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
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
CERN
European Organization for Nuclear Research
CERN accelerator complex
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

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
- 1000000 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
LHC Data Challenge
Collisions in the LHC

Proton - Proton  
Protons/bunch  $10^{11}$
Beam energy  7 TeV ($7 \times 10^{12}$ eV)
Luminosity  $10^{34}$ cm$^{-2}$ s$^{-1}$

Crossing rate  40 MHz

Collision rate  $10^7$-$10^9$

New physics rate  $\approx \,00001$ 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

4 Experiments

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

\[
\sigma_{\text{eff}} \approx \sigma_{\text{eff}}^0 \times \frac{s \Gamma_Z^2}{(s - m_Z^2)^2 + s^2 \Gamma_Z^2 / m_Z^2} \quad \text{with}
\]

\[
\sigma_{\text{eff}}^0 = \frac{12 \pi \Gamma_{\text{ee}} \Gamma_{\text{ff}}}{m_Z^2 \Gamma_Z^2}
\]

and

\[
\Gamma_{\text{ff}} = \frac{G_F m_Z^3}{6 \pi \sqrt{2}} \times (v^2 + a^2)
\]
- 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

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

Stored ~ 15 PB in 2010

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

Peak rate: 225TB/day

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

2 PB/month

HI

2010 Reprocessing

LHCb (compass)
CMS
ATLAS
ALICE

Disk Servers (GB/s)
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
WLCG
Worldwide LHC Computing Grid
WLCG sites:

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>8,076</td>
</tr>
<tr>
<td>Processors</td>
<td>13,802</td>
</tr>
<tr>
<td>Cores</td>
<td>50,855</td>
</tr>
<tr>
<td>HEPSpec06</td>
<td>359,431</td>
</tr>
<tr>
<td>Tape Drives</td>
<td>160</td>
</tr>
<tr>
<td>Tape Cartridges</td>
<td>45000</td>
</tr>
<tr>
<td>Tape slots</td>
<td>56000</td>
</tr>
<tr>
<td>Tape Capacity (TB)</td>
<td>34000</td>
</tr>
<tr>
<td>Disks</td>
<td>53,728</td>
</tr>
<tr>
<td>Raw disk capacity (TB)</td>
<td>45,331</td>
</tr>
<tr>
<td>Memory modules</td>
<td>48,794</td>
</tr>
<tr>
<td>RAID controllers</td>
<td>3,518</td>
</tr>
<tr>
<td>High Speed Routers (6.4 Tbps)</td>
<td>9</td>
</tr>
<tr>
<td>Ethernet Switches</td>
<td>500</td>
</tr>
<tr>
<td>10 Gbps ports</td>
<td>3000</td>
</tr>
<tr>
<td>Switching Capacity</td>
<td>15.36 Tbps</td>
</tr>
</tbody>
</table>

Mars 2011
CERN Tier0 LCG new network

- Tier2/3s
- Tier1s
- CERN Campus
- LHC Experiments

- Border routers
- Core routers
- Distribution routers
- Access switches
- Servers

- 170G aggregated
- 40G links
- 100G links
- 10G or 40G links
- 1G or 10G links

- x892 (max)
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
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
- 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

LHCOPN TOTAL Traffic (CERN -> Tier1s)

LHCOPN TOTAL Traffic Flow (Tier1s -> CERN)

Total to Tier1s: Avg 9.99 G, Max 33.64 G, Curr 5.72 G
Last update: Tue Feb 08 2011 13:27:49

Total from Tier1: Avg 5.59 G, Max 27.91 G, Curr 4.59 G
Last update: Tue Feb 08 2011 13:27:49
Monitoring

LHCOPN

FR-CCIN2P3
TW-ASGC
CA-TRIUMF
US-FNAL-CMS

NDGF
NL-T1

INFN
IT-INFN-CNAF

DE-KIT
ES-PIC

CERN

Traffic Load
0%
25%
50%
75%
100%

3.52%
0.00%
0.00%
0.00%
0.00%
0.00%
0.00%
0.00%

US-T1-BNL

LHCOPN Dashboard

Current status

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

Legend: OK Deviation from Baseline Critical Unknown

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>CA-TRIUMF</th>
<th>CH-CERN</th>
<th>DE-KIT</th>
<th>ES-PIC</th>
<th>FR-CCIN2P3</th>
<th>IT-INFN-CNAF</th>
<th>NDGF</th>
<th>NL-T1</th>
<th>TW-ASGC</th>
<th>US-FNAL-CMS</th>
<th>US-T1-BNL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Driving the change

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

Ian Bird, WLCG project leader
Change of computing model (ATLAS)

Evolution of Computing the ATLAS Model
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

- TX
- Aggregation Network
- TX
- Aggregation Network
- TX
- Aggregation Network
- TX
- Aggregation Network
- TX
- Aggregation Network

- continent
- continent
- continent

- diamond: distributed exchange point
- circle: single node exchange point
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
CERN Investigation of Network Behaviour and Anomaly Detection

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

data sources

collectors

storage

analysis
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
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

- Sample-based SNORT evaluation engine
- Statistical Analysis engine
- Traffic Profiles
- New baselines
- New Signatures
- Anomaly Alerts
- Rules
- Translation
Statistics and Signatures

Statistical and signature-based anomaly detection
Openlab project: WIND
Wireless Infrastructure Network Deployment

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
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

Ryszard Jurga (CINBAD)
Milosz Hulboj (WIND)
Sebastien Ceuterickx (WIND)
Vlad.Lapadatescu (WIND)
Artur Barczyk (LHCONE)
Thank you