



Fault Isolation and service quality assurance

in a 10GbE redundant grid infrastructure

Nikos Trikoupis


Infrastructure and Operations

CERN IT/CS

Agenda

- Requirements and challenges for Monitoring in the CERN network
- Fault Isolation essentials
- Service Quality Assurance

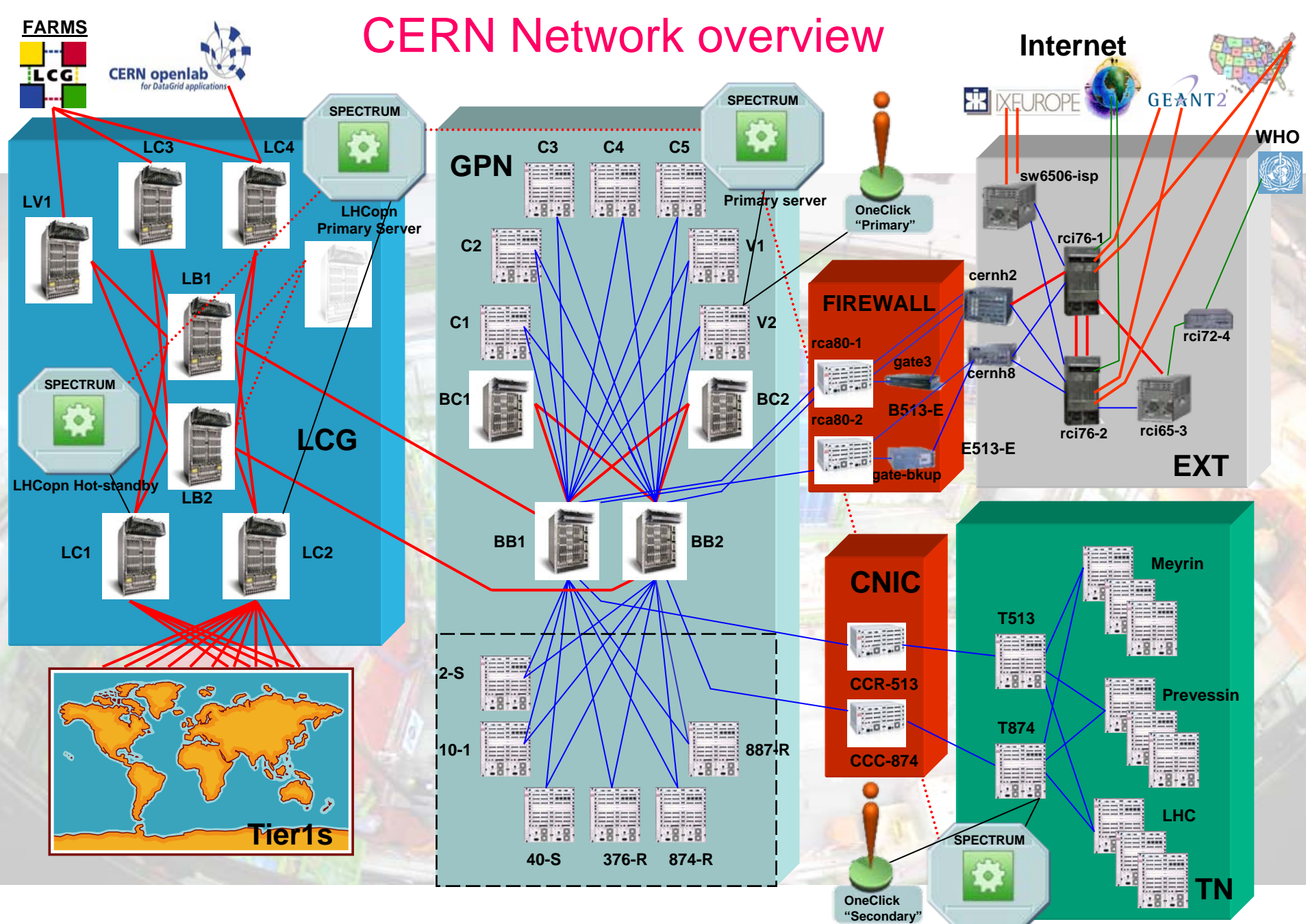
Introduction

- 
- CERN's “business processes” rely on network services availability.
 - Our mission is to deliver and manage an infrastructure **reliable** but **capable of sustaining a high rate of change**.
 - Being the LCG Tier-0 network provider is a huge responsibility.

Facts and Challenges

- A collaborative environment with highly complex applications
- Network redesign: multi-10GbE core, 10GbE to the farms
- The 10G WAN PHY standard allows for WAN connectivity at LAN speeds and the use of the same management tools
- For the first time, the barrier between Campus and WAN disappears.

CERN Network overview



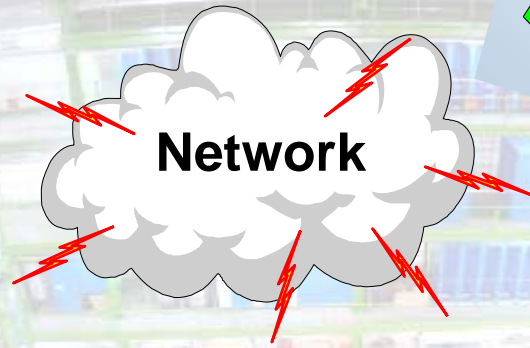
Pressures on Network Management

Increased Demand

More and Higher-Speed Bandwidth Choices



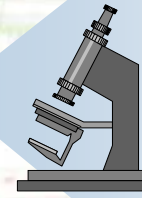
Network



Chaotic, Unpredictable Traffic Patterns



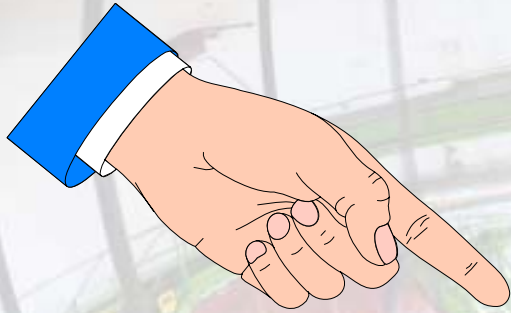
Restrictions in budget and personnel



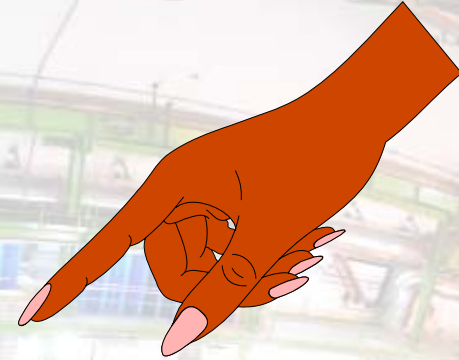
New Types of network and user equipment



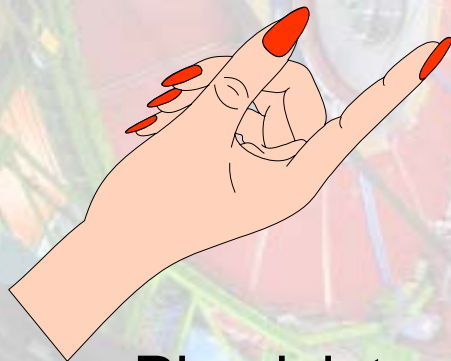
Managing the Network Foundation



Technical Services



System Managers



Physicists

**It's the
Network's Fault!**



Application Developers

Defending the Network

Faults may occur, but:

- They have to be detected (before users do it) as quickly as possible.
- The cause of the fault has to be identified so that corrective action may be taken.
- This task has to be performed by operations on a 24x7 basis.
- Time To Repair must be reduced
- Prioritize faults based on **impact**

The size and complexity of the network infrastructure dictates the use of automated network management tools

Network Root Cause Analysis

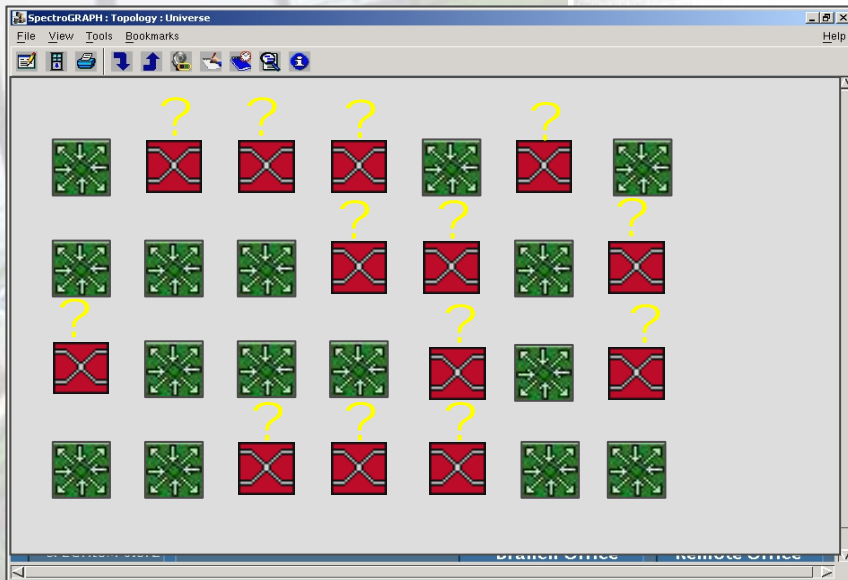
Requirements

- Event filtering, deduplication, suppression and correlation
- Automated network discovery and update
- Accurate Layer 2 and Layer 3 network topology, including redundancy and routing protocols



Without Network Root Cause Analysis

Traditional Procedure



Which to fix?

How to prioritize?

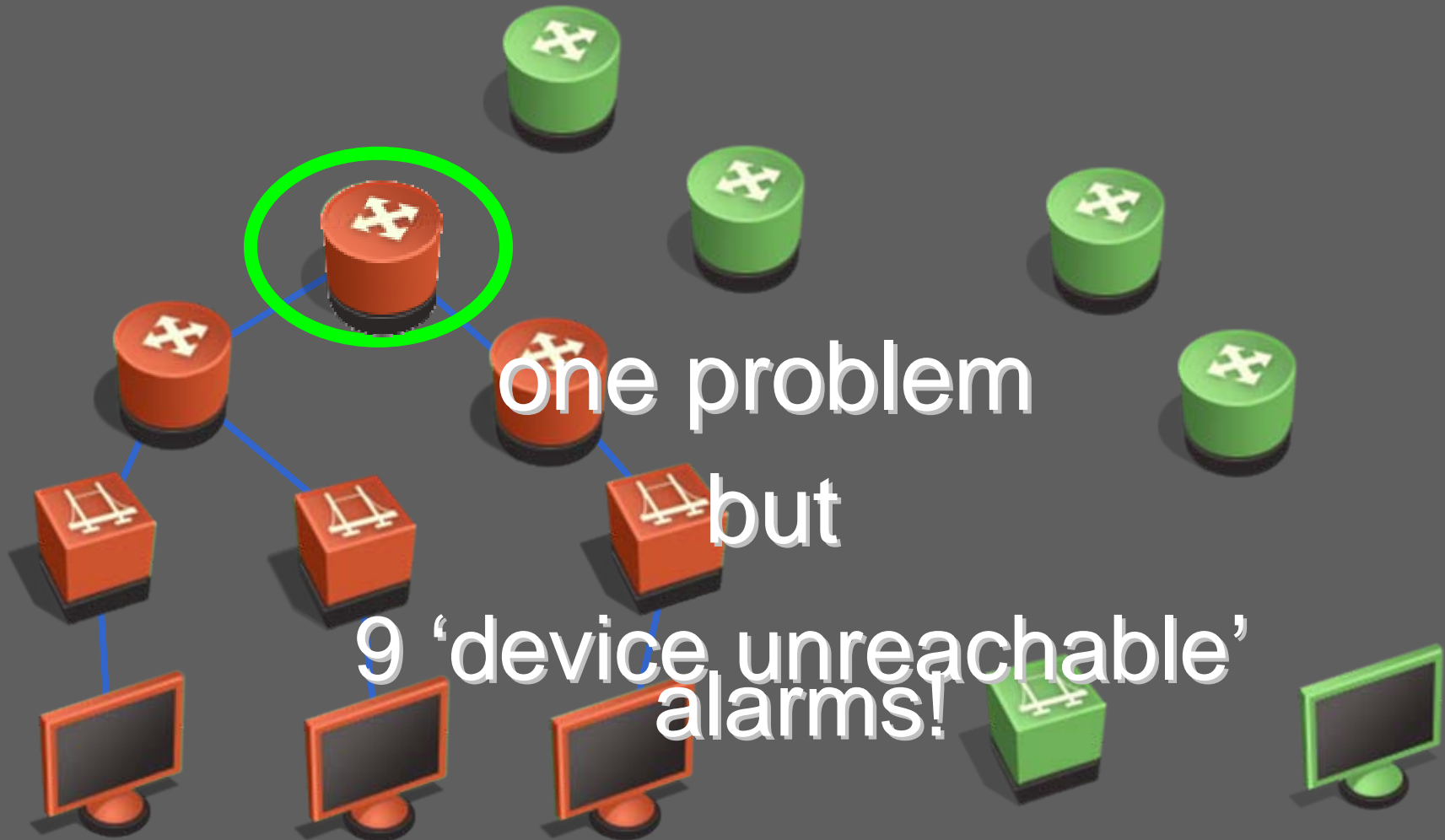
Event Message	Time	IP Address	Severity	Device Type
Device connection lost	1:20	10.1.1.20	Critical	Switch
Device connection lost	2:20	10.1.1.21	Critical	Router
Device connection lost	3:20	10.1.1.22	Critical	Hub
Device connection lost	4:20	10.1.1.23	Critical	Hub
Device connection lost	5:20	10.1.1.24	Critical	Switch
Device connection lost	6:20	10.1.1.25	Critical	PC
Device connection lost	7:20	10.1.1.26	Critical	Server
Link Failure	0:00	10.1.3.27	Critical	Switch
Device connection lost	3:20	10.1.3.28	Critical	Router
Device connection lost	10:20	10.1.3.29	Critical	Hub
Link Failure	0:00	10.1.3.30	Critical	Hub
Link Failure	12:20	10.1.3.31	Critical	Switch
Device connection lost	0:00	10.1.3.32	Critical	PC
Device connection lost	14:20	10.1.3.33	Critical	Server
SNMP Failure	15:20	12.11.10.1	Major	Switch
Hard disk full	16:20	12.11.10.2	Major	Server
Backup power supply failure	17:20	12.11.10.3	Major	Hub
Unauthorized SNMP access detected	18:20	12.11.10.4	Major	Router
Service Failure	19:20	12.11.10.5	Major	Service
SLA will be breached in 30 minutes	20:20	12.11.10.6	Major	SLA
Link utilization above normal burst rates	21:20		Minor	Link
New root bridge detected	22:20		Minor	Bridge
Link up trap registered on interface 1	23:20		Minor	Router

Pinpointing the root cause

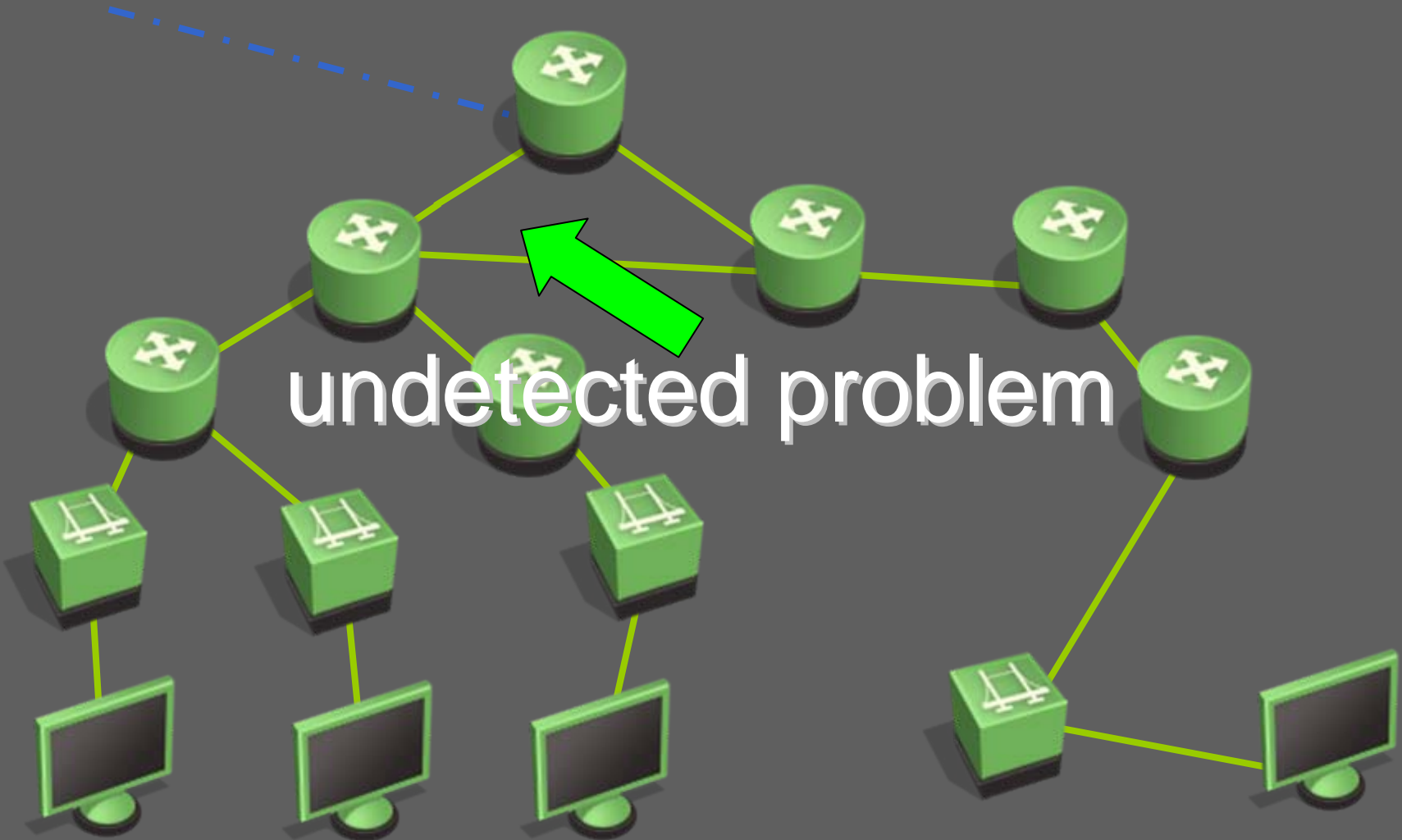
An alarm displayed in CERN's alarm manager is the result of a fault isolation process, **Root Cause Analysis**.

- ONE alarm displayed for one problem
- symptomatic faults suppressed
- operational procedures are followed to complete the troubleshooting process

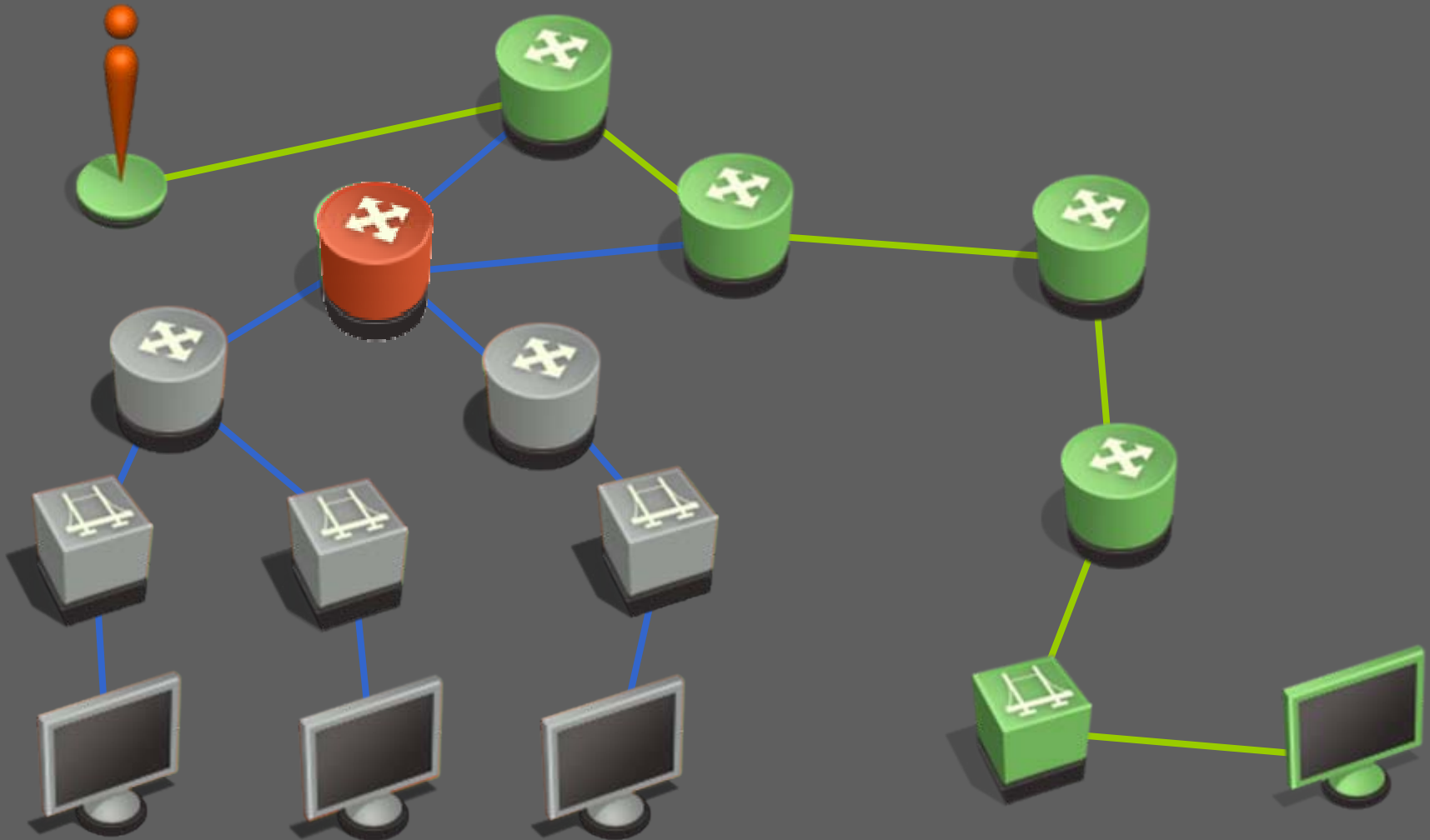
A failure scenario: device fault



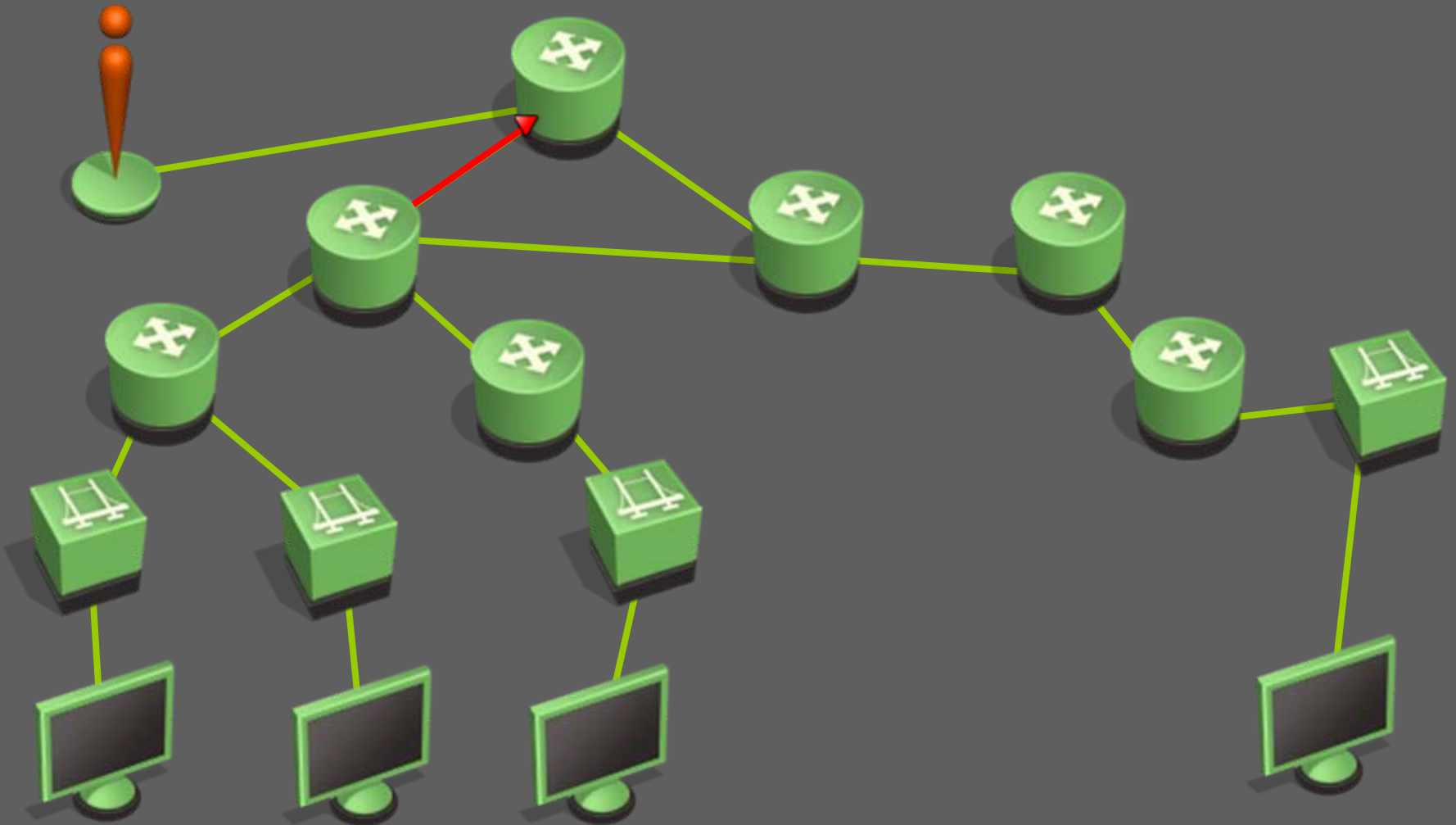
A failure scenario: Loss of redundancy



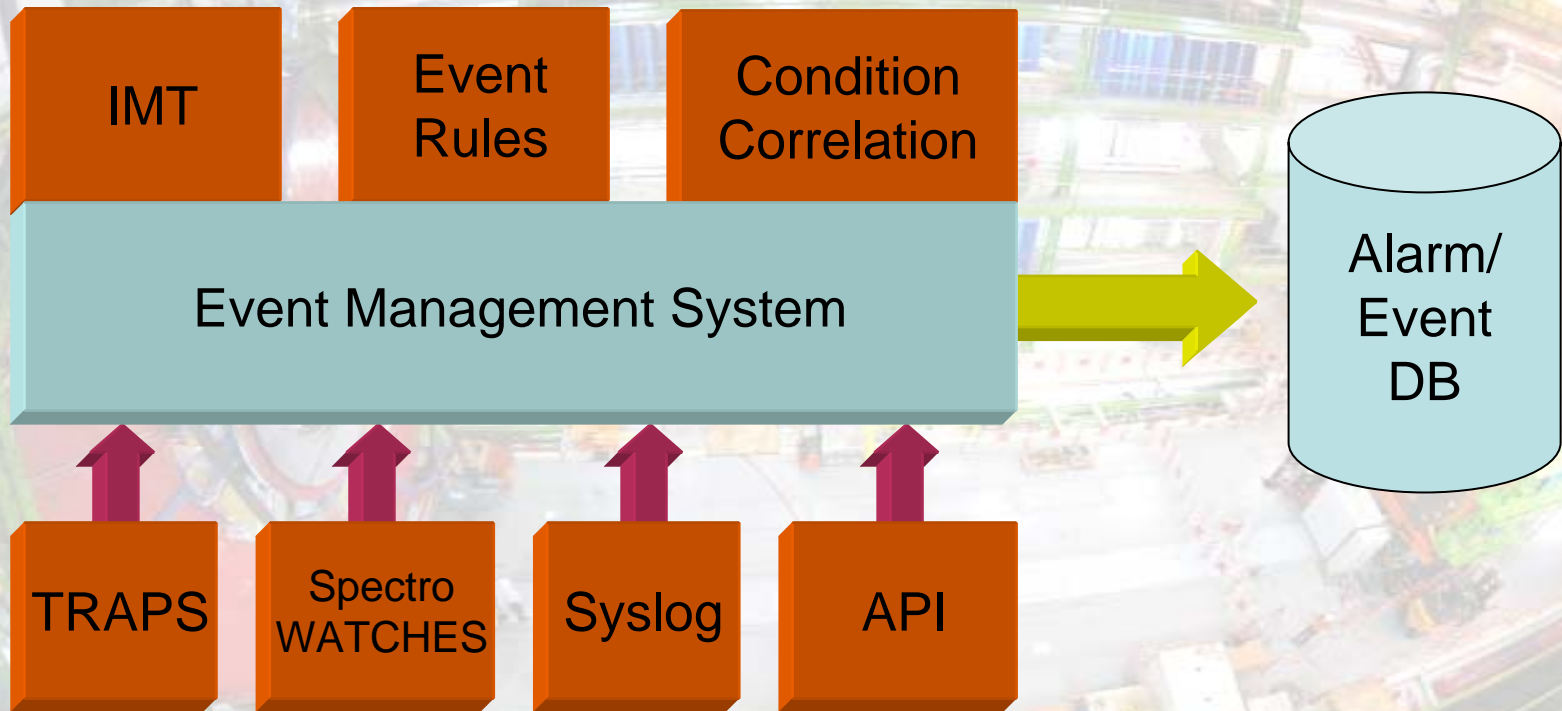
SPECTRUM in the device failure scenario



SPECTRUM in the loss of redundancy scenario



Distinguishing events and meaningful alarms



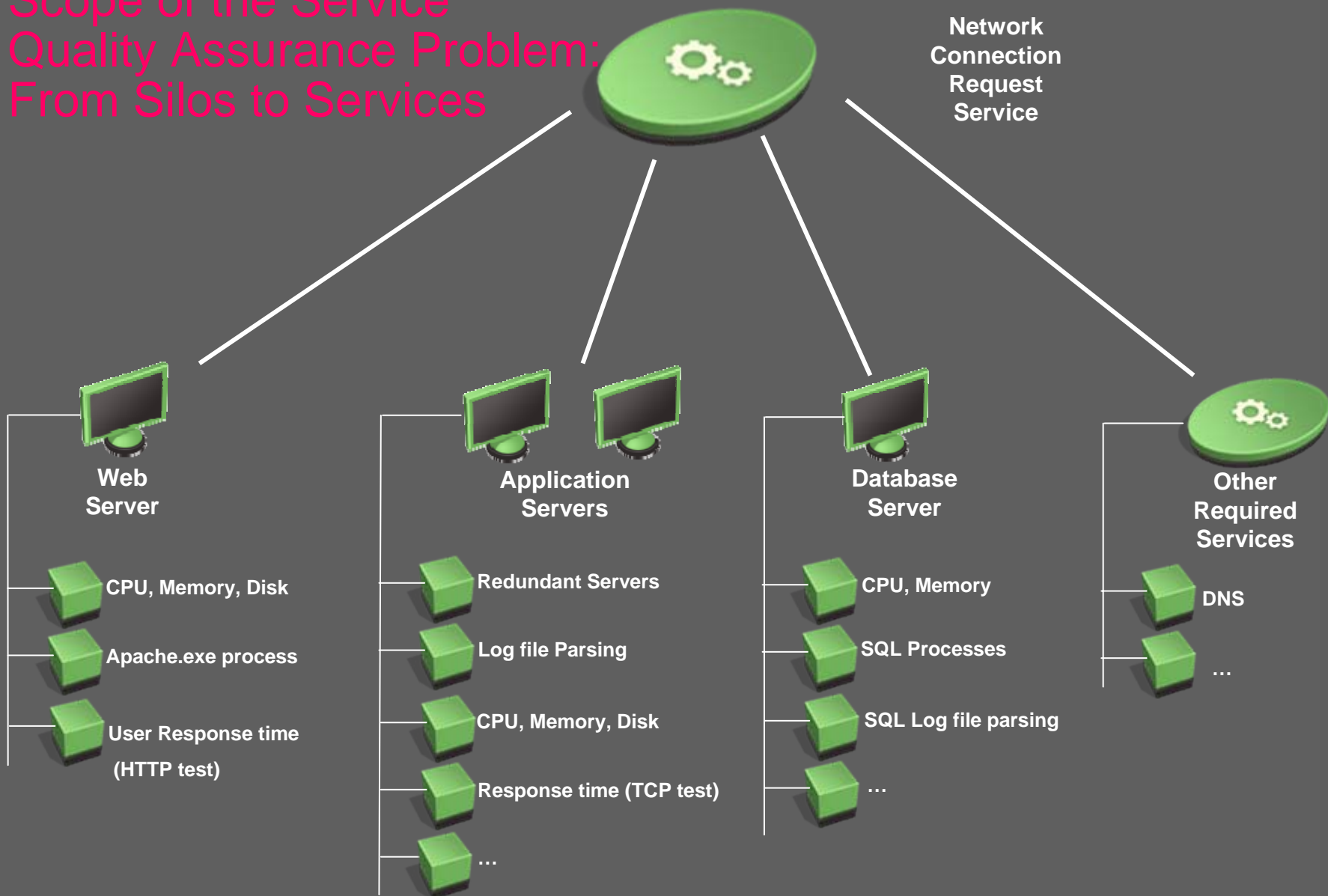
~25.000 events a day (30 Nov 06)

Event Management System allows configuration, creation and control of traps, events and alarms

To do the job, our monitoring system must:

1. **understand topology and relationships.**
2. **work across multiple-vendor and technology solutions.**
3. **distinguish between a plethora of events and meaningful alarms.**
4. **quickly pinpoint the root cause and suppress all symptomatic faults.**
5. **help prioritize based on impact.**
6. **be fault-tolerant.**

Scope of the Service Quality Assurance Problem: From Silos to Services



Availability and Performance Monitoring Best Practices

- Don't fall in the trap of collecting all possible data available
- Try to focus on key metrics that are indicators of total end-to-end service quality.
- Automate!

Create Service - OneClick

Name * CRM

Criticality * High

Select Policy

Name

Security String

Select Value Map

Name Port Status New... Copy... Edit...

Attribute: Port Status (0x10f1b)

Default Reason: Bad Port Status

Attribute Value	Service Health ▲	Root Cause Reason
down	▼ Down	The port is down
disabled	▼ Down	The port is disabled
unreachable	▼ Down	The port is unreachable
up	▲ Up	

Filter: Displaying 4 of 4

Select Rule Set

Name Low Sensitivity New... Copy...

- When a Average Greater Than
- When a Average Greater Than Or Equal
- When a High Sensitivity
- When a Low Sensitivity
- When a Percentage
- When a Redundancy
- Sum Greater Than
- Sum Greater Than Or Equal

< ||| >

OK Cancel

Filter: Displaying 0 of 0

* indicates a required field

Create Cancel

Service Dashboard

- The tool to provide real time service status and statistics
- Allows at-a-glance understanding of
 - How well the services are running
 - Problems and status
 - Transparency towards users and IT management
- Status and Statistics exported to PerfSonar, MonaLisa as well as other databases and alarm systems.

- Summary
 - Current Service Status
 - Current “Customer” Status

- General Details
 - MTTR
 - MTBF
 - % Uptime, Downtime, Degraded

- Outage Details
 - Duration
 - Cause
 - Troubleshooter
 - Impact

All Services	
Service	Radius Servers
Status	Up
Monitoring Policy	Condition Redundancy
Policy Rules	When all resources are Down then the service is Down. When all resources are Degraded then the service is Degraded. When all resources are Slightly Degraded then the service is Slightly Degraded. When any 1 resource(s) are Down then the service is Slightly Degraded.

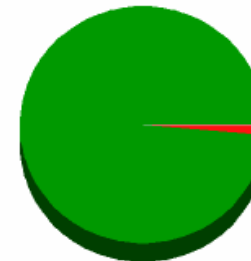
Current Outages | **Outage History**

Outage History

Outage Summary for the last 30 days

Total Up Time	30 Days + 13:28:37
Total Down Time	0 Days + 09:44:11
Total Degraded Time	0 Days + 00:00:00
Total Slightly Degraded Time	0 Days + 00:21:42
Total Maintenance Time	0 Days + 00:00:00
Total Loss Of Management Time	0 Days + 00:25:30
Total Defunct Time	0 Days + 00:00:00
Total Initial Time	0 Days + 00:00:00

■ % Up Time 98.586%
■ % Down Time 1.309%
■ % Slightly Degraded Time 0.049%
■ % Loss Of Management Time 0.057%



Recent Outages

Service Health	Duration	StartTime	End Time
Loss Of Management	0 Days + 00:25:30	2006-11-29 18:00:57	2006-11-29 18:26:27
Slightly Degraded	0 Days + 00:17:13	2006-12-04 12:30:07	2006-12-04 12:47:20
Down	0 Days + 01:07:42	2006-12-04 12:47:20	2006-12-04 13:55:02
Down	0 Days + 08:36:29	2006-12-04 15:31:21	2006-12-05 00:07:50
Slightly Degraded	0 Days + 00:02:51	2006-12-05 00:07:50	2006-12-05 00:10:41
Slightly Degraded	0 Days + 00:01:02	2006-12-05 08:08:04	2006-12-05 08:09:06
Slightly Degraded	0 Days + 00:00:36	2006-12-05 08:09:58	2006-12-05 08:10:34

IT/CS Service Management

All Services	
Service	LHCOPN_T0-T1
Status	Up
Monitoring Policy	Service Health Percentage
Policy Rules	When 75% of the resources are Down then the service is Down. When 50% of the resources are Down then the service is Degraded. When 25% of the resources are Down then the service is Slightly Degraded.
Resources	Up CERN-BNL_Link
	Up CERN-CNAF_Link
	Up CERN-FNAL_Link
	Up CERN-GEANT_Link
	Up CERN-IN2P3_Demo_Link
	Up CERN-RAL_Link
	Up CERN-SARA_Link
	Up CERN-TAIPEI_Links
	Up CERN-TRIUMF_Link
	Up IN2P3-CERN_Link
	Slightly Degraded US_LHCNET

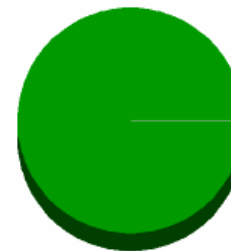
[Current Outages](#) [Outage History](#)

Outage History

Outage Summary for the last 30 days

Total Up Time	30 Days + 23:34:30
Total Down Time	0 Days + 00:00:00
Total Degraded Time	0 Days + 00:00:00
Total Slightly Degraded Time	0 Days + 00:00:00
Total Maintenance Time	0 Days + 00:00:00
Total Loss Of Management Time	0 Days + 00:25:30
Total Defunct Time	0 Days + 00:00:00
Total Initial Time	0 Days + 00:00:00

■ % Up Time 99.943%
■ % Loss Of Management Time 0.057%



IT/CS Service Management

All Services	
Service	US_LHCNET
Status	Slightly Degraded
Monitoring Policy	Condition High Sensitivity
Policy Rules	When any 1 resource(s) are Down then the service is Down. When any 1 resource(s) are Degraded then the service is Degraded. When any 1 resource(s) are Slightly Degraded then the service is Slightly Degraded.

Current Outages | **Outage History**

Outage History

Outage Summary for the last 30 days

Total Up Time	9 Days + 18:01:40
Total Down Time	0 Days + 03:43:55
Total Degraded Time	0 Days + 00:02:49
Total Slightly Degraded Time	21 Days + 01:46:06
Total Maintenance Time	0 Days + 00:00:00
Total Loss Of Management Time	0 Days + 00:25:30
Total Defunct Time	0 Days + 00:00:00
Total Initial Time	0 Days + 00:00:00

- % Up Time 31.455%
- % Down Time 0.502%
- % Degraded Time 0.006%
- % Slightly Degraded Time 67.980%
- % Loss Of Management Time 0.057%



Recent Outages

Service Health	Duration	StartTime	End Time
Down	0 Days + 00:20:20	2006-11-05 11:49:44	2006-11-05 12:10:04
Slightly Degraded	0 Days + 00:02:09	2006-11-05 12:10:04	2006-11-05 12:12:13
Down	0 Days + 00:02:41	2006-11-05 12:16:34	2006-11-05 12:19:15
Down	0 Days + 00:04:50	2006-11-05 13:19:35	2006-11-05 13:24:25

IT/CS Service Management

All Services	
Service	US_LHCNET
Status	Slightly Degraded
Monitoring Policy	Condition High Sensitivity
Policy Rules	When any 1 resource(s) are Down then the service is Down. When any 1 resource(s) are Degraded then the service is Degraded. When any 1 resource(s) are Slightly Degraded then the service is Slightly Degraded.

[Current Outages](#) [Outage History](#)

Outage History

Outage Summary for the last 30 days

Total Up Time	9 Days + 17:58:25
Total Down Time	0 Days + 03:43:55
Total Degraded Time	0 Days + 00:02:49
Total Slightly Degraded Time	21 Days + 01:49:21
Total Maintenance Time	0 Days + 00:00:00
Total Loss Of Management Time	0 Days + 00:25:30
Total Defunct Time	0 Days + 00:00:00
Total Initial Time	0 Days + 00:00:00

- % Up Time 31.448%
- % Down Time 0.502%
- % Degraded Time 0.006%
- % Slightly Degraded Time 67.987%
- % Loss Of Management Time 0.057%



Recent Outages

Service Health	Duration	StartTime	End Time
Down	0 Days + 00:20:20	2006-11-05 11:49:44	2006-11-05 12:10:04

Affected Resources

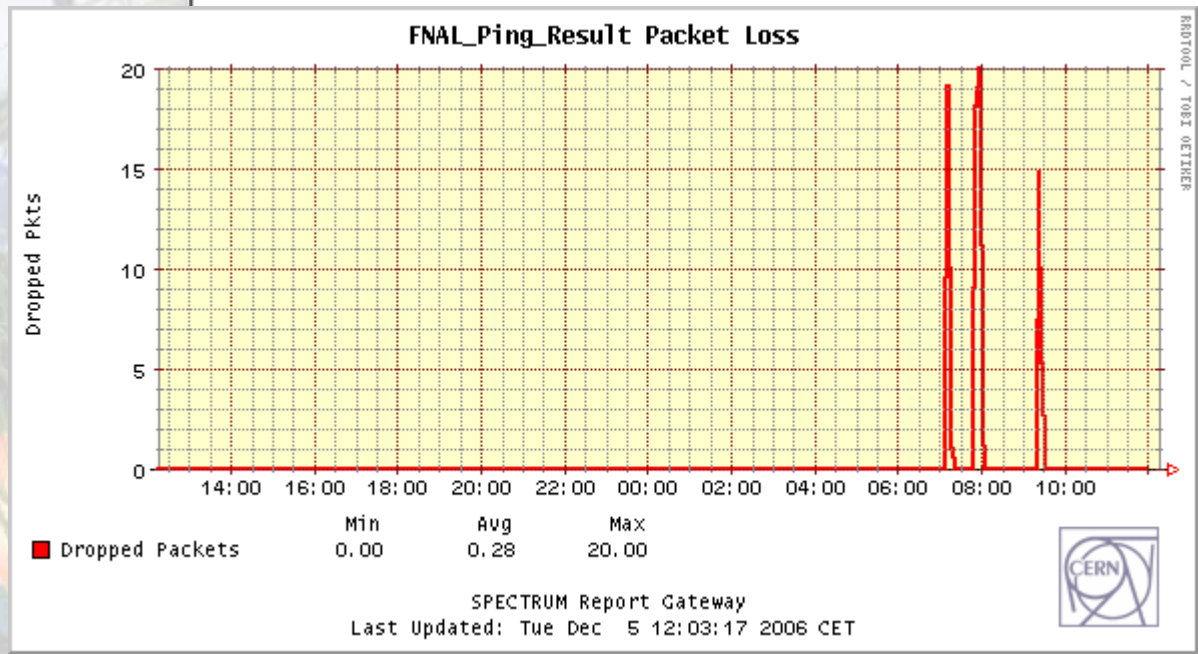
Name	Type	Condition
MANLAN Peering	Network	Suppressed
MANLAN Hosts	Network	Suppressed
as1-nyc	Rtr_Cisco	Suppressed
e600nyc	Force10Rtr	Critical
x424nyc	GnSNMPDev	Suppressed

IT/CS Service Management

All Services					
Service	CERN-FNAL_Link				
Status	Up				
Monitoring Policy	Service Health High Sensitivity				
Policy Rules	When any 1 resource(s) are Down then the service is Down. When any 1 resource(s) are Degraded then the service is Degraded. When any 1 resource(s) are Slightly Degraded then the service is Slightly Degraded.				
Resources	<table border="0"> <tr> <td>Up</td> <td>FNAL_Links</td> </tr> <tr> <td>Up</td> <td>FNAL_RTM</td> </tr> </table>	Up	FNAL_Links	Up	FNAL_RTM
Up	FNAL_Links				
Up	FNAL_RTM				

[Current Outages](#)
[Outage History](#)

Outage History



Name	Type	Condition
FNAL_RTM	SM_AttrMonitor	Degraded
FNAL 2 Ping	RTM_Test	TIMEOUT

Conclusions

- Monitoring is essential for robust network operation.
- Root Cause Analysis enables quick reactions to faults.
- Transparent, real time reporting and information exchange demonstrates service quality and **gains the trust** of users and collaborators.
- Focus on collecting and storing relevant data.



Thank you!

Q & A

<http://cern.ch/monitoring>