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.

Historically there are several strategies to achieve proper level of reliability. For instance, redundancy in tandem with hot-swappable hardware, allows you to decrease potential downtime of the system. But redundant solutions are expensive and also may lower overall performance of the system, because of additional software and hardware requirements.

Another way to improve durability of IT-assets is predictive failure analysis (PFA). This is cost-effective and proven method in reducing system crashes. The design goal of PFA is to guarantee proactive prediction of an impending failures month before it takes place. Number of available PFA tools and connected researches clearly show that predictive analysis actually able to protect hardware from potential failures.

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.

  PFA should not be another “headache” for IT-administrator. It’s critical for PFA software to be transparent in deployment and exploitation.

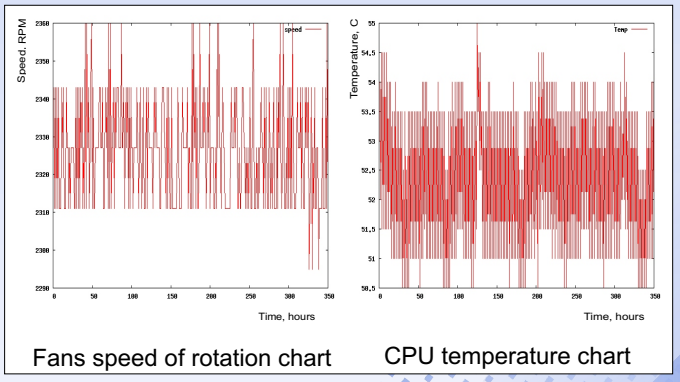
- Proactive failure prediction.

  Key feature of PFA solution is timely forecast of future system crashes. For instance, existing S.M.A.R.T. technology for hard disk drives generates an alert up to 48 hours before failure actually occurs, and even this number may not be sufficient from the end-user viewpoint.

- Rational use of system resources.

  PFA “agents” should not consume too many CPU cycles and other resources in order to provide comfortable conditions of regular work for users of subscribed clusters. That’s why we have chosen to implement resource-intensive tasks of analysis being on central server rather than on client-side. Use of secured connections and reliable data bases.

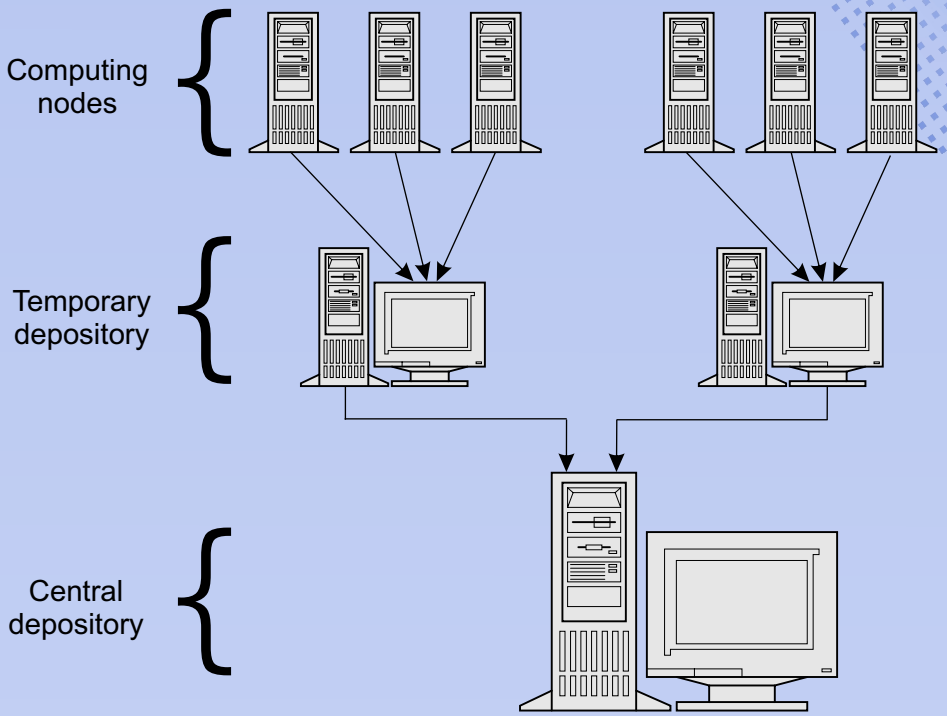
- It’s a question of vital importance to use protected and safety connections and repositories in order to secure valuable information about IT-assets of subscriber.



Fans speed of rotation chart

CPU temperature chart

Russian Academy of Sciences
Program Systems Institute



The collector collects data from nodes, such as CPU temperature, voltage, speed of fans.

Temporary depository is responsible for storing PC health snap-shots from monitoring nodes.

The central depository collects information from temporary depositories of servers subscribed to failure prediction service.

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 depository of the system. Third step is storing collected data in central depository. It’s responsible for archiving time history and providing prediction service to subscribed clusters.

The graphics on the top exposures the key components and data flow of PFA.

In our implementation we use system task scheduler to periodically start agents on computational nodes of cluster to collect and transfer primary sensor information to central depository.

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.

We investigate different approaches to statistical analysis of collected sensor information.

We believe that even simplistic analysis (like extrapolating trends) will produce results, sufficient to improve reliability of computational clusters.

Conclusions

One of the major trends of late in the HPC world is 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.

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