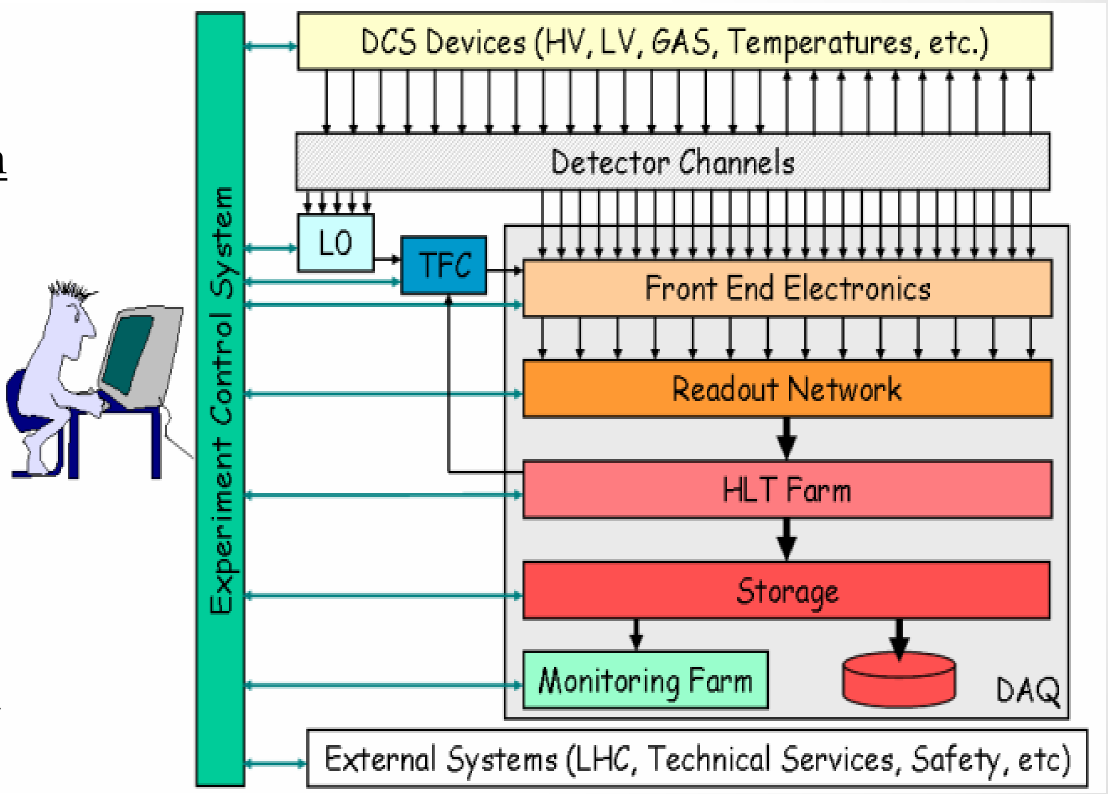
Control System Virtualization for the LHCb Online System

ICALEPCS – San Francisco

Enrico Bonaccorsi, (CERN) enrico.bonaccorsi@cern.ch
Luis Granado Cardoso, Niko Neufeld (CERN, Geneva)
Francesco Sborzacchi (INFN/LNF, Frascati (Roma))

LHCb & Virtualization

- Completely isolated network
  - Data Acquisition System
  - Experiment Control System
- Why do we virtualize
  - Improve manageability
  - High Availability
  - Hardware usability
    - Better usage of hardware resources
    - Move away from the model “one server = one application”



What we are virtualizing

- Around 200 control PCs running WinCC OA
  - 150 linux
    - Red Hat / CentOS / Scientific Linux 6
  - 50 windows
    - Windows 2008 R2

- Web Servers
- Gateways
  - Linux SSH and NX
  - Windows terminal services
- Common infrastructure servers
  - DHCP, DNS, Domain Controllers, …

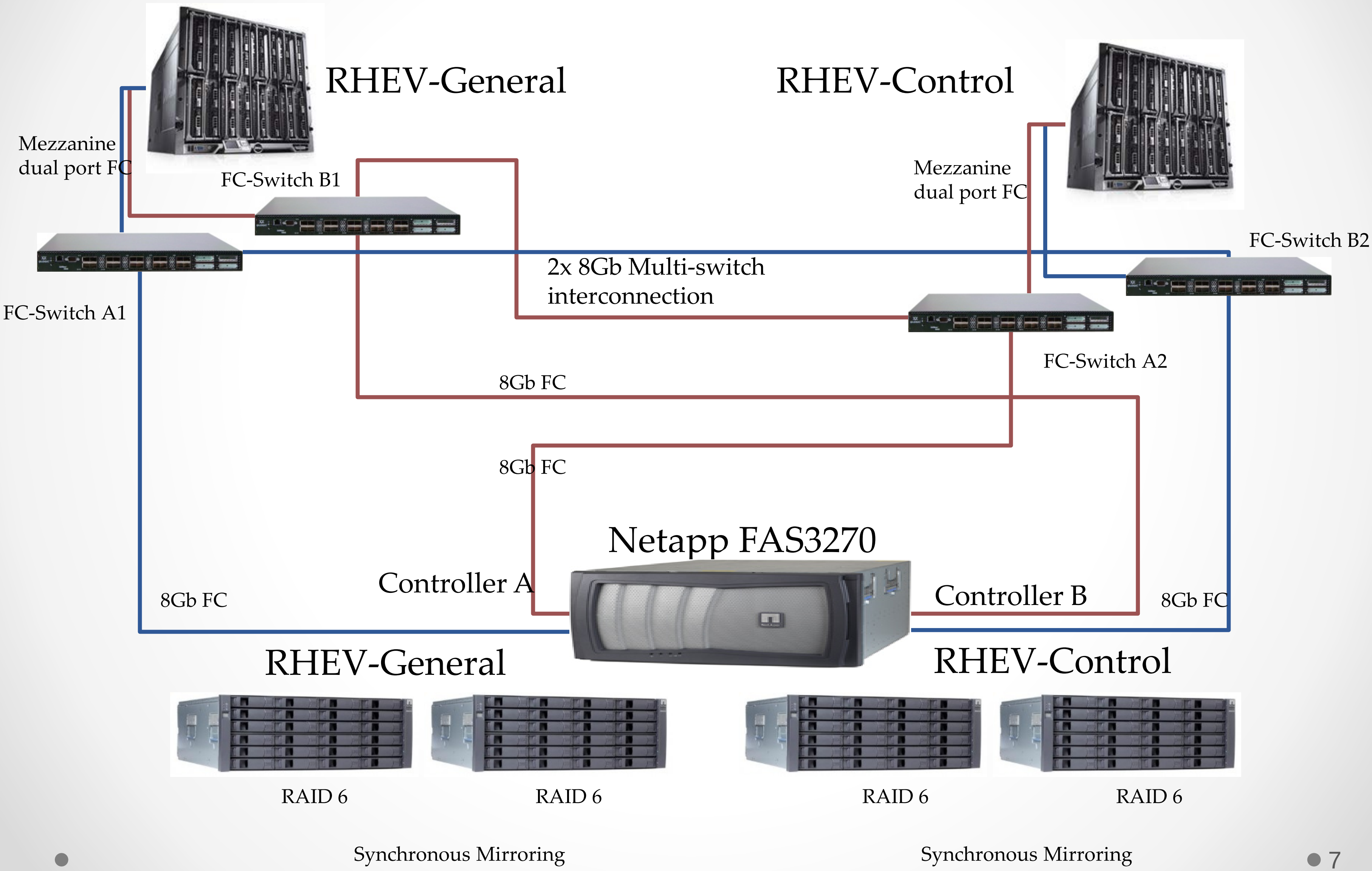
Current virtualization infrastructure

- 20 blade servers distributed in two chassis in different racks
- 4 x 10 Gb/s Ethernet switches
- 4 x 8 Gb/s Fiber channel (FC) switches
- 2 x NetApp 3270 accessible via FC and iSCSI
  - Hybrid storage pool: SSD + SATA
- 2 independent clusters
  - General Purpose Cluster (DNS, DHCP, Web services, ..)
  - Control Cluster (Dedicated to the control system)

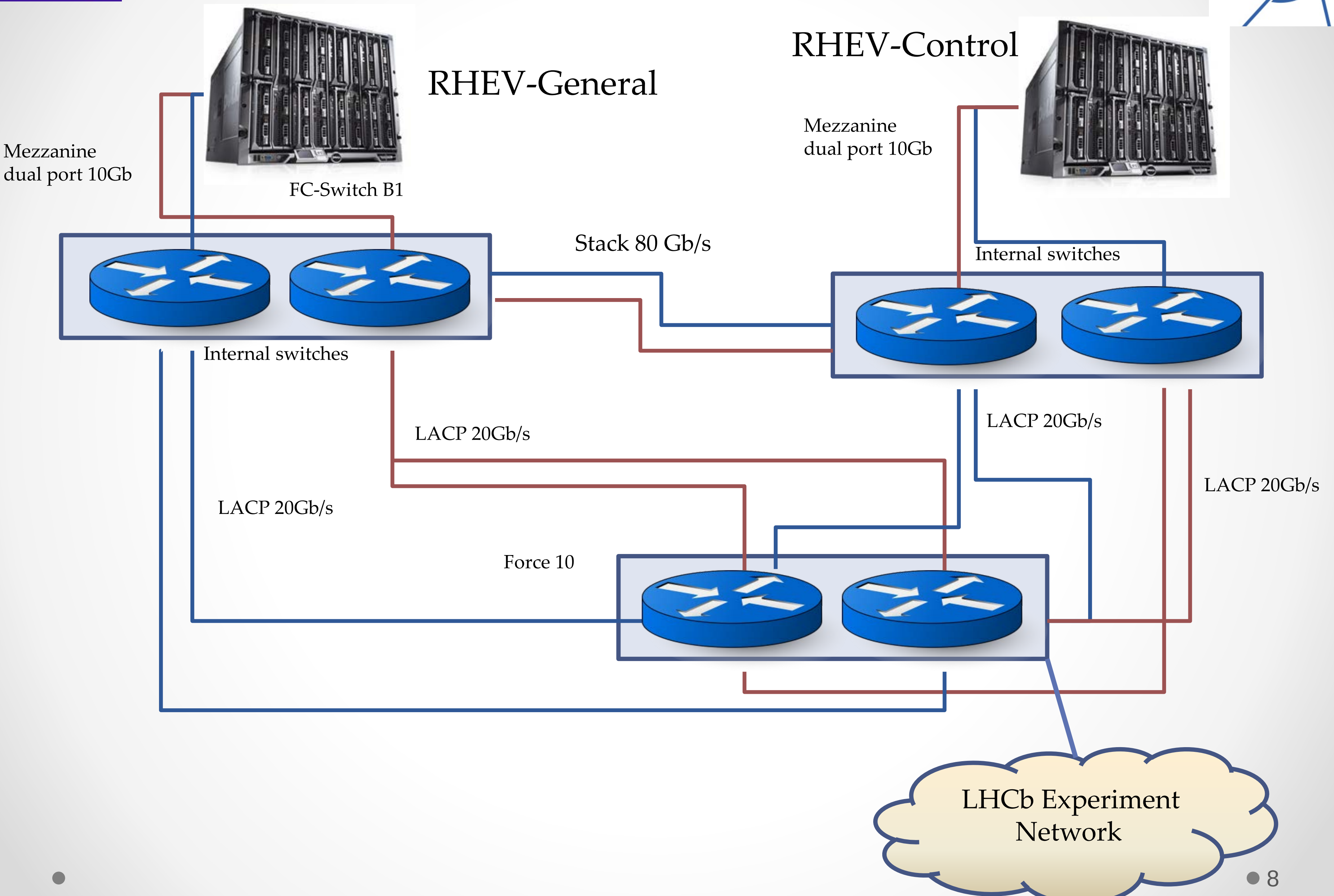
Shared Storage

- Crucial component of the virtualization infrastructure
- Required for high availability
- Performance is a key point
- We want to guarantee a minimum of 40 random IOPS per VM
    - The equivalent experience of using a laptop with a 5400 RPM HDD

Storage area network



Ethernet network



Resources optimization

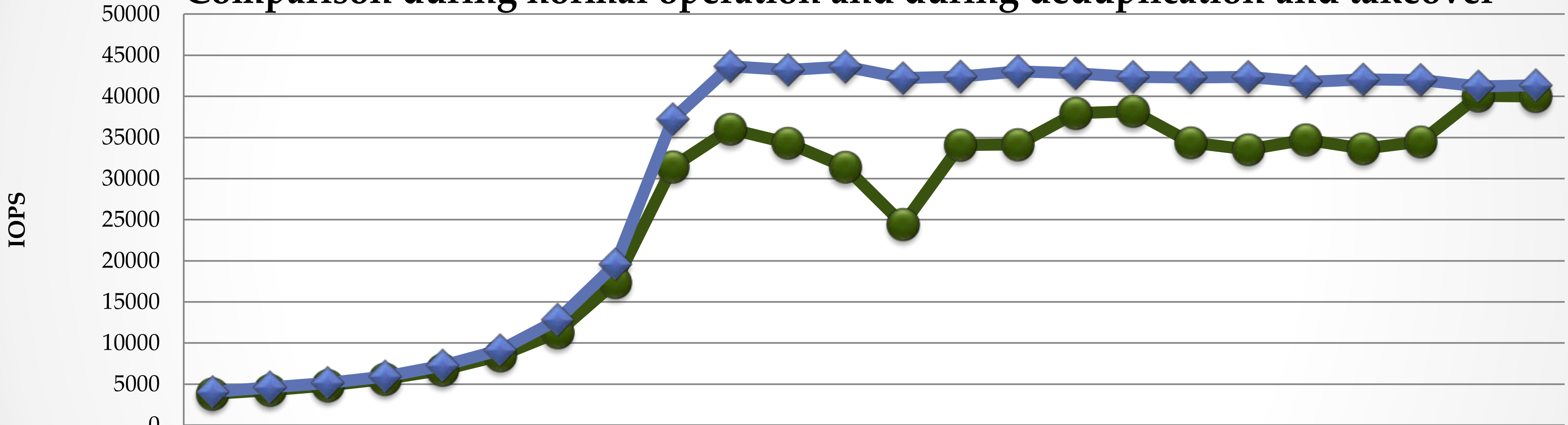
High performance hardware is expensive

- Storage deduplication
  - Eliminates duplicates of repeated data
  - Currently saving 67% of used storage space
  - Provides improvements in terms of IOPS (less data to cache!)
- Kernel Shared Memory
  - Maximize the usage of memory
  - Merge the same memory pages allowing overcommitting of memory without swapping

Benchmarks

- Blade Poweredge M610
  - 2 x E5530 @ 2.4GHz (8 real cores + Hyper Threading)
  - 3 x 8 GB = 24GB RAM
  - 2 x 10Gb network interfaces
  - 2 X 1Gb network interfaces
  - 2 X 8Gb fiber channel interfaces
- Storage
  - 4 X 8Gb Fiber channel switches
  - SSD pool + SATA
  - Deduplication ON
- Network
  - 4 X 10Gb Ethernet switches
  - 4 X 1Gb Ethernet switches
- Limits:
  - Average of 15 VM per Server

Netapp 3270, random Reading 4k + random Writing 4k, 215VMs, 200MB/VM
Comparison during normal operation and during deduplication and takeover



	5.00	10.00	15.00	20.00	25.00	30.00	35.00	40.00	45.00	50.00	55.00	60.00	65.00	70.00	75.00	80.00	85.00	90.00	95.00	100.00	105.00	110.00	115.00	120.00
dedutake	3823	4248	4831	5629	6755	8475	11302	17384	31449	36033	34322	31420	24463	34086	34168	37944	38182	34376	33529	34731	33632	34497	40054	40000
noded	4069	4596	5150	5915	7239	9145	12819	19572	37255	43662	43223	43614	42222	42407	43057	42798	42369	42295	42374	41767	42062	41981	41199	41351

Storage (random)
IOPS=45K
Throughput=153MB/s writing, 300MB/s reading
Latency= ~10ms

Network
Throughput = 5.37 Gb/s
Latency = 0.15 ms for 1400B

Testing the infrastructure: Pessimistic SCADA workloads

- 150 WinCC OA Projects (WINCC001 .. WINCC150)
  - 1 project per VM
  - Each project is connected to other 5 projects
    - The two previous and after projects (according to the numbering
    - The master project
  - Each project has 1000 datapoints created for writing
  - Each project performs dpSets locally and on the connected projects
  - Number of DPs to be set and rate are settable
    - Each period the dps are selected randomly from the 1000 dps pool and set

Testing the infrastructure: Pessimistic SCADA workloads (2)

- 1 Master Project (WINCC001)
  - This project connects to all other projects
  - Has System Overview installed for easier control of the whole system
    - FW version for PVSS 3.8 – produces a couple of errors but the PMON communication with the other projects works just fine

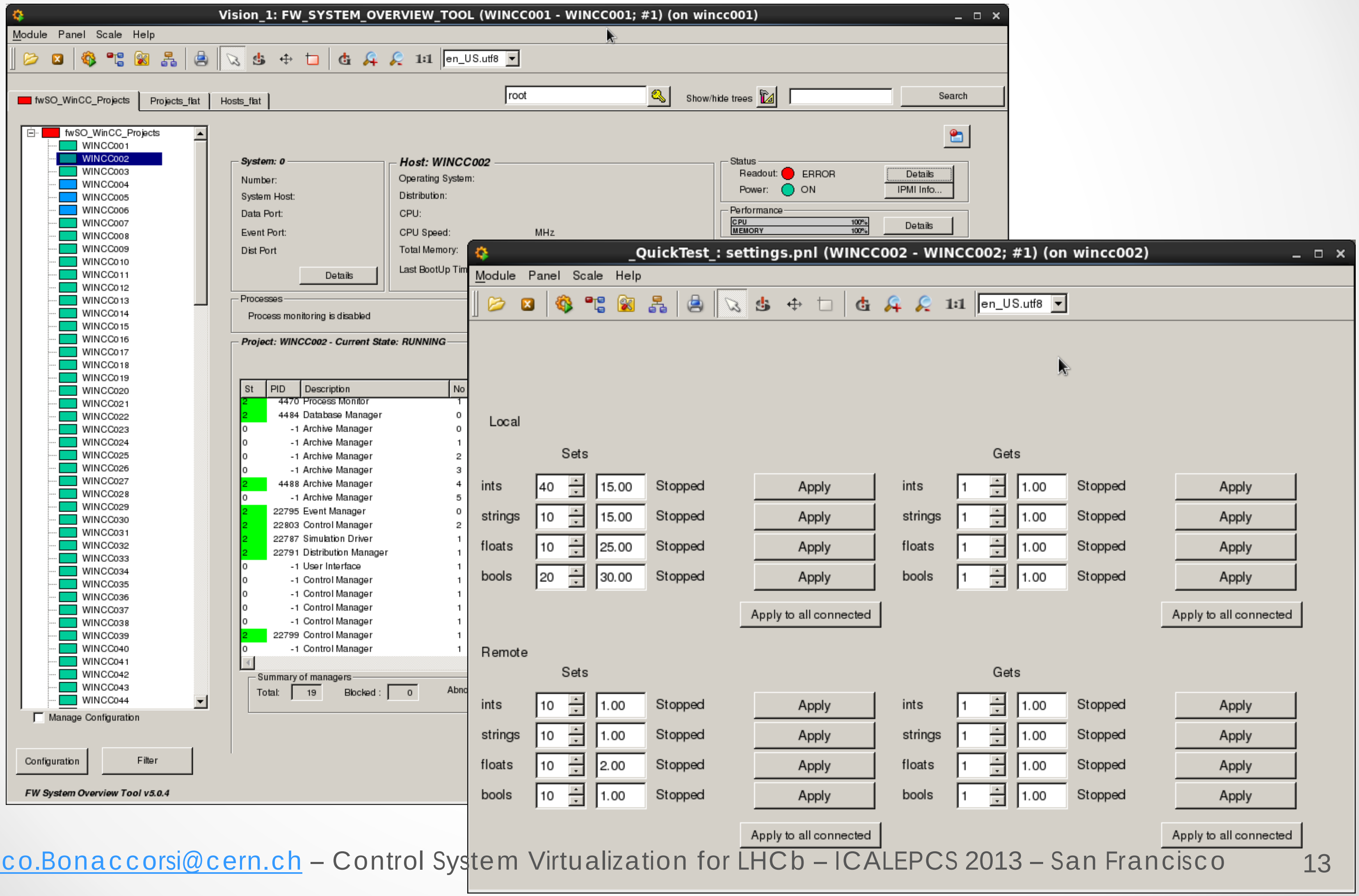
Results Summary

- At the end of each "run" period, logs are collected and analysed for problems
  - PVSS_II.log, WCCOActrlNN.log are "grepped" for possible issues ("disconnect", "connect", "queue", "pending", "lost", …)
- Plots are also produced by calculating the rate from the dpSets timestamp (only local dpSets)

Date	Local Rate*	Remote Rate*	Total*	CPU (%)	Comment
18.12.2012	1200	100	1700	85	All OK
20.12.2012	1200	0	1200	35	All OK
09.01.2013	1200	1000	5210	85	All OK
14.01.2013	1600	1400	7250	93+	Problems with 1 project (multiple disconnections/connections)**
17.01.2013	1600	50	1850	50-60	Decreased for live migration tests
*dpSets per Second					

** WINCC006, after some period, started disconnecting/connecting to WINCC005 and WINCC007 indefinitely. Problem was fixed by restarting the projects WINCC004 and WINCC008 which also connect to WINCC006.

- Unattended long term test:
  - all VMs and projects performed stably
  - One instance had to be live migrated to solve some issues related to the real server
- We run the same tests that have been done for real machines by the CERN industrial control group (EN/ICE) and we obtained very similar results



Vision_1: FW_SYSTEM_OVERVIEW_TOOL (WINCC001 - WINCC001; #1) (on wincc001)
Module Panel Scale Help
en_US.utf8
fwSO_WinCC_Projects
Projects_flat
Hosts_flat
root
Show/hide trees
Search
WINCC001
WINCC002
WINCC003
WINCC004
System: 0
Number:
System Host:
Data Port:
Event Port:
Dist Port
Details
Host: WINCC002
Operating System:
Distribution:
CPU:
CPU Speed:
MHz
Total Memory:
Status
Readout: ERROR
Power: ON
Details
IPMI Info...
Performance
CPU
MEMORY
Processes
Process monitoring is disabled
Project: WINCC002 - Current State: RUNNING
St PID Description
4484 Database Manager
Archive Manager
22795 Event Manager
22803 Control Manager
22787 Simulation Driver
22791 Distribution Manager
User Interface
Summary of managers
Total: 19
Blocked: 0
Manage Configuration
Configuration
Filter
FW System Overview Tool v5.0.4
QuickTest: settings.pnl (WINCC002 - WINCC002; #1) (on wincc002)
Local
Sets
Gets
ints
strings
floats
bools
Stopped
Apply
Apply to all connected
Remote

Summary and outlook

- Virtualization of LHCb ECS
  - Reduce hardware
  - Achieving High Availability
- Storage the key component of the infrastructure
- Realistic SCADA workload emulator
  - Indispensable in the evaluation of many commercial storage systems
- Resources optimizations
- Performance results
- ----------------------
- Migration of all Control PCs to VMs should be completed by Q4 2013

Backup slides

Hypervisors

- Essentially 3 choices:
  - Kernel based Virtual Machine (KVM)
    - Currently used in LHCb
    - Open source
  - VMWare:
    - Most advanced even if closed source
    - Too expensive for us
  - Hyper-V Core R2 and System Center Virtual Machine Manager (SCVMM)
    - Almost for free (license needed for SCVMM)

Capacity planning