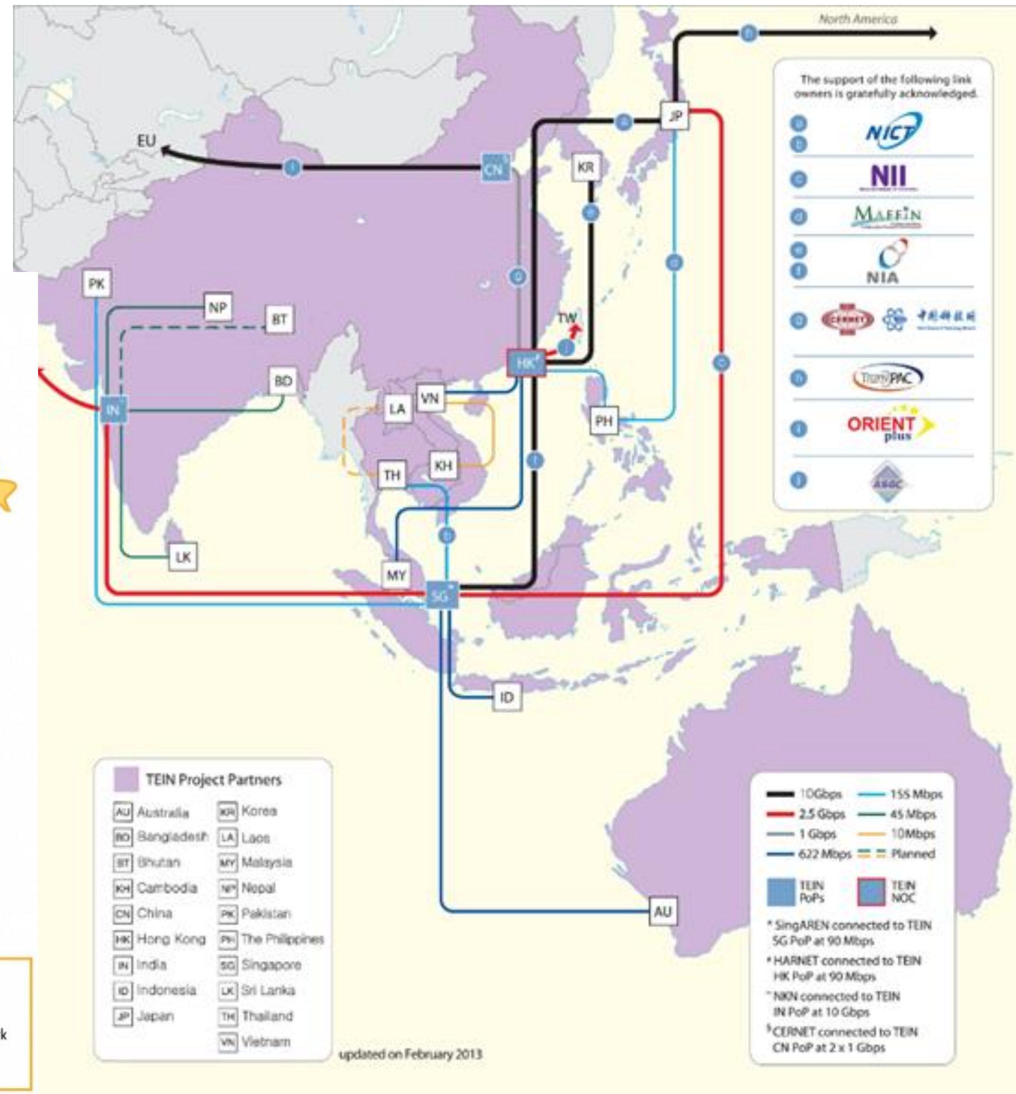
Collaborating Indian Institutes at LHC (14 or more)

- **TIFR, Mumbai as National Facility** WLCG Site
- Variable Energy Cyclotron Centre (VECC, Kolkata) WLCG Site
- Bhabha Atomic Research Centre (BARC, Mumbai)
- Delhi University
- Saha Institute of Nuclear Physics (SINP, Kolkata)
- Punjab University
- Indian Institute of Technology, Bombay (IITB, Mumbai)
- Indian Institute of Technology, Madras (IITC, Chennai)
- Raja Ramanna Centre for Advanced Technology (RRCAT, Indore)
- Indian Institute of Technology Bhubaneswar (IITBBS)
- Institute for Plasma Research (IPR, Ahmedabad)
- National Institute of Science Education and Research (NISER, Bhubneshwar)
- Vishva-Bharti University (Santiniketan, WB)
- Indian Institute of Science Education and Research, Pune

(More then 90 active users from these institutes have accounts at T2_IN_TIFR and TIFR Tier III)

India-WLCG network

Indian LHC traffic was not structured



Routing anomalies

[root@cmst3ui2 ~]# traceroute -I 192.65.184.73
traceroute to 192.65.184.73 (192.65.184.73), 30 hops max, 60 bytepackets
1 172.16.11.252 (172.16.11.252) 2.353 ms 2.642 ms 2.883 ms
2 172.16.0.254 (172.16.0.254) 0.580 ms 0.574 ms 0.560 ms
3 vpn2.saha.ac.in (14.139.193.1) 0.945 ms 0.969 ms 0.968 ms (NKN)
4 10.118.248.93 (10.118.248.93) 0.951 ms 0.954 ms 0.954 ms (NKN Private core)
5 * * *
6 * * *
7 10.255.221.34 (10.255.221.34) 31.804 ms 31.700 ms 31.694 ms (NKN Private core)
8 115.249.209.6 (115.249.209.6) 37.302 ms 37.541 ms 37.541 ms (RCOM – Andhra)
9 * * *
10 * * *
11 62.216.147.73 (62.216.147.73) 46.588 ms 46.577 ms 46.556 ms (UK)
12 xe-0-0-0.0.pjr03.ldn001.flagtel.com (85.95.26.238) 186.642ms 174.002 ms 173.979 ms
13 xe-5-2-0.0.cji01.ldn004.flagtel.com (62.216.128.114) 187.455ms 187.519 ms 187.693 ms
14 80.150.171.69 (80.150.171.69) 295.319 ms 293.396 ms 293.372ms (Germany)
15 217.239.43.29 (217.239.43.29) 305.694 ms 309.873 ms 306.677ms (Deutsche Telekom AG)
16 e513-e-rbrxl-1-ne1.cern.ch (192.65.184.73) 221.143 ms 230.249 ms 215.501 ms (CERN)

Resolved now

Solution ………. Steps taken at TIFR

- Understanding the practical implementation of entire LCG network ecosystem enabled us to take some major upgrades.

  - Creating a LCG VRF In India on NKN network connecting all partner institutes
  - Routing VRF traffic on TIFR-CERN P2P link, significantly improving latency and throughput.
  - Connecting to LHCONE network directly instead of via CERNLite.
  - GEANT upgraded backbone link between CERNLite router (where Indian link terminates at CERN to GEANT POP) to 10G
  - Enabled jumbo frames (9000 bytes) on NKN L3VPN in India to TIFR to LHCONE

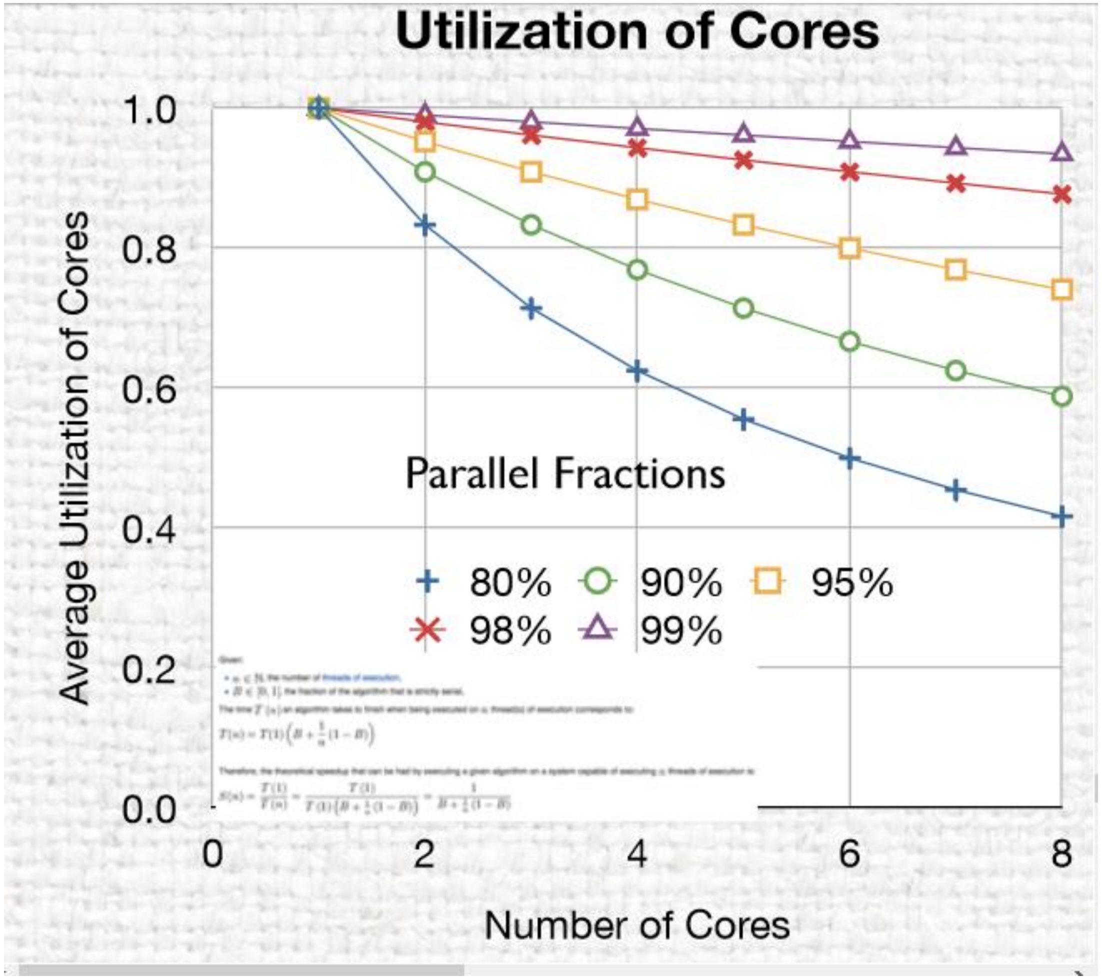
CPU overhead is reduced significantly and efficiency is improved

Initiatives duly acknowledged by collaborating institutes.

To keep 8 cores 95% busy need 99.2% of our code to run in parallel
Even quick running modules will bottleneck threading

Dimensions of performance

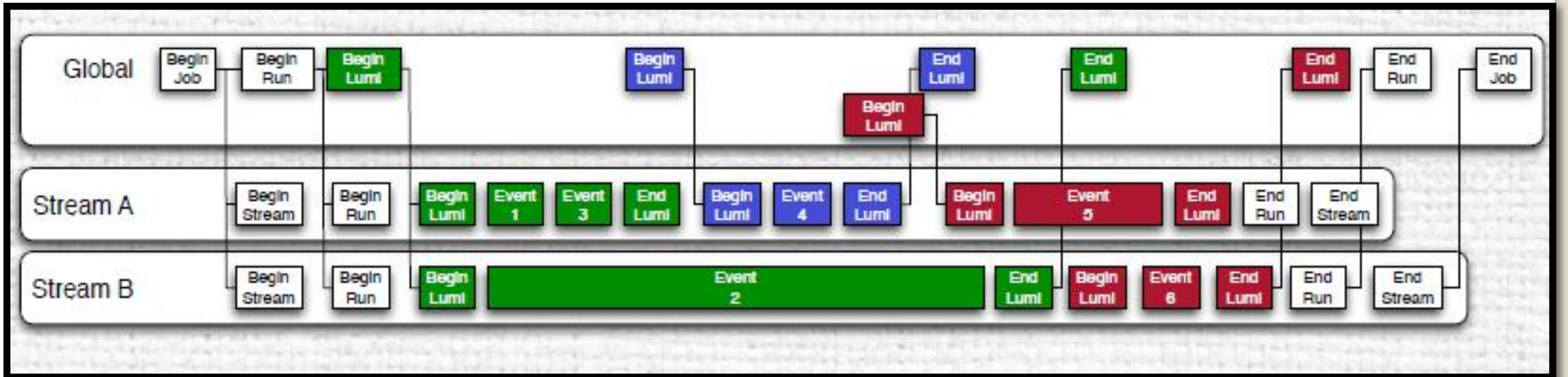
- Vectors
- Instruction Pipelining
- Instruction Level Parallelism (ILP)
- Hardware threading
- Clock frequency
- Multi-core
- Multi-socket
- Multi-node



multi-threaded framework

The design allows many different levels of concurrency

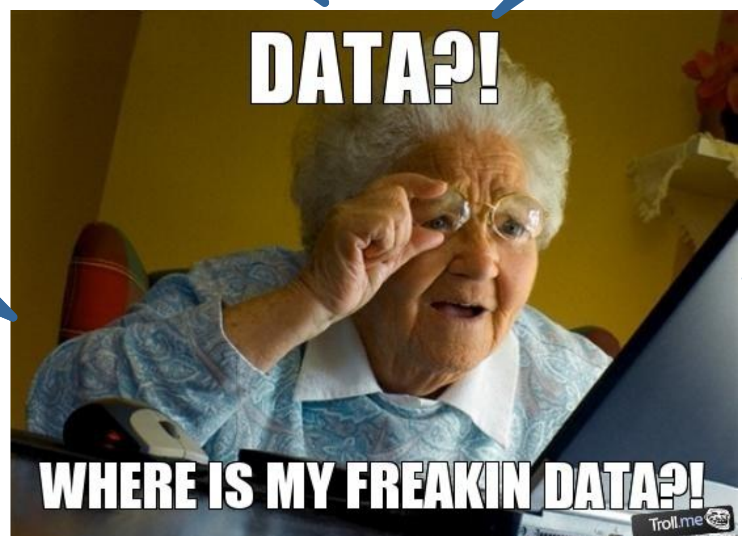
- Extensive use of Intel TBB
  - Events, modules and sub-module
    - Thread-unsafe code is allowed via 'One' module variety
  - Framework guarantees serialization
    - Tools to find thread-safety issues have been developed
  - First performance results show that 99.3% of our reconstruction application can run in parallel
    - memory consumption is no longer a problem
    - network load is way down



Which data ?

Where is it ?

Do I have Permission to access that data ?



For my analysis where do I store it ?

Data storage technology ?

Client configuration for accessing the data ?

AAA

CMS new architecture for data access, emphasizing the following three items:

Reliability: No I/O error unless no CMS site can server the file

Transparency. Automatic catalogue lookups, redirections and reconnections

Usability: Natively integrate with CMS application frameworks (CMSSW, ROOT …..)

Global: Any data Any where, Any time

For users**:**

- Once he knows what dataset he wants to use for analysis
  xrdfs cms-xrd-global.cern.ch locate /store/path/to/file
  (Universal LFN data store of CMS)
- As long as you do not get the message "*No servers have the file"* it is safe for you to use the AAA service

https://twiki.cern.ch/twiki/bin/view/Main/CmsXrootdArchitecture

AAA (XROOTD) at T2_IN_TIFR

- T2_IN_TIFR is a CMS T2 at TIFR, Mumbai with ~ 1 PB of DPM storage.
- Well connected with LHCONE where last years WLCG traffic crossed 1 PB
- At TIFR we implemented XROOD for access and fallback in 2013 Feb
- We actively participated in testing and implementation of XROOTD on DPM and helped in tracking down many bugs experienced in due course.

We have also setup a regional redirector for India–CMS T3 users and Indian institutes.

CMS Remote Analysis Builder

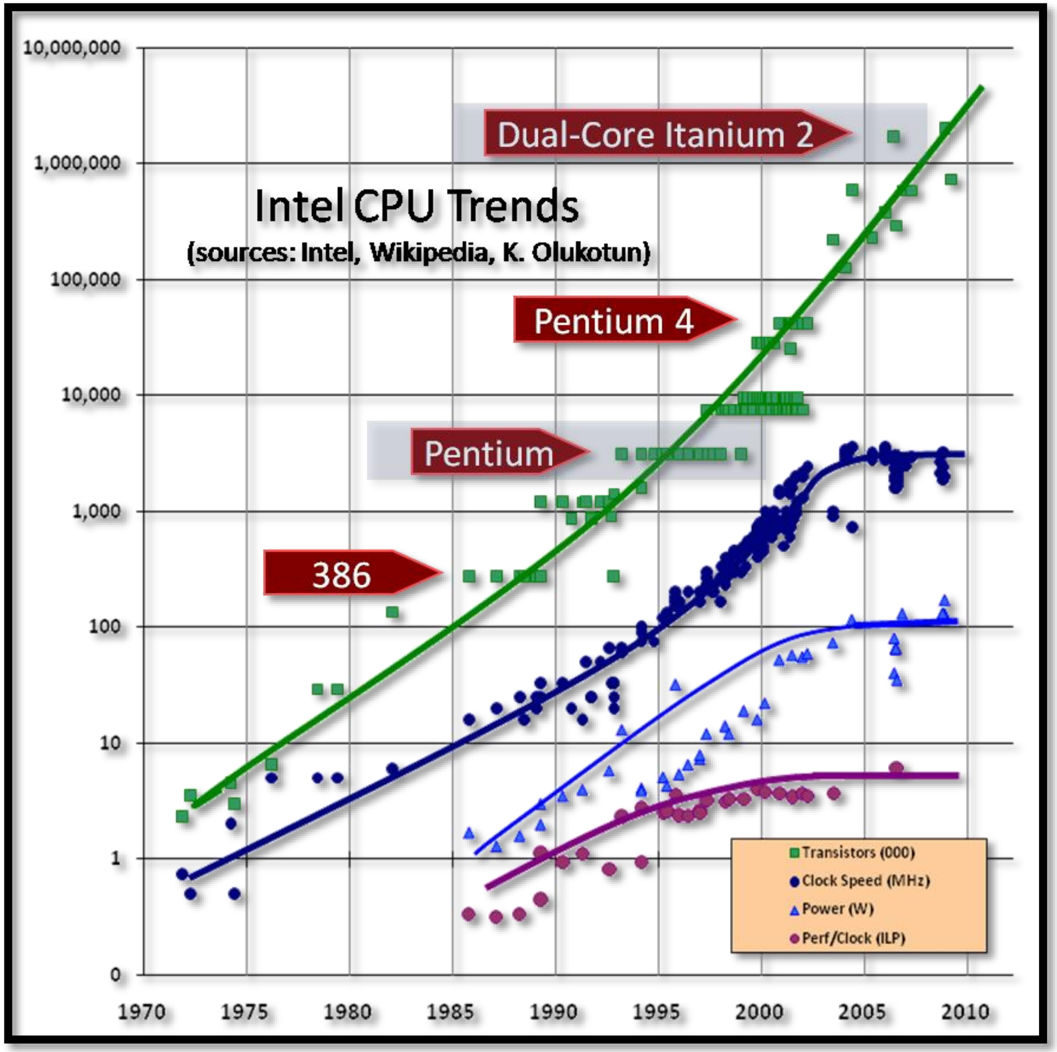
- Significant improvements from earlier version of CRAB
- New grid submission tool enables option to ignore data locality, i.e., use AAA
- Tested by artificially forcing jobs to run with remote access
- During CSA14: 20k cores in production, 200k jobs/day, average of 300 users/week (**TIFR Participated in the exercise**)
- Improves handling of read failures and monitoring
- Python implementation wrapping cpp modules.

Continuous upgrades … … ….

- Operating system and underlying services.
- Middleware from gLite to EMI1 > EMI2 > EMI3
- Implementation of numerous new services.
- Storage migration and consistency checks
- Automation framework
- Virtualization of services for increasing the efficiency
- Creation of test beds
- Tier-III upgrade

Future

• Keeping up with Moore's law



Thank you

Questions ??