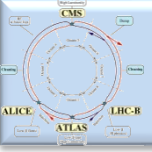
VIRTUALISATION

ALICE, ATLAS, CMS & LHCB
JOINT WORKSHOP ON *DAQ@LHC*

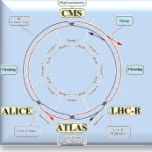
- Introduction
- Why virtualise?
- Advantages of the abstraction layer
- Understand the limit
- Infrastructures overview
- Virtualisation in the present
- Virtualisation in the future
- Conclusions

Enrico Bonaccorsi
on behalf of
ALICE, ATLAS, CMS, LHCb Virtualisation



Introduction

- Virtualisation:
  - ★ in computing, is a term that refers to the various techniques, methods or approaches of creating a virtual version of something, such as a virtual hardware platform, operating system, storage device, or network resources
    - http://en.wikipedia.org/wiki/Virtualization
- Hardware virtualisation:
  - ★ Hides the physical characteristics of a computing platform from users, instead showing another abstract computing platform
- Host:
  - ★ Physical server that runs the VMs
- Guest:
  - ★ Virtual machine running on a physical server

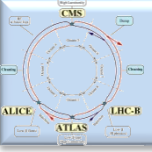


Why virtualise?

- Cut down costs
  - ★ EX. Between 300 and 600 CHF per VM at LHCb
- CPUs: from single core to multi cores to many cores
- Mitigate server sprawl abandoning the model “one server -> one application”
  - ★ Optmise resource usage, less servers, save energy
  - ★ Manage the complexity of the data center
- Server consolidation and improved resource utilization
  - ★ Bring many workloads on a single machine- reduce the idle time of servers
- Faster deploy of new server
  - ★ Clone a gold image, deploy from templates or from existing virtual machine
- Isolate application
  - ★ Providing an abstraction layer between HW and SW
  - ★ Reduce vendor dependencies
- Increase availability
  - ★ If a component fail the VMs are moved or restarted somewhere else
- Virtual labs & Testing

Advantages of the abstraction layer

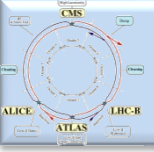
- Snapshot
  - ★ Is the state of a virtual machine, and generally, its storage devices, at an exact point in time
  - ★ You can revert the state of a VM to a previous state stored in a snapshot
- Migration
  - ★ A snapshot can be moved to another host machine
  - ★ VM is temporarily stopped, snapshotted, moved, and then resumed on the host
- Failover
  - ★ Allows the VM to continue operations if the host fails – live migrating on another host or restarting if live migration is not possible
- Storage live migration
  - ★ Allows the VM to continue operations while its virtual drive is moving to another storage



Understand the limit: Virtualisation is not magic

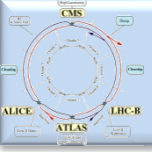
- Abstracting hardware does not increase hardware resources
  - ★ Each server has finite resources, in terms of:
    - ➢ CPU
    - ➢ Memory is limited (even if it could be virtually increased by KSM and/or swapping on SSDs)
    - ➢ Network -> do not underestimate latency and throughput
    - ➢ Storage -> do not underestimate maximum IOPS, throughput

- Capacity planning is difficult but it is fundamental to achieve good results:
  - ★ Don't pretend what the HW can't do
  - ★ What are the available HW resources?
  - ★ How many machines will use the same infrastructure?
  - ★ Storage? How many random IOPS per VM?
  - ★ What about network usage?
  - ★ Make your system able to manage peak loads
    - ➢ A VM with high IO can severely impact the others



Infrastructures overview

	ATLAS	CMS	LHCB
Hypervisor	XEN & KVM	KVM	KVM
Management SW	• LibVirt • OpenStack for Sim@p1	• LibVirt • OpenStack	• RHEV • LibVirt • Evaluating OpenStack
Current number of VM	~35 ~11 testbed	10 LibVirt 1300 OpenStack	~40 ~200 testbed
Number of foreseen VMs at end of LS1	~1800-openstack	~1300 (maybe more)	~300
Number of VMs per Hypervisor	6-8 VMs	1 VM	~15 VMs
Storage backend (Problems with high I/O?)	• Local drives • NFS, ISCSI for TDAQ Testbed • Evaluating NetApp	• Local SATA • Evaluating GlusterFS	• Shared storage: FC & iSCSI based on NetApp
Average Network Bandwidth per VM under peak load		1Gb/s	500Mb/s



Virtualization in the present

ALICE

- none

ATLAS

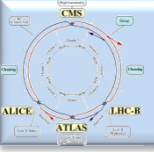
- gateways
- domain controllers
- few windows services
- development web servers
- core Nagios servers
- Puppet and Quattor servers
- one detector machine
- public nodes

CMS

- domain controllers
- Icinga workers and replacement server
- few detector machines

LHCb

- web services
- infrastructure services
  - ★ DNS, Domain Controller, DHCP, firewalls
  - ★ always a tandem for critical systems: one VM, one real
- few control PCs



Virtualization in the future

- Virtualization is a very fertile playground
  - ★ Everyone thinking how to exploit
- Offline software (analysis and simulation) will run on virtual machines on the ATLAS and CMS HLT farms
  - ★ OpenStack is used for management

ALICE

- Control Room PCs
- Event Builders

LHCb

- general login services
  - ★ gateways and windows remote desktop
- all control PCs
  - ★ PVSS, Linux, Windows, specific HW issues (CANBUS)

ATLAS

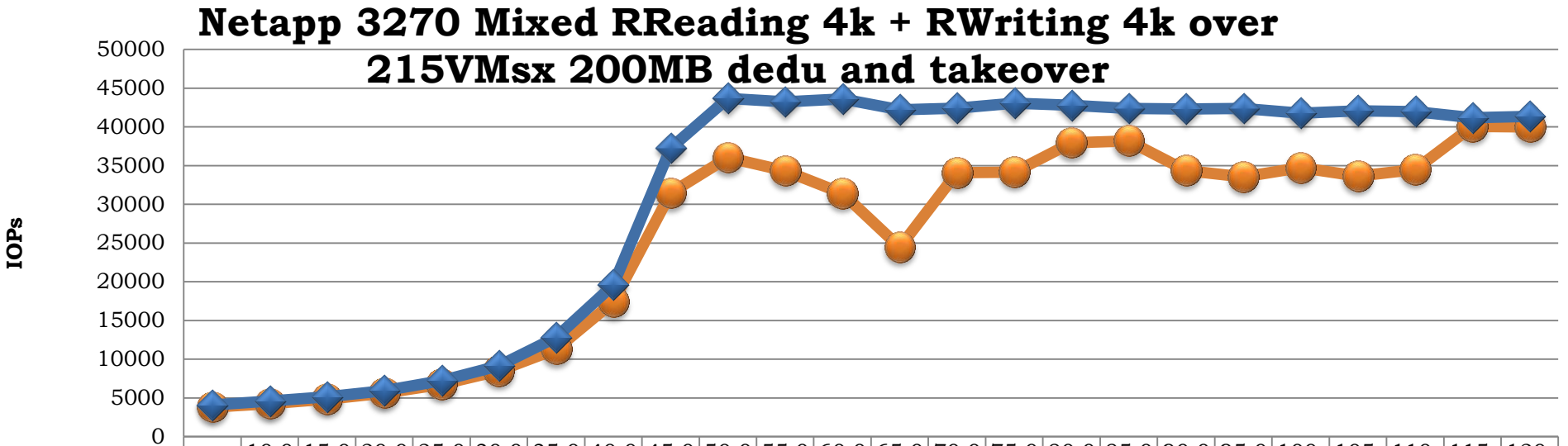
- DCS windows systems

CMS

- servers
  - ★ DNS, DHCP, Kerberos, LDAP slaves
- DAQ services

Benchmark – LHCb VM storage backend & Network

- 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



	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	1130	1738	3144	3603	3432	3142	2446	3408	3416	3794	3818	3437	3352	3473	3363	3449	4005	4000
noded	4069	4596	5150	5915	7239	9145	1281	1957	3725	4366	4322	4361	4222	4240	4305	4279	4236	4229	4237	4176	4206	4198	4119	4135

Storage (random)

IOPS=45K

Throughput=153MB/s

Latency= ~10ms

Network

Throughput = 5.37 Gb/s

Latency = 0.15 ms for 1400B

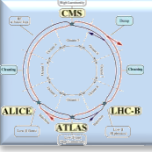
WinCC benchmark in virtual environment: Results Summary

- 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					

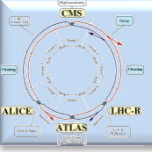
** WINCC006, after some period, started disconnecting/connecting to WINCC005 and WINCC007 indefinitely. Problem was fixed by restarting the projects WINCC004 and WINCC008 which also connect to WINCC006.

- Globally, WinCC seemed to perform stably. Only one instance gave some issues which were able to be resolved.
- Check twiki for more info: https://lbtwiki.cern.ch/bin/view/Online/VirtualizationWinCCTest



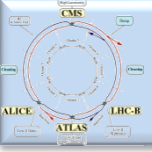
Issues

- VMs Storage slow
  - ★ Check paravirtualisation
  - ★ Lack of IOPS is normally the cause
    - ➢ Solution: Provide enough resources, some tuning can be done but workload should be redistributed or storage backend should be upgraded (IOPS)
  - ★ Maximum number of IOPS could drastically decrease if filesystem is not aligned
  - ★ Filesystem sector size vs disk/array block size
  - ★ Tuning (see backup slide)
- VMs Network slow:
  - ★ Check paravirtualisation
  - ★ Large Receive Offload (LRO) should be disabled in the hypervisor
  - ★ Flow control
  - ★ Provide enough resources
- Time
  - ★ VMs does not see every tick
  - ★ Solved with guest agents – worst case with ntpdate
- PCI, USB & live migration
  - ★ USB could be used over IP but stability must be tested
  - ★ PCI cards make less easy live migration



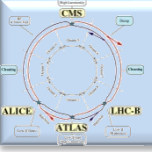
Conclusions

- Experiments are looking more and more at virtualisation
- Virtualisation can provide a solution to the server sprawl phenomenon with the consolidation of several operating systems on a single server
  - ★ Reduce the number of physical server to be managed
  - ★ Reduce the hardware maintenance costs
- Virtualisation increase manageability and efficiency
- Use cases may be different depending on the experiment
  - ★ Different implementations may be required
    - ➢ Ex. Shared storage vs Local storage
    - ➢ "1 VM per Server" vs "Many VMs per Host"
  - ★ Almost all experiments are looking forward to a more cloudy infrastructure
  - ★ OpenStack & virtualisation are common points for which experiments could share knowledge and experience
- Capacity planning is fundamental
- virtualise the DAQ?
  - ★ 1 VM per host?

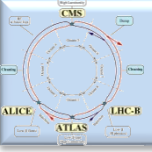


Thanks

- Sergio Ballestrero
- Franco Brasolin
- Olivier Chaze
- Marc Dobson
- Ulrich Fuchs
- Niko Neufeld
- Diana Scannicchio
- Francesco Sborzacchi

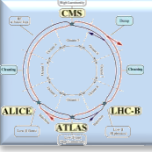


Backup slides



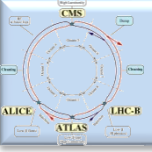
VMs Tuning

- Use paravirtualization
- Mount filesystems with noatime,nodiratime
- Change scheduler to NOOP in VMs
  - ★ kernel /vmlinuz-2.6.18-194.el5 ro root=/dev/VolGroup00/LogVol00 elevator=noop
  - ★ for i in `ls -d /sys/block/vd*`; do echo noop > $i/queue/scheduler; done
- Change scheduler to ANTICIPATORY in the HOSTS
- Cache DNS requests
  - ★ Use nscd
- Disable ipv6
  - ★ echo 'alias net-pf-10 off' >> /etc/modprobe.d/blacklist_ipv6
- Use SSDs, Hybrid drives or tiered storage
- Move metadata away from data
  - ★ Ex. Using LVM



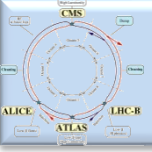
Other Issues

- Hardware Compatilbity
  - ★ Fiber Channel example -> qlogic firmware
  - ★ Force 10 VLAN tag example -> move to a routing environment -> stability at the cost of latency
  - ★ Intel E5000 series – ACPI – HyperV rare bug
- Filesystems timeouts
  - ★ Read only filesystem if waiting for I/O is excessive



WinCC Setup

- 150 WinCC 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



WinCC Setup

- 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
  - ★ Rates of dpSets different for this project only (as it connects to all the others)

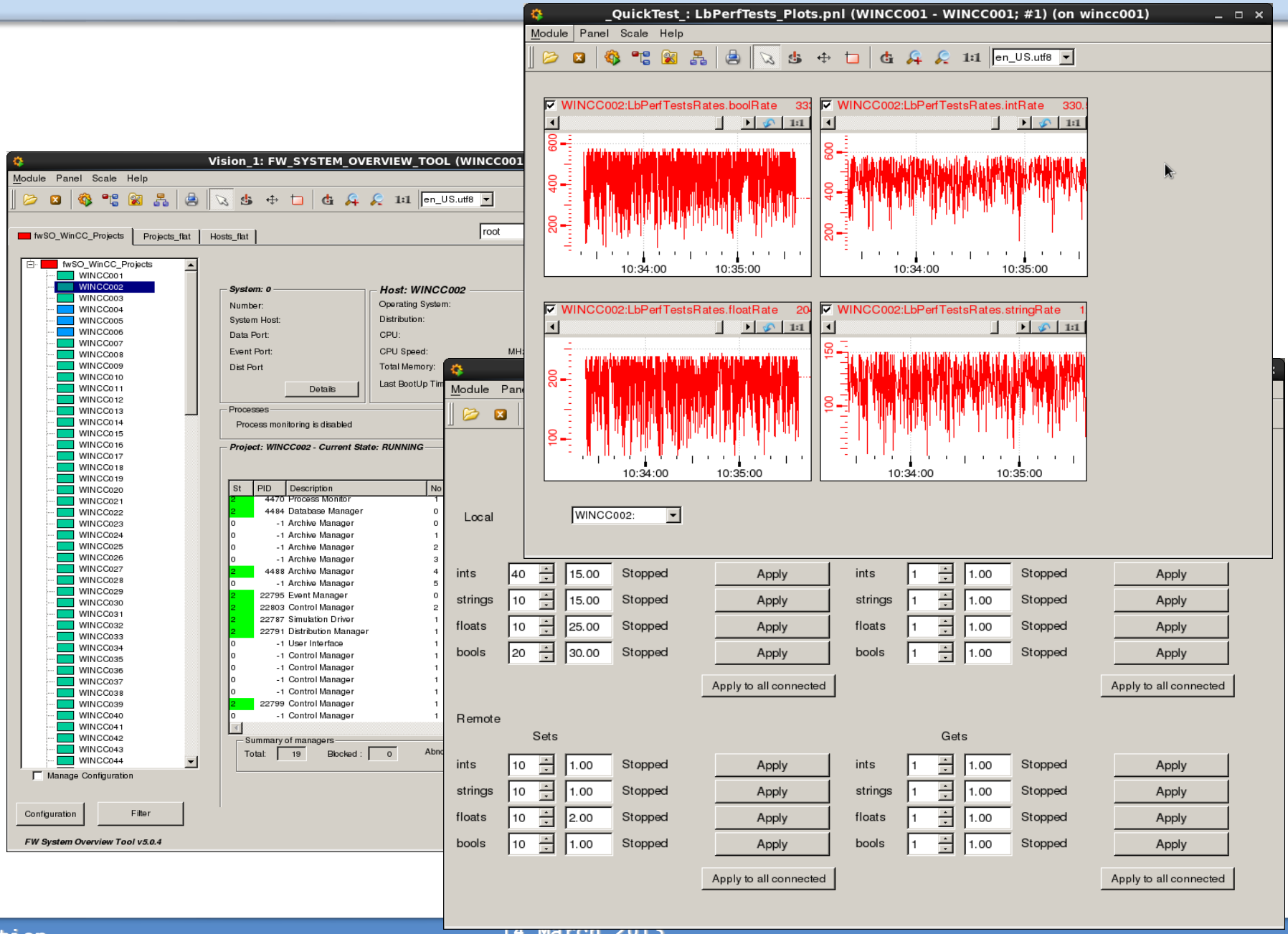
WinCC Results Summary

- 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.

- Globally, WinCC seemed to perform stably. Only one instance gave some issues which were able to be resolved.
- Check twiki for more info: https://lbtwiki.cern.ch/bin/view/Online/VirtualizationWinCCTest


QuickTest: LbPerfTests_Plots.pnl (WINCC001 - WINCC001; #1) (on wincc001)
Module Panel Scale Help
WINCC002:LbPerfTestsRates.boolRate
WINCC002:LbPerfTestsRates.intRate
WINCC002:LbPerfTestsRates.floatRate
WINCC002:LbPerfTestsRates.stringRate
10:34:00
10:35:00
Vision_1: FW_SYSTEM_OVERVIEW_TOOL (WINCC001
fwSO_WinCC_Projects
Projects_flat
Hosts_flat
System: 0
Host: WINCC002
Details
Processes
Process monitoring is disabled
Project: WINCC002 - Current State: RUNNING
Summary of managers
Manage Configuration
Configuration
Filter
FW System Overview Tool v5.0.4
Local
WINCC002:
Remote
Sets
Gets
ints
strings
floats
bools
Stopped
Apply
Apply to all connected