Russian Academy of Sciences

Program Systems Institute

Cluster Hardware Monitoring for Failure Predictive Analysis

Introduction

One of the major trends of late in the HPC world is that an average number of computational nodes, gathered to create a supercomputer, grows very rapidly. These HPC devices are designed to be effective and reliable tool for scientific and engineering calculation. However, there is a problem providing reliability of a device, comprised of commodity hardware. Despite all the advances in manufacturing technology system failures are inevitable for a computational cluster of moderate size of hundreds of nodes. Consequences of such failures may lead to computational errors, data losses and in general, nothing can be done until the system is operational.

Our research focuses on developing intelligent PFA solution interacting with number of components of HPC clusters. Analyzing history of sensor data, such software could provide reliable forecast of future state of cluster.

Requirements to PFA Software

Our experience in developing prototype of PFA allows formulating requirements to such solutions:

- Ease of installation and setup.
- Proactive failure prediction.
- Rational use of system resources.
- Use of secured connections and reliable data bases.

PFA Design

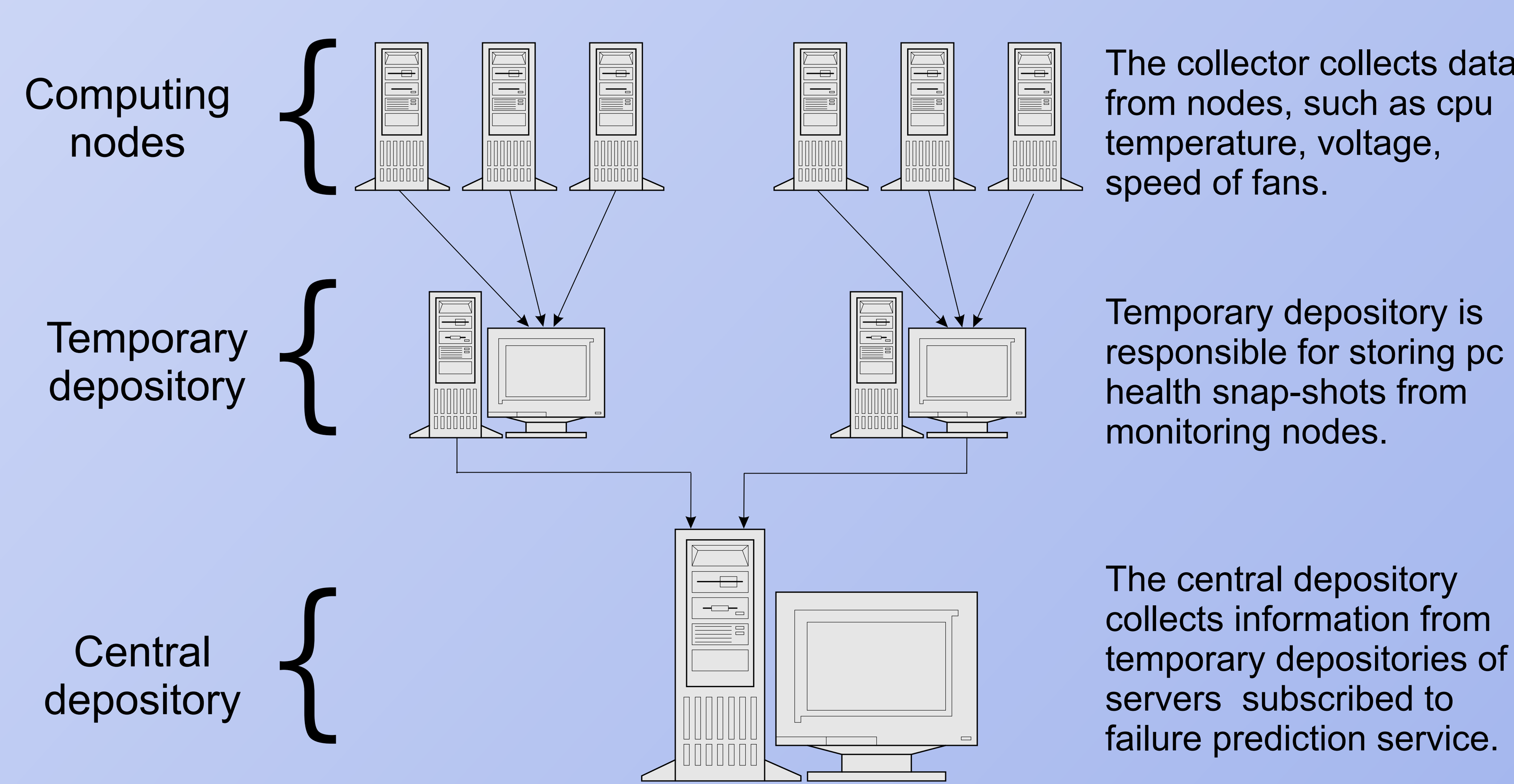
Our PFA solution is built upon client-server architecture and as simple as 1-2-3.

First step is *extracting* primary information. Subscribed clusters run special program, called "collector", responsible for retrieving information of node health from sensors.

On the next step, sensor information is *transferred* to "front-end" node of cluster. Once a day (or other period) data are sent forward, to the central repository of the system.

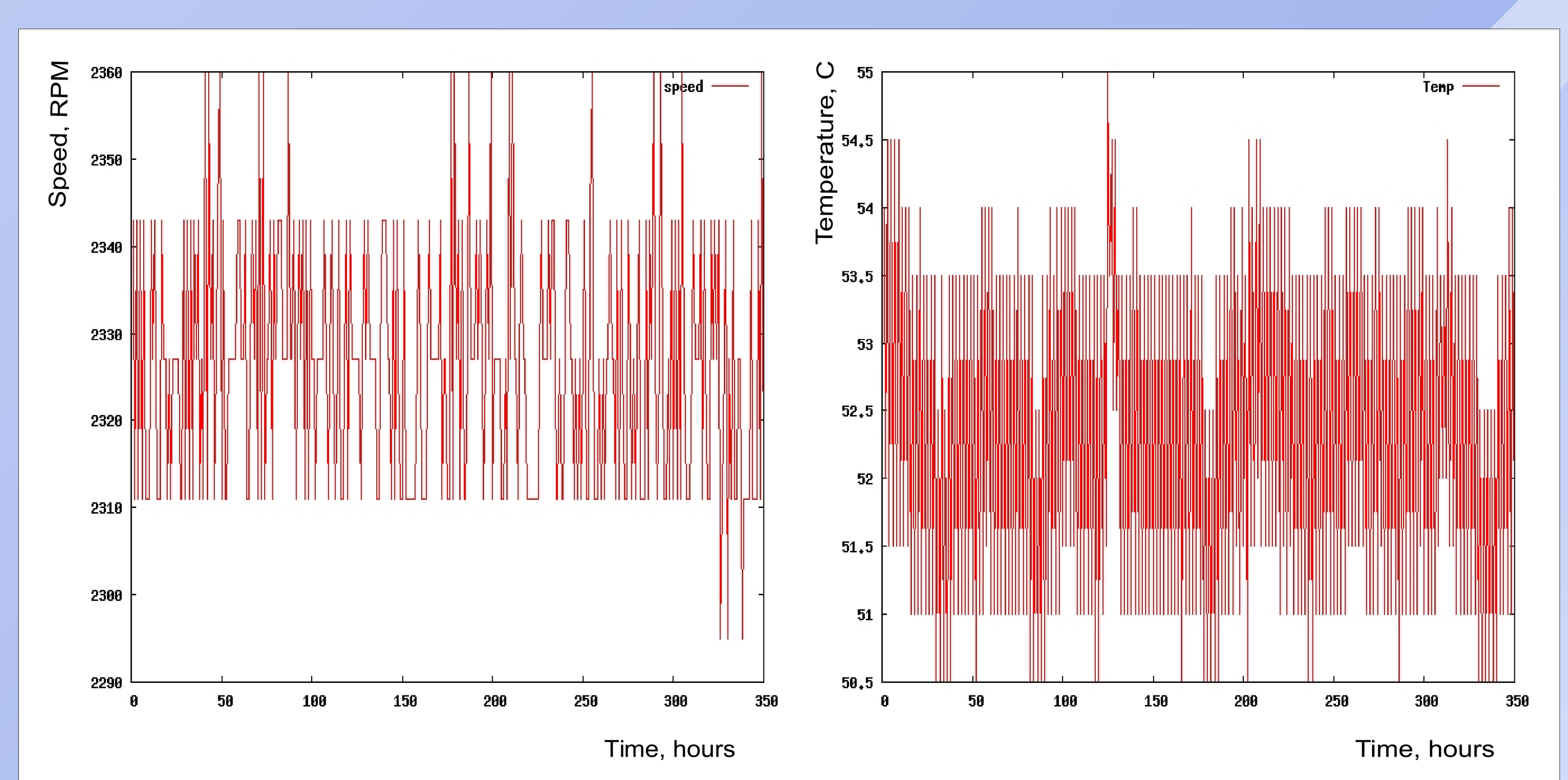
Third step is *storing* collected data in central depositary. It's responsible for archiving time history and *providing prediction service* to subscribed clusters.

The chart on the right shows the key components and data flow of PFA.



Results

Developed prototype of PFA currently is able to retrieve sensors information about actual CPU temperature, voltages, speed of fans and network i/o errors. During PFA installation number of static parameters describing computational nodes is being stored in data base. These parameters, such as CPU model, motherboard and chipset model, may be used during analysis of time history to provide more accurate forecasts.



Fans speed of rotation chart — CPU temperature chart

Failure predictive service is an effective maintenance solution, protecting HPC clusters from data losses during inevitable hardware crashes. It serves two purposes: to reduce MTBF of computational nodes and at the same time to reduce TCO of cluster.

Although our research is primarily targeted to HPC clusters, we believe that failure predictive service will be in demand in many others cyberinfrastructure components, such as storage area networks, server farms and even personal computers.

ADDRESS

Research Center for
Multiprocessor Systems
Program Systems Institute
Russian Academy of Sciences

Pereslavl-Zalessky
Yaroslavl Region
Russia, 152020
Tel/Fax: +7 (48535) 98064
E-mail: abram@botik.ru
Web-site: http://www.botik.ru/PSI/RCMS