GRIDBYTE ™
MULTI-DIMENSIONAL THINKING

✉ 29911 Niguel Road—Box 6298
Laguna Niguel, CA 92607-6298
☎ 949.916.0799
☎ 800.634.0796
email: info@gridbyte.com

Title: Muiti-dimensional solutions in live network
White Paper: GridByte-CS2007-03
Created by: Sam O. George/GRIDBYTE™
Release Date: 2007-APR-30
URL: http://www.gridbyte.com/Resources/Documents/GridByte-CS2007-03.pdf
Keywords: Compute infrastructure stabilization

Case Outline: Client was a SMB with 30 employees. The site contained 27 compute servers, two fileservers, and 40 client workstations mostly running Linux. The servers and workstations are made by at least 10 different manufacturers. Circuit simulations and verification jobs on various compute servers failed to complete on a regular basis because the servers crashed; these failures caused engineers to lose critical simulation data and days of productivity. One critical $20K fileserver, named BTFS4, crashed at least once per week, even after motherboard and raid-disk replacements. BTFS4 is a raid server containing eight disks. The client reported two to three server crashes per week. The client was not satisfied with the explanations from the IT contractor that things were, "… normal given the demanding nature of IC simulations." Additionally, the client was dissatisfied that the problems persisted even after they had made critical repairs and motherboard replacements. The client wanted to be sure that the problems would be fixed by solutions that focused on rebuilding the raid container and/or replacing the raid controller.

GridByte™ was retained to diagnose and stabilize the client's compute infrastructure. During the problem-solving phase, the client wanted minimum downtime since the lost engineering productivity affected their bottom line and signaled to their customers that they could not deliver on promised deadlines.

Solution Brief: GridByte™ solution used a wide set of performance monitoring tools to identify and diagnose the problems. Server crashes had two causes: (1) BTFS4 fileserver power supply unit (PSU) was not correctly rated to provide required power at elevated temperatures. (2) The compute server crashes were the result of the root partitions filling up thereby causing systems to become unstable/unresponsive.

Solution Detail: During triage, the first priority is stabilization, while also monitoring a host of server performance metrics. We had to determine whether problems were related to load or to device configuration. While selected compute servers could be taken off-line for extensive testing, the critical fileserver BTFS4 could not be removed from the production network easily. Notably, the client had critical licenses node-locked to specific MAC addresses. Also, the infrastructure placed critical design data on this fileserver.

GridByte™ began by quickly adding a temporary fileserver, FS-temp, and mirrored all data. Within 24 hours, we replicated the BTFS4 file-system and successfully transferred its mount-points to FS-temp. This action secured critical design data without loss of productivity. This work is critical because various CAD tools store fully resolved pathnames in various setup files. We still needed to maintain BTFS4 because of license resource issues. Although most data would be safe, a BTFS4 crash would bring down production for some length of time due to loss of access to licenses.

Bringing FS-temp online reduced the frequency of crashes on BTFS4 to one in two weeks. But, what was the problem with BTFS4? Since we had taken all design data off the BTFS4 disks, we loaded performance monitoring software and initiated a disk I/O stress-test cycle on a Friday. Within 24 hours, BTFS4 crashed and came back online on its own. Our monitoring applications did not show anything abnormal immediately before the crash. In fact, we explored various debug options relating to memory, memory slots and other areas.

Because environmental factors are always issues in electronics equipment debug, we installed in-case and closet temperature monitoring equipment. The temperature inside the case was about 121°F consistently. The closet temperature varied from 84°F to 105°F—warm, but not necessarily catastrophic. With graduated stress tests, the in-case temperature began to rise. Full stress-testing brought the in-case temperature to 146°F. BTFS4 crashed again. This time, temperature data correlated with high disk access. So temperature was the culprit—but why? Disk faults? Faulty power supply unit? Poor case ventilation?

The client was still unwilling to renegotiate their license agreement to rehost their licenses because it involved large costs and fees for unpaid maintenance. A second proposed solution was to create a slave license server but that was not accepted by the client. BTFS4 had to be worked on while limiting downtime.

Although the case fans appeared adequate, the eight raid disks increased the in-case temperature to 146°F under high stress. We decided to replace the power supply unit as part of a diagnostic check because the client had a similar model in the server room. In doing so, we found the real culprit. It turns out that the PSU was not correctly sized to power eight raid disks. In low-stress operation with minimal read/write access, the PSU was barely adequate. Under high stress, the disks exhibited various failure modes. In addition, the PSU could not support high disk I/O. Replacing the PSU with a correctly rated device resolved all BTFS4 issues (QED—almost).

Then, we shifted focus to analyzing the performance metrics from the 27 computation servers. Over three weeks we had all crash logs and dynamic performance data with samples taken every 10 minutes. Data analysis of live production networks allows GridByte™ to understand many things. Notably, we ask, what is the nature of infrastructure failures? In this situation—even though it seemed counter-intuitive, we quickly drilled down to a systemic problem. The server crashes were random and were not localized along any of several dimensions—i.e., sub-clusters, load average, memory utilization, disk I/O, bus access or specific node name.

Identifying the systemic cause of random server crashes took a few more days. It turned out, that the partition scheme of the compute servers was indeed causing the problem. Moreover, the problems were specifically due to how poorly some CAD vendors utilized the */tmp* and */var/tmp* directories. Each compute server had one 36GB disk. Each disk had three partitions: *root* (*/*), *swap*, */boot*. */tmp*, */var/tmp* and */local* were sub-directories of *root*. As engineers dumped large simulation data to */local* the size of the *root* partitions grew rapidly. To compound the problems, the CAD tools wrote a lot of data to */tmp*. We could see that on most of the compute servers, just prior to crash, the *root* partition filled up leaving no space for kernel or other writes. In all these cases, the servers would crash and reboot. After reboot, the data in */tmp* directories were cleaned-up (consistent with normal server operation), so the server would return to operation after the crash and function almost normally for a few days.

Our diagnostics showed that the servers were not optimally partitioned. At the very least, they did not have some of the basic partitions routinely recommend by GridByte™. In a staged fashion, all 27 compute servers and 40 workstations were taken off-line and all disks were re-partitioned for greatest fault tolerance. By creating unique partitions, we eliminated system instabilities due to root partition filling up. In other words, the engineers could still write as much data as they

Notice: Our case studies highlight how GridByte™ applies multi-dimensional thinking methodologies, consisting of engineering, mathematics and problem solving combined with business logic, to address client needs. We focus on situations. We do not include any identifying client information for competitive and privacy reasons. We also do not discuss specific brands or products.

needed to */local* partitions without affecting system stability. If */local* was full, the engineers' current simulations would die—however the system would not crash.

Conclusion: This case study highlights how GridByte™ uses multi-dimensional thinking to solve difficult problems. For this client, the savings are estimated at $200K. They did not buy a new fileserver and they eliminated productivity losses due to frequent server crashes.

Citation: Sam O. George, "Multi-dimensional solutions in live network," http://www.gridbyte.com/Resources/Documents/GridByte-CS2007-03.pdf, GridByte™, 30 April 2007.

Notice: Our case studies highlight how GridByte™ applies multi-dimensional thinking methodologies, consisting of engineering, mathematics and problem solving combined with business logic, to address client needs. We focus on situations. We do not include any identifying client information for competitive and privacy reasons. We also do not discuss specific brands or products.