LEMON – Monitoring in the CERN Computer Centre

Helge Meinhard / CERN-IT

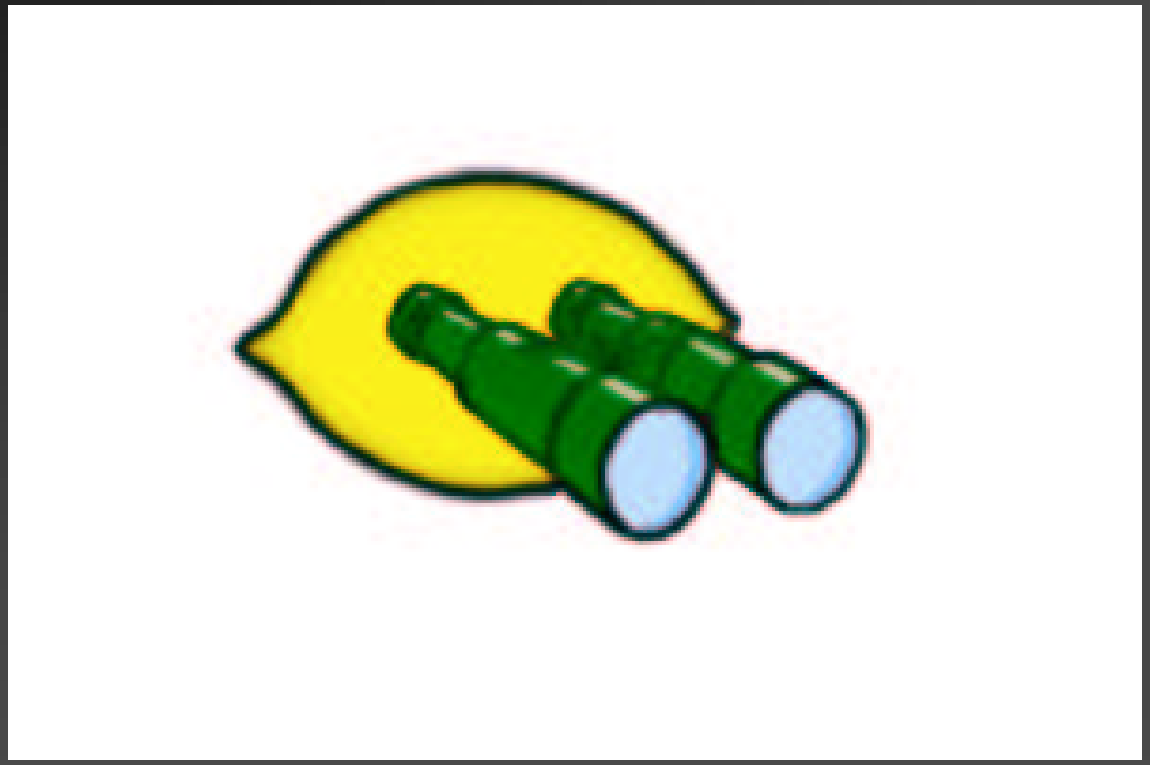
LCG Project Execution Board

23 September 2003

Outline

- LEMON – Monitoring
  - Overview
  - MSA and sensors
  - MSA → repository transport
  - Repository
  - Query API and displays
  - Local recovery, derived metrics, QoS
- LEAF – Advanced fabric management
  - Hardware management system
  - State management system
  - Fault tolerance, HSM operator interface

Lhc Era MONitoring



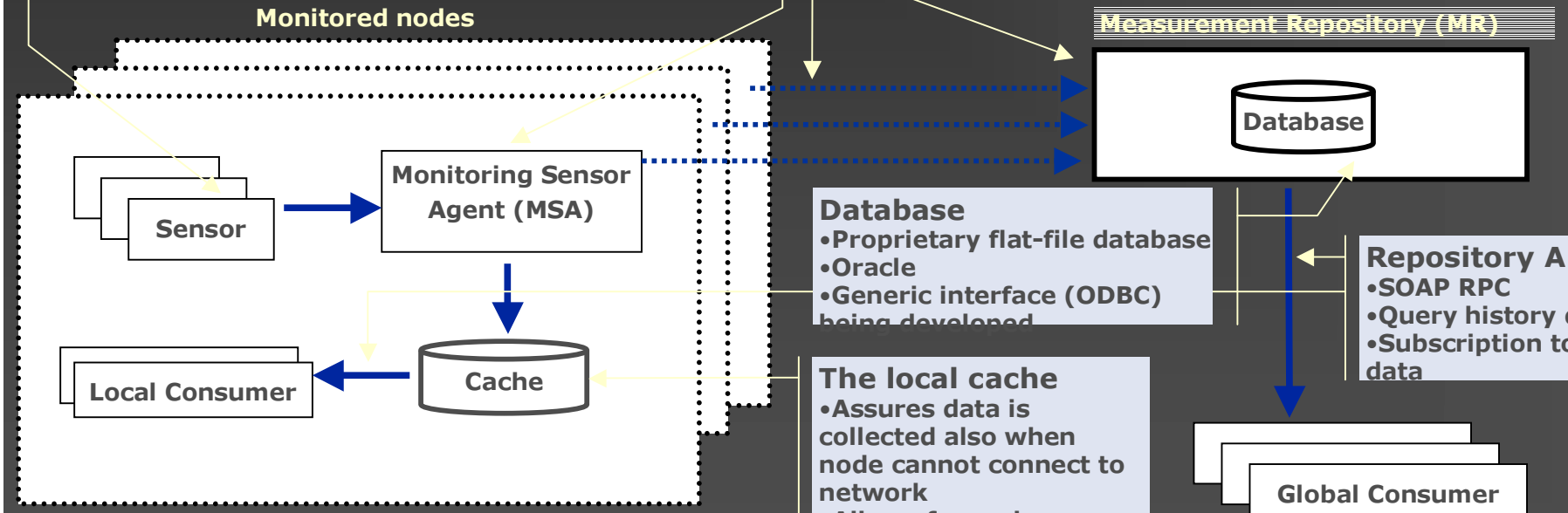
LEMON Overview

Monitoring Sensor Agent
- Calls plug-in sensors to sample configured metrics
- Stores all collected data in a local disk buffer
- Sends the collected data to the global repository

Transport
- Transport is pluggable.
- Two protocols over UDP and TCP are currently supported where only the latter can guarantee the delivery

Plug-in sensors
- Programs/scripts that implement a simple sensor-agent ASCII text protocol
- A C++ interface class is provided on top of the text protocol to facilitate implementation of new sensors

Measurement Repository
- The data is stored in a database
- A memory cache guarantees fast access to most recent data, which is normally what is used for fault tolerance correlations

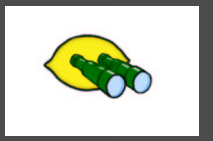


Database
- Proprietary flat-file database
- Oracle
- Generic interface (ODBC) being developed

Repository API
- SOAP RPC
- Query history data
- Subscription to new data

The local cache
- Assures data is collected also when node cannot connect to network
- Allows for node autonomy for local repairs

MSA and Sensors



- MSA stable for almost 2 years and in production on GNU/Linux machines at CERN for ~ 18 months
  - Increasing functionality over the period
- Sensors deployed on GNU/Linux to provide performance and exception metrics for HW, OS and application-specific items (eg. batch schedulers)
  - 150 metrics defined
  - Installed on almost 20 different clusters, 1500 nodes. 80-120 metrics collected per node depending on the cluster.
- Now developing sensors to collect other information, specifically from disk & tape servers
  - Much code already exists; need to bring measurements under the Lemon framework and collect metrics centrally

MSA → Repository Transport

- EDG/WP4 specific UDP based protocol in production.
  - Potential concern about routers dropping UDP packets, but no problem today
  - No security: anybody can inject any value into the repository
- TCP version of EDG/WP4 specific protocol tested
  - Some concern about load of multiple permanently open sockets on repository → proxies developed
  - Required for security (but nothing tested here) and also to resend metrics after network failure
- Need work to interface to SNMP world
  - E.g. for routers and switches
  - SNMP has been tested successfully for input to PVSS repository
  - Could be implemented as additional sensor, too

LEMON Repository (1)

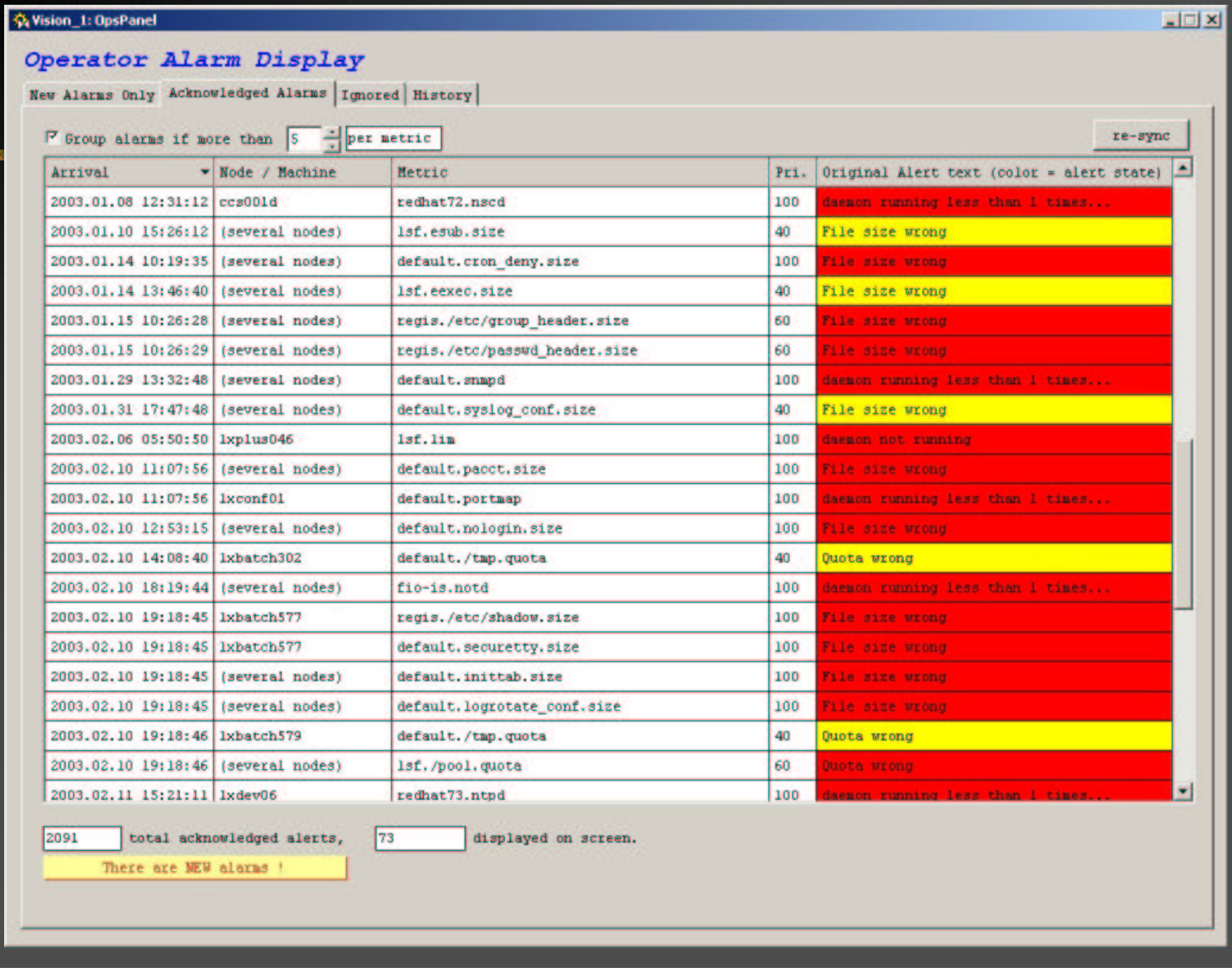
- Oracle-based repository required for long term storage of metrics
  - Needed to understand detailed (node to node) performance issues over required timescales (days…months)
    - Don’t want compression of data
- Oracle-based EDG/WP4 repository (OraMon) in production
- Alternative approach: PVSS-based repository, in production

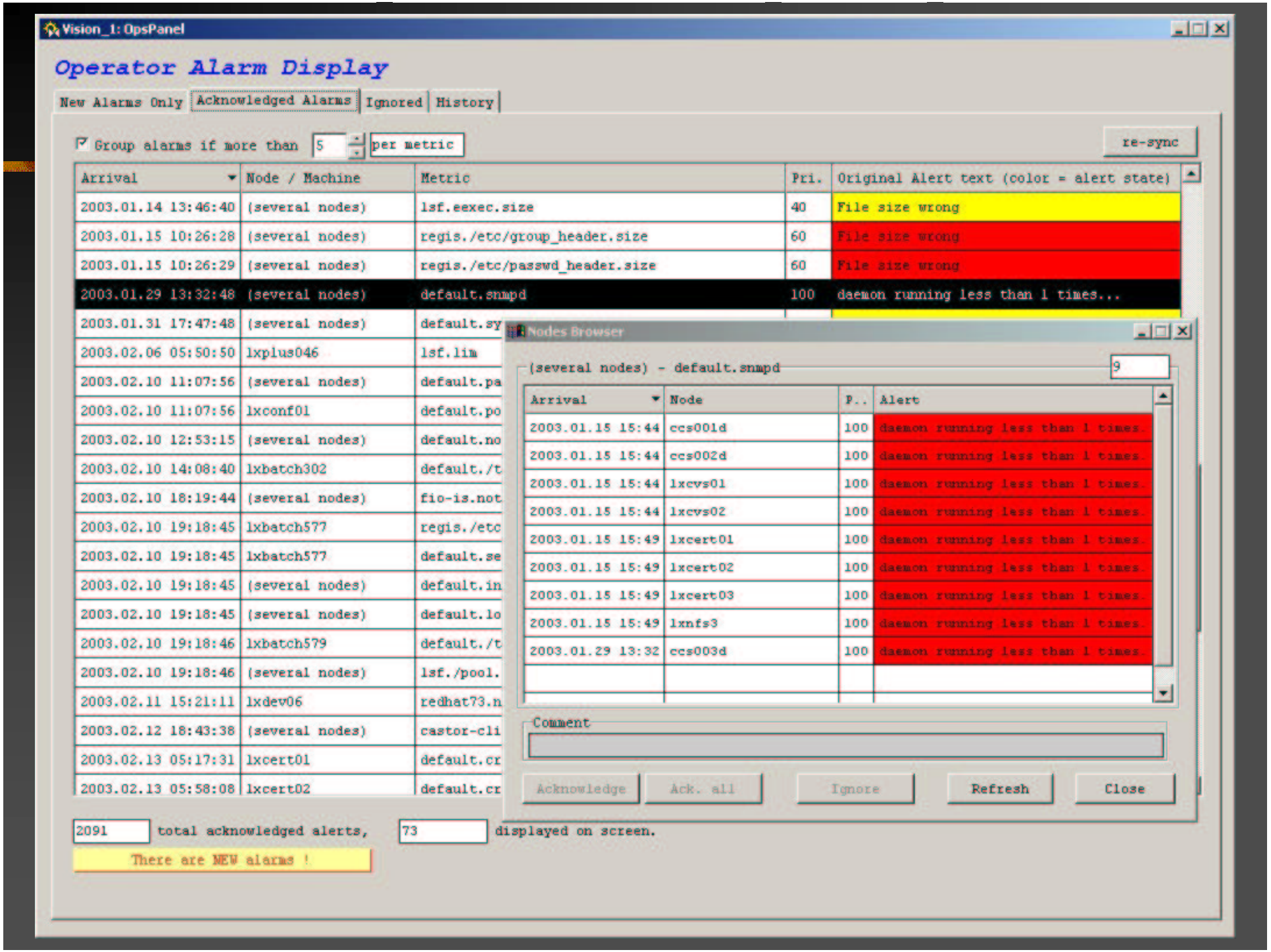
LEMON Repository (2)

- These two alternatives compared earlier this year
- Found that both systems can do the job, and are both necessary to fully address our requirements
  - Native Oracle archive for PVSS, promised for end-2003, is potentially interesting given our requirements…
- Decision (June): have all clients feed data both to OraMon repository and to PVSS
- Scaling limitations seen with both repositories already; very likely that the final (2006) system will run on multiple repositories of the same kind
  - Imminent deployment of smoothing at MSA level will alleviate this problem

Query API & Displays (1)

- EDG/WP4 defined an API to retrieve metrics
  - Implementations exist for Oracle repository (in C) and for direct extraction from PVSS (in C and Perl)
  - Perl (for OraMon) and command line implementations still to be done
- Operator and synoptic displays created for PVSS
  - Scales well with number of client machines so far
  - Used by bat. 513 operators in test mode for > 6 months (production system based on old SURE system)


Vision_1: OpsPanel
Operator Alarm Display
New Alarms Only
Acknowledged Alarms
Ignored
History
Group alarms if more than 5 per metric
re-sync
Arrival
Node / Machine
Metric
Pri.
Original Alert text (color = alert state)
2003.01.08 12:31:12 ccs001d redhat72.nscd 100 daemon running less than 1 times...
2003.01.10 15:26:12 (several nodes) lsf.esub.size 40 File size wrong
2003.01.14 10:19:35 (several nodes) default.cron_deny.size 100 File size wrong
2003.01.14 13:46:40 (several nodes) lsf.eexec.size 40 File size wrong
2003.01.15 10:26:28 (several nodes) regis./etc/group_header.size 60 File size wrong
2003.01.15 10:26:29 (several nodes) regis./etc/passwd_header.size 60 File size wrong
2003.01.29 13:32:48 (several nodes) default.snmpd 100 daemon running less than 1 times...
2003.01.31 17:47:48 (several nodes) default.syslog_conf.size 40 File size wrong
2003.02.06 05:50:50 lxplus046 lsf.lim 100 daemon not running
2003.02.10 11:07:56 (several nodes) default.pacct.size 100 File size wrong
2003.02.10 11:07:56 lxconf01 default.portmap 100 daemon running less than 1 times...
2003.02.10 12:53:15 (several nodes) default.nologin.size 100 File size wrong
2003.02.10 14:08:40 lxbatch302 default./tmp.quota 40 Quota wrong
2003.02.10 18:19:44 (several nodes) fio-is.notd 100 daemon running less than 1 times...
2003.02.10 19:18:45 lxbatch577 regis./etc/shadow.size 100 File size wrong
2003.02.10 19:18:45 lxbatch577 default.securetty.size 100 File size wrong
2003.02.10 19:18:45 (several nodes) default.inittab.size 100 File size wrong
2003.02.10 19:18:45 (several nodes) default.logrotate_conf.size 100 File size wrong
2003.02.10 19:18:46 lxbatch579 default./tmp.quota 40 Quota wrong
2003.02.10 19:18:46 (several nodes) lsf./pool.quota 60 Quota wrong
2003.02.11 15:21:11 lxdev06 redhat73.ntpd 100 daemon running less than 1 times...
2091 total acknowledged alerts, 73 displayed on screen.
There are NEW alarms !


Vision_1: OpsPanel
Operator Alarm Display
New Alarms Only
Acknowledged Alarms
Ignored
History
Group alarms if more than 5 per metric
re-sync
Arrival
Node / Machine
Metric
Pri.
Original Alert text (color = alert state)
2003.01.14 13:46:40 (several nodes) lsf.eexec.size 40 File size wrong
2003.01.15 10:26:28 (several nodes) regis./etc/group_header.size 60 File size wrong
2003.01.15 10:26:29 (several nodes) regis./etc/passwd_header.size 60 File size wrong
2003.01.29 13:32:48 (several nodes) default.snmpd 100 daemon running less than 1 times...
2003.01.31 17:47:48 (several nodes) default.sy
2003.02.06 05:50:50 lxplus046 lsf.lim
2003.02.10 11:07:56 (several nodes) default.pa
2003.02.10 11:07:56 lxconf01 default.po
2003.02.10 12:53:15 (several nodes) default.no
2003.02.10 14:08:40 lxbatch302 default./t
2003.02.10 18:19:44 (several nodes) fio-is.not
2003.02.10 19:18:45 lxbatch577 regis./etc
2003.02.10 19:18:45 lxbatch577 default.se
2003.02.10 19:18:45 (several nodes) default.in
2003.02.10 19:18:45 (several nodes) default.lo
2003.02.10 19:18:46 lxbatch579 default./t
2003.02.10 19:18:46 (several nodes) lsf./pool.
2003.02.11 15:21:11 lxdev06 redhat73.n
2003.02.12 18:43:38 (several nodes) castor-cli
2003.02.13 05:17:31 lxcert01 default.cr
2003.02.13 05:58:08 lxcert02 default.cr
Nodes Browser
(several nodes) - default.snmpd
9
Arrival
Node
P..
Alert
2003.01.15 15:44 ccs001d 100 daemon running less than 1 times.
2003.01.15 15:44 ccs002d 100 daemon running less than 1 times.
2003.01.15 15:44 lxcvs01 100 daemon running less than 1 times.
2003.01.15 15:44 lxcvs02 100 daemon running less than 1 times.
2003.01.15 15:49 lxcert01 100 daemon running less than 1 times.
2003.01.15 15:49 lxcert02 100 daemon running less than 1 times.
2003.01.15 15:49 lxcert03 100 daemon running less than 1 times.
2003.01.15 15:49 lxnfs3 100 daemon running less than 1 times.
2003.01.29 13:32 ccs003d 100 daemon running less than 1 times.
Comment
Acknowledge
Ack. all
Ignore
Refresh
Close
2091 total acknowledged alerts, 73 displayed on screen.
There are NEW alarms !

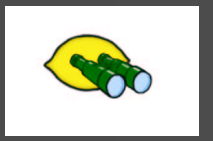
Query API & Displays (2)

- Display situation for EDG/WP4:
  - Java/swing alarm displays for small-scale clusters in development
  - A Java-based time series display exists but does not work for all browsers
  - Some simple web displays also exist (table based display of metric values extracted with the API)
  - Clarification about future directions needed

Local Recovery Actions

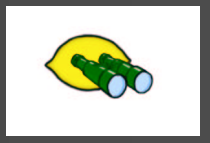
- Two parts:
  - a ***framework*** to interface to the repository
    - This “subscribes” to metrics and is alerted when metric values meet a defined condition
  - ***Actuators***
    - invoked as necessary by framework to take action (e.g. restart a daemon)
- A Framework has been developed by Heidelberg as part of EDG/WP4
  - Successfully invokes actuators in simple cases
  - More complex cases still to be tested
  - Work needed at CERN to implement actuators but lower priority than…

Derived Metrics



- Created upon values read from the repository and, perhaps, elsewhere, typically combining metrics from different nodes
  - e.g. maximum [system load|/tmp usage|/pool usage] on batch nodes running jobs for [alice|atlas|cms|lhcb]
  - Sampled regularly, and/or triggered by value changes
- Investigating whether EDG Framework is useful (and whether we wish to use it…)
  - Alternative implementation: Just another sensor

Quality of Service



- Essentially, a derived metric
  - per-application combination of system parameters that affect performance of the application on a node or on the overall service
- Initial BARC work was for a CPU bound application
  - Wall clock time increases with system load. Little affected by anything else except swapping
  - Now looking at other applications, in particular with I/O requirements

LEMON Status

- Sensors, MSA and OraMon and PVSS repositories are running in production, and running well
  - Status of, e.g. lxbatch, nodes has much improved since initial sensor/MSA deployment
  - Solaris port working
- Still much work to do, though, notably for displays, derived metrics and local recovery


LEAF

LEAF Components

LEAF

- HMS Hardware Management System
- SMS State Management System
- Fault Tolerance

HMS

LEAF

- HMS tracks systems through steps of HW life cycle; defined, implemented and in production:
  - Install
  - Move
  - Retire
  - Repair (vendor call)
- HMS was used (inter alia) to manage the migration of systems to the vault
- Some developments/improvements needed as other systems develop

SMS

LEAF

- Use cases:
  - “Give me 200 nodes, any 200. Make them like this. By then.”
  - “Take this sick node out of lxbatch”
- Tightly coupled to Lemon to understand current state, and to CDB to record changes of desired state, which SMS must update
- Analysis & requirements documents discussed. Led to architecture prototype and definition of interfaces to CDB
- Aim for initial SMS driven state change by the end of the year

Fault Tolerance

LEAF

- Requires actions such as reallocating a disk server to experiment A “because one of theirs has failed and they are the priority users at present”
- Could also
  - reallocate a disk server to experiment B because this looks a good move given the LSF/Castor load
  - reconfigure LSF queue parameters at night or weekend based on workloads and budgets
- Essentially just a combination of derived metrics and SMS
  - Concentrate on building the basic elements for now and start with simple use cases when we have the tools

HSM Operator interface

LEAF

- Ought to be a GUI on top of an intelligent combination of Quattor, Lemon and LEAF
- Want to be able to (re)allocate disk space (servers) to users/applications as necessary
- These are CDB and SMS issues
  - But strong requirements on Castor and the stager
    - Many of the necessary stager changes planned for next year
  - In the meantime, concentrate on extending CDB to cover definition of, e.g., staging pools

Conclusion

- FIO have gone a long way in the direction of fabric automatisation
  - Still a long way to go
    - Some areas still sketchy
- We have started to see enormous benefits already