Fixing On-Call When Nobody Thinks it’s (too) Broken

Tony Lykke
Hudson River Trading
Jan 2012 to Oct 2018
71,317 high urgency pages
Average: 201/week
Worst month: 2,327

Nov 2018 to Feb 2019
1,015 high urgency pages
Average: 56/week
What This Talk is (not)

IS:

● Guide to reducing pager noise

IS NOT:

● Explanation of why you should
What This Talk is (not)

IS:

● General advice you can (hopefully) apply regardless of platform

IS NOT:

● Technical deep-dive into Nagios and PagerDuty
A Brief History

Atlassian (Sydney)
  Service Ops (SOPS)
  Jira SRE

Hudson River Trading (NYC)
  Trade Systems SRE
HRT at a Glance

Automated trading firm

>5% of daily US equities volume

Milliseconds are slowwww

Everything is quantifiable

~300 people across 10 offices
Trade Systems Overview

Thousands of machines

Tens of colos

Hundreds of pages per week
Me: “You don’t have to get paged so much”

Them:
Why?

“That’s just how it’s always been”

“Each of our colos is configured slightly differently, so it’s hard to set good thresholds”

“Reducing noise that much just isn’t realistic without Big Corp™ resources”

“It’s better than it used to be”
9 Really Hard Steps to Reduce Pager Noise

1) Understand your audience
2) Understand the problem
3) Understand the system
4) Devise a game plan
5) Get permission (optional)
6) Lay the groundwork
7) Fix the lowest hanging fruit
8) Communicate, communicate, communicate
9) GOTO 7
Step 1: Understand Your Audience

Stockholm syndrome + alert fatigue = ???

Culture of alerts meaning things are working

Disagreement that this “problem” is fixable

Priorities not aligned

Preference for conservative approaches
Step 2: Understand the Problem

Dashboard / Pagerduty Trends

weekly page count - sysstrade

current period incident count - sysstrade

51,908

average incidents per week - sysstrade

Average Pages per Week

197.485
Step 2: Understand the Problem

- Approximate US market hours
- Daily data consolidation
Step 3: Understand the System

Nagios 3

30 satellites, 40,000+ checks

200,000 lines of configuration

Huge amounts of stale and duplicated config

Manually deployed via rsync

Adding or changing checks tedious and error-prone
Step 4: Devise a Game Plan

- Doesn’t have to be comprehensive
- Low risk and high impact changes first
- Communicate the plan and ask for feedback
- Listen to the data
Step 4: Devise a Game Plan

Satellites

Use templates to simplify

Already exists, make smarter

Nagios®

Master

Python

pd
Step 5: Get Permission (Optional)

Ask for forgiveness instead of permission?

Use this if you've not asked.

How long does it take to ask questions?

Be annoyingly persistent.
Overview

Nagios on shmonitor6 was not successfully processing all check results, and hence dropping some or all notifications for SysTrade, for around 21 hours. Nagios → PagerDuty alerts processing alerts slower than Nagios was attempting to send them, resulting in a Nagios backlog which manifested as forked processes.

- Normal Case
- Force a Deploy
- Rolling Back a Release
- Disabling Automatic Deployments
- Deploying From a Branch
- Deploying When Bamboo is Unavailable
Step 6: Lay the Groundwork

Neglect creates technical debt
Step 6: Lay the Groundwork

Neglect creates technical debt

Standardise config

```
> git shortlog -s -n | rg -n Tony
26:  5   Tony Lykke
> git shortlog -s -n | rg -n Tony
  7:  765   Tony Lykke
```

Templated 15k lines of duplicate config

Automated build and deploy
Step 7: Fix the Lowest Hanging Fruit

Minimise:
- Risk
- Time

Maximise:
- Impact
Step 7: Fix the Lowest Hanging Fruit
Step 7: Fix the Lowest Hanging Fruit

- Approximate US market hours
- Daily data consolidation
Step 7: Fix the Lowest Hanging Fruit
“What’s it feel like to be a hero?”
Step 8: Communicate, Communicate, Communicate

Blog posts

RFCs

Documentation

Comprehensive pull request messages

Documented rollout plans

???
Step 8: Communicate, Communicate, Communicate

Silence made people uncomfortable

Decade of alerts being ok

Needed to wean our team off the noise
Step 8: Communicate, Communicate, Communicate

Tuesday, February 26th

pagerduty_nagios  APP: 4:30 PM
Downgrading mysqld-slave on shadbwrite3 until nagios re-notifies because Warning level alerts are not urgently actionable

Wednesday, February 27th

pagerduty_nagios  APP: 1:04 PM
Created group for os_health on shabizdev5 with host_multiple_services because there are three or more services down on shabizdev5, containing shabizdev5:os_health, shabizdev5:ssh, shabizdev5:time
Adding ntpd on shabizdev5 to host_multiple_services for shabizdev5
Adding disk on shabizdev5 to host_multiple_services for shabizdev5

pagerduty_nagios  APP: 3:02 PM
Silencing clockdrift on lsetradeg10 for 9 hours, 51 minutes and 49 seconds because out of trading hours clockdrift alerts on trade machines are not actionable
Step 9: GOTO 7

THAT WAS PERFECT.
LET'S DO IT AGAIN.
Then vs Now

Satellites

Nagios®

Nagios®

Nagios®

Master

Nagios®

Python

pd®
Then vs Now

Satellites

Nagios

Nagios

Nagios

Master Nagios

Drop Downgrade Group

pd
Ten Second Review

- Gather data
- Review data
- Make a plan
- Communicate
- ???
- Talk at SREcon about it
Sound Fun? We’re Hiring

Come talk to us at our booth! We’re at: Booth 41, Northside Ballroom

Visit our careers page: hudson-trading.com/careers

hi@tonylykke.com

@tl on SREcon Slack