

CASE STUDY: ADOPTING SRE



TOM LIMONCELLI, SRE

STACK EXCHANGE, INC. @YESTHATTOM



Tom's SRE CREDS

- Google SRE from 2006-2012
- Worked mostly on internal infrastructure and virtualization services (Ganeti) as well as external services like Blog Search.
- Joined Stack Exchange, Inc. in June 2012.

Who is Tom Limoncelli?

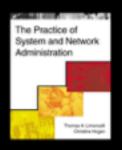
Google SRE 2006-2013.

Author, blogger, speaker.

Sysadmin since 1988. Worked at Bell Labs and many smaller companies

SRE at Stack Exchange, Inc.

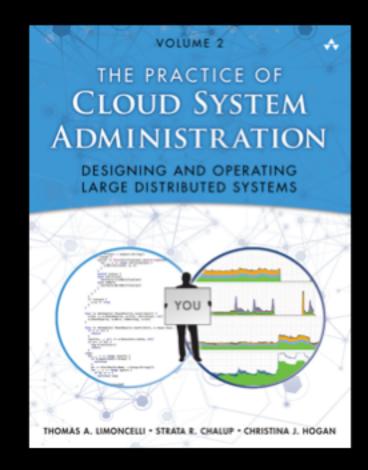
serverfault.com / stackoverflow.com





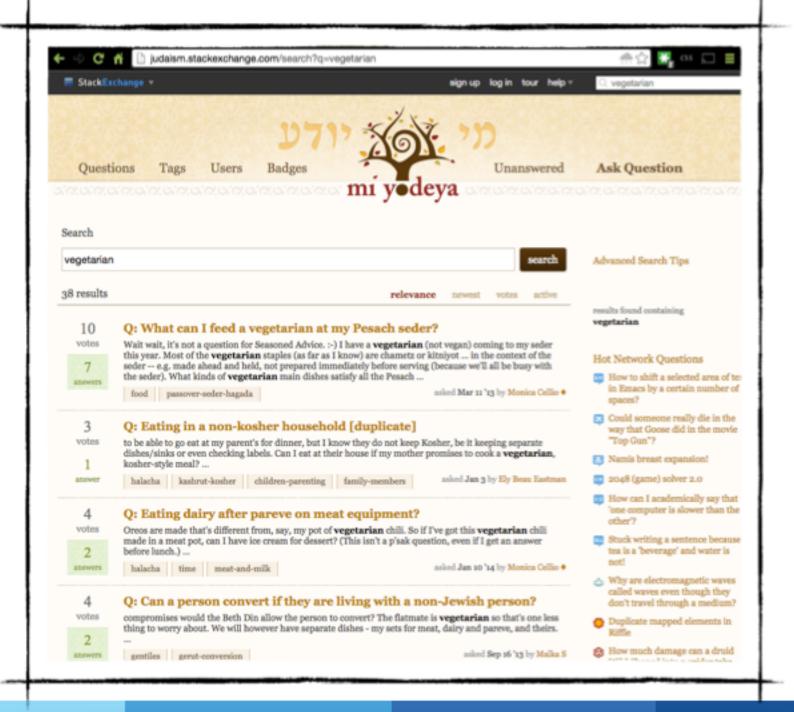


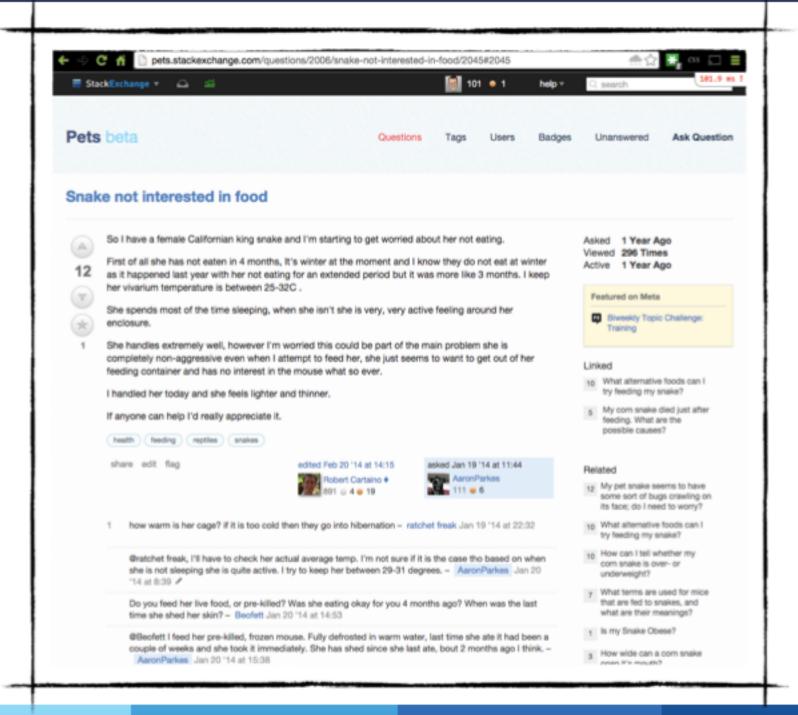


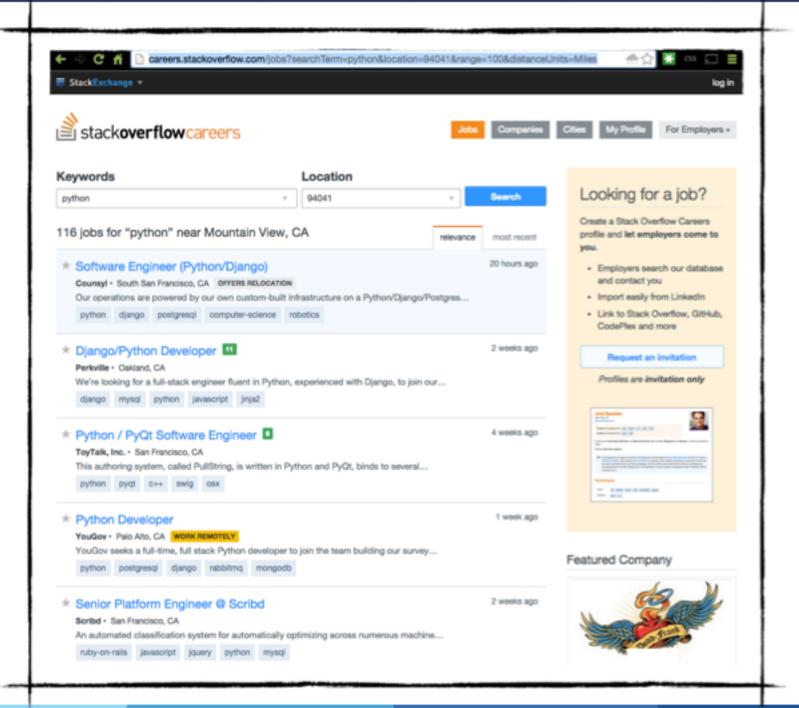


WHAT IS STACK EXCHANGE, INC?

- 140+ Question and Answer Web sites:
 - StackOverflow (coders)
 - Serverfault (sysadmins)
 - Over 130 more: Grammar, Math, Physics, Poker, Pets, Aviation, each major religion
- Stack Overflow Careers
 - Helping the best developers and sysadmins find jobs that don't suck.









STACK EXCHANGE IS...

- Smaller than you'd think:
 - Dozens of machines, not hundreds
- Mixed Windows and Linux
 - Rendering and DB: C#, IIS, SQLServer
 - Linux: Cache, Load Bal, Sec, Search, +

ASSESSMENT SO FAR (I.E. HOW DO WE STACK UP?)

SITE RELIABILITY PRACTICES (BEN TREYNOR'S SRECON14 KEYNOTE)

- 1. Hire only coders.
- 2. Have an SLA for your service.
- Measure and report performance against the SLA.
- Use Error Budgets and gate launches on them.
- Have a common staffing pool for SRE and Developers.
- 6. Have excess Ops work overflow to the Dev team.
- Cap SRE operational load at 50 percent.
- Share 5 percent of Ops work with the Dev team.
- Oncall teams should have at least eight people at one location, or six people at each of multiple locations.
- 10. Aim for a maximum of two events per oncall shift.
- 11. Do a postmortem for every event.
- Postmortems are blameless and focus on process and technology, not people.

2013

- YES 1. Hire only coders.
- NO 2. Have an SLA for your service.
- NO 3. Measure and report performance against the SLA.
- NO 4. Use Error Budgets and gate launches on them.
- IRR 5. Have a common staffing pool for SRE and Developers.
- IRR 6. Have excess Ops work overflow to the Dev team.
- NO 7. Cap SRE operational load at 50 percent.
- IRR 8. Share 5 percent of Ops work with the Dev team.
- Oncall teams should have at least eight people at one location, or six people at each of multiple locations.
- YES 10. Aim for a maximum of two events per oncall shift.
- NO 11. Do a postmortem for every event.
- NO 12. Postmortems are blameless and focus on process and technology, not people.

2015

- YES 1. Hire only coders.
- 2. Have an SLA for your service.
- 3. Measure and report performance against the SLA.
- NO 4. Use Error Budgets and gate launches on them.
- YES 5. Have a common staffing pool for SRE and Developers.
- IRR 6. Have excess Ops work overflow to the Dev team.
- NO 7. Cap SRE operational load at 50 percent.
- IRR 8. Share 5 percent of Ops work with the Dev team.
- Oncall teams should have at least eight people at one location, or six people at each of multiple locations.
- YES 10. Aim for a maximum of two events per oncall shift.
- YES 11. Do a postmortem for every event.
- 12. Postmortems are blameless and focus on process and technology, not people.

Management/budget policies that put value on operations:

5. Have a common staffing pool for SRE and Developers.

Handle Outages/Incidents Awesomely

- 10. Aim for a maximum of two events per oncall shift.
- 11. Do a postmortem for every event.
- Postmortems are blameless and focus on process and technology, not people.

SLA-Driven Operations & Monitoring

- 2. Have an SLA for your service.
- 3. Measure and report performance against the SLA.
- 4. Use Error Budgets and gate launches on them.

Control Operations Overload

- 1. Hire only coders.
- 6. Have excess Ops work overflow to the Dev team.
- Cap SRE operational load at 50 percent.
- Share 5 percent of Ops work with the Dev team.
- Oncall teams should have at least eight people at one location, or six people at each of multiple locations.

...THE EASY ONES...



MANAGEMENT/BUDGET DYNAMICS:

Have a common staffing pool for SRE and Developers.



HANDLE OUTAGES/INCIDENTS AWESOMELY

- Aim for a maximum of two events per oncall shift.
- 11. Do a postmortem for every event.
- 12. Postmortems are blameless and focus on process and technology, not people.

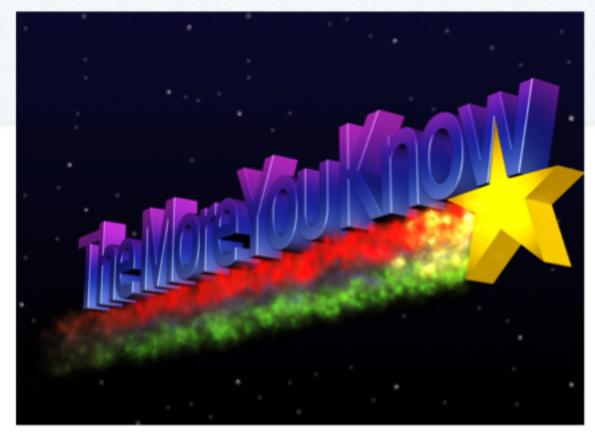


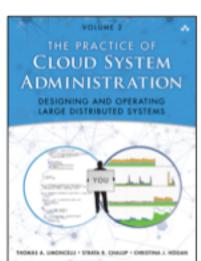
WHY IMPORTANT?

- Postmortems are how we learn from outages.
- Each outage educates the team.
- The team then educates the entire company.
- The net result: The organization gets smarter over time.

 We must create an environment where people have time to PM every event. I dunno about anybody else, but I really like getting these post-mortem reports. Not only is it nice to know what happened, but it's also great to see how you guys handled it in the moment and how you plan to prevent these events going forward. Really neato. Thanks for the great work:)

---Anna





- Template in Appendix D
- Web search [John Allspaw Postmortem]

...THE HARDER ONES...

SLA-DRIVEN OPERATIONS & MONITORING

- 2. Have an SLA for your service.
- Measure and report performance against the SLA.
- Use Error Budgets and gate launches on them.





DEFINING SLAS IS DIFFICULT BECAUSE...

- ...they are complex
- ...it means talking with product management

 Important because SLAs are fundamental to managing a service.

MONITORING

- Collect a time-series of measurements.
 - NOT "up/down" monitoring.
 - Example:

```
03:04:12:00 server1.cpu.loadave 3.1 03:04:17:00 server1.cpu.loadave 3.0 03:04:22:00 server1.cpu.loadave 2.9 03:04:27:00 server1.cpu.loadave 0.1
```

- Up/down monitoring == bad
 - You get alerted when the system is down.
 - The outage is "built in"
- Time-series == good:
 - Alerting when "sick", not "down".
 - Prevents outages
 - Long-term history needed for planning.



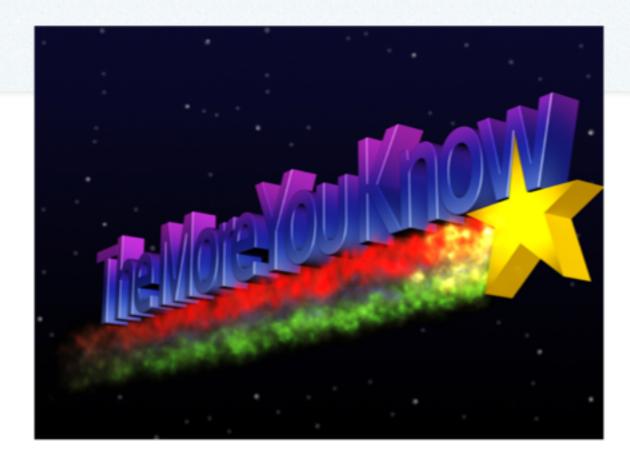
STATUS

- Dissatisfied with existing monitoring tools.
- So we built our own: <u>www.bosun.org</u>
 - The first monitoring system with an IDE for designing complex alerts!
 - Uses OpenTSDB for storage
 - Rewrote agent in Go, works w/Linux+Win
 - Can use Graphite as a back-end (NEW!)



STATUS (2)

- We made it easy for Devs to specify what is collected and define alerts.
 - They think its fun!
 - Found+fixed many previously unknown performance issues.
 - Other departments have made more interesting rules than we have.
- SLAs: In process of defining.



- Time series rules
- www.bosun.org



SRE IS CONCERNED WITH...





SLA

speed and performance security capacity planning software upgrades hardware upgrades availability

You don't know when to stop optimizing and/or spending.

You know the gaps.
You can prioritize.
You know when to stop!

IT ISN'T A SERVICE IF IT ISN'T MONITORED. IF THERE IS NO MONITORING THEN YOU'RE JUST RUNNING SOFTWARE.

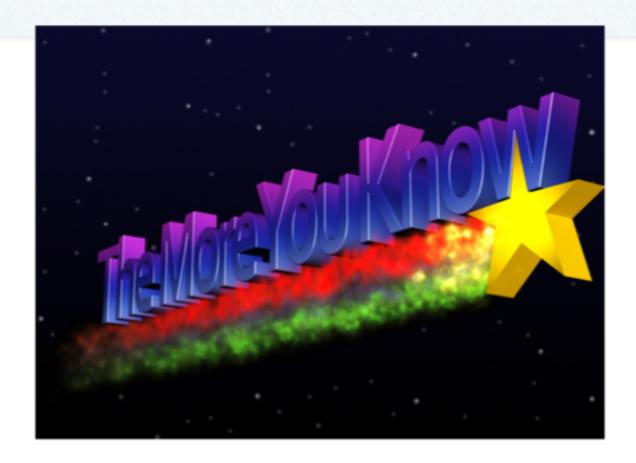
IT ISN'T A SERVICE IF IT ISN'T MONITORED AGAINST AN SLA. IF THERE IS NO SLA-BASED MONITORING THEN YOU'RE JUST RUNNING SOFTWARE.

SRE IS SLA-DRIVEN SYSTEM ADMINISTRATION



EXAMPLE: CDN DEPLOYMENT

- Enable proxy-style CDN for all content.
 Better TCP handshake, compression, improvement over HTTP.
 - Ooops! Performance got worse!
 - Bug fix involved vendor code overhaul
 - Why didn't other customers notice this?
 - Isn't anyone doing before&after comparison?
 - Measure twice, cut once!



CONTROL OPERATIONS OVERLOAD

CONTROL OPERATIONS OVERLOAD

- Hire only coders.
- Have excess Ops work overflow to the Dev team.
- Cap SRE operational load at 50 percent.
- Share 5% Ops work w/ the Dev team.
- Oncall teams.. 6-8 ppl/location

ENDING TRANSACTIONAL SYSTEM ADMINISTRATION



TRANSACTIONAL SYSTEM ADMINISTRATION

- Users make requests
- Sysadmins fulfill requests
- Bad because:
 - Wrong relationship dynamic
 - Scales by hiring more people (yuck!)



SERVICE ADMINISTRATION

- You don't process transactions.
- You run the system that processes transactions
- Better because:
 - Scales
 - SLAs
 - Better service for customers
 - Less boring for the sysadmins



EXAMPLE 1

- We didn't create monitoring rules for Devs.
- We created and run a system that lets Devs be self-sufficient

- Better because:
 - Devs have better knowledge of internals
 - "Workforce multiplier"

EXAMPLE 2 (VERY "ENTERPRISE-Y")

- SREs shouldn't install and deliver laptops to users.
- SREs should maintain the system that:
 - Knows when a new laptop is needed:
 - Tracks HR database for new employees
 - Emails users when eligible for refresh
 - A portal for special ordering machines
 - Automates OS installation for use by a "laptop delivery crew" in each building. (Non-technical clerks and IT coordinator)
 - Does capacity planning, generates purchase orders, etc. for the purchasing/finance group.
- Better because:
 - OMG do I have to explain?



STATUS

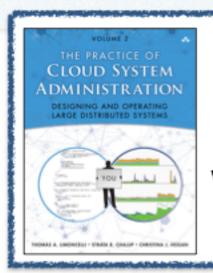
- Stack Exchange SRE:
 - Got rid of the SRE "ticket system"
 - Dedicated SRE "point person(s)" on each development team
 - SRE-related work is scheduled and assigned to SRE in a cooperative manner (agile-style).
- SREs have more time for service development, optimization, projects, etc.



IN CONCLUSION

- SLAs SLAs SLAs
- End transactional system administration
- Spread this philosophy to the enterprise
- Try www.Bosun.org
- Q&A

ADDITIONAL RESOURCES



The Practice of Cloud System Administration

www.informit.com/TPOSA
Discount code TPOSA35

- Ch 7: Defines SRE
- Ch 14: Postmortems
- Ch 15: Drills/DiRT/Game Days
- Ch 16-17: Monitoring
- Ch 18: Capacity Planning
- Ch 19+20+App A: SLAs/KPIs

Our SRE team has a blog: blog.serverfault.com

Follow me on Twitter: @YesThatTom

We're hiring in NYC! stackexchange.com/work-here

I'll be the speaker at these Meetups this week:

Tuesday, 6pm — Large-Scale Production Eng Meetup (at URL Cafe, Yahoo! Sunnyvale) Thursday, 7:30pm sharp! — BayLISA (at LinkedIn office, Sunnyvale)



IN CONCLUSION

- SLAs SLAs SLAs
- End transactional system administration
- Spread this philosophy to the enterprise
- Try www.Bosun.org
- Q&A