Large CPU-Farm Implementation in a HEP experiment with Tight Constraints (June 2005)

L. Brarda, B. Gaidioz, D. Ruffinoni, P. Gavillet and G. Decreuse, CERN

Abstract—The LHCb event filter farm will contain ~ 2000 dual-CPU computers, located in a 80 square meters room underground at the experiment pit. To implement such a large farm in such a small volume, a lot of issues have to be addressed. This paper describes the powering, cooling and packaging solutions developed by LHCb for its event filter farm.

Index Terms—Computers, Computer power supplies, Cooling

I. INTRODUCTION

Building a 2000 computers farm is not an easy task. In the LHCb experiment case, there are several additional constraints because, with an economic aim, the experiment wishes to reuse most of the already existing infrastructure previously used by the Delphi experiment on the site of installation. These constraints are mostly the reuse of one of the 80 m2 Delphi counting room to host the farm and the reuse of the 19 inch / 59 U Delphi racks. On the other hand, the existence of the cooling water network and electrical distribution panels are clear advantages.

Because of the computer density, we quickly made several choices: use 1U rackable computers (keeping in mind that our solution must be compatible with blade systems) and water-cooling (traditional air cooling is impossible).

We also wanted to be able to replace each computer in the farm by another that could be of a different model or a different brand.

We will now detail what we learned from our tests and the solutions we found on all the issues.

II. TEST SETUP

The test setup is composed of two racks fully equipped with cooling equipment and power distribution. They have a standard width of 19 inches, are 59 U high and 900 mm deep.

The first rack contains forty six 1U computers of mixed type, one 100Mbit 48 ports control switch and two 1 Gigabit 24 ports data switches. The second rack contains forty four 1U computers of the same type (Dell PowerEdge SC1425), one 100Mbit 48 ports control switch (3COM 4250T) and one 1 Gigabit 48 ports data switch (3COM 3870).

The measuring equipment is composed of a Keithley Model 2000 digital multimeter, a Keithley Model 7001 multi-channel scanner linked via GPIB to a PC measuring station running a PVSS [2] application.

For temperature measurement, PT100 probes are connected to the multi-channel scanner itself connected to the digital multimeter.

An Infratek 106A power analyzer is used for manual power measurements (30 A maximum current). See ref. [3] for the equipment used for the power supplies characterization.

III. POWERING

A. Computer power supply characterization

To have a better understanding of the powering issues, measurements were made by the CERN electrical service on different types of 1U computers [3], [4].

The total harmonic current distortion reaches up to 30.7 % of the nominal current which will result in neutral currents below 100% of the nominal current (if the load is balanced between the three phases), which looks acceptable.

The inrush current at power on can be higher than 60 times the nominal current for a few cycles. In practice, it is impossible to precisely predict the total inrush current of large quantities of PCs, even if the inrush current of a single PC is known. Inrush currents up to 400...700A per rack can be expected.

B. Power consumption

We measured with a power analyzer (Infratek 106A) the consumption of our different PCs in three different states: Off, Idle and full charge (~100 % CPU).
TABLE I

<table>
<thead>
<tr>
<th>PC type</th>
<th>Off (W)</th>
<th>Idle (W)</th>
<th>Full charge (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elonex Mono CPU 2.4 GHz</td>
<td>6.4</td>
<td>55.25</td>
<td>96</td>
</tr>
<tr>
<td>Melrow Mono CPU 2.4 GHz</td>
<td>4.5</td>
<td>59</td>
<td>100</td>
</tr>
<tr>
<td>Melrow Dual CPU 2.8 GHz CW2</td>
<td>8.5</td>
<td>134</td>
<td>241</td>
</tr>
<tr>
<td>Melrow Dual CPU 2.8 GHz WV2</td>
<td>12</td>
<td>118</td>
<td>225</td>
</tr>
<tr>
<td>Melrow Dual CPU 3.06 GHz CW2</td>
<td>8</td>
<td>127</td>
<td>269</td>
</tr>
<tr>
<td>Dell SC1425 Dual CPU 2.8 GHz</td>
<td>174</td>
<td>276</td>
<td></td>
</tr>
</tbody>
</table>

For the rack with 46 computers of mixed types, this sums up to around 6900 W at full charge. For the rack with 44 Dell SC1425 computers, this sums up to over 12 000 W.

C. Chosen powering scheme

After the measurements, we made some practical tests (CERN/TS/EL) to define the electrical characteristics (e.g. number of power lines, type of circuit breakers) that will ensure a safe powering of a rack full of PCs in one go.

As we already have tri-phase electric distribution racks (composed of several 63 A tri-phase TDM drawers) in our computing room, we choose the following scheme (Fig 1): The power cable from each TDM drawer goes to three racks. In each rack, we have a power distribution box with a time relay and six circuit breakers (two per phase) which then feeds six IEC 320 power distribution units to which the computers are finally plugged. The circuit breakers in the TDM drawers are of C curve type and the ones in the power distribution boxes are of D curve type (10 to 14 times the nominal current for about 1s). The delay for the power up of each rack is configurable so we power the racks consecutively.

D. Tests

Following the choice of powering scheme, two prototypes of the distribution box have been made. They were installed on the two racks in the test setup. A C curve circuit breaker was installed in the TDM drawer.

Several power on tests were made with delays from 5 seconds to 20 seconds between the 2 racks. We powered on successfully the installation from the power feed of the TDM drawer each time (no circuit breaker drop).

IV. COOLING

The most obvious challenge was the cooling of the farm. At CERN, we had some experience with horizontal heat exchangers in electronic racks. As rackable computers have an horizontal front to back airflow, we could not use horizontal heat exchangers, so we started searching on the market for vertical heat exchangers. As the three other LHC experiments had nearly the same concerns, an informal working-group was setup.

A first set of tests [1] were made, with a heat exchanger giving 7 kilowatts (kW) of cooling power, and showed that the concept was good. As the cooling power was too low, we designed a new heat exchanger with a company specialized in cooling (specifications in Table II). We setup a new lab for the LHCb farm prototype, with all our specificities taken into account. We first installed one of our racks with 30 single CPU computers and 16 new dual-CPU computers. The prototype of new heat exchanger is integrated in the back door (Fig. 3). With that setup (and with air-conditioner stopped), we did some measurements to check that the heat-exchanger was working as desired. The estimated power consumption for the 46 computers is 6900 Watts. We calculate the power extracted by the water from the difference of temperature of the cooling water (see [1]).

In Fig. 2, we see that there is more power going into the water than the computers electrical power and that the temperature at the rear is lower than the temperature in front of the rack. Hence, the heat exchanger is working well in this configuration where we have more cooling power than electrical power consumed.
Table II
Heat exchanger specifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooling power</td>
<td>10 kW</td>
</tr>
<tr>
<td>Water flow</td>
<td>2 m³/h</td>
</tr>
<tr>
<td>Water in temperature</td>
<td>15 ºC</td>
</tr>
<tr>
<td>Max water out temperature</td>
<td>20 ºC</td>
</tr>
<tr>
<td>Max air out temperature</td>
<td>20 ºC</td>
</tr>
</tbody>
</table>

We bought 44 more powerful Dual-CPU computers and setup another rack with them. We used the final version of the heat exchanger (only minor difference with the prototype like the 15 cm deeper door to be able to put longer computers). The total electrical power for this new rack exceed 12 kilowatts at full charge (100 %CPU). The preliminary results (Fig. 4) show that, with a 14 ºC cooling water, the heat exchanger cannot fully absorb all that electrical power: the temperature at the back of the rack is higher than the front. More tests will be needed but it seems that the maximum power that the heat exchanger can transfer to water will be around 12 kilowatts.

V. PACKAGING

A. Computer rack-mounting

The main problem in the integration of the 1U computers was the mounting of these computers in the rack. Most 1U computers are sold with sliding rails specific to each 1U chassis. As we do not want to be tied to a specific chassis (or to a specific manufacturer), and want to be able to replace any faulty computer with one of our spares, even if they are not of the same type, these sliding rails were not the solution.

We saw in the different models we already acquired that there were big differences between the different chassis dimensions.

With the electronic chassis we are used to, we were using corner angles to support the chassis. The problem is that some of the 1U computer chassis are exactly 1U high, so there is no place between 2 chassis to put anything. We had the idea to add 2 millimeters to (nearly) each U in the rack. This will make us use 23 U to put 22 (1U) computers, but in these 2 mm, we could put the corner angles.

To add these 2 mm to each unit, we fixed four horizontal supporting bars (Fig. 5, Fig. 6) on the normal rack mounting holes, drilled and taped to have the rack occupation as showed on Fig. 7: 4 real units left at the bottom of the rack, 22 new units (in 23 real units) for one subfarm, 5 real units in the middle of the rack for the network switches and the network patch panel, then again 22 new units for the second subfarm and 4 real units for the electrical distribution box and its cables. It’s how the second rack of our test is setup.

On these supporting bars, we screw 1mm inox corner angles. On Fig. 8, you can see a detail of the rack, one supporting bar and two corner angles.
B. Cabling

Power cables are directly wired from the power bars to the computer’s power supply.

For the network cabling inside the rack, we are using standard network cables of different length, with one color for the control network (yellow) and another one for the data network (green). They are fixed with plastic bindings on a Cablofil structure (Fig. 9).

VI. HARDWARE MONITORING AND SAFETY

The chassis and CPU fan speed and temperature will be read via IPMI (Intelligent Platform Management Interface) or the lm_sensor package, depending on the computer mainboard. Heat exchanger fans speed will be read by fieldbus nodes. All these informations will be treated by the Farm Control System (part of the whole Experiment Control System).

Fire detection sensors will be put between rack rows (Heat exchanger side) and will be handled by the CERN Safety Alarm Monitoring (CSAM). In case of fire detection, the
whole farm power will be cut. In addition, a fire neutralization system is being evaluated.

In Ref. [1], we have studied the rack behavior in case of fan and/or water failure and it showed that the Detector Control System (DCS) and the Detector Safety System (DSS) of the experiment can safely handle such a situation which is not immediately critical.

VII. CONCLUSION

We have found cost effective solutions to the powering, cooling, packaging & cabling of 19 inch racks housing up to 40-50 PCs which can be adapted to any farm. Commercial solutions largely similar to ours are being announced for the implementation of farms in computing centers.

REFERENCES