The High Performance Archiver for the LHC Experiments

Manuel Gonzalez Berges
CERN, Geneva (Switzerland)

Outline

- Context
- Archiving in PVSS
- PVSS Client
- Database Server
- Conclusions

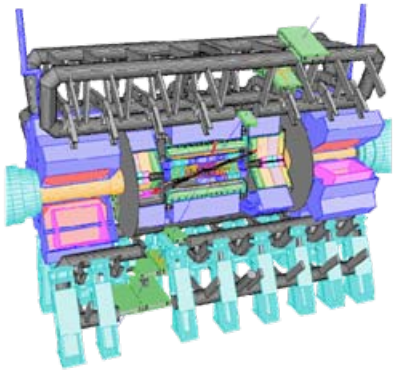
Controls

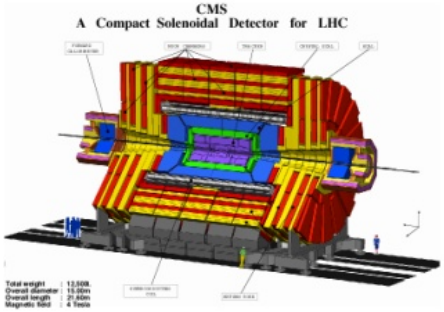
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The LHC Experiments

ALICE

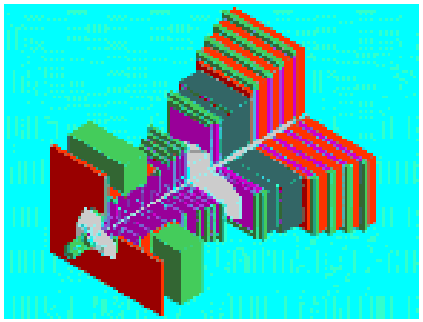
ATLAS





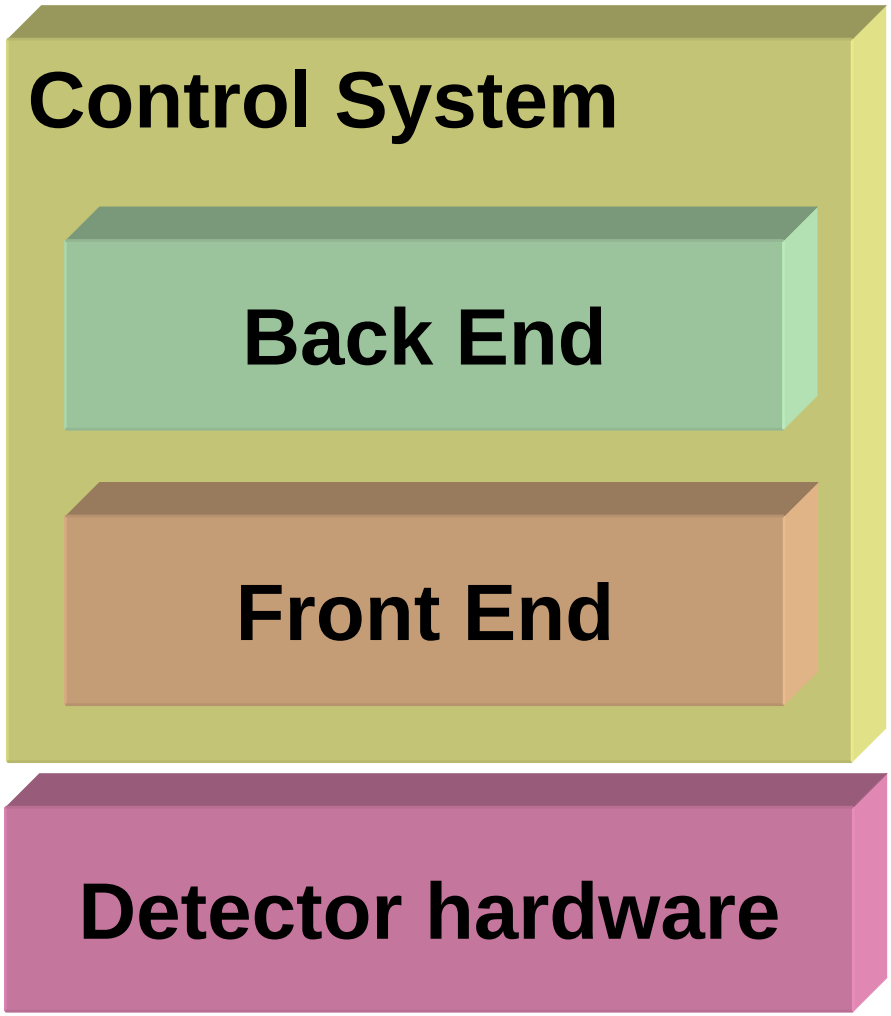
CMS

LHCb



- Worldwide collaborations
- One order of magnitude bigger than previous generation (~1 million i/o)
- Complex operation
  - Many subdetectors and subsystems
- Lifetime of 20 years
- Common approach to controls (Joint Controls Project – JCOP)
- Currently installing and commissioning

LHC experiment control system

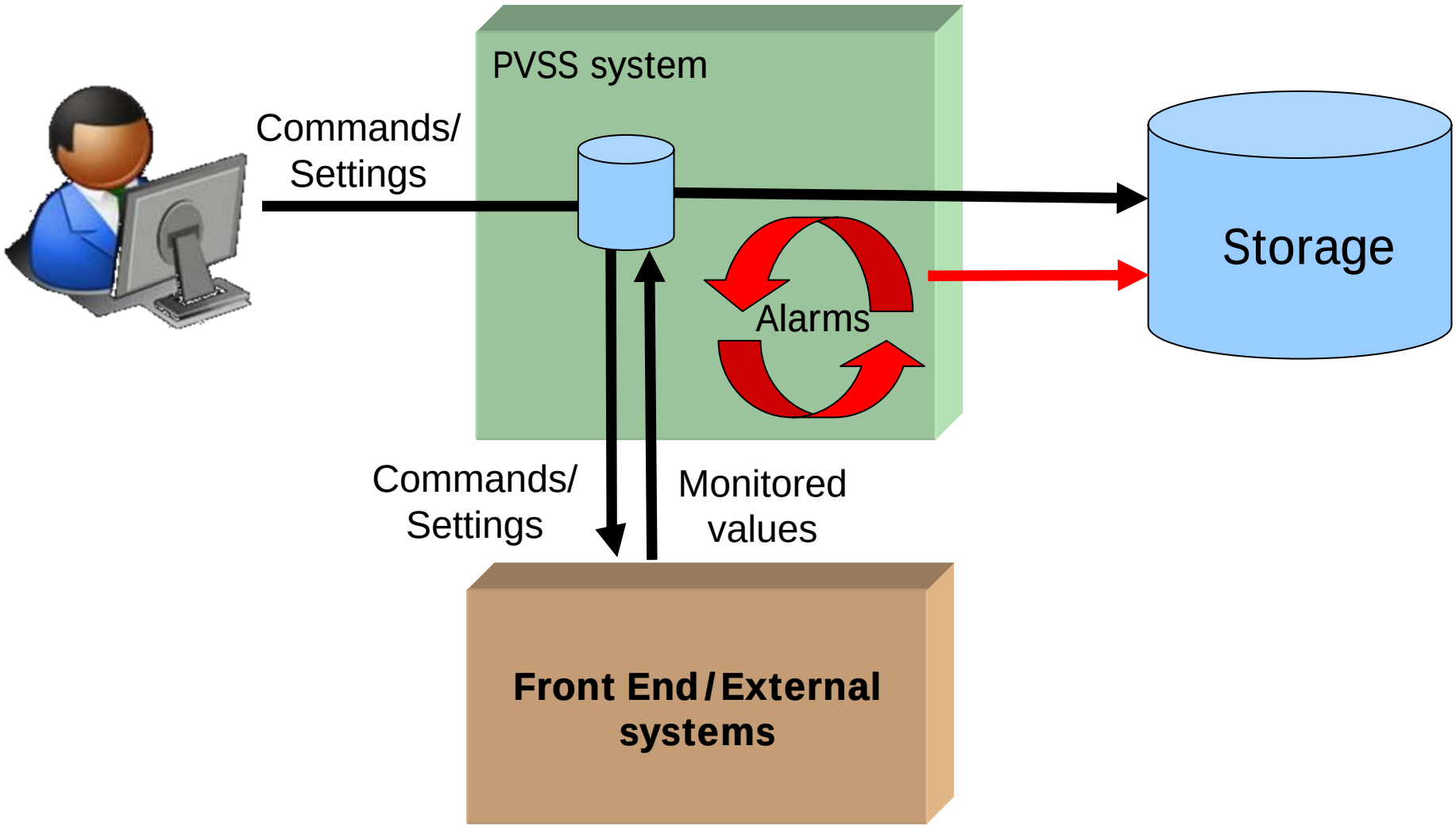


- Back End
  - Linux/Windows PCs (~150)
  - PVSS + JCOP Framework
  - Main functions
    - Graphical user interfaces
    - Operation as a set of Finite State Machines (FSM)
    - Alarm handling
    - Interface to external systems
    - **Archiving**
- Front End
  - Several technologies
    - PCs, PLCs, Embbeded computers, etc
  - Main functions
    - Data acquisition
    - Filtering
    - Real time loops & FSM
    - Interlocks

LHC experiments archiving

- Purpose
  - Debugging of the control system
  - Operation
  - Physics offline analysis (conditions data)
- Requirements
  - Data storage
    - Single computer peak rate of 2000 changes/s
    - Full application sustained rate of ~150 000 changes/s (~150 computers at 1000 changes/s)
  - Data retrieval
    - Optimize known common queries
      - Get a set of values for a specified time range
      - Snapshot at a given time
    - Other queries will come with usage

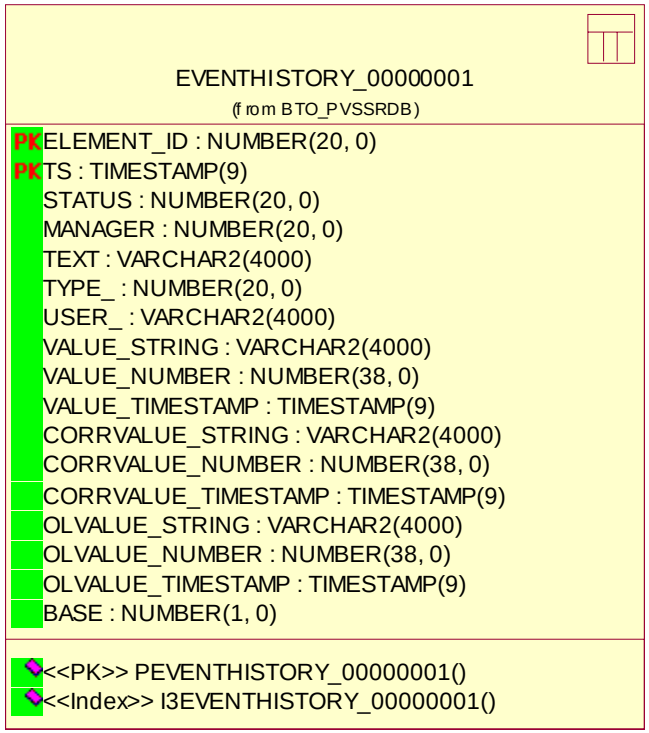
Archiving in PVSS

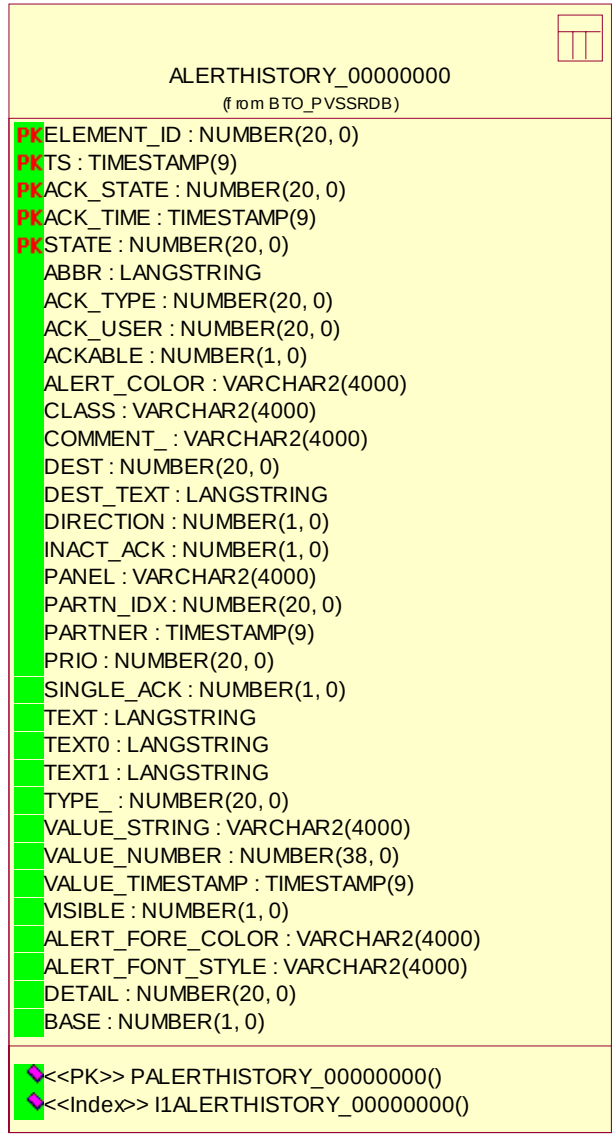


File vs Database archiver in PVSS

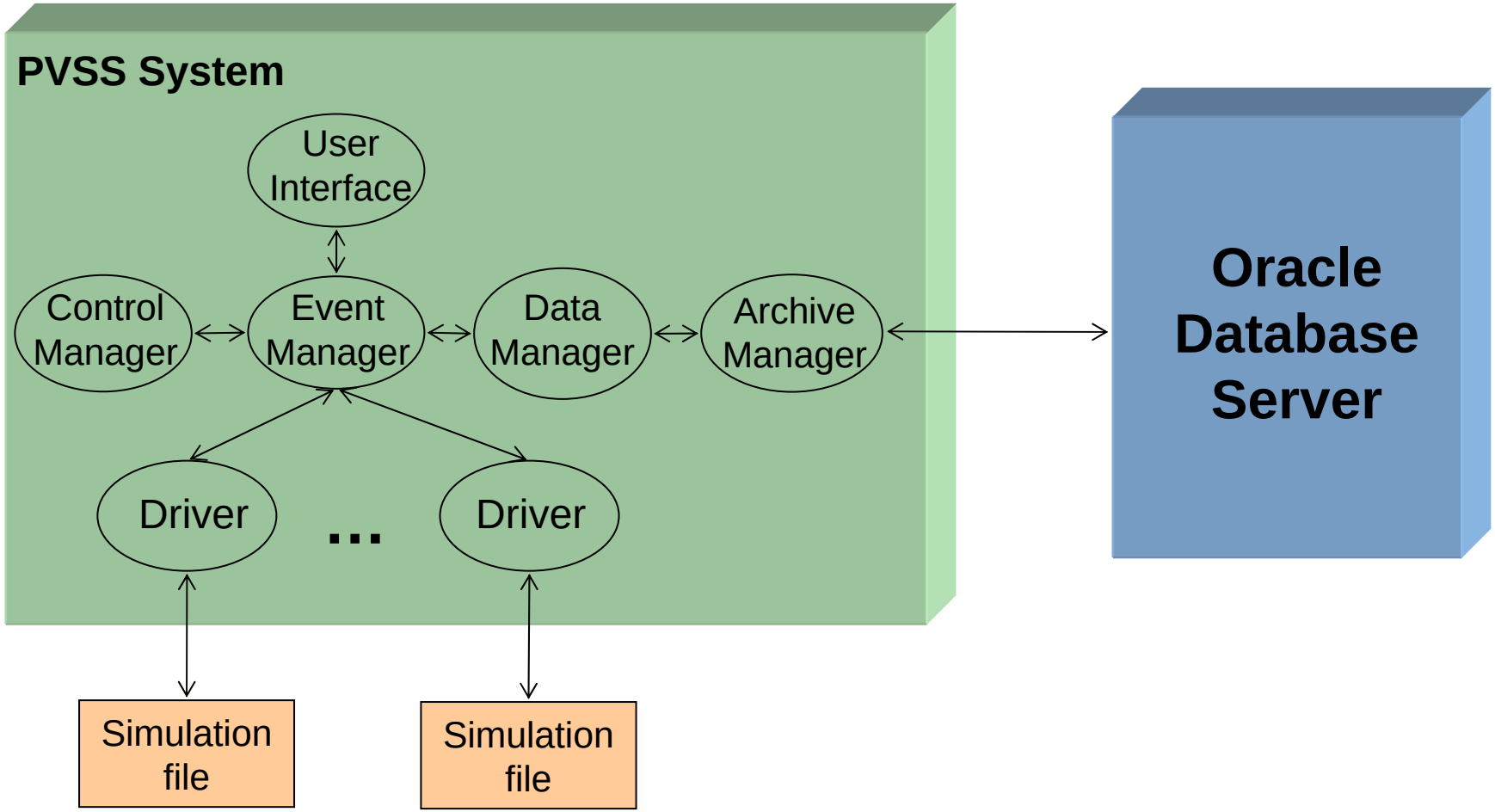
- File archiver
  - Local hardisk of each computer
  - Issues with managing many big files
  - Proprietary format
- Database archiver
  - Centralized server
  - Relational database
  - Initially developed for some specific customers
    - Performance far from the extreme requirements of the LHC experiments
    - Close collaboration ETM – CERN to improve it
      - CERN expertise in databases
      - Facilities to test on very large systems

Tables


EVENTHISTORY_00000001
(from BTO_PVSSRDB)
PK ELEMENT_ID : NUMBER(20, 0)
PK TS : TIMESTAMP(9)
STATUS : NUMBER(20, 0)
MANAGER : NUMBER(20, 0)
TEXT : VARCHAR2(4000)
TYPE_ : NUMBER(20, 0)
USER_ : VARCHAR2(4000)
VALUE_STRING : VARCHAR2(4000)
VALUE_NUMBER : NUMBER(38, 0)
VALUE_TIMESTAMP : TIMESTAMP(9)
CORRVALUE_STRING : VARCHAR2(4000)
CORRVALUE_NUMBER : NUMBER(38, 0)
CORRVALUE_TIMESTAMP : TIMESTAMP(9)
OLVALUE_STRING : VARCHAR2(4000)
OLVALUE_NUMBER : NUMBER(38, 0)
OLVALUE_TIMESTAMP : TIMESTAMP(9)
BASE : NUMBER(1, 0)
<<PK>> PEVENTHISTORY_00000001()
<<Index>> I3EVENTHISTORY_00000001()

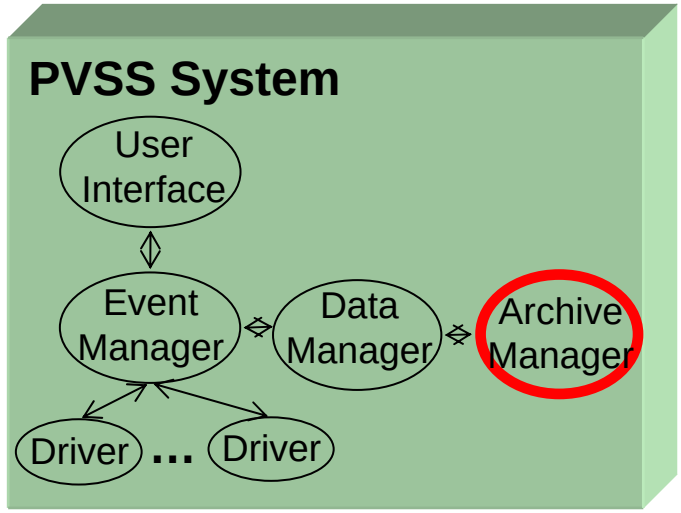

ALERTHISTORY_00000000
(from BTO_PVSSRDB)
PK ELEMENT_ID : NUMBER(20, 0)
PK TS : TIMESTAMP(9)
PK ACK_STATE : NUMBER(20, 0)
PK ACK_TIME : TIMESTAMP(9)
PK STATE : NUMBER(20, 0)
ABBR : LANGSTRING
ACK_TYPE : NUMBER(20, 0)
ACK_USER : NUMBER(20, 0)
ACKABLE : NUMBER(1, 0)
ALERT_COLOR : VARCHAR2(4000)
CLASS : VARCHAR2(4000)
COMMENT_ : VARCHAR2(4000)
DEST : NUMBER(20, 0)
DEST_TEXT : LANGSTRING
DIRECTION : NUMBER(1, 0)
INACT_ACK : NUMBER(1, 0)
PANEL : VARCHAR2(4000)
PARTN_IDX : NUMBER(20, 0)
PARTNER : TIMESTAMP(9)
PRIO : NUMBER(20, 0)
SINGLE_ACK : NUMBER(1, 0)
TEXT : LANGSTRING
TEXT0 : LANGSTRING
TEXT1 : LANGSTRING
TYPE_ : NUMBER(20, 0)
VALUE_STRING : VARCHAR2(4000)
VALUE_NUMBER : NUMBER(38, 0)
VALUE_TIMESTAMP : TIMESTAMP(9)
VISIBLE : NUMBER(1, 0)
ALERT_FORE_COLOR : VARCHAR2(4000)
ALERT_FONT_STYLE : VARCHAR2(4000)
DETAIL : NUMBER(20, 0)
BASE : NUMBER(1, 0)
<<PK>> PALERTHISTORY_00000000()
<<Index>> I1ALERTHISTORY_00000000()

Single System Archiving (I)

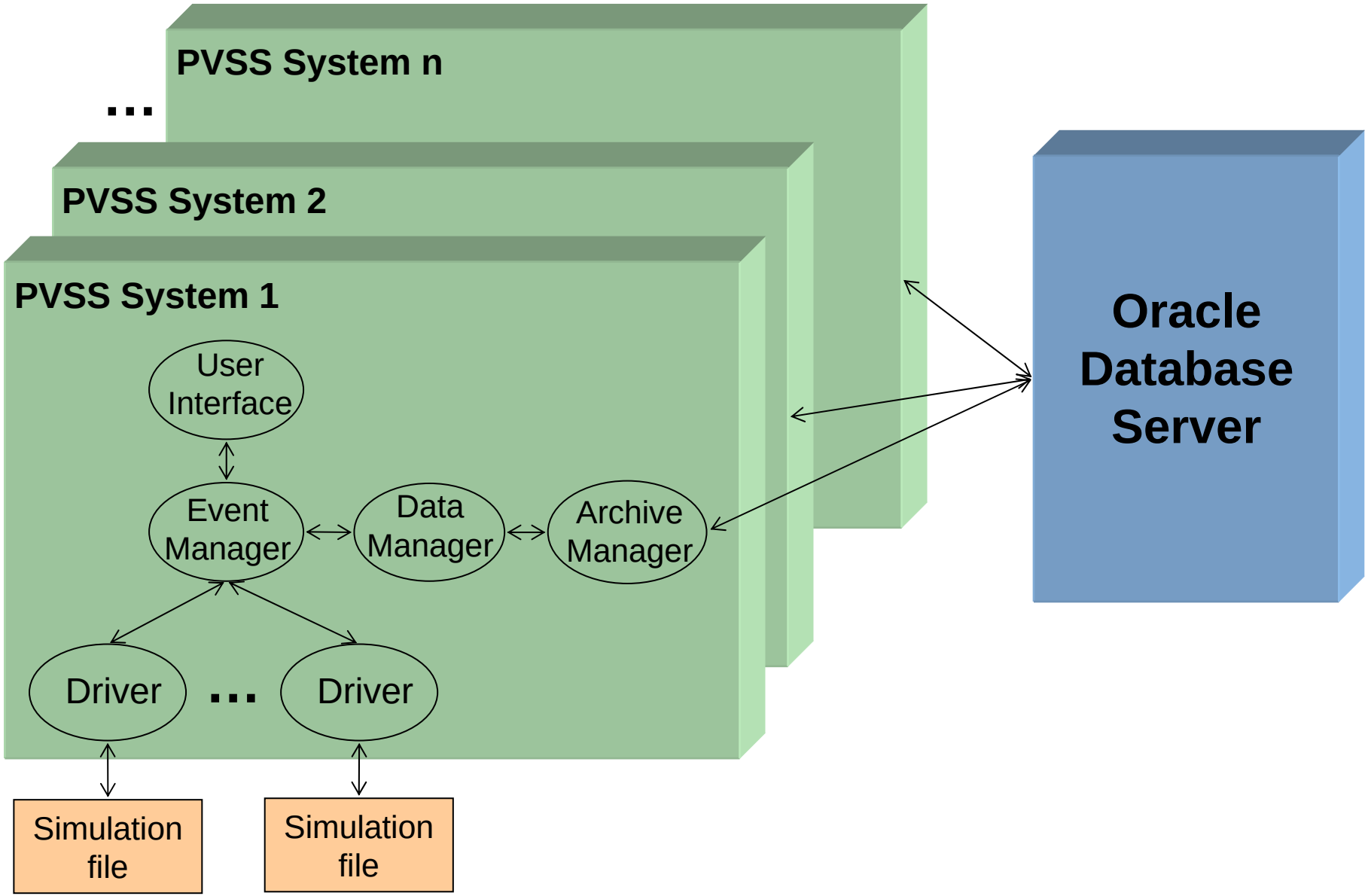


Single System Archiving (II)

- Test results
  - ~100-300 changes/s depending on setup
  - Bottleneck in Archive Manager
  - Generic interface to database
  - Value changes sent one by one
- Improvements
  - Use of Oracle native libraries (OCCI)
  - Bulk insertion
    - Client sends blocks to DB
    - A function inside the DB inserts the data into the history tables
  - Reduce number of connections to DB
    - One permanent for storage/retrieval
    - One temporary for configuration
  - Possible to store 2000 changes/s continuously



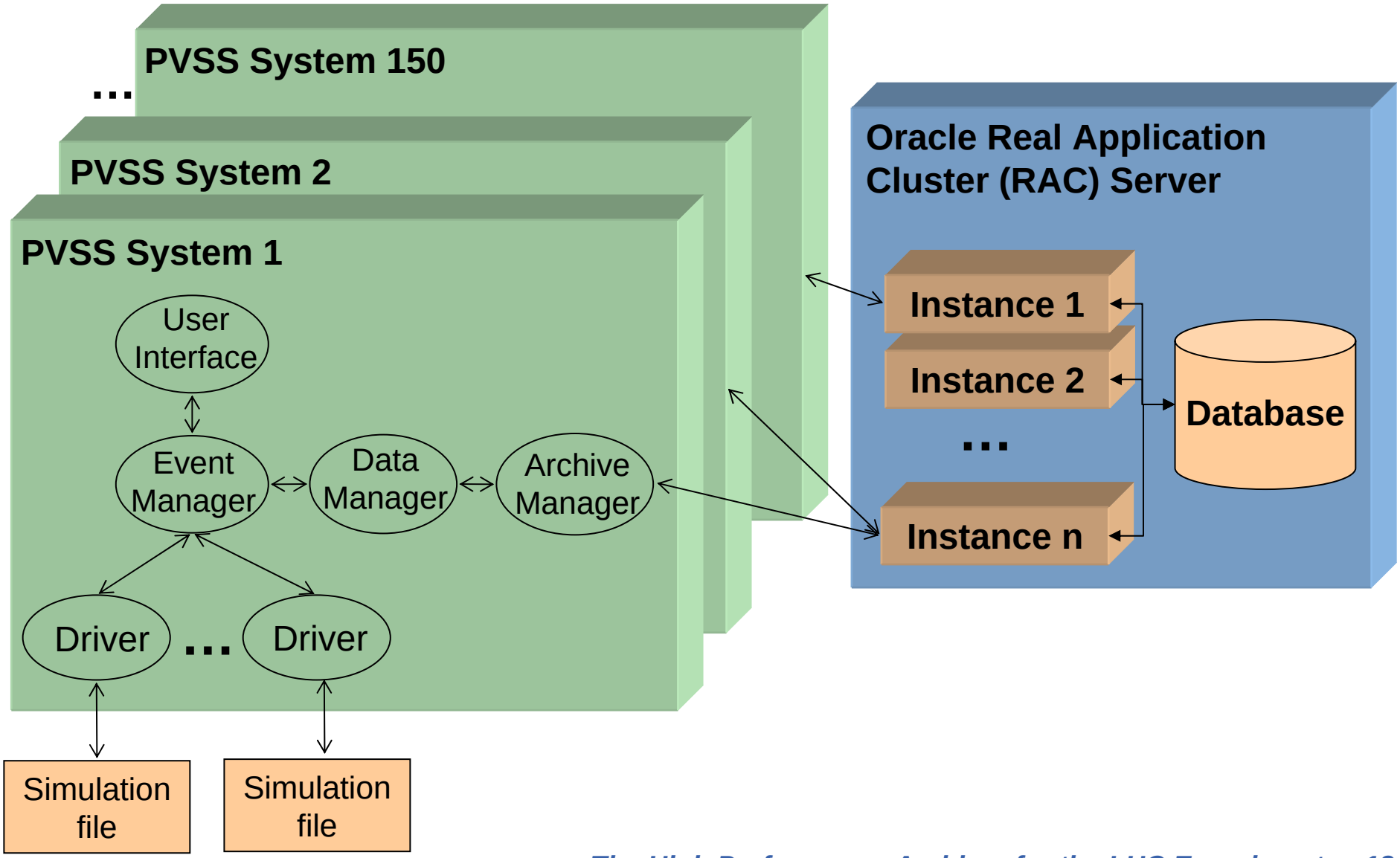
Test load on the database server



Performance of Oracle server

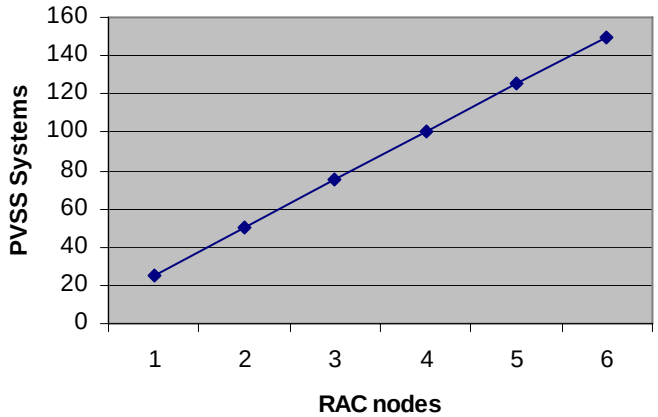
- Oracle server hardware
  - PC with two Xeon processors, 3 GHz, 4 GBytes RAM, Red Hat Enterprise 4
- Tests results
  - Server can handle ~20-30 systems each at 1000 changes/s
  - High CPU load
- Options for improvement
  - Server with better hardware
    - High cost, limited improvement
  - Clustered server
    - Low Cost (PCs)
    - Redundancy -> high availability
    - Easy to upgrade by adding more nodes
    - Issues with scalability

Test setup with RAC server



Test results with RAC

- Initial test results
  - 4 server nodes could handle ~ 50 systems
  - Does not scale linearly
    - The nodes interfere with each other
- Main improvements
  - Direct path insertion
    - Constraints disabled during insertion
    - Can be inefficient in space usage
  - History table partitioned per PVSS client
    - Isolates clients
- Final performance
  - Scalable server
    - 6 nodes could handle 150 clients at 1000 changes/s each
    - Possible to add more nodes
  - Issue with allocation of new storage space (tablespace creation)



Outlook

- Values
  - Insertion
    - Prototype for issue with new space allocation
      - Space preallocated in advance or with background job
  - Queries
    - Continue with optimization of known queries
    - Develop an API for external programs
- Alarms
  - Bottlenecks identified
  - ETM working on them
  - Not extreme requirements
- Buffering to disk when DB not reachable

Conclusions

- LHC experiments archiving requirements are met
  - Improvements included in standard PVSS version since earlier this year
  - Issues with retrieval and alarms understood
    - New version expected for the end of the year
- Fruitful collaboration ETM-CERN
  - Benefit from CERN expertise on databases
  - Many solutions were explored with real size test bench
  - Maintenance guaranteed by the company
- The archiver is currently in use for the commissioning of the LHC experiments

Acknowledgements

- The following people have been involved in the work presented:
  - CERN
    - Chris Lambert
    - Eric Grancher
    - Laura Fernandez
    - Luca Canali
    - Milosz Hulboj
    - Nilo Segura
    - Pior Golonka
    - Svetozar Kapusta
  - ETM
    - Ewald Sperrer
    - Ronald Putz
  - Oracle
    - Lothar Flatz

Questions?