Hunting for hardware changes in data centres

M. Coelho dos Santos, I. Steers, I. Szebenyi, A. Xafi, O. Barring, E. Bonfillou
European Organization for Nuclear Research (CERN), 1211 Geneva 23, Switzerland
miguel.coelho.santos@cern.ch

Abstract. With many servers and server parts the environment of warehouse sized data centres is increasingly complex. Server life-cycle management and hardware failures are responsible for frequent changes that need to be managed. To manage these changes better a project codenamed "hardware hound" focusing on hardware failure trending and hardware inventory has been started at CERN. By creating and using a hardware oriented data set - the inventory - with detailed information on servers and their parts as well as tracking changes to this inventory, the project aims at, for example, being able to discover trends in hardware failure rates.

1. Introduction
In the early 2000s, when developing the computing infrastructure and facilities for the Tier-0 of the Worldwide LHC Computing Grid[1], CERN's data center started being populated by large quantities of commodity x86 based servers running Linux, currently around eight thousand. With the massification of the World Wide Web, commercial companies such as Google, Facebook and Amazon have, since, greatly surpassed the number of servers CERN manages.

Managing a large population of commodity x86 servers has challenges regarding inventory and failure trending. In the last five years there have been some studies on failure trending of components of commodity x86 server among which the contributions from Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber on Disk Failures [2][3] and DRAM errors [4]. It is critical to not only know what servers and components are deployed but also how and when they fail.

The configuration management database used at CERN – Quattor[5] – has information on the deployed hardware. That together with the data recorded for server component replacement intervention has allowed for calculating failures rates [6][7].

2. Goals
The goal of this project, codename Hardware Hound, is to improve the current inventory by recording changes to each server, keeping all server and component history, improve the quality of data and allow for continuous monitoring and reporting of the relevant metrics with high level of detail, e.g. age of each component, or allow for grouping by different characteristics such as model, firmware level, etc.

3. The inventory
The new hardware inventory is a document database (CouchDB), containing one document per server. Each document contains a very detailed description of the hardware in the server and related history. There are seven main areas being tracked: BMC, ATA, NETWORK, NIC, RAM, CPU and BIOS.
Each of these contains high levels of detail, as seen by the operating system, including serial numbers, firmware levels, model, etc.

Existing inventory solutions, such as OCS inventory, were considered but found lacking the level of detail and historical record keeping capabilities required. Being difficult to port the existing level of detail into an existing inventory tool it was considered more efficient to store the data in a simple document database.

The server Purchase Order ID – from here on referred to as ‘Server Vintage’ – and the server serial number identifies each server document.

4. The collector
The data collector is a Python script that uses existing libraries that wrap several CLI’s and utilities and allows for querying them. The collector’s job is to find out the current hardware setup on the system and compare it to the previous state. If there are changes then it works out the details of the change, formats the change and then append it to the document’s changelog. Information is sent in JSON format over https.

5. The Displays
A series of displays have been developed. These displays have been built as a django app running on the same nodes as the CouchDB cluster.

The aim is to report basic statistics on the hardware deployed on the data centre, for example number of CPUs, memory modules, disk, network interface cards, RAID cards, etc.

Besides basic statistics the displays report on annualized failures rates – AFR[8]. One of the important features is to be able to display failures and population by component age and server vintage.

Figure 1. Hardware Inventory diagram
6. Deployment
The Server Vintage and the serial number are stored locally on the server, they are stored on the FRU of the BMC and do not change during the server lifetime. New servers are purchased with these data already recorded in the FRU. A campaign took place to record these data on already deployed servers.

When the collection script runs it queries the FRU for both values, if for some reason the values are not there then the script exits cleanly. The log files produced by the collection script are forwarded on to Splunk[9] which allows us to monitor the incoming logs and check for certain events. This allows us to monitor problems such as missing or duplicate serial numbers.

7. Initial results
So far the inventory contains the history of around seven thousands servers and their full history since the beginning of 2012.

Older historical data has been imported but it relied on reports that should have been gathered on every intervention to replace a server component. These reports have not been triggered consistently according to the analysis of the data; therefore older imported data appear to be, for some server vintages, of lower quality.

With the inventory data currently collected it is possible to generate reports such as the distribution of CPU types (figure 2.) as well as of other components.

![Figure 2. CPU type distribution](image)

Historical data allow us to see, for example, the age of the disk population by server vintage on a given moment (figure 3.) as well as the evolution of the disk population of a given server vintage over time. Regarding failures it is now possible to see failures by disk age and server vintage (figure 4.), which we believe is important to estimate the quality and health of server vintages.

![Figure 3. Disk Population by disk Age, by Server Vintage at the end of December 2011](image)
8. Conclusion

With these developments the system is now capable of tracking changes in great detail. As data are recorded we expect to identify systematic problems in server vintages being purchased and do identify ‘old age’ in components already in production.

References

[1] LCG - Worldwide LHC Computing Grid, http://lcg.web.cern.ch/LCG/Another reference
[2] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz Andre; Barroso. 2007. Failure trends in a large disk drive population. In Proceedings of the 5th USENIX conference on File and Storage Technologies (FAST ’07). USENIX Association, Berkeley, CA, USA, 2-2.
[3] Bianca Schroeder and Garth A. Gibson. 2007. Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you?. Trans. Storage 3, 3, Article 8 (October 2007).
[4] Bianca Schroeder, Eduardo Pinheiro, and Wolf-Dietrich Weber. 2009. DRAM errors in the wild: a large-scale field study. In Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems (SIGMETRICS ’09). ACM, New York, NY, USA, 193-204.
[5] http://quattor.sourceforge.net/
[6] Olof Barring, Wayne Salter. 2011 Hardware failures at CERN, HEPiX Fall Workshop, http://indico.cern.ch/contributionDisplay.py?contribId=23&confId=138424
[7] Miguel Coelho dos Santos, Dennis Waldron, "Observations made while running a multi-petabyte storage system," MSST, pp.1-4, 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010.
[8] J. G. Elerath. AFR: problems of definition, calculation and measurement in a commercial environment. In Proc. of the Annual Reliability and Maintainability Symposium, 2000
[9] http://www.splunk.com/