

BlueHeat

Lessons Learned -- Next Steps

| | |
|--|-----------|
| OVERVIEW | 2 |
| CURTAIN OPENS | 2 |
| ISSUES | 3 |
| END USER EXPERIENCING SLOWNESS | 3 |
| <i>Narrative.....</i> | <i>3</i> |
| <i>Lessons Learned.....</i> | <i>5</i> |
| <i>What We Didn't Do.....</i> | <i>5</i> |
| <i>Next Steps.....</i> | <i>5</i> |
| ANTI-VIRUS PROBLEMS..... | 6 |
| <i>Narrative.....</i> | <i>6</i> |
| <i>Lessons Learned.....</i> | <i>6</i> |
| <i>Next Steps.....</i> | <i>6</i> |
| BACKUP PERFORMANCE | 6 |
| <i>Narrative.....</i> | <i>6</i> |
| <i>Lessons Learned.....</i> | <i>7</i> |
| <i>What We Didn't Do.....</i> | <i>7</i> |
| <i>Next Steps.....</i> | <i>7</i> |
| LACP FUNCTIONALITY | 7 |
| <i>Narrative.....</i> | <i>7</i> |
| <i>Lessons Learned.....</i> | <i>8</i> |
| <i>Next Steps.....</i> | <i>8</i> |
| HEAP ERRORS..... | 8 |
| <i>Narrative.....</i> | <i>8</i> |
| <i>Lessons Learned.....</i> | <i>8</i> |
| <i>Next Steps.....</i> | <i>8</i> |
| TESTING TITAN OS PATCHES BEFORE INSTALL | 8 |
| <i>Lessons Learned.....</i> | <i>9</i> |
| <i>Next Steps.....</i> | <i>9</i> |
| SDS PATH NOT FOUND/OTHER SDS JOB FAILURES..... | 9 |
| <i>Lessons Learned.....</i> | <i>9</i> |
| <i>Next Steps.....</i> | <i>9</i> |
| HIGH NUMBER OF VIRTUAL VOLUMES AND CIFS SHARES/PRIORITY OF DRIVE REBUILDS | 9 |
| <i>Lessons Learned.....</i> | <i>9</i> |
| <i>Next Steps.....</i> | <i>9</i> |
| CURTAIN CLOSES..... | 10 |
| LEARNING | 10 |
| BUGS | 10 |
| TRADE-OFFS | 10 |
| SUMMARY | 10 |
| ACKNOWLEDGEMENTS | 10 |

OVERVIEW

The BlueHeat project arose during the first few days of December 2005, as an effort to better understand a range of issues around the mass storage unit 'Indigo', a BlueArc Titan. The deliverables include an analysis of why Indigo behaves as it does and what we might do to modify that behavior to better serve our needs.

This document describes the findings to date of this project, summarizes lessons learned, and offers an outline for how we might proceed.

These are the issues on which the project has focused to-date.

- End User Experiencing Slowness
- Antivirus Problems
- Backup performance
- LACP functionality
- Heap Errors
- Testing Titan OS Patches before install
- SDS Path Not Found/Other SDS job failures
- High number of Virtual volumes and CIFS shares
- Priority of Drive Rebuilds

CURTAIN OPENS

By the end of November 2005, we had accumulated a number of issues with Indigo, including performance issues when anti-virus was enabled, loss of data after a spontaneous reboot, intermittent reboots (heap errors), intermittent SDS job failures, LACP negotiation failures, and slow performance during backups (using the new CommVault/ADIC solution). We believe that the end-user community was being inconvenienced (or worse) by the frequent Sunday evening outages, during which we would reboot Indigo or otherwise take it off-line.

During this period, we had upgraded Indigo's OS, upgraded firmware, migrated Clinical from a SATA-based tray to a Fibre Channel-based tray, explored various theories, and invoked the local BlueArc team, looking for answers and resolutions. BlueArc support kept requesting 'diagnostic dumps', which we kept sending to them. BlueArc support talked about "not enough spindles ... latency ... slow SATA drives ... ten year old Ethernet switches ... saturated Ethernet switches ..."

On the morning of Wednesday 11/30/2005, end-users experienced wide-spread performance issues with Indigo, and this issue catapulted itself to the top of the list of woes. After this event, we convinced ourselves that SDS's "Tidal" jobs could induce Indigo to repeat this situation.

Four days later, we held our first BlueHeat analysis session, during the Sunday 12/4 outage window, and the following Monday 12/5/2005, we held our BlueHeat project kick-off meeting. Two weeks after that, on Sunday 12/18/2005, we held our second analysis session. During this

period, we ran one analysis session and one planning session per week ... we characterized the cause of the wide-spread performance issue (aka *NVRAM*) ... managed to replicate the problem during the day ... and figured out, temporarily at least, how to avoid repeating it. At this point, we 'stood down' from our high-intensity effort and turned to a list of lower-intensity issues, developed during the project kick-off meeting.

ISSUES

End User Experiencing Slowness

NARRATIVE

The BlueArc team identified our network infrastructure as a possible source of some of these issues, describing our gear as "ten years old".¹ And while age may, or may not, be a good thing, the BlueArc team suggested that the blocking design of the switches servicing the access layer might be leading to substantial packet loss and thus contributing to performance issues.

As far as the switch serving the CommVault/ADIC solution was concerned, this was an unlikely model -- that switch has a non-blocking design: even if it were fully-loaded with gigabit ports, all running a line-rate, the backplane should be incapable of dropping packets. However, the switch servicing Indigo itself has an oversubscribed, aka 'blocking', design, and we knew from looking at counters on the Ethernet switch that Indigo's ports were reporting packet loss, though not on the scale needed to cause all these issues, and so I claimed that this was a reasonable path of inquiry.

On further investigation, I also noticed that Indigo's Fibre Channel ports were reporting errors, so I developed another model, in which the Fibre Channel switches were fragging packets on their way to and from the trays and thus causing performance issues.

Finally, I knew that Robert and Susan had been staring at various parameters on Indigo itself, some of which seemed to relate to parameters which could affect performance ... and some of which were meaningless to us. Gathering a bunch of those parameters was a bit of a stab in the dark, but seemed like a reasonable thing to do, particularly given how unsure we were of the cause: cast the net wide.

During the 12/4/2005 event, we replicated the problem and then gathered data from various points: from the point of view of a client, from Indigo and its Fibre Channel switches, and from the intervening Ethernet switch. From this data gathering session, we saw that Indigo would "fall silent" during the performance affecting events, emitting no further packets for periods extending to a minute. We verified that while Indigo's Ethernet switch port was dropping packets, it wasn't dropping packets during the periods when clients were experiencing problems. We verified that the Fibre Channel port error counters were not incrementing.² And we

¹ Actually, our gear was five years old (the switch supporting the CommVault/ADIC solution was six years old); and all switches were running modern (2005) operating systems.

² One particular error counter was incrementing, but that turned out to be unrelated to the issue.

identified NVRAM utilization as a key parameter in understanding when Indigo would fall silent.

The following week, we did not communicate this information to our SDS colleagues ... and they brought Indigo to its knees again, when they launched their “Tidal” jobs. Thereafter, SDS modified their jobs, inserting delay between each packet, pacing the work they were handing to Indigo and thus reducing the chance of repeating this event.

During the 12/18/2005 event, we validated our NVRAM exhaustion model and ruled out a number of additional factors.³

And in early January, we spent several days with Denis Kornilov, a third tier engineer from BlueArc, learning more about spindles and latency and other storage throughput factors. As a result of this, Susan re-engineered the SATA and ATA trays, increasing their throughput using various techniques.

Robert has developed a model for thinking about designing volumes, to harden the head against NVRAM exhaustion.⁴ Factors include usage patterns (size of files written, number of files written per second, number of spindles servicing a volume, type of spindle (SATA or FC, random vs sustained IO). However, even with a massively engineered volume, we can still develop usage patterns which would overwhelm NVRAM -- preventing NVRAM exhaustion is a mix of hardware engineering and throttling usage patterns. In response to this, Robert has wired our in-house monitoring systems -- trending, monitoring, and alarming -- to track NVRAM exhaustion. We can see from these tools that normal usage hardly touches NVRAM, peaking to perhaps 3% utilization. However, SDS jobs and disk replication jobs and both push NVRAM to exhaustion, under some circumstances.

We foresee using a three pronged approach to mitigating the risk of future NVRAM exhaustion:

- trending, monitoring, and alarming
- education of sys admins and power users
- designing future trays for optimal performance

We propose that we have educated the current collection of sys admins and power users on Indigo through various forums, including All IT Staff meeting and this session, on the dangers of writing lots of data to Indigo rapidly; we claim that all such sys admins and power users reside inside Center IT. However, notice that if we were to expand Indigo’s scope, to service additional groups, that we would need to remember to replicate this educational process.

³ In hindsight, I returned to the O’Reilly [Understanding SANs and NAS](#) book, which I had read during the Indigo product selection phase, and rediscovered the issue of NVRAM exhaustion, where I learned that this is an issue common to many NAS devices. When manufacturers want to harden their NAS device against data loss or corruption in the event of a power failure, they include ‘NVRAM’ ... all but the lowest-end NAS boxes include this capability and thus are vulnerable to NVRAM exhaustion. But of course, at the time, I didn’t have enough experience to absorb the significance of this.

⁴ See “System-Wide Performance Problems” for analysis.

LESSONS LEARNED

Develop our skills at telling a coherent story.⁵

Focus first on what people are feeling and needing ... don't let content distract you.

Use one of many standard analytical models⁶ when trouble-shooting problems, rather than winging it.

Analyzing problems can be costly and time-consuming.⁷

We can move fast when sufficiently motivated.

Loosely coupled divide & conquer: IMHO a key aspect to our data gathering and analytical success involved dividing duties amongst participants while remaining loosely coupled (communicating status with one another, sharing observations, sharing analysis).

Communication can keep us from tripping over our own feet.⁸

Understand your architecture ... this is an iterative process..

Own your tools ... particularly in multi-vendor environments, no single vendor can analyze your problem: you have to do it yourself.

Sometimes you have to touch the stove before you understand what "hot" means.

WHAT WE DIDN'T DO

We didn't analyze the size of the j4sr-x-esx packet loss effect on Indigo or other devices in j4-401.

The Fibre Channel switches are still reporting steadily incrementing counters of `swFCPortTooManyRdys` on all ports.⁹

NEXT STEPS

None.

⁵ Notice that BlueArc support identified the root cause of the performance issue right from the beginning ... but they were unable to tell a coherent story to us.

⁶ For example, Describe the observation, Develop a model, Sketch a hypothesis, Implement an experiment, Revise the model and/or hypothesis, Iterate until the observations are consistent with the model and hypothesis.

⁷ Roughly 80 hours of staff time spread across eight staff members spent just on these two data gathering Sunday evenings alone ... not to mention all the time spent planning for these events and analyzing the results.

⁸ Had we broadcast our understanding of the NVRAM issue directly after the 12/4/2005 data capture event, we might have prevented the service-affecting repeat of this experience a few days later.

⁹ According to the local Brocade sales team, this is likely a cosmetic bug in the NIC drivers facing these ports. Having heard their explanation, I support it. --sk

Anti-Virus Problems

NARRATIVE

We were wary of re-enabling anti-virus, because of the performance impact which this may have had in the past, and we discussed the idea of leaving anti-virus disabled permanently. When we tried re-enabling anti-virus protection, we discovered that it did not, in fact, affect end-user performance.¹⁰

LESSONS LEARNED

We value anti-virus protection and will go to significant effort to keep it functioning.

NEXT STEPS

None.

Backup Performance

[I need to write the analysis document to support this. --sk]

NARRATIVE

We analyzed various methods of backing up Indigo.

Using simple protocols, we verified that Indigo, Ernie (the backup server), and the network in between can deliver close to Gigabit Ethernet throughput.

The original backup method, the Share method, employs the CIFS protocol. Packet analysis reveals design choices in the CIFS protocol which reduce its performance to something around 50% of a Fast Ethernet pipe.

Packet analysis of the NDMP-over-IP method reveals an application choice (CommVault) to override the OS default TCP Window, dropping it from the megabyte at which we have set it to 16K, and this choice reduces performance to roughly 1.5 times the performance of a Fast Ethernet pipe.

We did not capture packets on the NDMP-over-FC method; however, measurements show that performance runs at a little under 3 times the performance of a Fast Ethernet pipe.

This analysis stressed the tools and techniques we possess, on account of the high rates of throughput delivered across extended periods of time across numerous files; I spent a lot of time struggling with our tool limitations and with my learning curve around how to tackle such problems.

¹⁰ See “BlueHeat Anti-Virus” presentation materials.

However disappointing the NDMP-over-FC performance may be, it is still sufficiently enthusiastic to reduce backup times dramatically, when compared to the original method.

LESSONS LEARNED

Establish a baseline, i.e. figure out what is 'normal', and expand from there.

The right tool, and the associated expertise, for the job is a good thing.

WHAT WE DIDN'T DO

We haven't analyzed why NDMP, over IP and over FC, doesn't come close to saturating the tape drive.

NEXT STEPS

Analyze NDMP to see how its performance could be increased.

LACP

NARRATIVE

Ideally, one purchases redundant heads on a BlueArc box, and this delivers a range of redundancy. However, we traded our second head for the ability to implement anti-virus, so we are searching for additional ways to deliver redundancy.

LACP, aka EtherChannel, allows one to bundle multiple NICs together in order to deliver greater throughput. The Titan can swallow more than 1 Gigabit of traffic per second, so from a Titan point of view, this makes sense. Our current network infrastructure cannot deliver more than 1 Gigabit of traffic per second to an end-station, so from a network point of view, this doesn't make sense.

However, BlueArc wraps some fancy capabilities around their LACP implementation, which allows us to use it to deliver redundant Ethernet connections to the redundant Ethernet switches in j4-401.

We spent a year or more messing with this, trying to get it to work, without success. Here again, BlueArc offered a model of "the network switches are ten years old" as an explanation. And again, I think here I would have been better served by *Focus first on what people are feeling and needing ... don't let content distract you.* After all, the software loads on these switches are vintage 2005, only a few months older than the OS running on Indigo itself.

In April 2006, we revisited the issue, and we now have it working ... the last step in the chain was to remove port-specific statements which instruct the Ethernet switch to permit the incoming

device to tag packets as belonging to an ‘auxiliary VLAN’ ... this is a function which we employ to offer VoIP traffic a higher class-of-service than commodity traffic.¹¹ For still unknown reasons, our Catalyst switches quit sending LACP packets when these ‘auxiliary VLAN’ statements are in place, although the management interface believes that it is trying. I have a ticket open with Cisco asking whether this is a bug or a feature. In-line sniffing tools gathered this information; the Catalyst SPAN function incorrectly showed the Catalyst emitting LACP responses.

LESSONS LEARNED

Sometimes, you gotta go to the source.

NEXT STEPS

TBD

Heap Errors

NARRATIVE

An internal BlueArc memory pool called a ‘heap’ gradually shrinks during normal operations. If it shrinks to a small enough value, Indigo reboots. To make this reboot predictable, we schedule weekly reboots, typically on a Sunday. BlueArc owns this issue and plans to release a patch or a version of code which fixes it.

LESSONS LEARNED

Humans write code with bugs in it.

NEXT STEPS

Install the patch or OS upgrade which fixes the issue.

Testing Titan OS Patches before install

How do we protect ourselves from a future upgrade which introduces more pain than we want? Having a ‘second tier’ Titan head would help; it could service lower-priority users (like Center IT) and could be upgraded first, prior to the ‘first tier’ Titan installation. However, buying another head for this purpose seems unlikely, given our size and finances. We haven’t come up with a solution for this.

¹¹ Thanks to Denis Kornilov for identifying the relevant Catalyst statement.

LESSONS LEARNED

Software contains bugs.

NEXT STEPS

???

SDS Path Not Found/Other SDS job failures

SDS has experienced a range of issues spanning Indigo's life span. Some of these issues involved admainsXX boxes talking to Indigo; some involved admainsXX boxes talking amongst themselves. On several occasions, Stewart Castaldi and I started to analyze them ... and then we were distracted by other issues. In hindsight, had we pushed this analysis to closure, we could have identified the NVRAM-exhaustion issue **before** end-users noticed. However, we didn't.

I can tell a plausible story which traces the source of some of these issues to NVRAM exhaustion; for others, I cannot (particularly the ones which don't involve Indigo).

The issues quit around the end of 2005 and have not returned.

LESSONS LEARNED

Stick-to-it-ness is a prerequisite for analyzing problems.

NEXT STEPS

If these problems resurface, analyze them and stick with the analysis until we understand the root cause.

High number of Virtual volumes and CIFS shares/Priority of drive rebuilds

These turned out to be red herrings, possible explanations for the performance problems we were seeing, but ones which we discarded in favor of the NVRAM exhaustion theory.

LESSONS LEARNED

In science, one develops and discards numerous hypotheses before arriving at one which sticks.

NEXT STEPS

None.

CURTAIN CLOSES

Looking back across the Indigo experience in general and the BlueHeat project in particular, I am going to finish by claiming that things happened pretty much the way we expected them to happen.

Learning

When we bought Indigo, we knew that we were taking a step forward into the brave new world of storage management, we knew that we had a lot to learn, and we knew that we would stumble along that path.

And we did. We learned that one of the issues relevant to NAS management is NVRAM exhaustion, which in turns depends on understanding the throughput delivered by differing volume architectures.

Bugs

When we bought Indigo, we knew we would encounter bugs ... we all have enough experience in this business to know that humans make mistakes, and that the people who design and build products are human.

And that's what we encountered: bugs.

Trade-offs

When we bought Indigo, we knew that we were trading uptime for data integrity, i.e. purchasing a single head in exchange for anti-virus. Life is full of trade-offs.

And that's what we got: a box with anti-virus protection but without head redundancy.

Summary

In summary, I claim that while the process may have been bumpy, nothing which happened was a surprise -- we encountered the type of issues we expected to encounter, and we resolved them.

ACKNOWLEDGEMENTS

I would like to acknowledge a range of contributors to this project.

Patrick Hirayama for gathering the data which demonstrated that the packet loss which Indigo's port on j4sr-a-esx was experiencing did not contribute significantly to the end user performance issues.

Rick Bawaan for wrestling with the in-line sniffer and for capturing Indigo-side traces.¹²

Ana Dos Santos for capturing key client-side traces, illustrating the “minute of silence” and for guiding us through the first data capture event.

Joseph Flahiff for keeping us on track and within scope.

Sean Kliger for developing the strategy for analyzing backup performance and for sanity-checking my analytical work throughout the project.

Estella McDermott for capturing additional Indigo-side traces.

Jason Burdullis for capturing additional client-side traces and for guiding us through the second data capture event.

Susan Way for juggling numerous tasks during the data capture sessions, for gathering enormous amounts of data around backups, for doggedly pursuing the LACP issue, and for engaging BlueArc tech support whenever needed.

Robert McDermott for wading through obscene amounts of data, finding the key NVRAM/Ethernet traffic connection, and for writing the AutoGraph tool and making it available to others.

Denis Kornilov for numerous contributions, directly to us during his training session and via e-mail, and for advocating for our issues within BlueArc itself.

The entire SOPS department for analytical support during various BlueHeat sessions.

APPENDIX

Conversions

To GB/hr into Mb/s, multiply by 2.275.

1 GB/hr equals how many Mb/s?

$(1 \text{ GB/hr} / 3600\text{s/hr}) = 1/3600 \text{ GB/s}$

$1/3600 \text{ GB/s} * 1024 \text{ MB/GB} = 1024/3600 \text{ MB/s}$

$1024/3600 \text{ MB/s} * 8 \text{ bits/Byte} = 8096/3600 \text{ Mb/s}$

$1 \text{ GB/hr} = 2.27555\dots \text{ Mb/s}$

¹² Rick, turns out these traces were, for the most part, good ... but a flaw in the hardware has it discarding one stream or the other from the few pages of packets.