CERN Internetwork

Openlab Summer 2011
CERN, 4th August 2011
edoardo.martelli@cern.ch

Summary

- CERN networks
- LHC Data Challenge
- WLCG
- LHCOPN
- LHCONE
- Openlab
- Conclusions

IT-CS
Communication systems

IT-CS

The IT-CS group is responsible for all communication services in use at CERN for data, voice and video

http://it-cs.web.cern.ch/it-cs/

IT-CS organization

Communication Systems

Group Leader	Jean-Michel Jouanigot
Deputy Group Leader	Frederic Chapron

CERN openlab

Group Management
DI

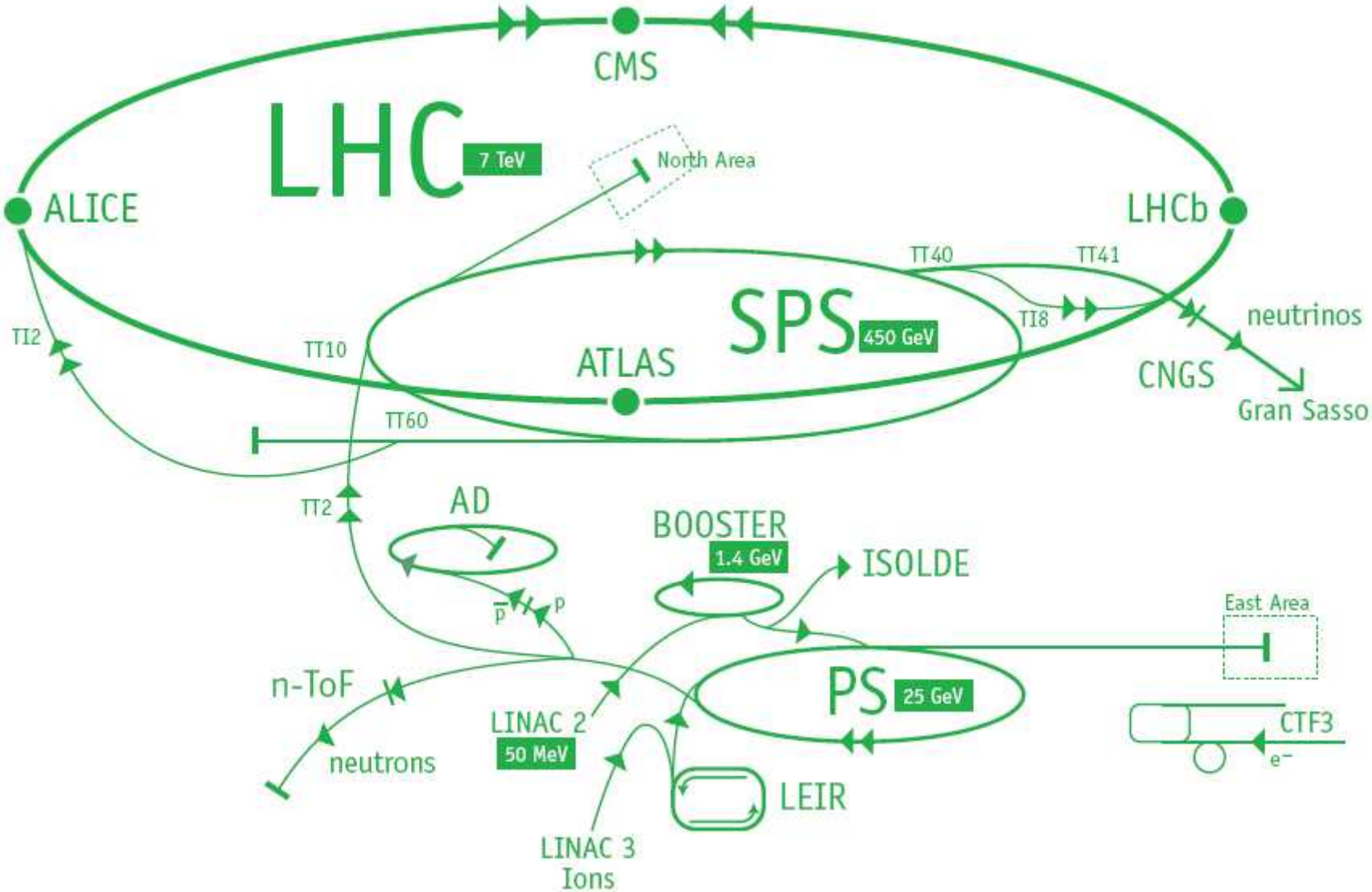
Ext. Net. & CIXP Manager	John Shade
WEB master	Tamara Smoliakova

Communications Deployment CD	Communications Engineering CE	Communications Support CS	Communications Tools CT
Section Leader	**Section Leader**	**Section Leader**	**Section Leader**
Eric Sallaz	Edoardo Martelli	Frederic Chapron	Jose Carlos Luna Duran
Marc Collignon	Gyorgy Balazs	Alberto Garcia Molero	Andreu Belmonte Pena
Maryse da Costa	Sebastien Ceuterickx	Italo Gard	Raul Garcia Martinez
Claude Dangoisse	David Gutierrez Rueda	Isaac Gonzalez Gonzalez	Milosz Marian Hulboj
Daniel Francart	Carles Kishimoto Bisbe	Anna Raczynska-Tartivel	Ryszard Erazm Jurga
Jean Simeoni	Pierre Le Corre	Rodrigo Sierra Moral	Vlad Lapadatescu
	Virginie Longo	Aurelie Pascal	Agnieska Orlewicz
	Mathew Paul Williamson	Jean-Pierre Puget	Miguel Filipe Santos Pinto
			Zbigniew Szymon Stanecki

CERN

European Organization for Nuclear Research

CERN accelerator complex

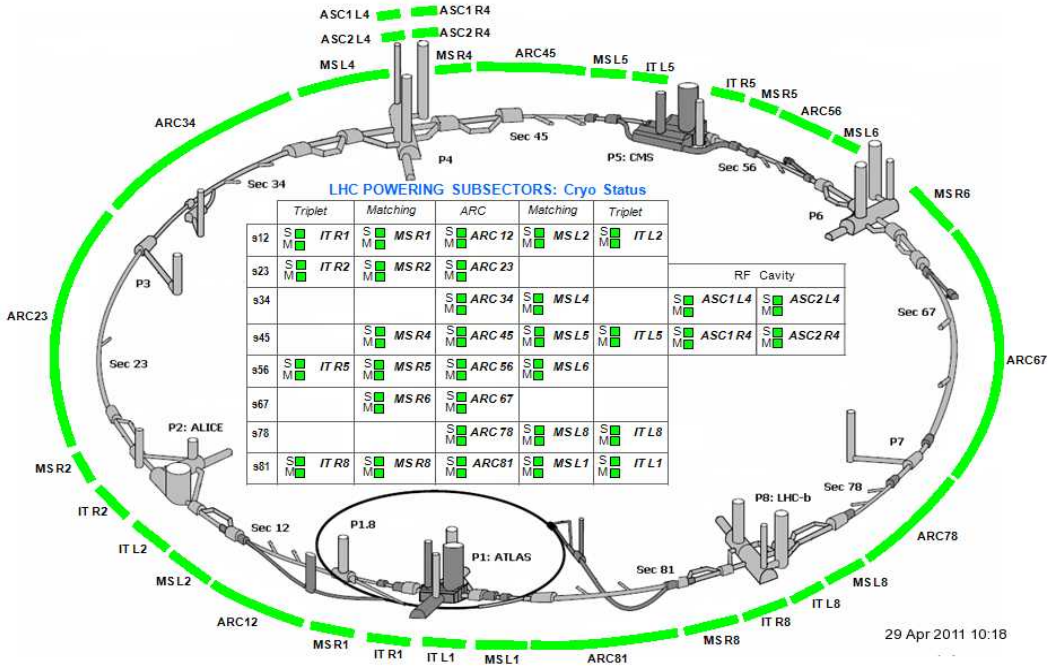


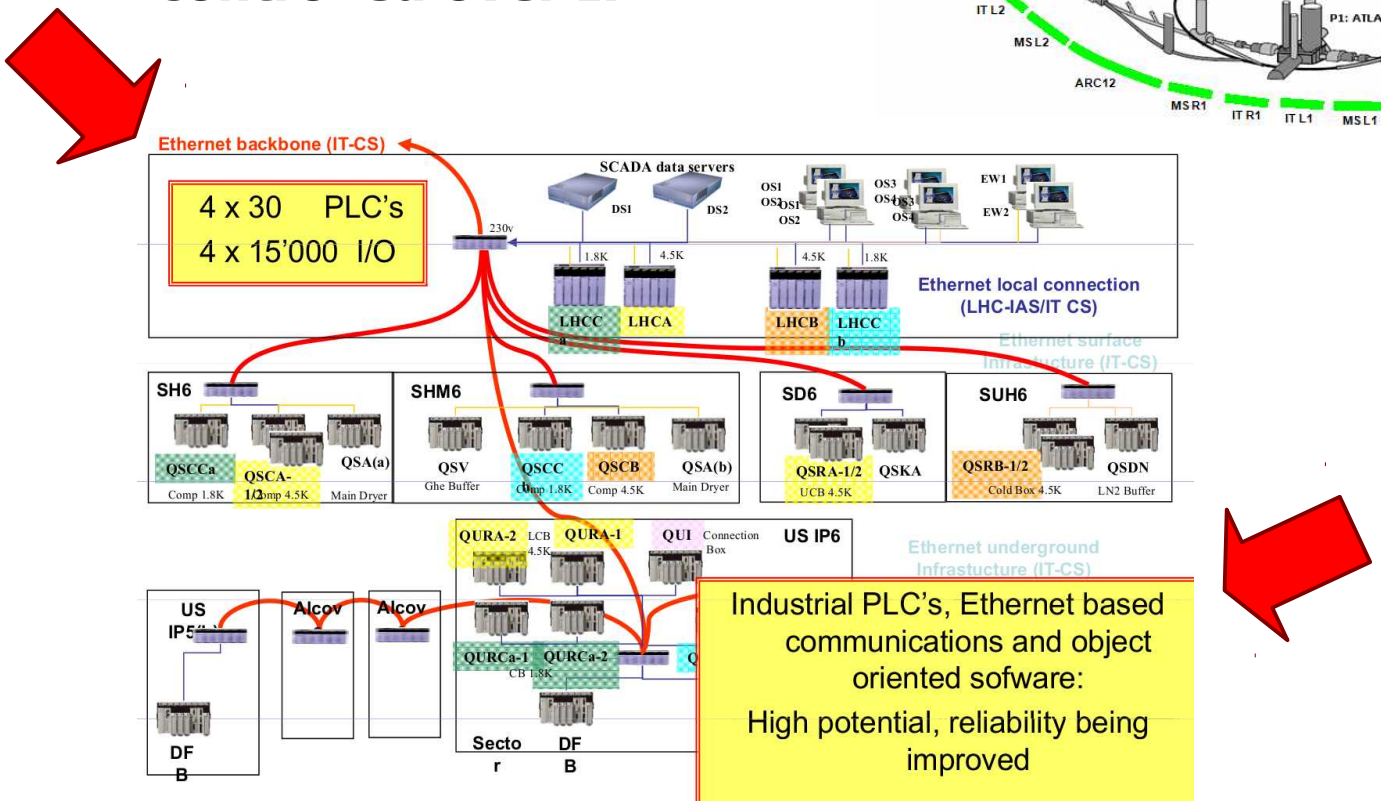
Physics over IP

Most of the CERN is controlled and managed over a pervasive
IP network

Cryogenics

27Km of pipes at -271.11° C by means of 700.000 litres of Helium: **controlled over IP**

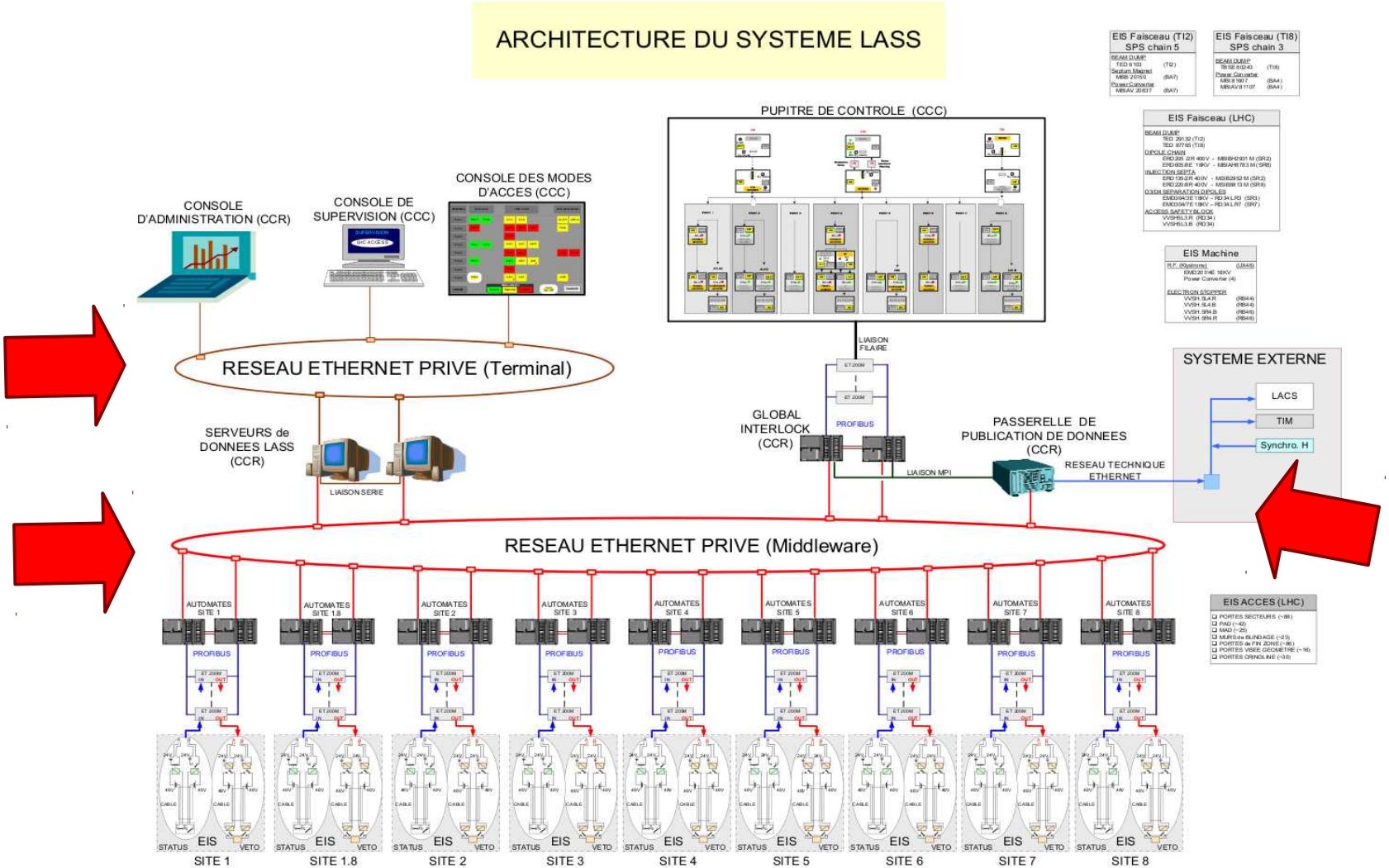




Source: http://te-dep-crg-oa.web.cern.ch/te-dep-crg-oa/te-crg-oa_fichiers/cryolhc/LHC%20Cryo_BEOP_lectures2009.pdf

Access control

Safety and Security: **made over IP**



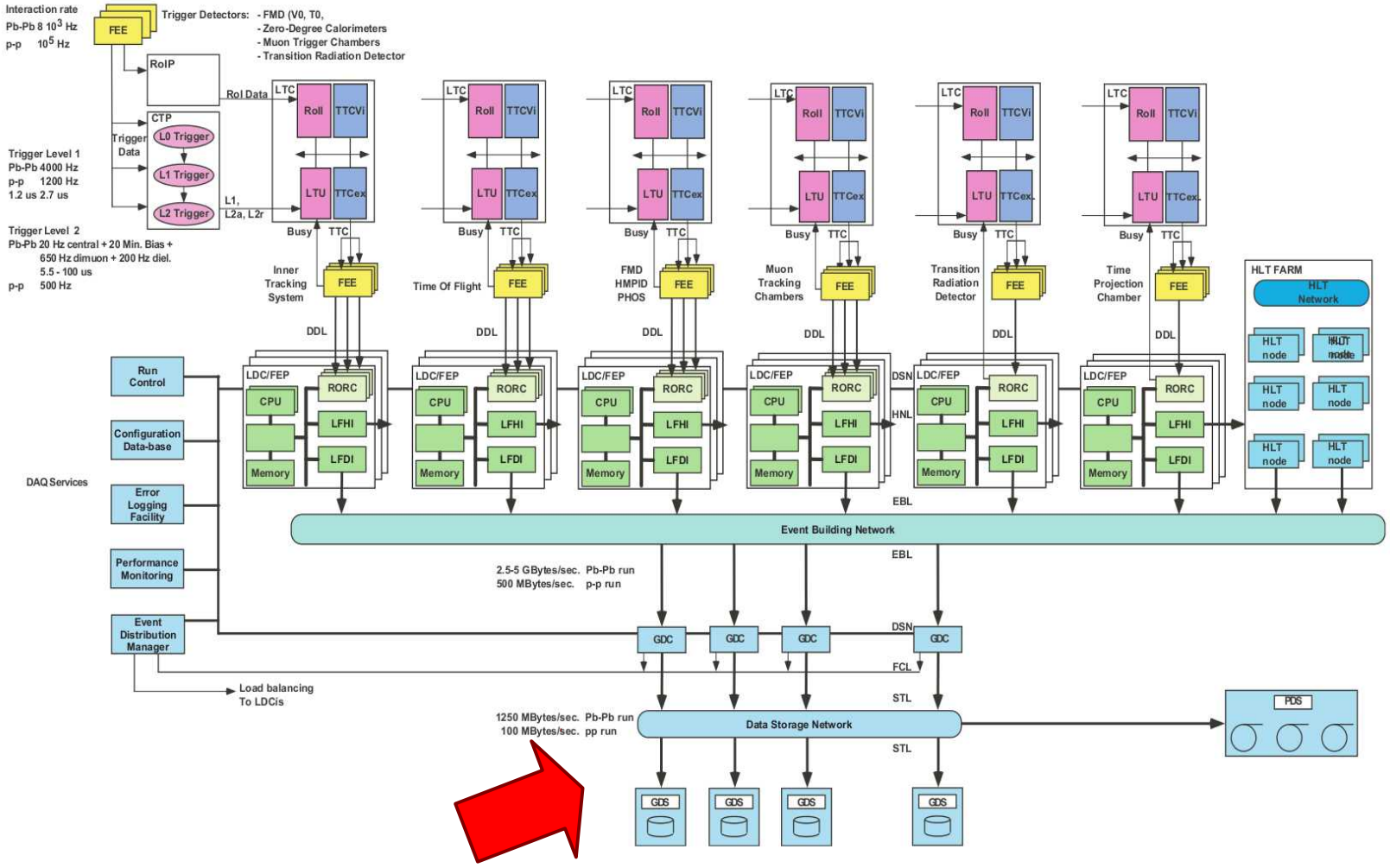
Source:https://edms.cern.ch/file/931641/1/LASS-LACS_IHM.pdf

Remote inspections

Remote inspection of dangerous areas: robots controlled and giving feedback over **WiFi and GSM IP networks**

DAQ: Data Acquisition

A constant stream of data from the four Detectors to disk storage

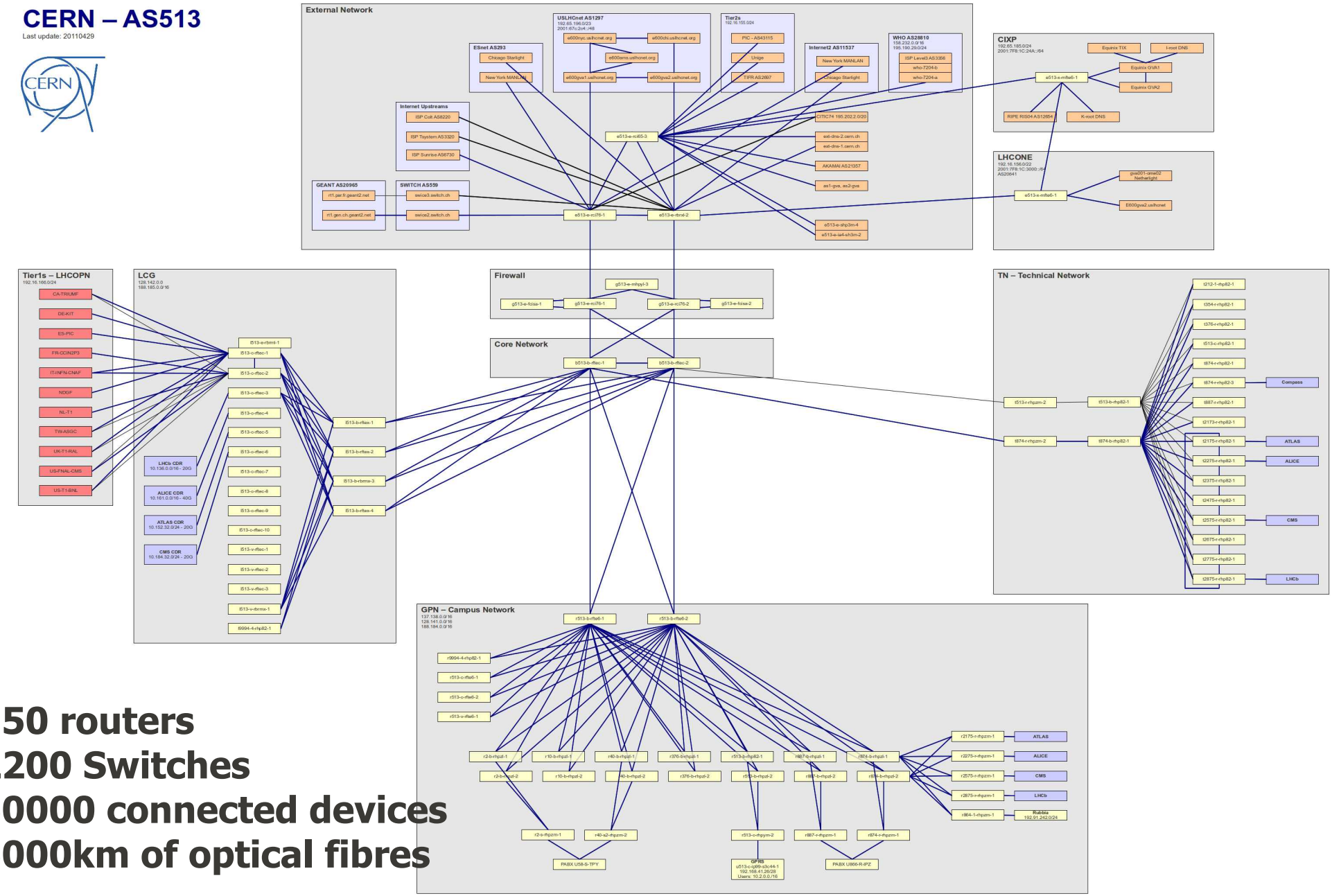


Source: http://aliceinfo.cern.ch/Public/Objects/Chapter2/DetectorComponents/daq_architecture.pdf

CCC: CERN Control Centre

The neuralgic centre of the accelerator: over IP

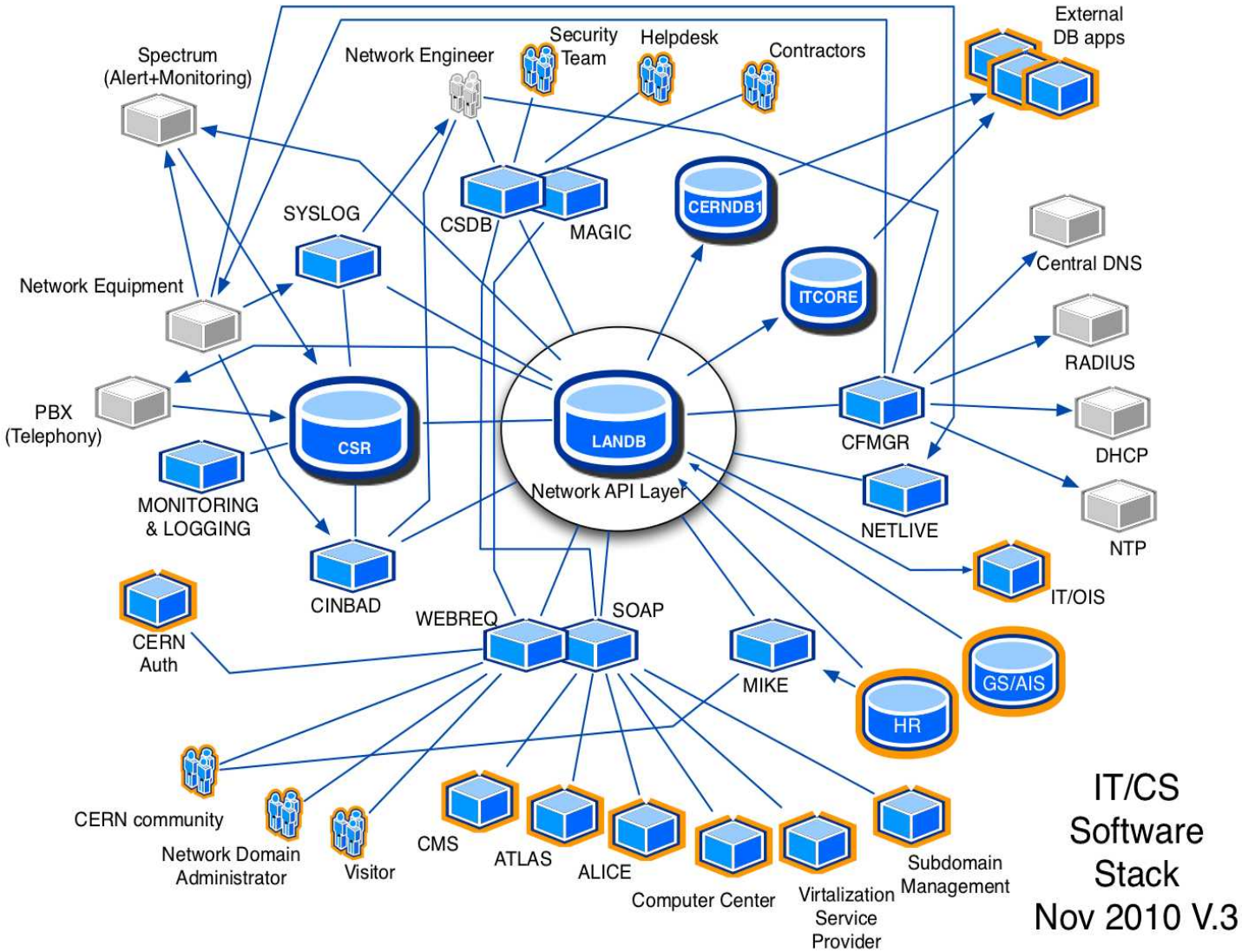
CERN network



- 150 routers
- 2200 Switches
- 50000 connected devices
- 5000km of optical fibres

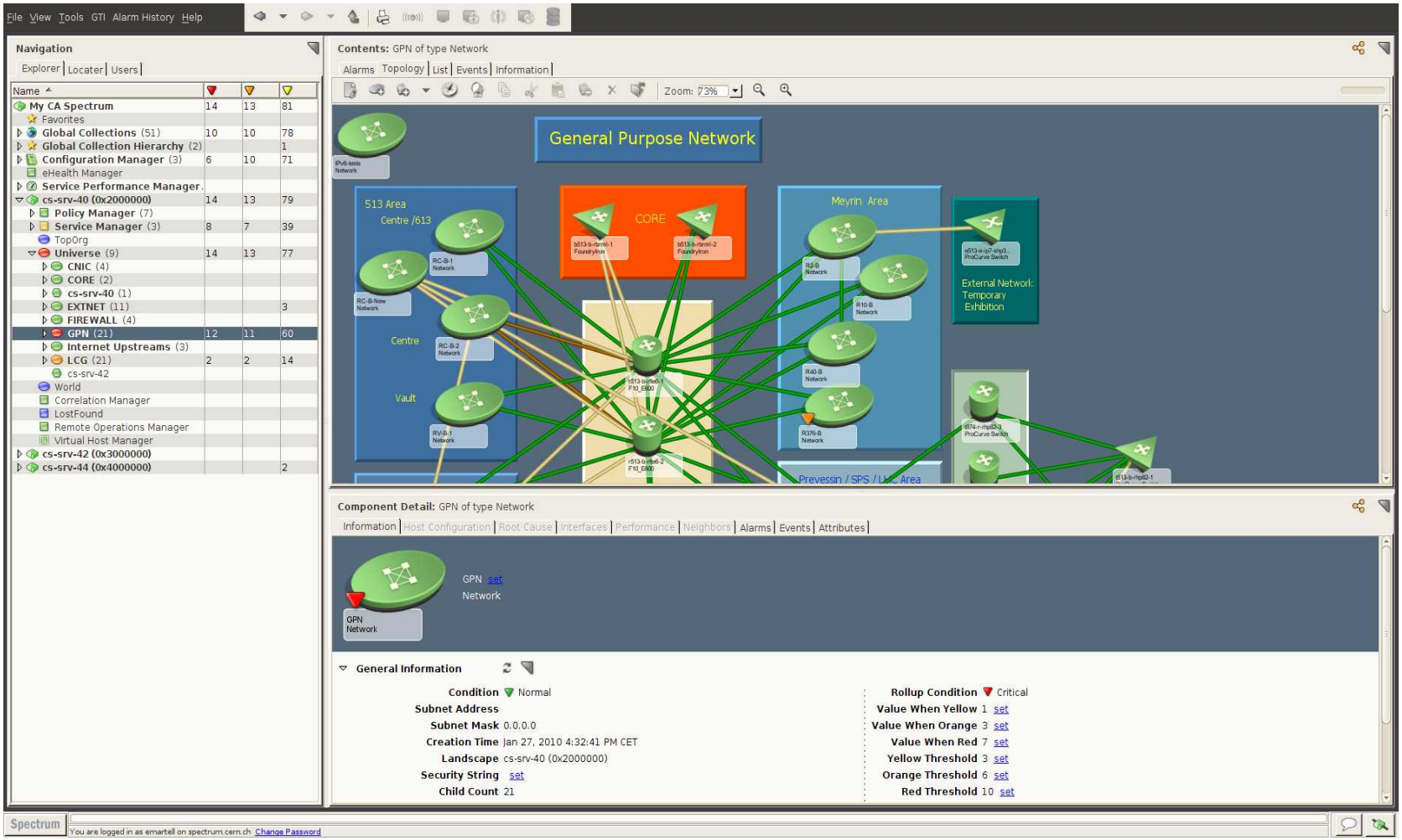
Network Provisioning and Management System

- 250 Database tables
- 100000 Registered devices
- 50000 hits/day on web user interface
- 1,000,000 lines of codes
- 10 years of development



Monitoring and Operations

The whole network is monitored and operated by CERN NOC (Network Operation Centre)



IPv6

Started deploying IPv6 dual stack network

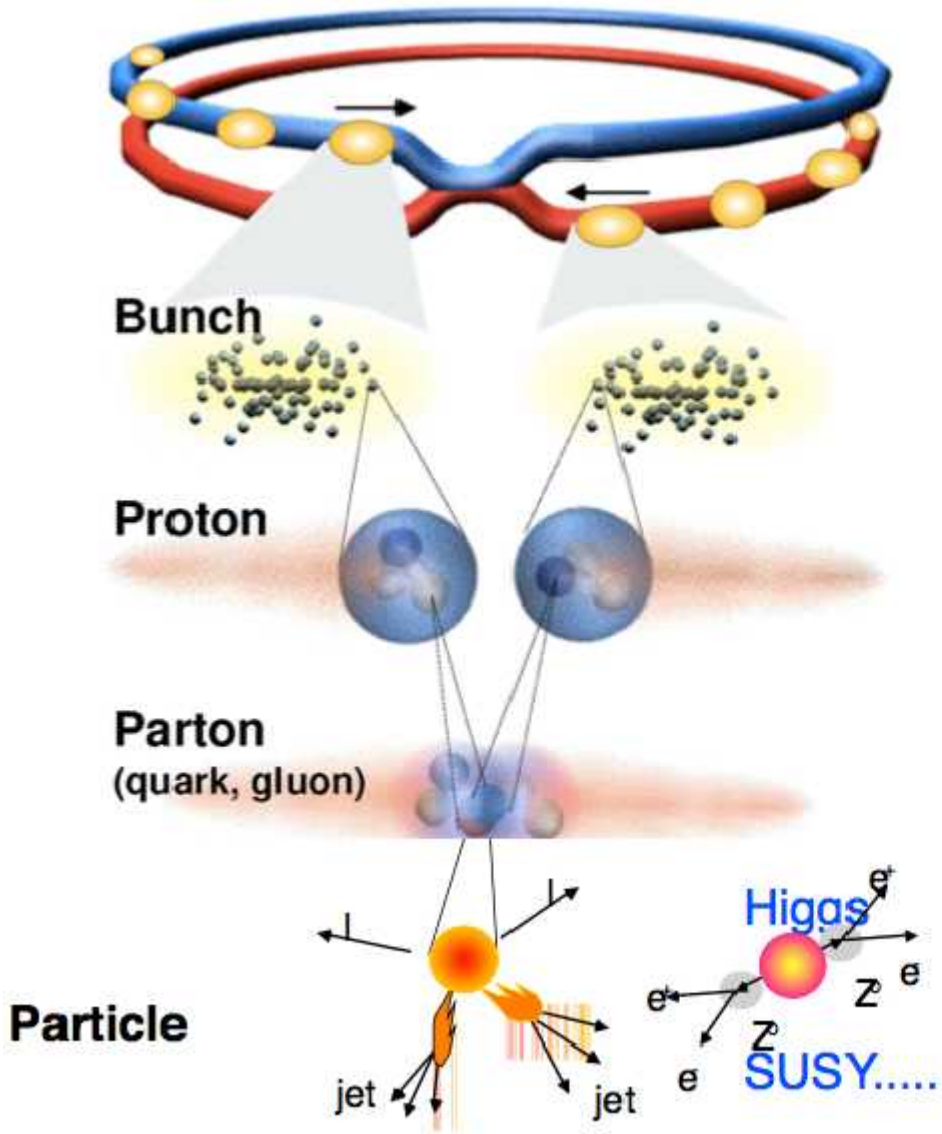
Soon available: testbed with dual stack Virtual Machines

More information:
http://cern.ch/ipv6

almost

LHC Data Challenge

Collisions in the LHC

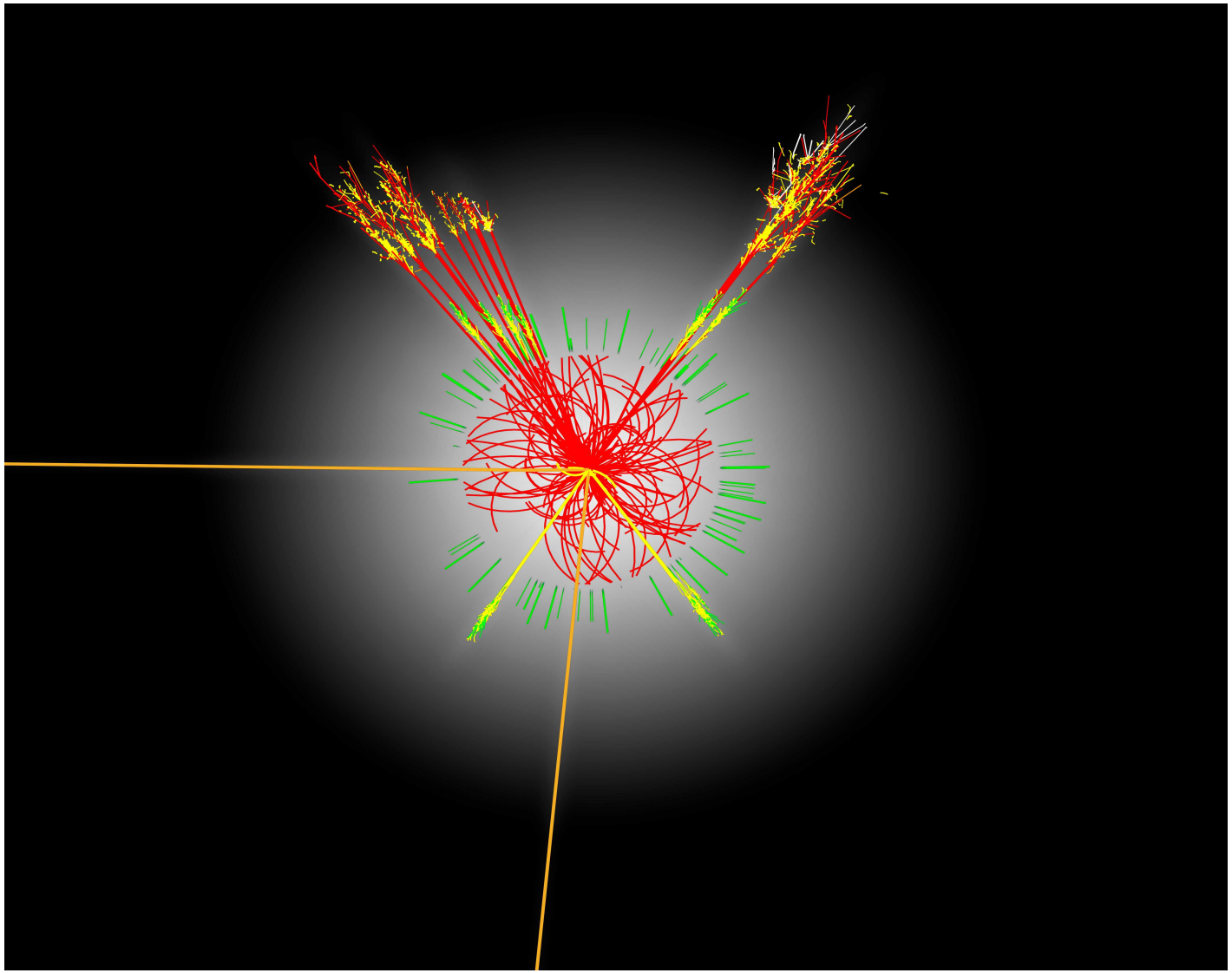


Proton - Proton	2808 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	$10^{34} cm^{-2} s^{-1}$
Crossing rate	40 MHz
Collision rate ≈	10^{7}-10^{9}

New physics rate ≈ .00001 Hz

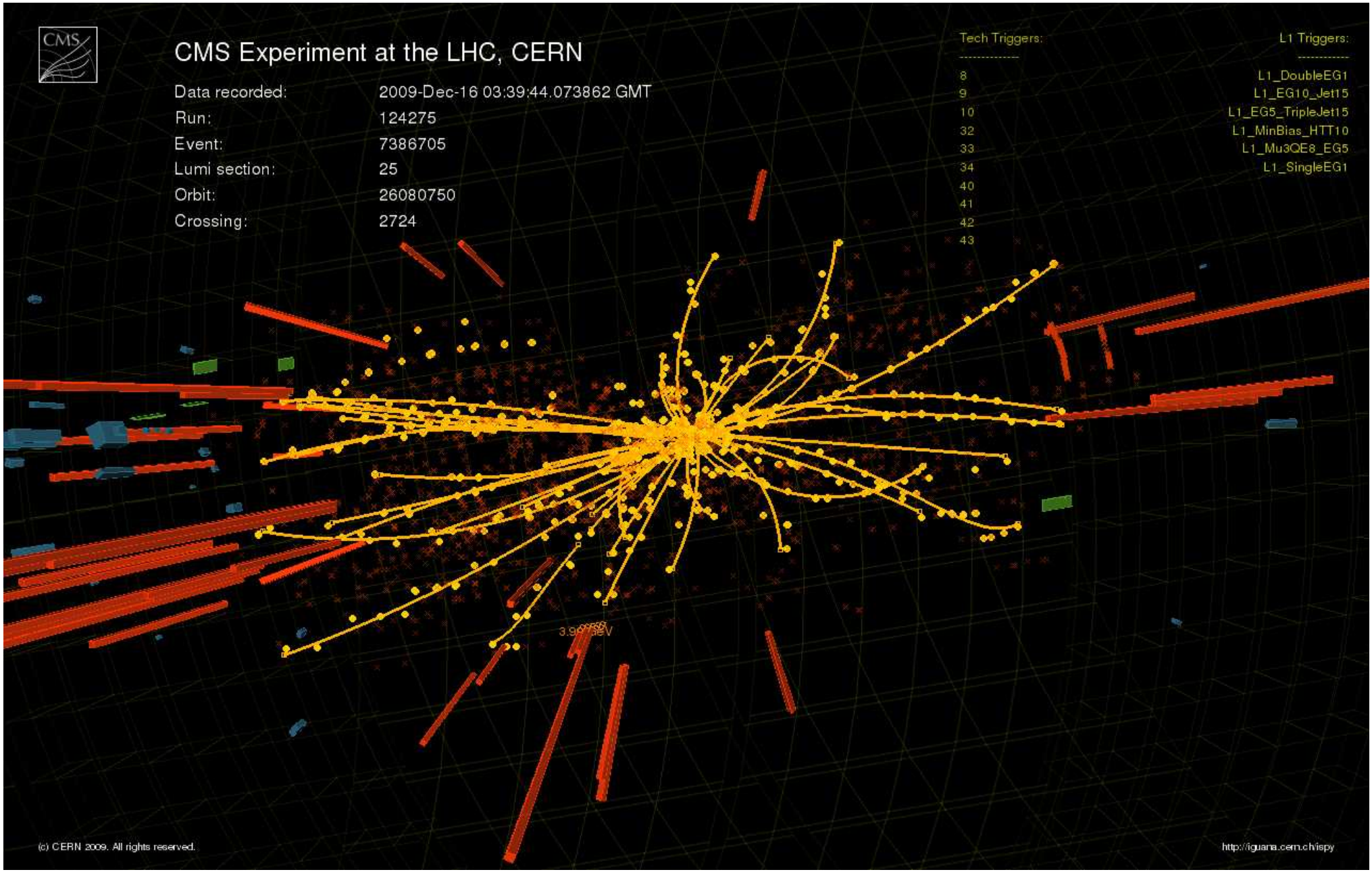
Event selection:
1 in 10,000,000,000,000

Comparing theory…

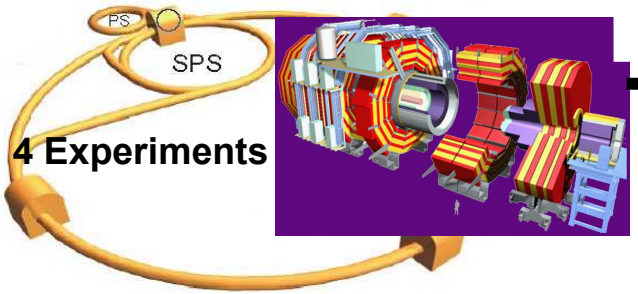


Simulated production of a Higgs event in ATLAS

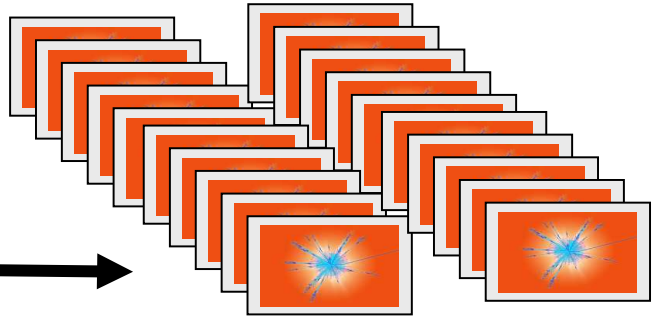
.. to real events



Data flow

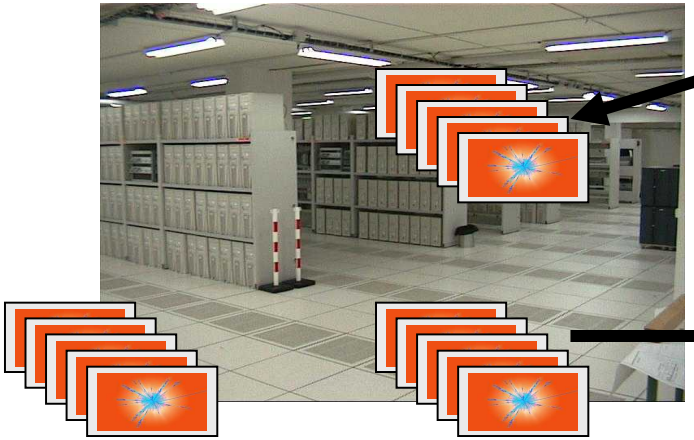


3 PBytes/s

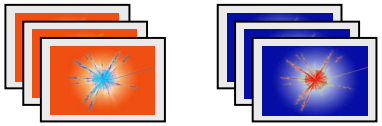


Filter and first selection

2 GBytes/s
to the CERN computer center



Create sub-samples



World-Wide Analysis

1 TByte/s ?
Distributed + local

Physics
Explanation of nature

10 GBytes/s

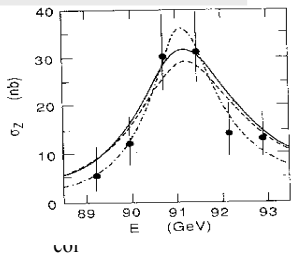
4 GBytes/s

Store on disk and tape

Export copies

$$\sigma_{f\bar{f}} \approx \sigma^0_{f\bar{f}} \times \frac{s\Gamma_Z^2}{(s - m_Z^2)^2 + s^2\Gamma_Z^2 / m_Z^2} \quad \text{with}$$

$$\sigma^0_{f\bar{f}} = \frac{12\pi}{m_Z^2}\frac{\Gamma_{ee}\Gamma_{ff}}{\Gamma_Z^2} \quad \text{and} \quad \Gamma_{ff} = \frac{G_F m_Z^3}{6\pi\sqrt{2}} \times (v_f^2 + a_f^2)$$



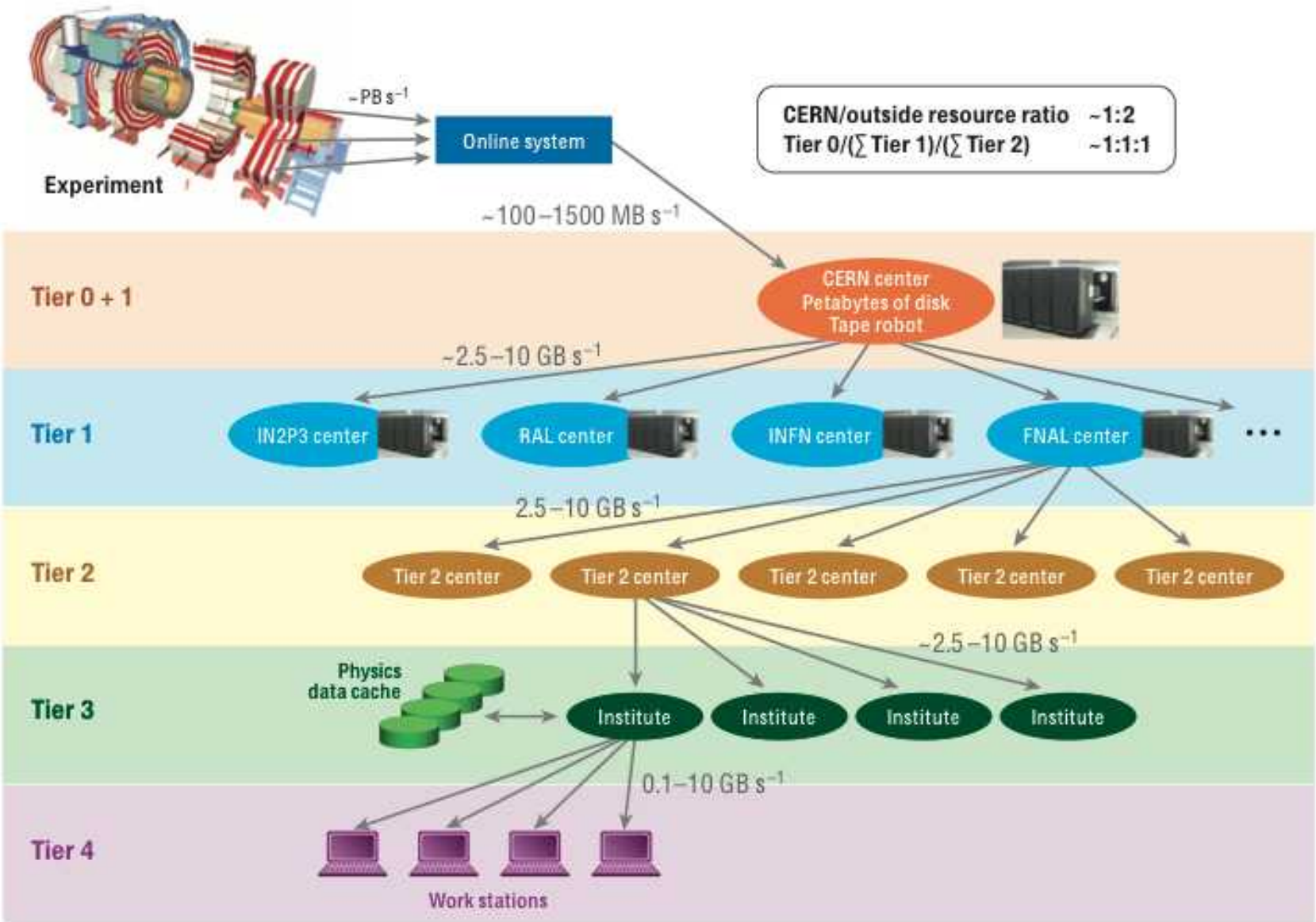
Data Challenge

- 40 million collisions per second

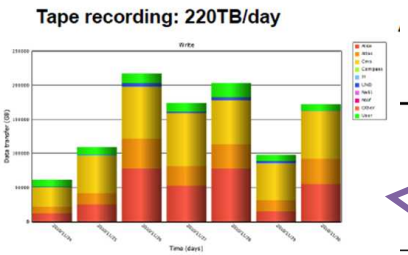
- After filtering, 100 collisions of interest per second

- 10^{10} collisions recorded each year
  = 15 Petabytes/year of data

Computing model



Last year data taking

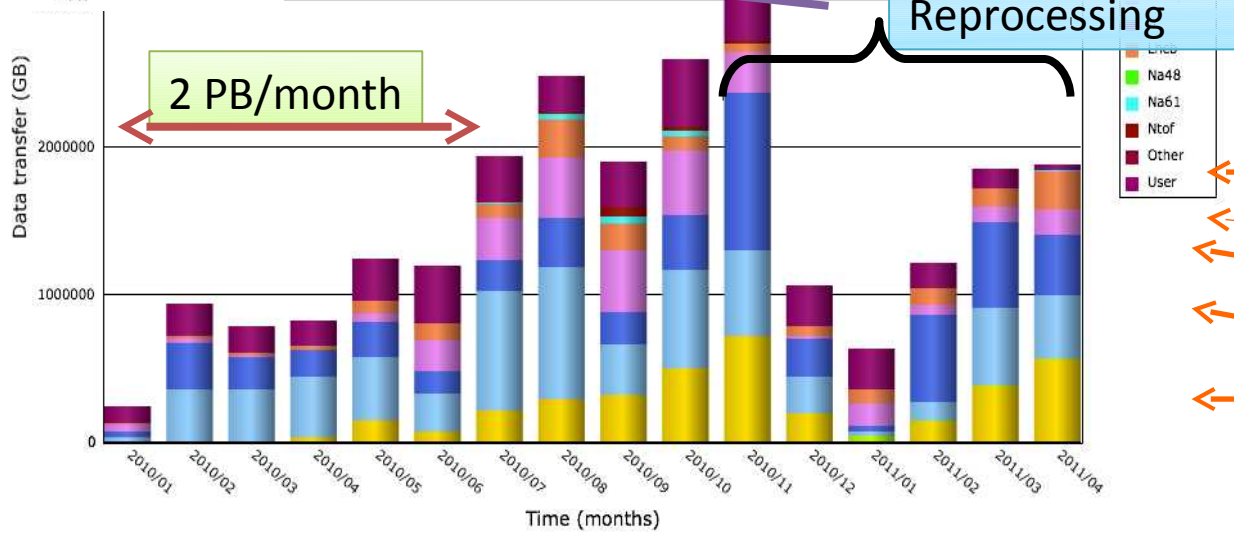


Data written to tape (GB/month): 2010-11

HI

2010 Reprocessing

2 PB/month

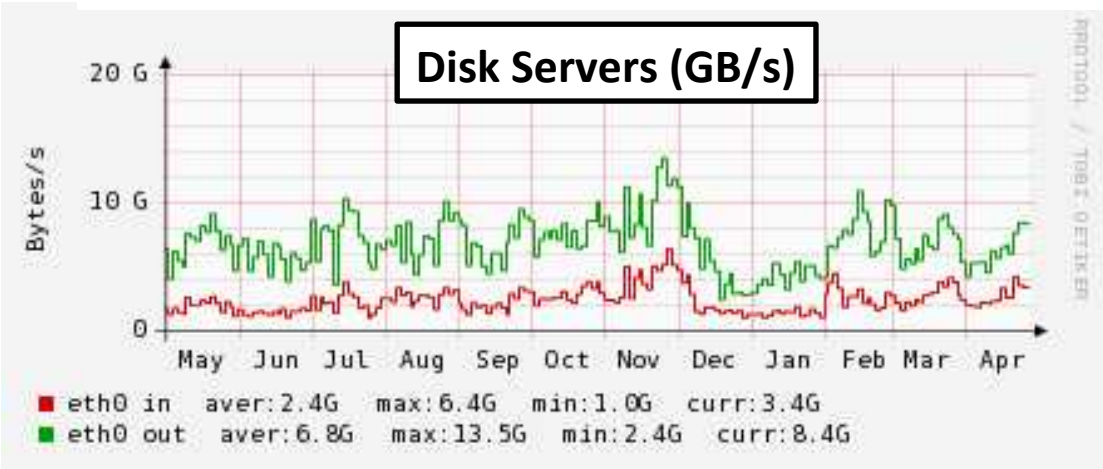


Stored ~ 15 PB in 2010

p-p data to tape at close to 2 PB/month

Peak rate: 225TB/day

LHCb
(compass)
CMS
ATLAS
ALICE



Tier 0 storage:

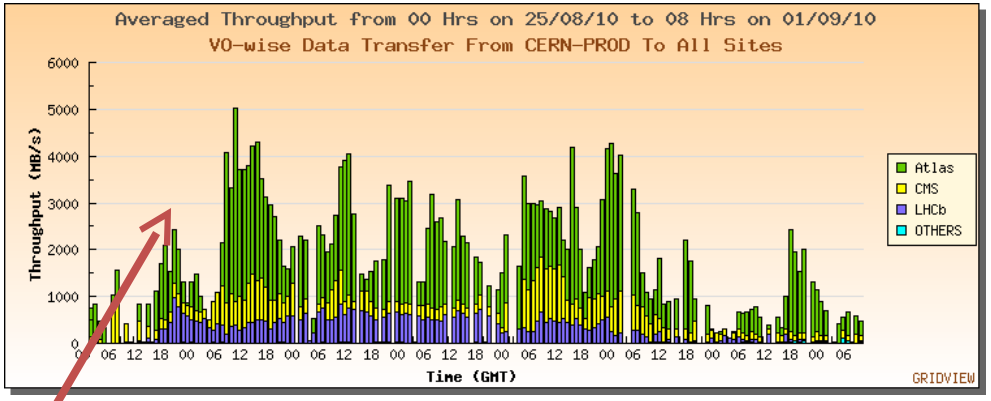
- Accepts data at average of 2.6 GB/s; peaks > 11 GB/s
- Serves data at average of 7 GB/s; peaks > 25 GB/s
- **CERN Tier 0 moves > 1 PB data per day**

Last year data transfers

LHC data transfers:
April 2010 – May 2011

Rates >> higher than planned/tested
Nominal: 1.3 GB/s
Achieved: up to 5 GB/s

World-wide: ~10 GB/s per large experiment



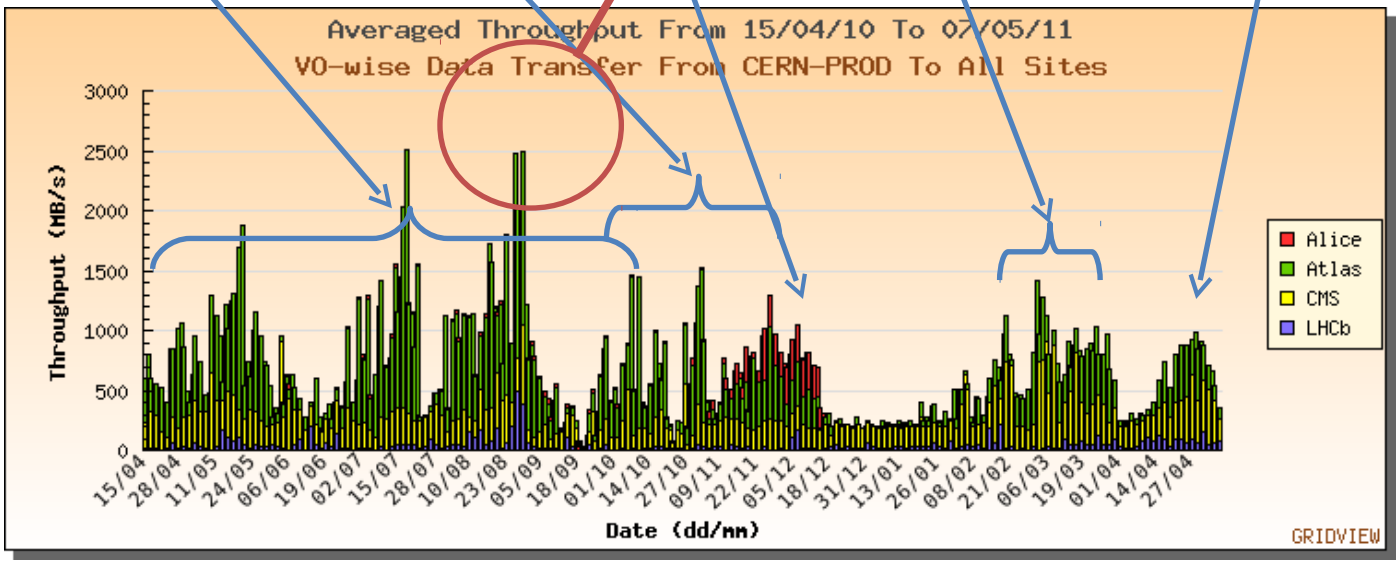
2010 pp data Tier 1s & re-processing

Re-processing 2010 data

ALICE HI data Tier 1s

CMS HI data zero suppression & FNAL

2011 data Tier 1s



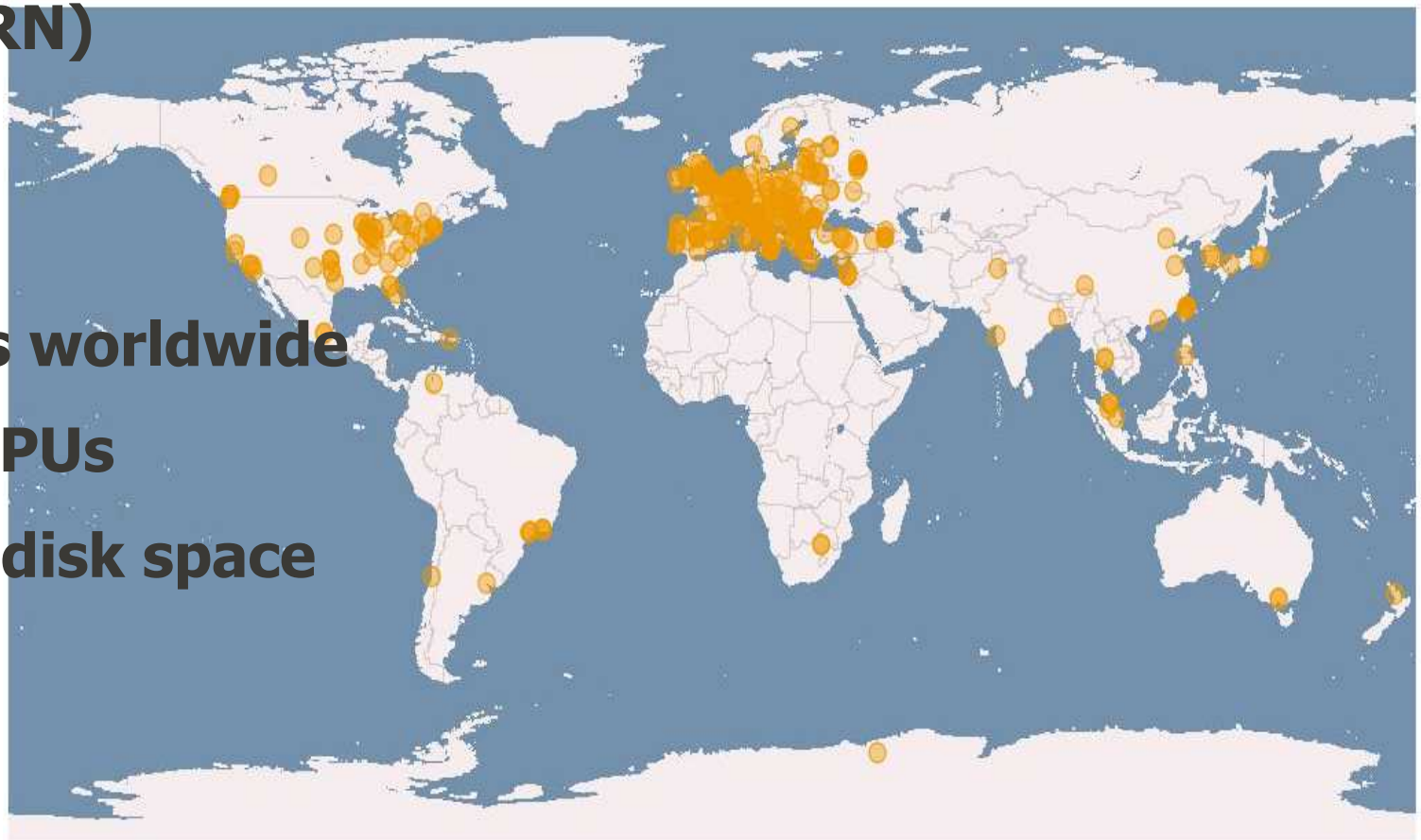
WLCG

Worldwide LHC Computing Grid

WLCG

WLCG sites:

- 1 Tier0 (CERN)
- 11 Tier1s
- 164 Tier2s
- >300 Tier3s worldwide
- ~250,000 CPUs
- ~ 100PB of disk space



CERN Tier0 resources

Servers	8,076
Processors	13,802
Cores	50,855
HEPSpec06	359,431

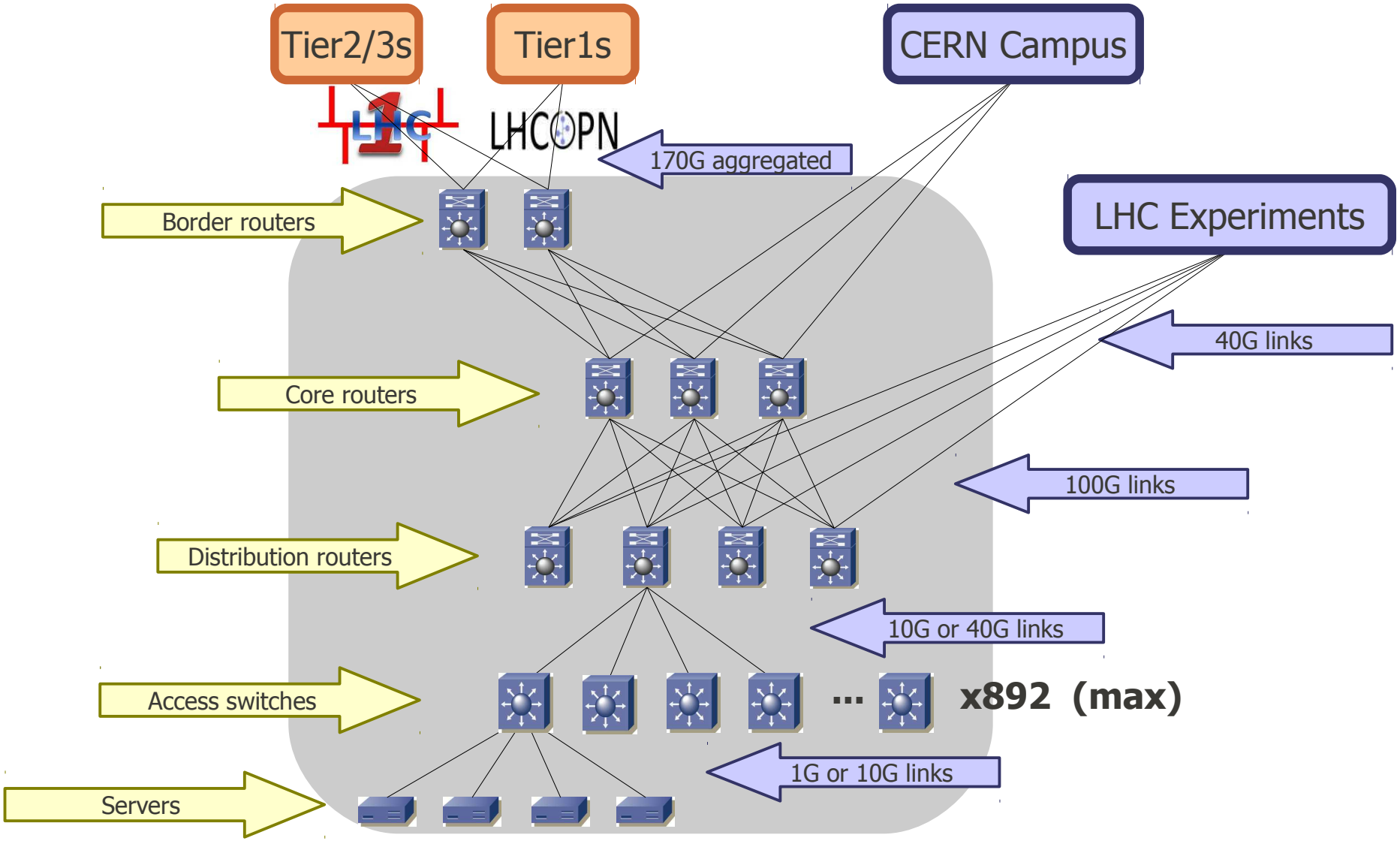
Disks	53,728
Raw disk capacity (TB)	45,331
Memory modules	48,794
RAID controllers	3,518

Tape Drives	160
Tape Cartridges	45000
Tape slots	56000
Tape Capacity (TB)	34000

High Speed Routers (6.4 Tbps)	9
Ethernet Switches	500
10 Gbps ports	3000
Switching Capacity	15.36 Tbps

Mars 2011

CERN Tier0 LCG new network



Trends

Virtualization mobility

Commodity Servers with 10G NICs

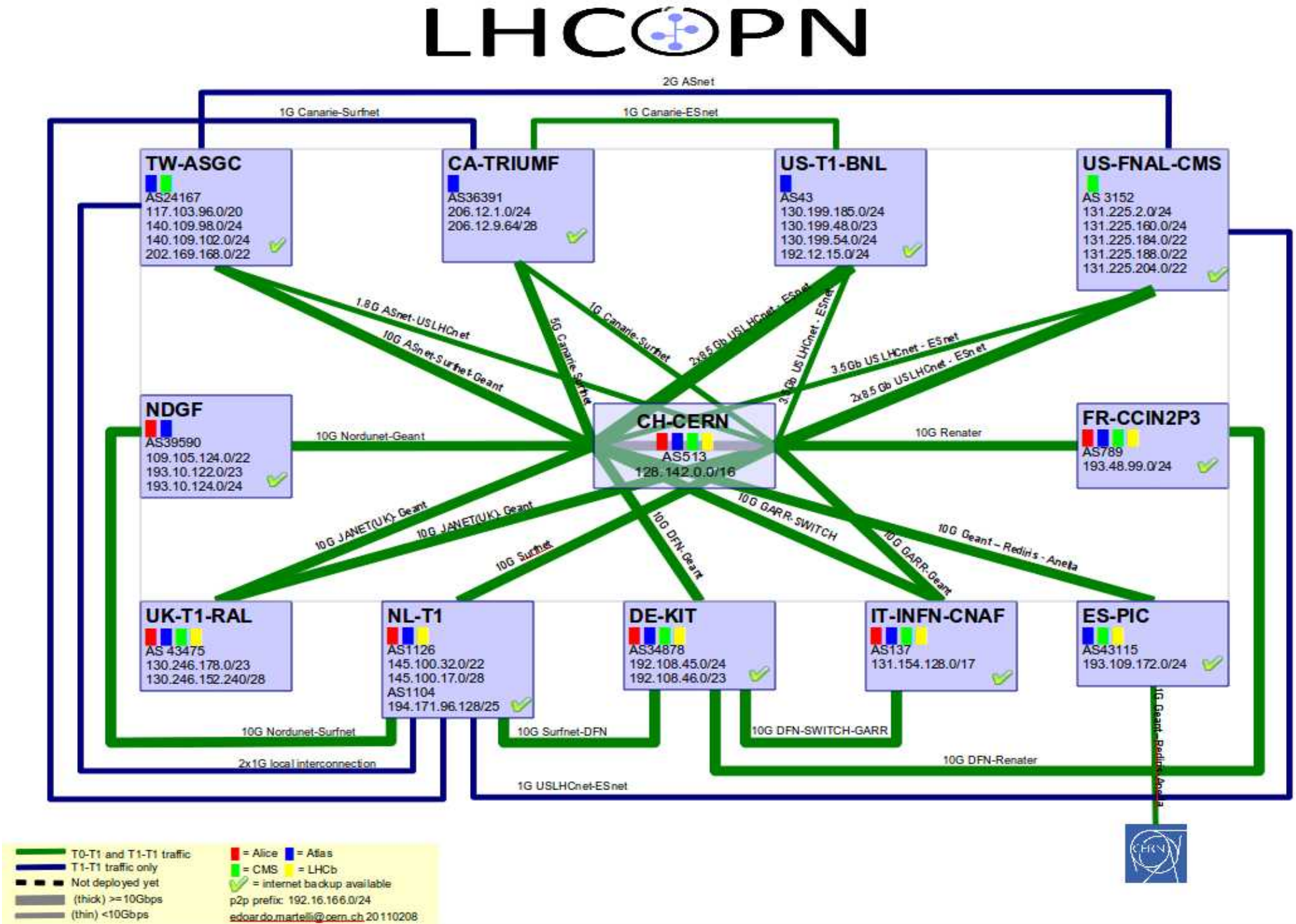
High-end Servers with 40G NICs

40G and 100G interfaces on switches and routers

LHCOPN

LHC Optical Private Network

Tier0-Tier1s network



A collaborative effort

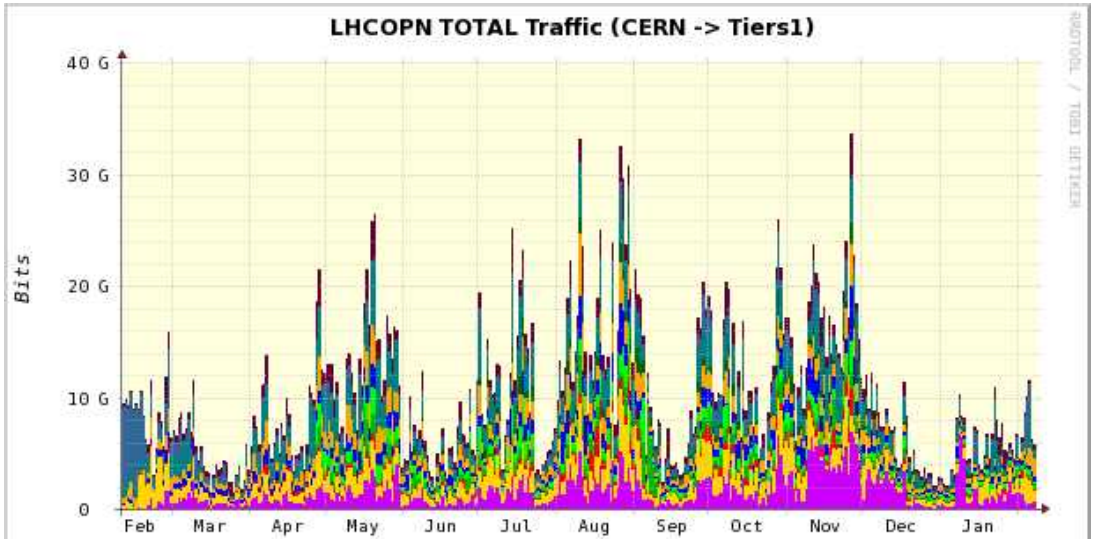
Designed, built and operated by the Tier0-Tier1s community

Links provided by the Research and Education network providers: Geant, USLHCnet, Esnet, Canarie, ASnet, Nordunet, Surfnet, GARR, Renater, JANET.UK, Rediris, DFN, SWITCH

Technology

- Single and bundled long distance 10G ethernet links
- Multiple redundant paths. Star+PartialMesh topology
- BGP routing: communities for traffic engineering, load balancing.
- QoS: T0-T1 traffic prioritized over T1-T1 traffic
- Security: only declared IP prefixes can exchange traffic.

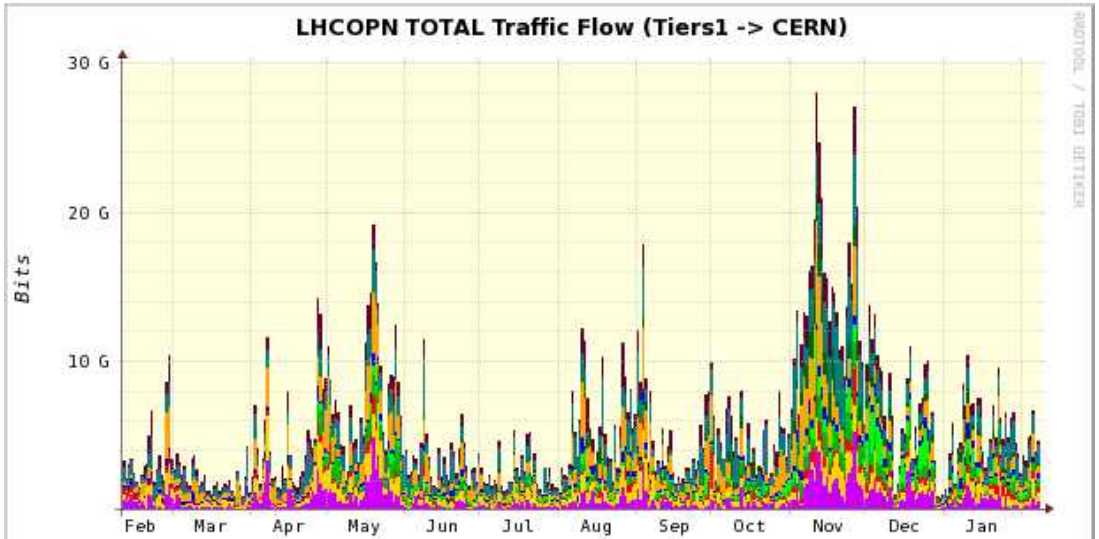
Traffic to the Tier1s



	Avg	Max	Peak	Curr
To DE-KIT	1.59 G	6.88 G	9.25 G	715.48 M
To IN2P3	1.47 G	5.58 G	9.46 G	974.59 M
To NDGF	328.26 M	2.71 G	6.86 G	185.13 M
To NLT1	635.37 M	6.22 G	9.46 G	375.54 M
To ASGC	754.37 M	4.11 G	7.93 G	275.32 M
To CNAF	797.86 M	4.98 G	N/A	464.67 M
To RAL	1.09 G	5.55 G	N/A	1.56 G
To TRIUMF	392.38 M	2.12 G	N/A	85.73 M
To BNL	1.60 G	6.81 G	N/A	542.31 M
To FNAL	889.71 M	7.43 G	N/A	216.81 M
To ES-PIC	876.92 M	4.21 G	N/A	332.33 M

Total to Tiers1: Avg 9.99 G Max 33.64 G Curr 5.72 G

Last update: Tue Feb 08 2011 13:27:49

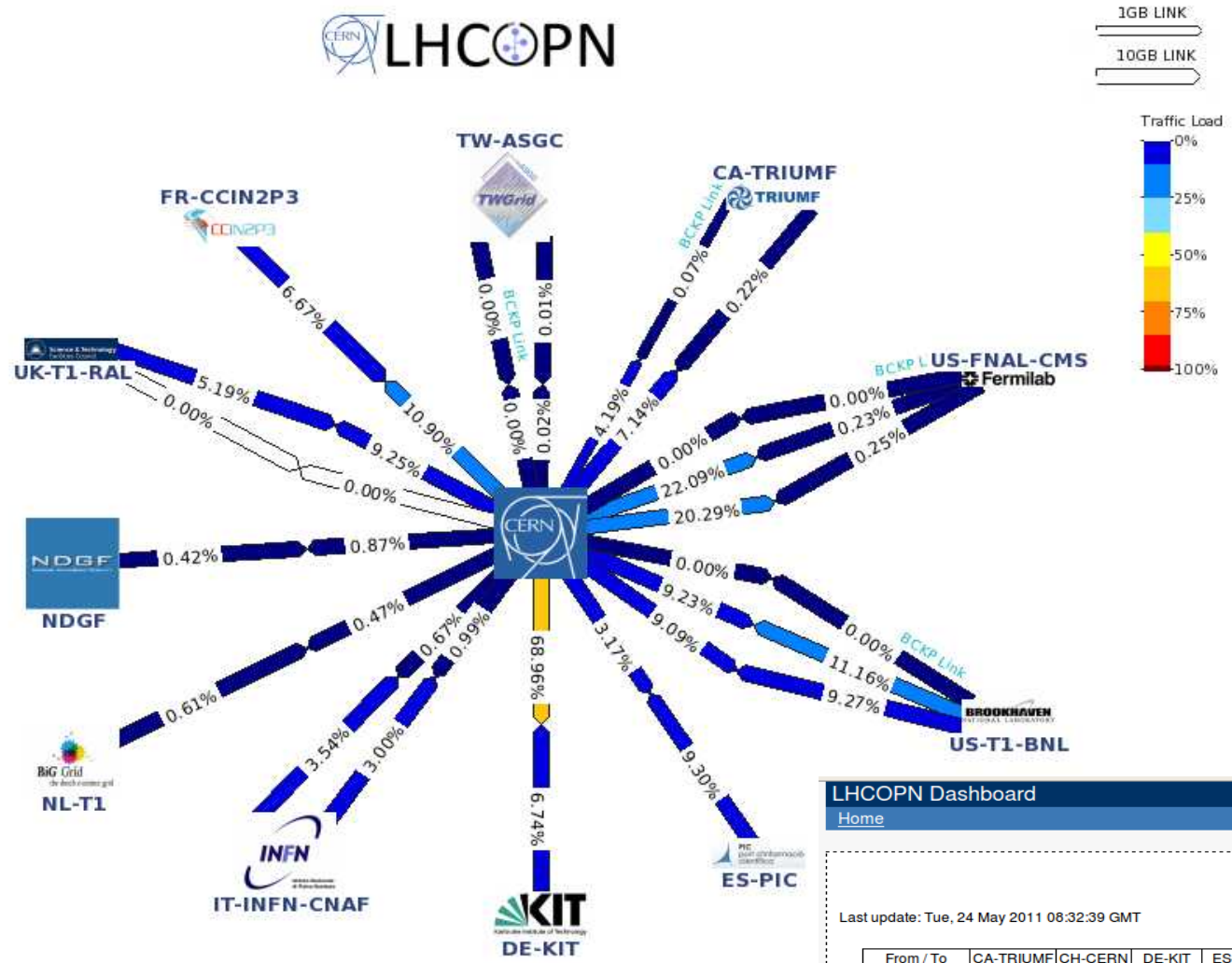


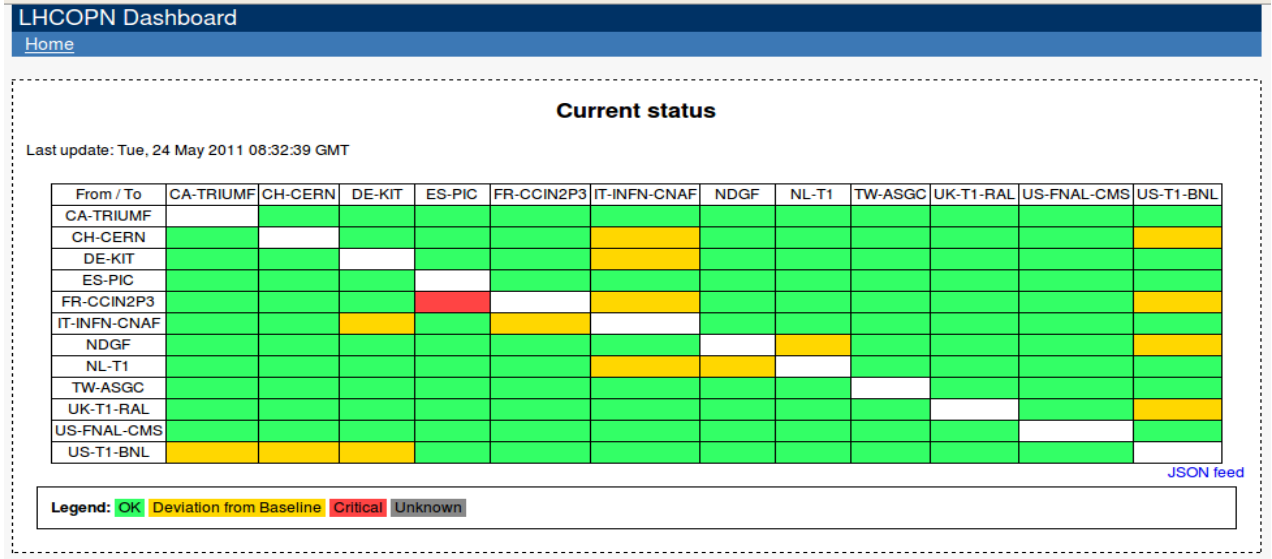
	Avg	Max	Peak	Curr
From DE-KIT	747.28 M	5.23 G	9.85 G	599.50 M
From IN2P3	612.35 M	3.53 G	8.11 G	616.14 M
From NDGF	237.68 M	2.66 G	7.34 G	323.12 M
From NLT1	447.93 M	2.14 G	7.68 G	515.64 M
From ASGC	371.52 M	2.98 G	6.89 G	113.53 M
From CNAF	232.78 M	1.10 G	N/A	314.21 M
From RAL	992.24 M	6.90 G	N/A	476.35 M
From TRIUMF	302.50 M	2.14 G	N/A	81.05 M
From BNL	905.30 M	5.00 G	N/A	593.45 M
From FNAL	330.77 M	2.32 G	N/A	758.72 M
From ES-PIC	599.30 M	4.06 G	N/A	198.82 M

Total from Tiers1: Avg 5.50 G Max 27.91 G Curr 4.59 G

Last update: Tue Feb 08 2011 13:27:49

Monitoring





LHCOPN Dashboard

Home

Current status

Last update: Tue, 24 May 2011 08:32:39 GMT

From / To	CA-TRIUMF	CH-CERN	DE-KIT	ES-PIC	FR-CCIN2P3	IT-INFN-CNAF	NDGF	NL-T1	TW-ASGC	UK-T1-RAL	US-FNAL-CMS	US-T1-BNL
CA-TRIUMF												
CH-CERN												
DE-KIT												
ES-PIC												
FR-CCIN2P3												
IT-INFN-CNAF												
NDGF												
NL-T1												
TW-ASGC												
UK-T1-RAL												
US-FNAL-CMS												
US-T1-BNL												

JSON feed

Legend: OK | Deviation from Baseline | Critical | Unknown

LHCONE

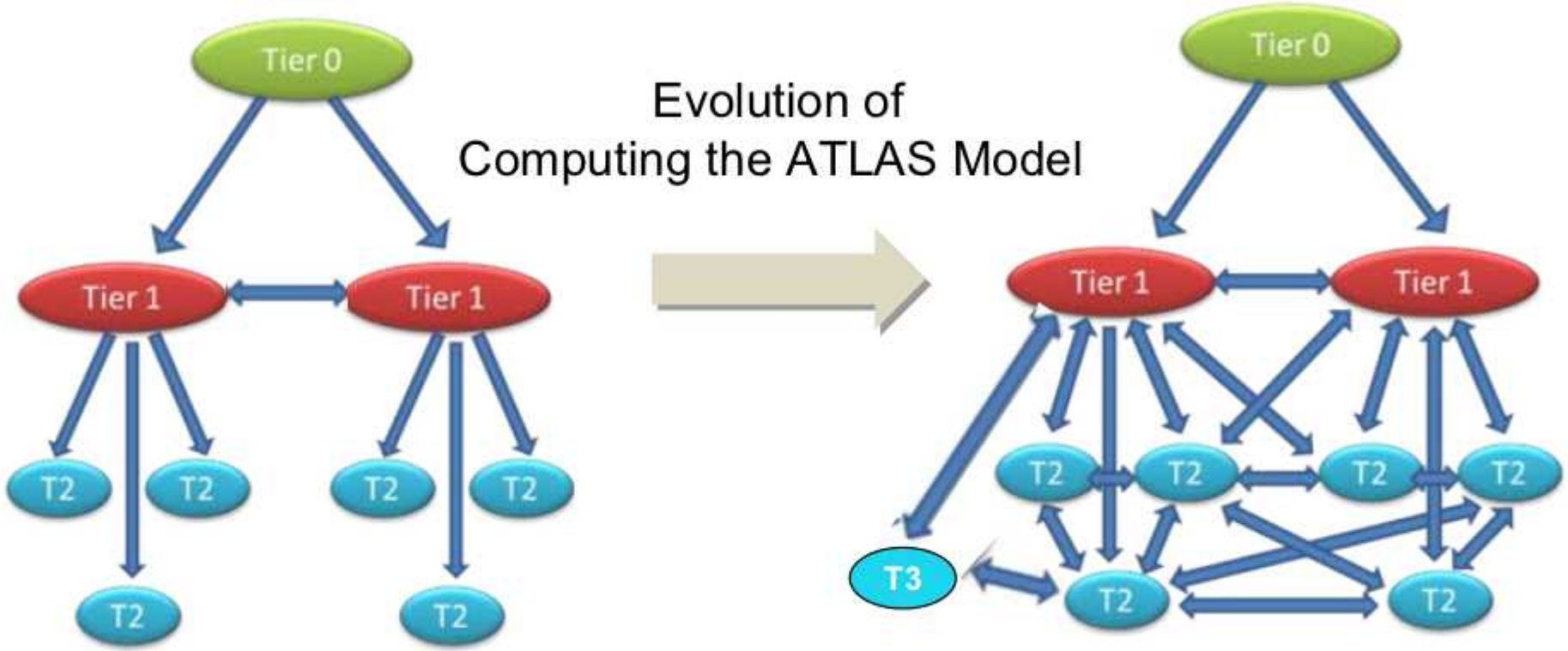
LHC Open Network Environment

Driving the change

“The Network infrastructure is the most reliable service we have”

Ian Bird, WLCG project leader

Change of computing model (ATLAS)



New computing model

"Network Bandwidth (rather than disk) will need to scale more with users and data volume"

"Data placement will be driven by demand for analysis and not pre-placement"

Ian Bird, WLCG project leader

New computing model

- Better and more dynamic use of storage
- Reduce the load on the Tier1s for data serving
- Increase the speed to populate analysis facilities

Needs for a faster, predictable, pervasive network connecting Tier1s and Tier2s

Requirements from the Experiments

- Connecting any pair of sites, regardless of the continent they reside
- Bandwidth ranging from 1Gbps (Minimal), 5Gbps (Nominal), 10G and above (Leadership)
- Scalability: sites are expected to grow
- Flexibility: sites may join and leave at any time
- Predictable cost: well defined cost, and not too high

Needs for a better network

- more bandwidth by federating (existing) resources
- sharing cost of expensive resources
- accessible to any TierX site

=

LHCONE

LHC Open Network Environment

LHCONE concepts

- Serves any LHC sites according to their needs and allowing them to grow

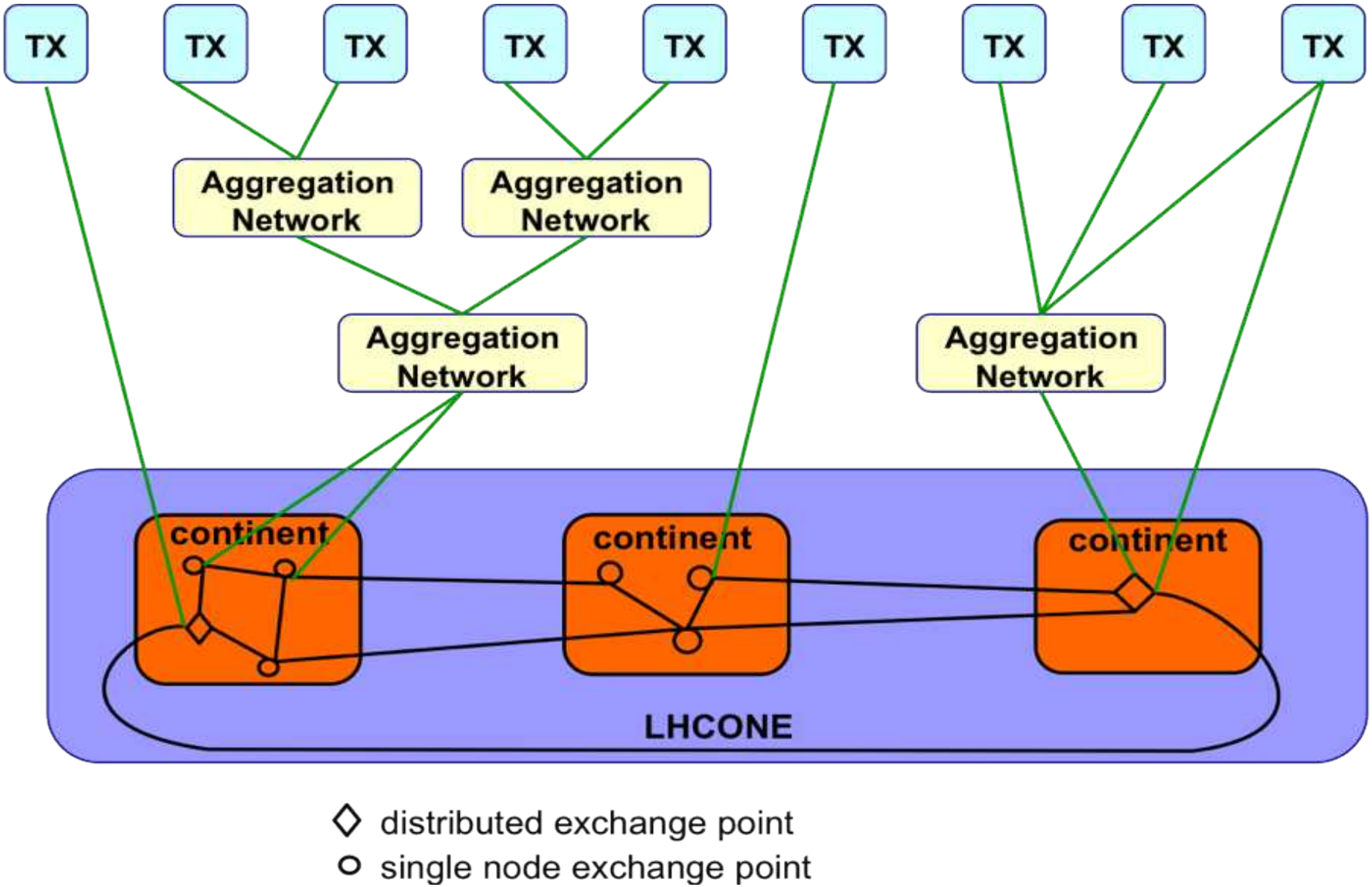
- A collaborative effort among Research & Education Network Providers

- Based on Open Exchange Points: easy to join, neutral

- Multiple services: one cannot fit all

- Traffic separation: no clash with other data transfer, resource allocated for and funded by HEP community

LHCONE architecture

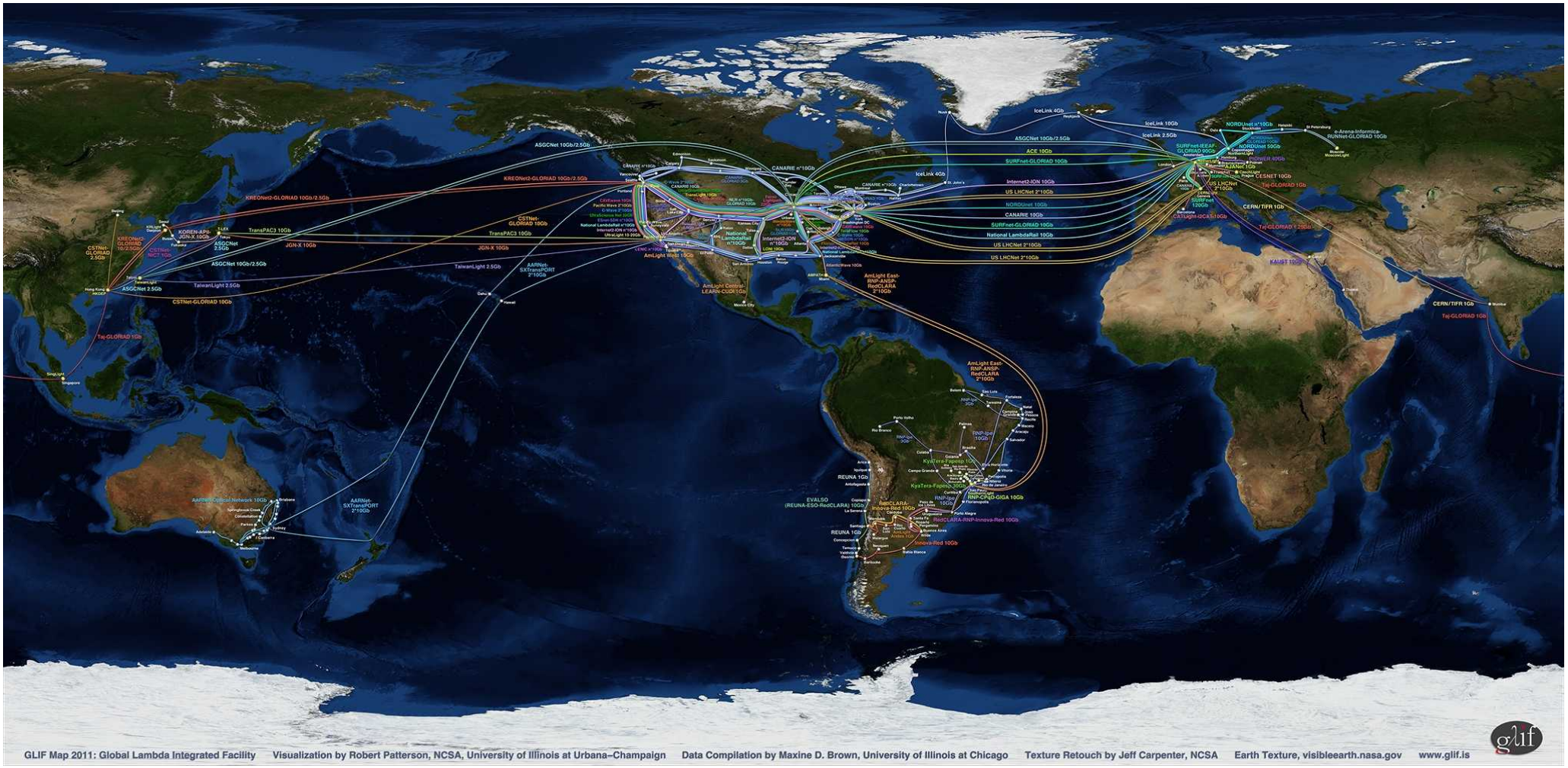


LHCONE building blocks

Based upon these building blocks:
- Single node exchange points
- Continental/regional distributed exchange points
- Interconnect circuits between exchange points

LHCONE is made up of the combination of exchange points and distributed exchange points. These exchange points, and the links in between, collectively provide LHCONE services and operate under a common LHCONE policy

The underlying infrastructure



LHCONE services

- Shared VLAN
- Dedicated VLANs
- Lightpaths
- Monitoring

Service: Shared VLAN

- A single VLAN reaching all the locations
- Any TierX can join the shared VLAN and get IPv4 and IPv6 addresses from common subnets: any-to-any reachability
- Routing policies up to the TierX
- Route Server service available to simplify routing configuration (one server per continent)

Based on major Internet Exchange Point model (AMSIX, DECIX...)

Service: Dedicated VLANs

- Layer 2 VLANs connecting a restricted number of TierXs
- More secure
- No guaranteed bandwidth

Service: Lightpaths

- Point-to-point links connecting pair of TierXs
- Guaranteed bandwidth
- Dynamically provisioned

Service: Monitoring

- A distributed monitoring system to ensure the healthiness of the system
- To be defined yet

Governance

- LHCONE is a community effort
- All stakeholders involved: TierXs, Network Operators, LHC Experiments, CERN.
- Exact roles and responsibilities not yet defined

Challenges for LHCONE Operators

- Coordination among “competitors”

- Ensure stability, reliability and performance of a large system not centrally controlled

- Develop a common provisioning system

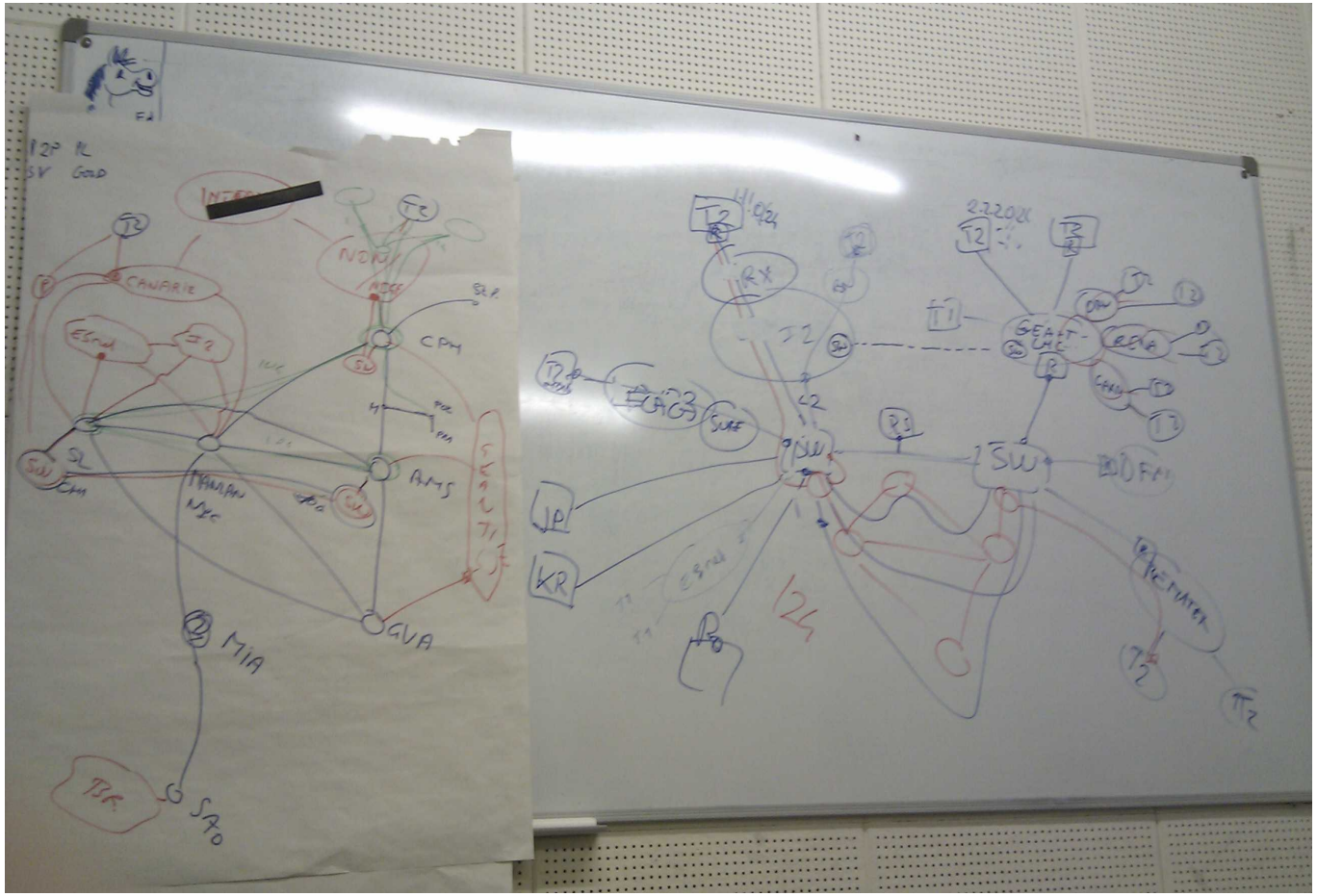
Opportunities

- Raise awareness of networking needs at TierXs
- More capacity to be provided by Network Operators
- Foster collaborations among Network Operators and among Network Users
- An application for already developed solutions
- New technologies to apply

On going

- Few prototypes are taking shape

- Soon time to stitch them together and start production



Openlab project: CINBAD

CINBAD

CERN **I**nvestigation of **N**etwork **B**ehaviour and **A**nomaly **D**etection

Project Goal

Understand the behaviour of large computer networks (10’000+ nodes) in High Performance Computing or large Campus installations to be able to:

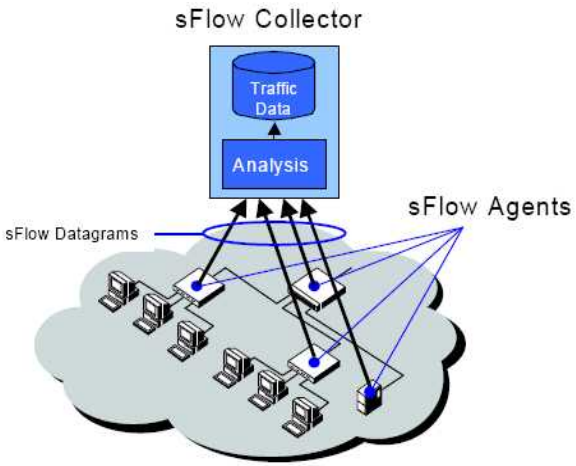
- detect traffic anomalies in the system
- perform trend analysis
- automatically take counter measures
- provide post-mortem analysis facilities

Data source: sFlow

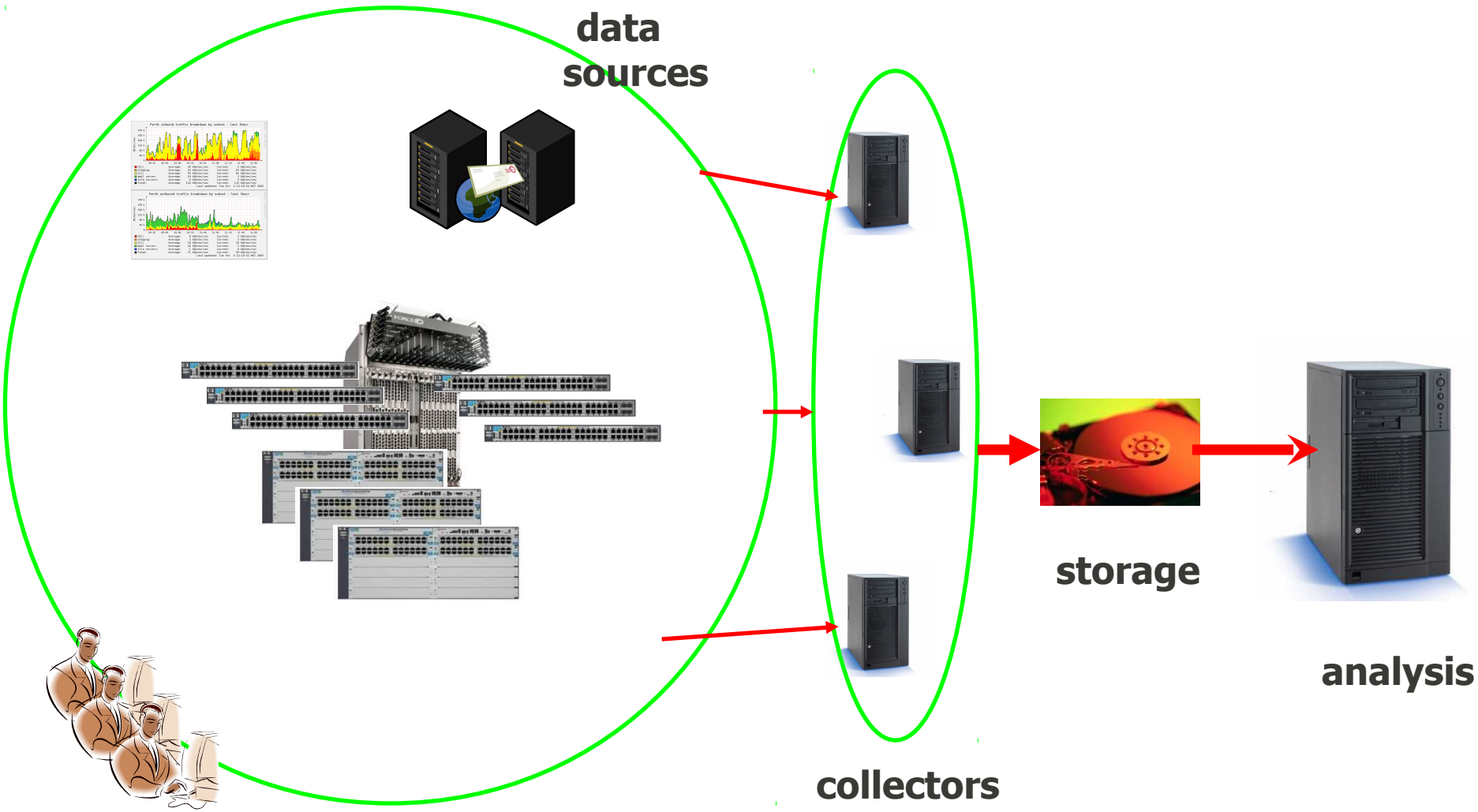
Based on packet sampling (RFC 3176)

1-out-of-N packet is sampled by an agent and sent to a collector

- packet header and payload included (max 128 bytes)
- switching/routing/transport protocol information
- application protocol data (e.g. http, dns)
- SNMP counters included



CINBAD Architecture

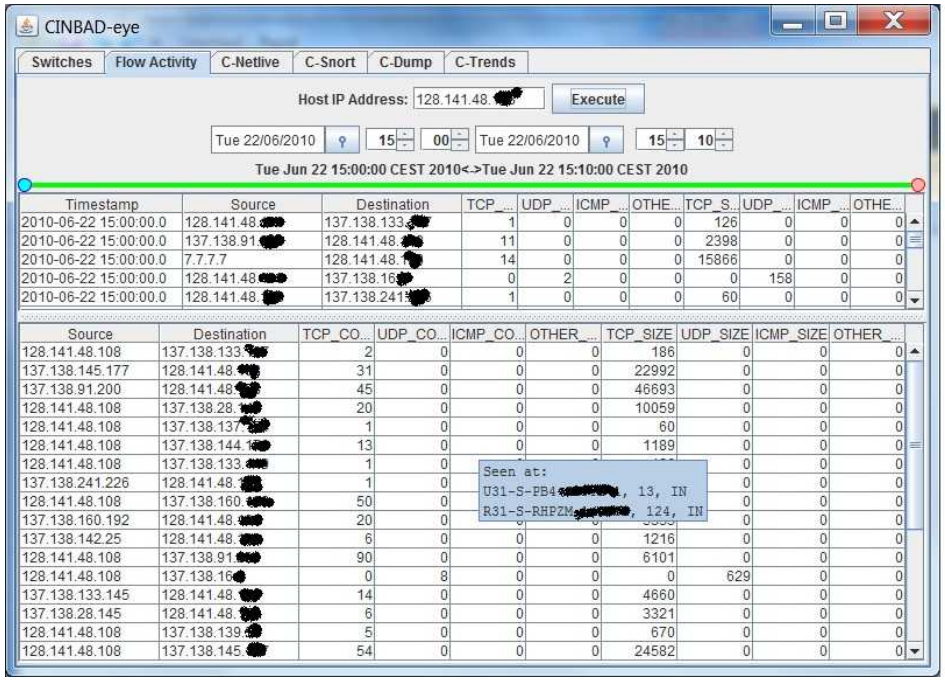


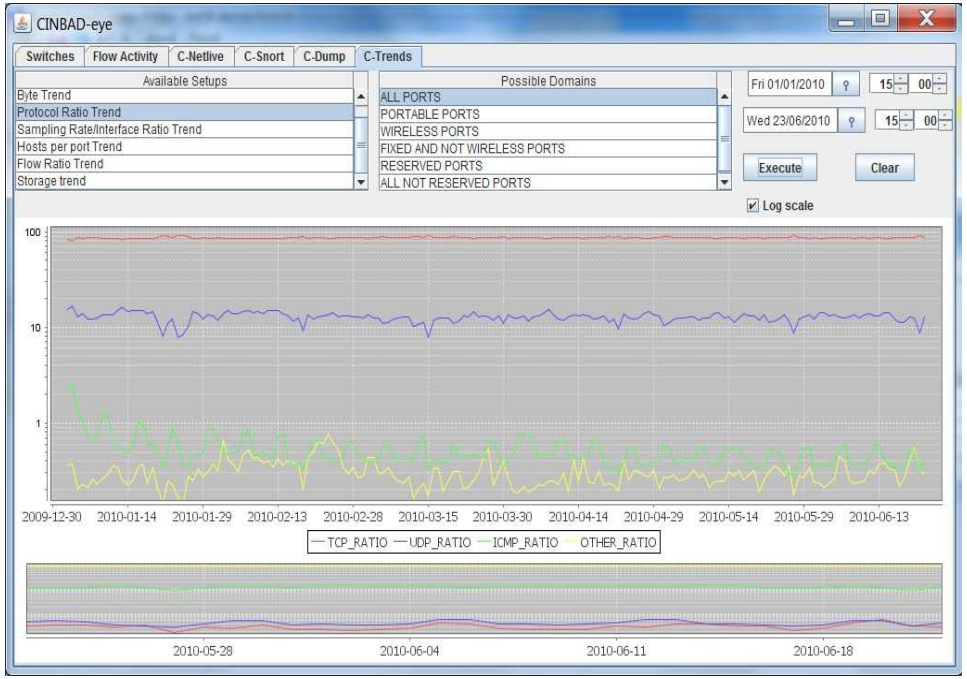
sFlow data collection

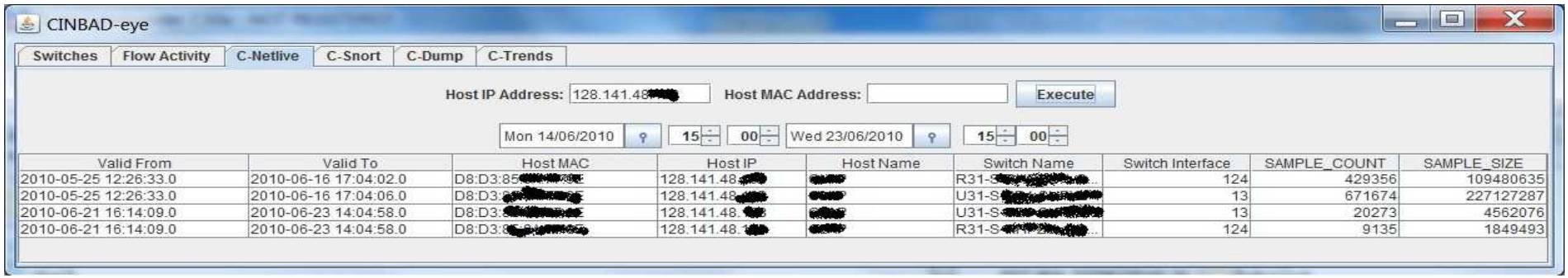
Current collection based on traffic from ~1000 switches
- 6000 sampled packets per second
- 3500 snmp counter sets per second
- 100GB per day

CINBAD-eye

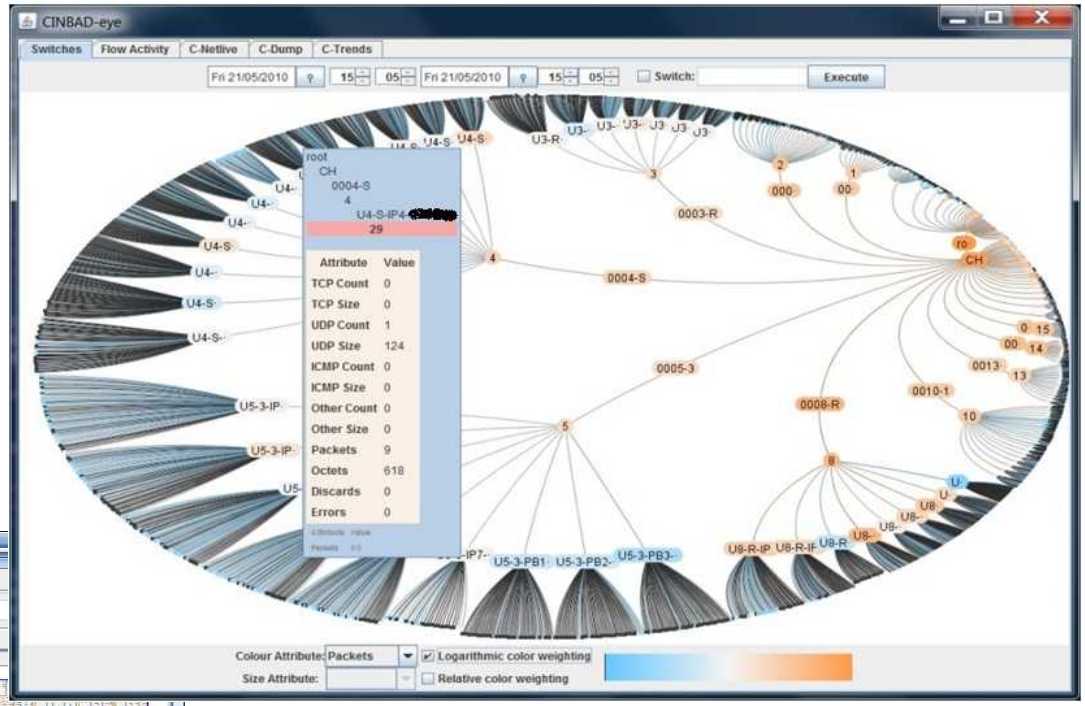
- Host activity and connectivity
- Traffic trends

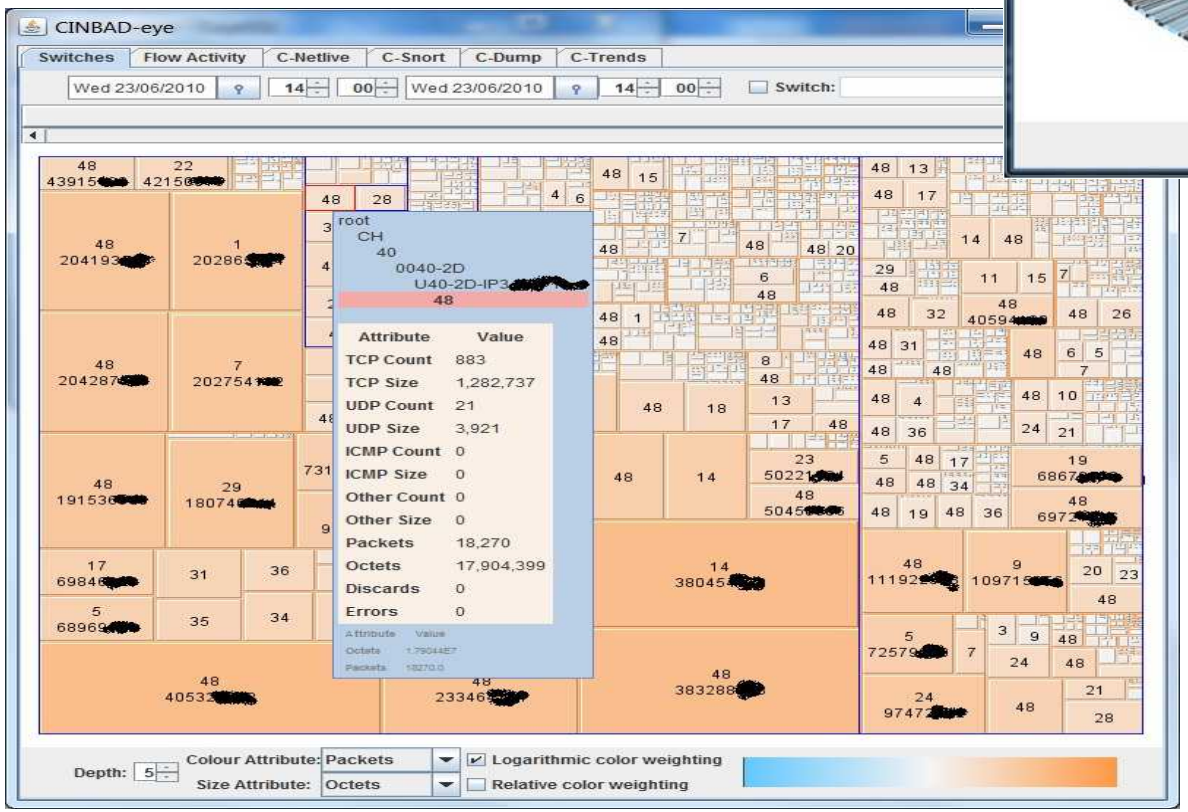






CINBAD-eye


CINBAD-eye
Switches
Flow Activity
C-Netlive
C-Dump
C-Trends
Fri 21/05/2010
15
05
Fri 21/05/2010
15
05
Switch:
Execute
root
CH
0004-S
4
U4-S-IP4-
29
Attribute Value
TCP Count 0
TCP Size 0
UDP Count 1
UDP Size 124
ICMP Count 0
ICMP Size 0
Other Count 0
Other Size 0
Packets 9
Octets 618
Discards 0
Errors 0
Colour Attribute: Packets
Size Attribute:
Logarithmic color weighting
Relative color weighting


CINBAD-eye
Switches
Flow Activity
C-Netlive
C-Snort
C-Dump
C-Trends
Wed 23/06/2010
14
00
Wed 23/06/2010
14
00
Switch:
root
CH
40
0040-2D
U40-2D-IP3-
48
Attribute Value
TCP Count 883
TCP Size 1,282,737
UDP Count 21
UDP Size 3,921
ICMP Count 0
ICMP Size 0
Other Count 0
Other Size 0
Packets 18,270
Octets 17,904,399
Discards 0
Errors 0
Depth: 5
Colour Attribute: Packets
Size Attribute: Octets
Logarithmic color weighting
Relative color weighting

Anomaly detection

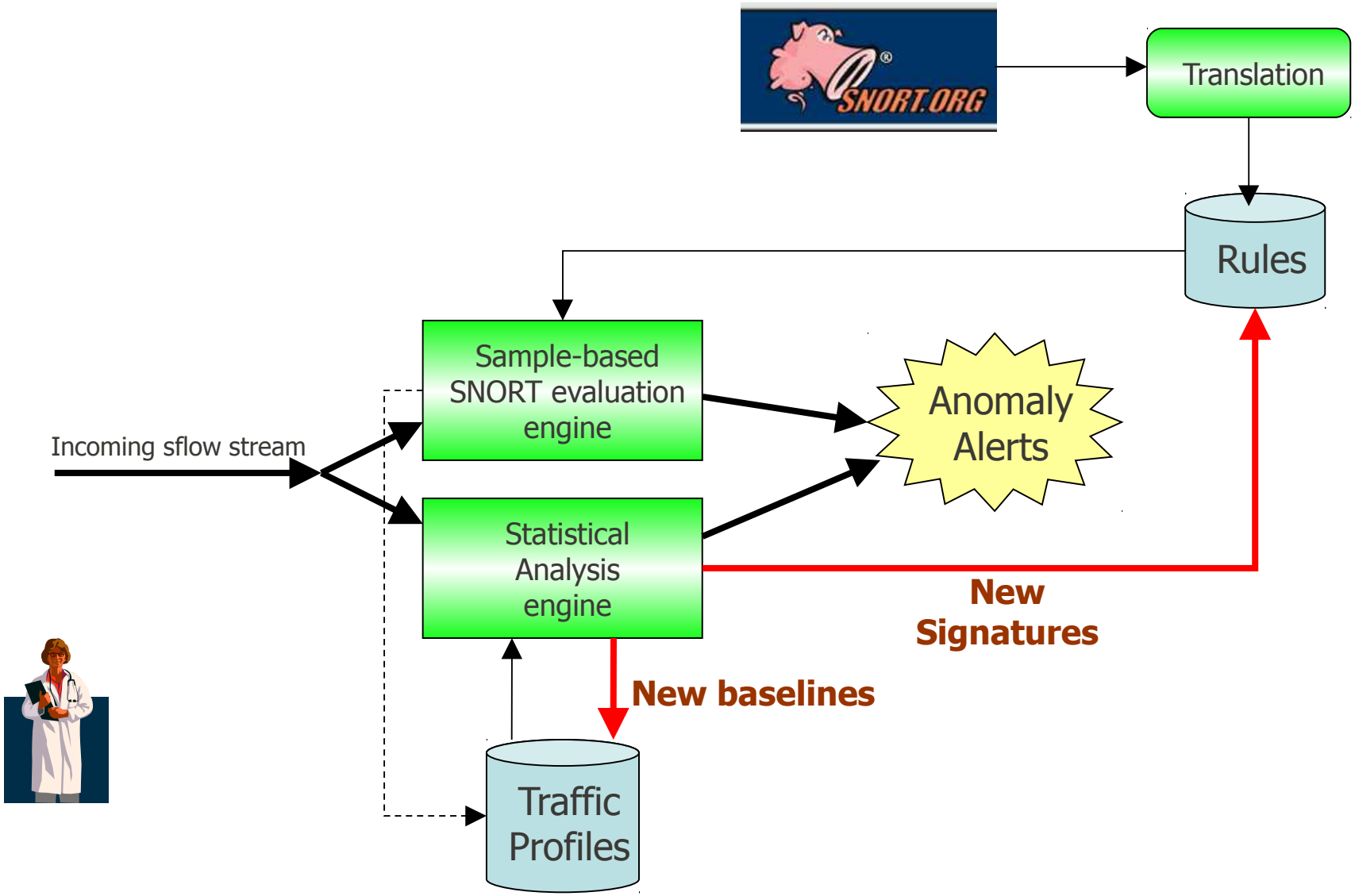
Statistical analysis methods

- detect a change from "normal network behavior"
- can detect new, unknown anomalies
- poor anomaly type identification

Signature based

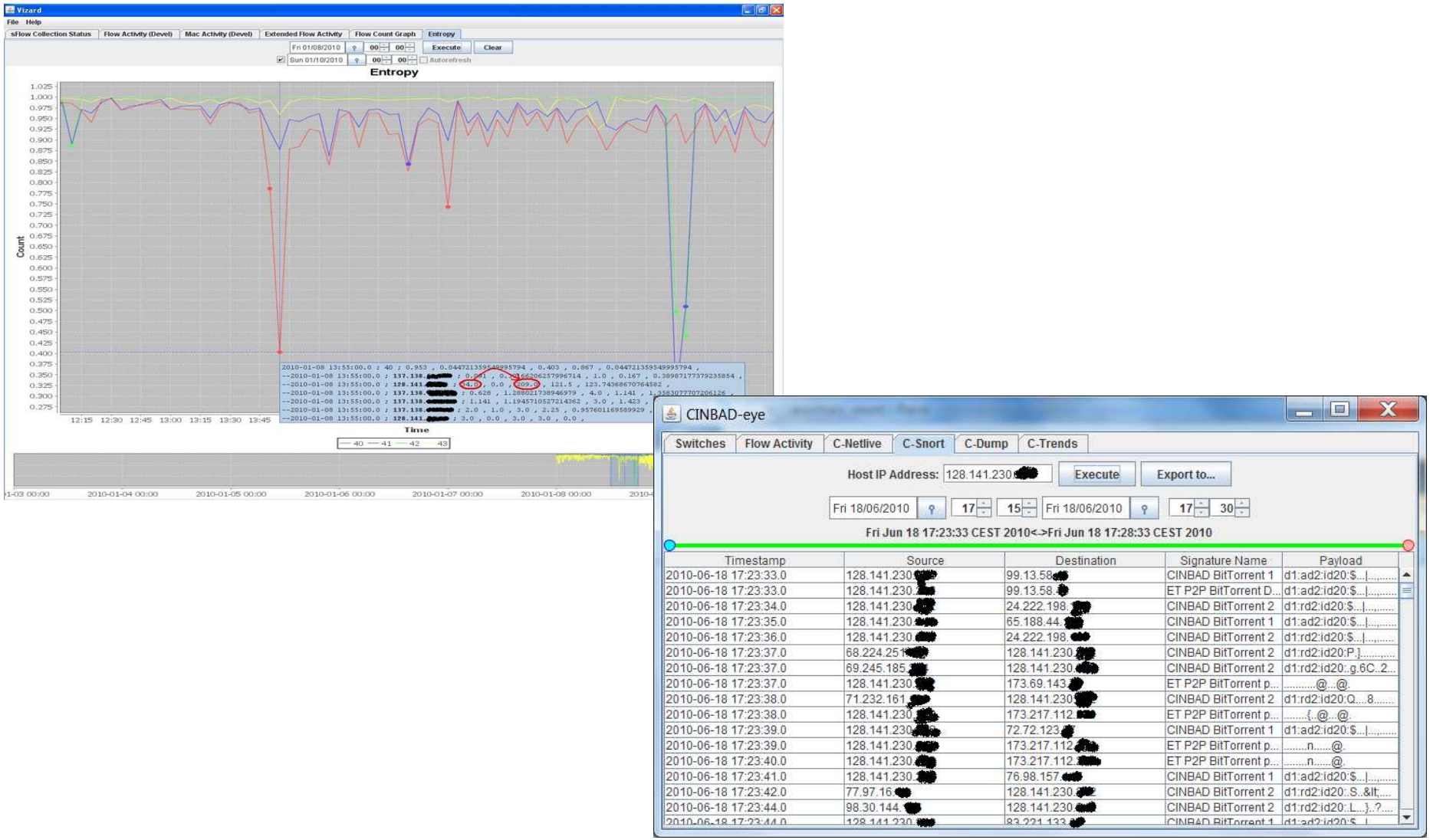
- SNORT ported to work with sampled data
- performs well against known problems
- tends to have low false positive rate
- does not work against unknown anomalies

Synergy from flows and signatures



Statistics and Signatures

Statistical and signature-based anomaly detection



Openlab project: WIND

WIND

Wireless **I**nfrastructure **N**etwork **D**eployment

Project Goals

- Analyze the problems of large scale wireless deployments and understand the constraint
- Simulate behaviour of WLAN
- Develop new optimisation algorithms
- Verify them in the real world
- Improve and refine the algorithms
- Deliver algorithms, guidelines, solutions

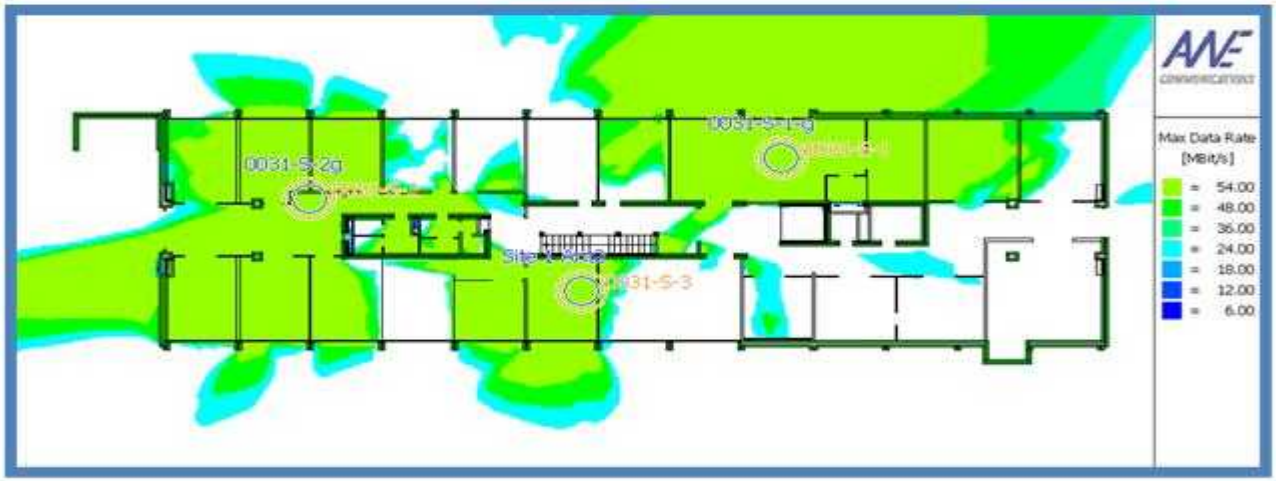
Needs

WLAN deployments are problematic

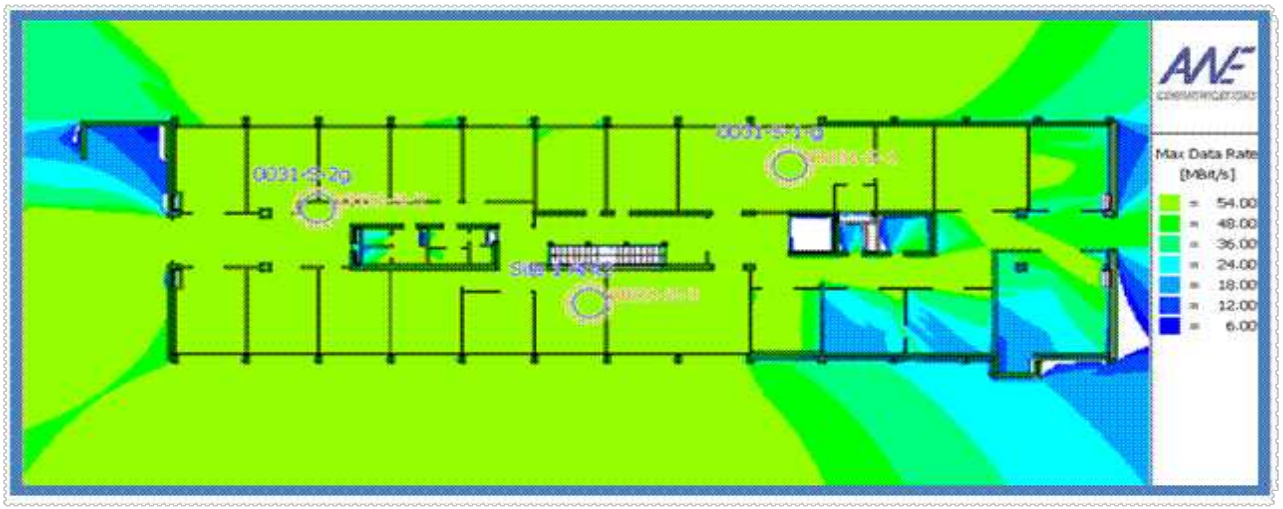
- Radio propagation is very difficult to predict
- Interference is an ever present danger
- WLANs are difficult to properly deploy
- Monitoring was not an issue when the first standards were developed
- When administrators are struggling just to operate the WLAN, performance optimisation is often forgotten

Example: Radio interferences

Max data rate in 0031-S: The APs work on the same channel



Max data rate in 0031-S: The APs work on 3 independent channels



Next steps

Extend monitoring and analysis tools

Act on the network
- smart load balancing
- isolating misbehaving clients
- intelligent minimum data rates

More accurate troubleshooting

Improve future network design

Conclusions

Conclusions

- The Data Network is an essential component of the LHC instrument
- The Data Network is a key part of the LHC data processing and will become even more important
- More and more security and design challenges

What's next

SWAN: Space Wide Area Network :-)

Credits

Ryszrard Jurga (CINBAD)

Milosz Hulboj (WIND)

Sebastien Ceuterickx (WIND)

Vlad.Lapadatescu (WIND)

Artur Barczyk (LHCONE)

Thank you