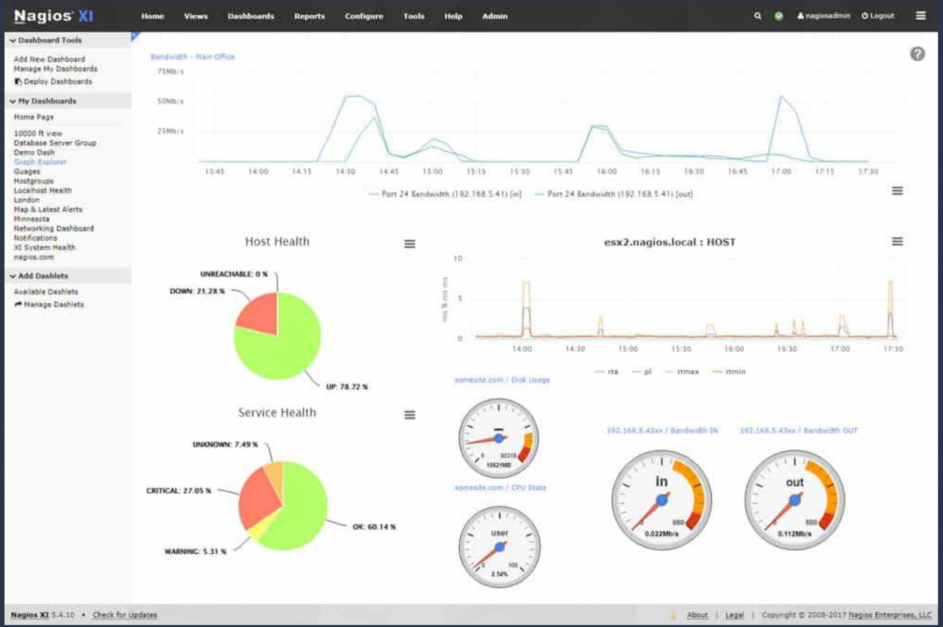
Smoke-detectors in large scale production systems

Abhijeet Mishra

Last9

Life of an ~~SRE~~, ~~Devops~~, Sysadmin 15 years ago

- Tools like Nagios, Zabbix
- When stuff broke, the reasons were fairly common to figure out
- Static thresholds were good enough
- No autoscaling, no K8s and more importantly - no YAML :)

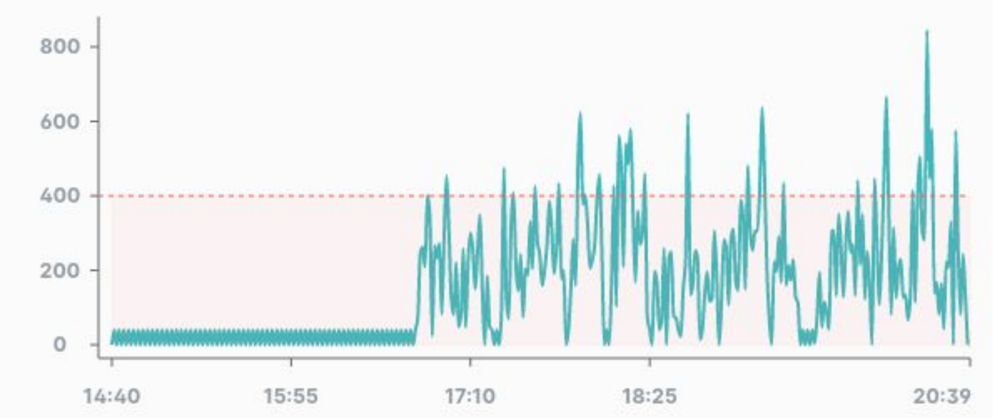


Fast forward to now...

- Autoscaled applications
- Cloud native K8s workloads
- Functions as a service
- Apes wearing gold chains

Are static thresholds enough?

- What is a good number of 5xx errors?
- How many container restarts for a K8s container is too many?
- How slow is slow enough?

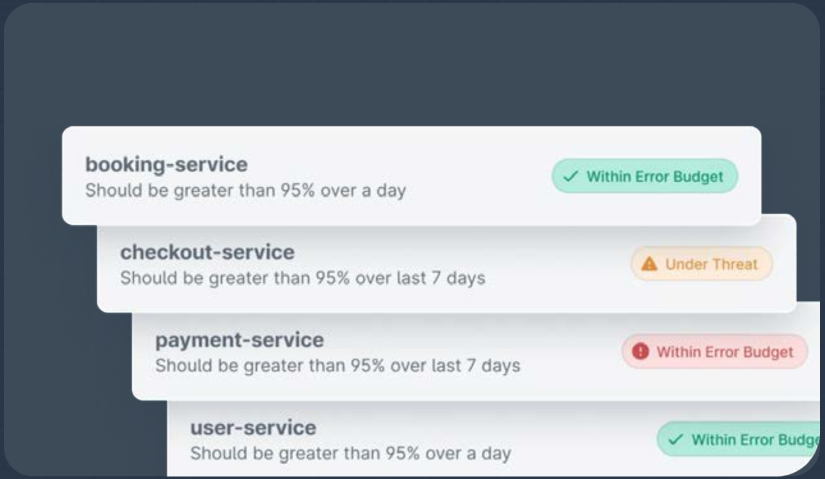


Context is king

- Increasingly complex systems
- Harder to isolate faults
- No sitting ducks

Aren’t SLOs enough?

- Act as lagging indicators
- Request based vs window based SLOs
- Rolling window SLOs and dynamic alerting go hand in hand



booking-service
Should be greater than 95% over a day
✓ Within Error Budget

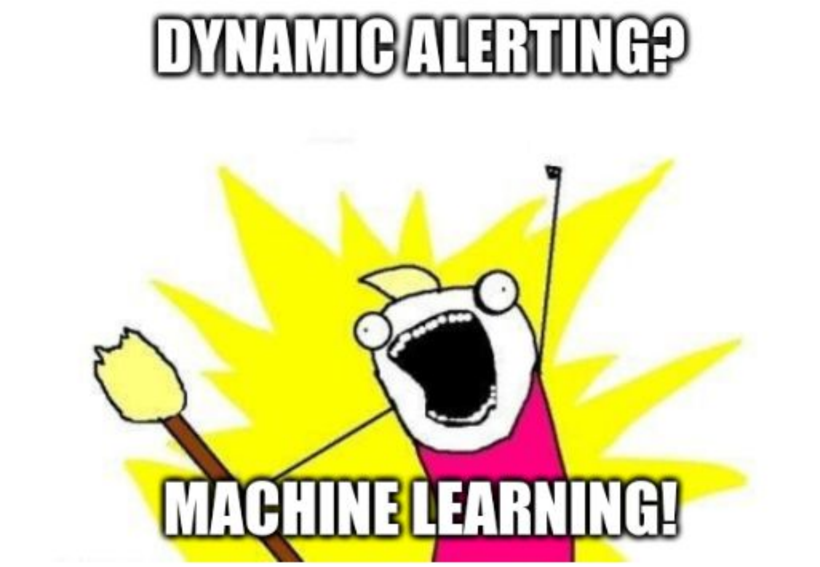
checkout-service
Should be greater than 95% over last 7 days
⚠ Under Threat

payment-service
Should be greater than 95% over last 7 days
! Within Error Budget

user-service
Should be greater than 95% over a day
✓ Within Error Budg

Dynamic alerting means...

- Machine learning! Or not.
- Cost intensive
- High number of models to be maintained
- High school math to the rescue

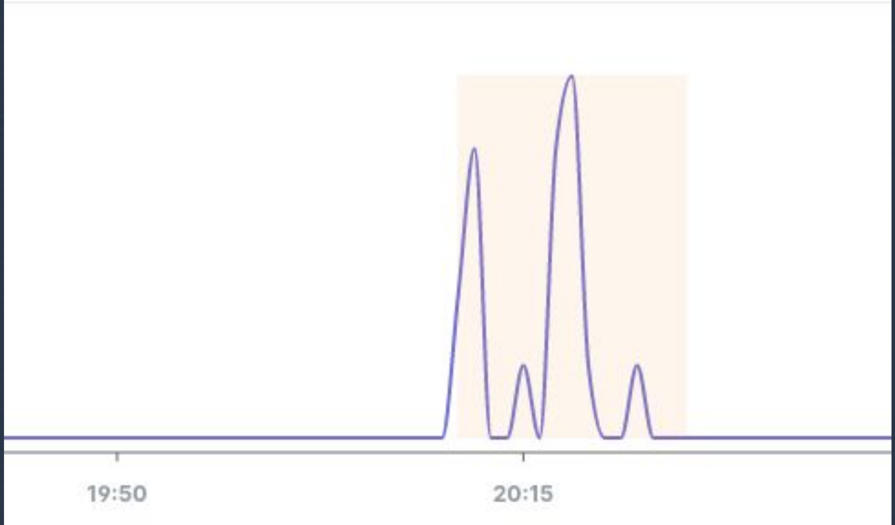


What does a smoke detector do?

- Raise an alarm before dashboards catch fire
- Anomaly = glitch in the matrix?
- Any perceived deviation from normal system behavior based on:
  - Rate
  - Amplitude/Spikes
  - Time of the day

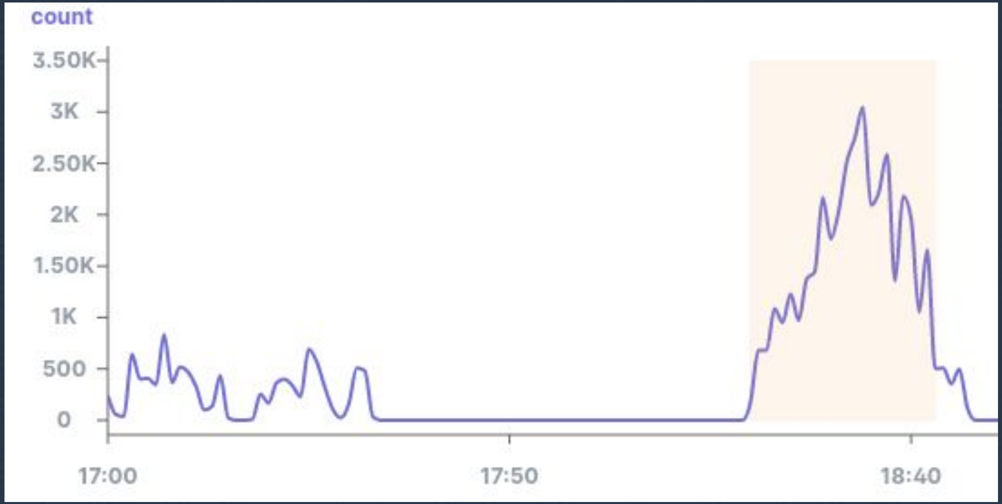
Detecting rate

- Can be used to flag metrics which have increased or decreased suddenly
- Can be measured using standard deviation across a sliding window
- Will lead to alert fatigue in cases where metric fluctuations are expected



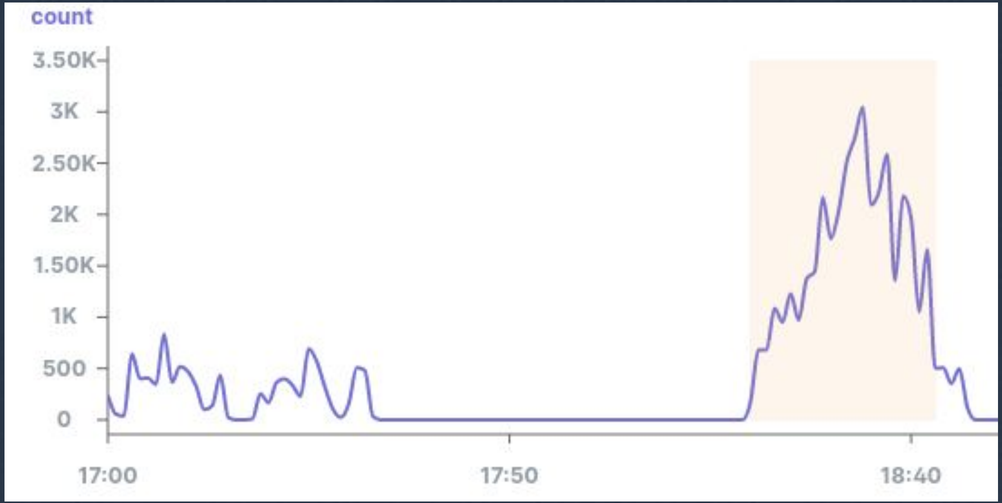
Spike Detection

- Can be used to flag unusually high values
- Percentile based cutoffs on historic data
- Will lead to alert fatigue in cases where the spikes are acceptable or even expected



Seasonality

- If the metric follows a loosely defined pattern depending on the time of the day, month or year
- Compute bounds using percentiles using data observed in relevant time periods in the past
- How short/long should the time frame be?



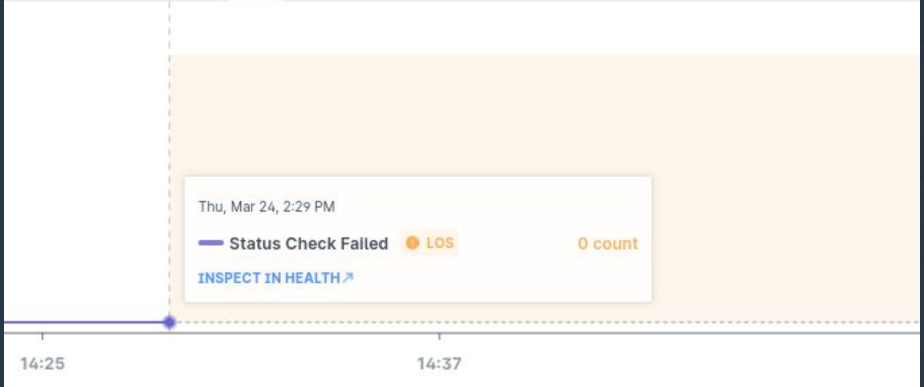
Last9

Multidimensional alerting

- Just one reason might not be enough
- Add more context- only an anomaly if one or more characteristics are unusual
- Can be used to gauge the severity of the situation

Loss of Signal

- What happens when the metrics/logs stop coming in?
- Is it because of 0 traffic or because the pipelines went down?
- Monitor how frequently metrics where observed in the past
- If the time gap is unusually high, send an alert



Conclusion

- Dynamic alerting is subjective and difficult
- There is no silver bullet
- No one size fits all approach

Thank You

in /mishra-abhijeet