CAPACITY PLANNING:

Telling the future without a crystal ball

HELLO!

I am Evan Smith

and I'm an SRE with Hosted Graphite.

HostedGraphite

You can find me at @TheJokersThief

Or iamevan.me if you're really desperate

WHY BOTHER?

- Anticipate sharp growth
- Only spend as much as you actually need
- **Avoid 3AM pages**

CAPACITY PLANNING OBJECTIVE

The goal should be to drive the system to the appropriate level of risk for the lowest cost.

A CASE STUDY:

"We've been running our authentication service for three years, how do we possibly start planning capacity now?"

– Generrico Shoppe

INTENTS & SLOs

What should this service accomplish?

1

- Understanding **intent** begins with helping define a **Service Level Objective**

The Good

"I want 50 cores in clusters X, Y, and Z for service Foo."

This and the following examples come from The SRE Book: Chapter 4 - Service Level Objectives

The Hard

"I want to meet service Foo's demand in each geographic region, and have N + 2 redundancy."

CASE STUDY: The Ugly

"I want to run the Authentication Service at 5 nines of reliability. It's gotta be up always."

CASE STUDY:

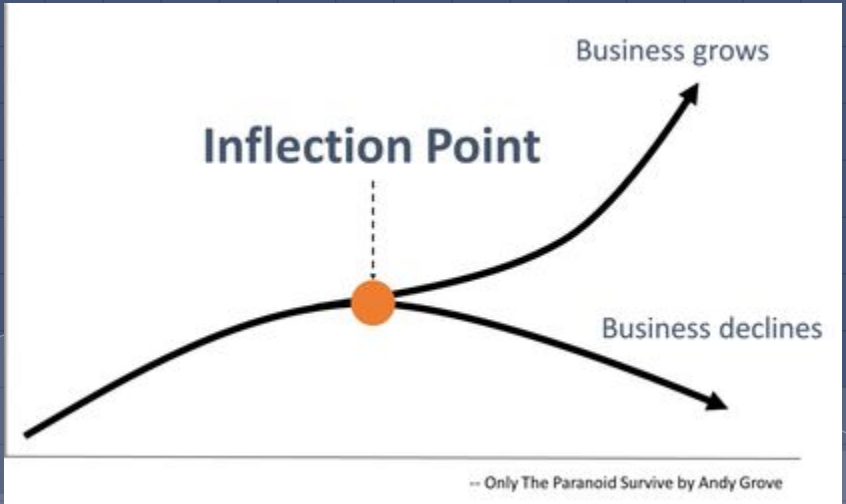
- What if it costs €10,000 to give you 5 nines?
  - *9 hrs/yr down (3 nines) is actually fine*
- Where is most of your business?
  - *Heaviest users in US and UK, we could use different SLOs per-region (lower SLOs outside of US/UK)*

SERVICE TRIGGERS

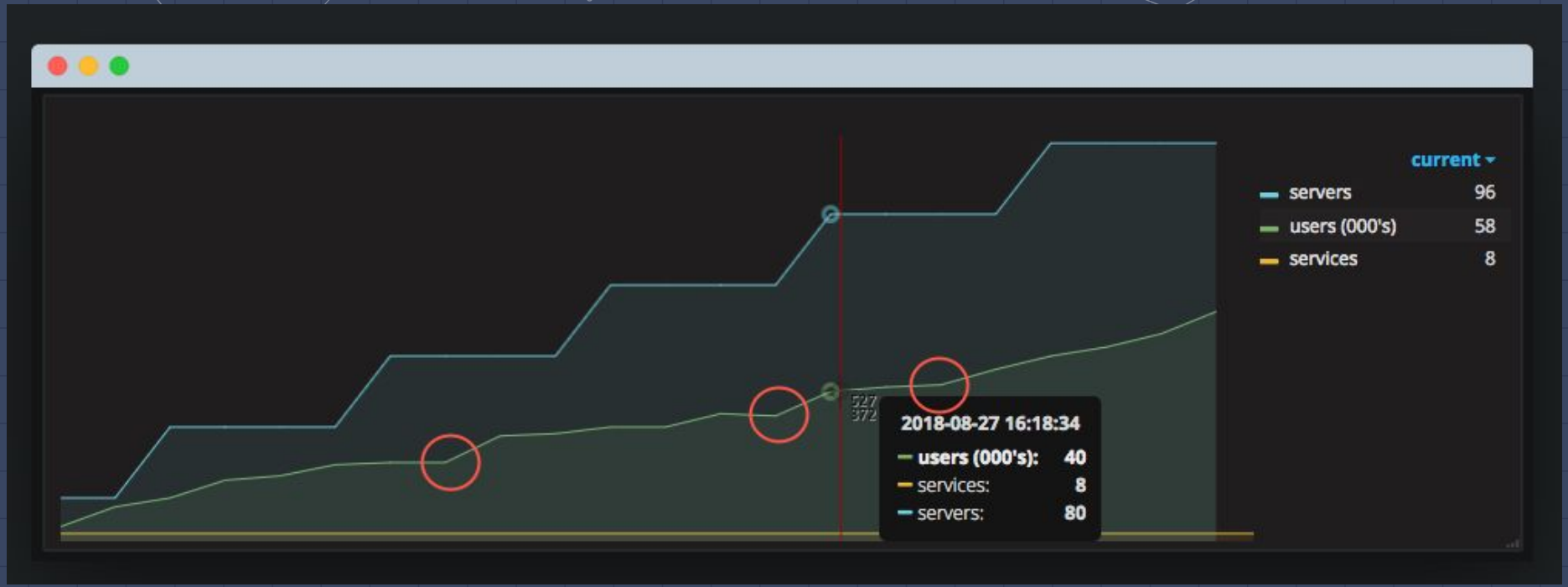
Which metrics move the needle?

2

- Find **driver metrics**
- Look at historical **inflection points**



CASE STUDY:


current
servers 96
users (000's) 58
services 8
527
372
2018-08-27 16:18:34
users (000's): 40
services: 8
servers: 80

CASE STUDY:

- Authentication is driven proportional to:
  - Number of users
  - Number of linked services
- Plotting capacity against the driver metrics, we discover:
  - **<num services> * <num users>** accurately describes load
  - Every 64,000 units, they increase capacity by 16 servers

ACTIONABLE INSIGHTS

When should capacity change?

3

- It sounds obvious but **document everything**
  - process
  - findings
  - assumptions
  - graphs and metrics
- **If a tree falls in the woods, it doesn't exist until it's documented**

Make insights **actionable**

Always provide easy access to **context**

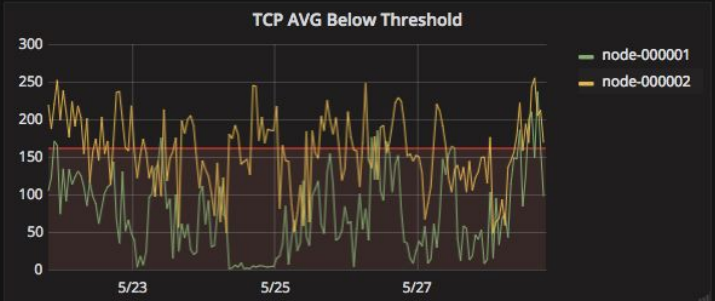
Information

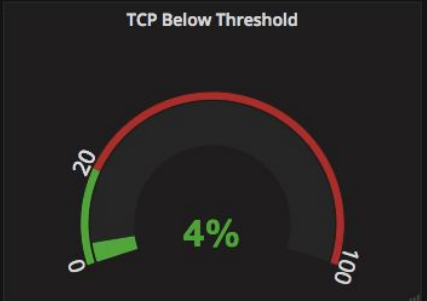
TCP Capacity Plan:

Capacity Plan Outline:

- **Increase** when percentage of nodes under threshold is above 20%

This graph shows total percentage for 8 cores - 800% max. To ensure we stay below 80% CPU Usage, we want idle time to be at least 160% for each host.

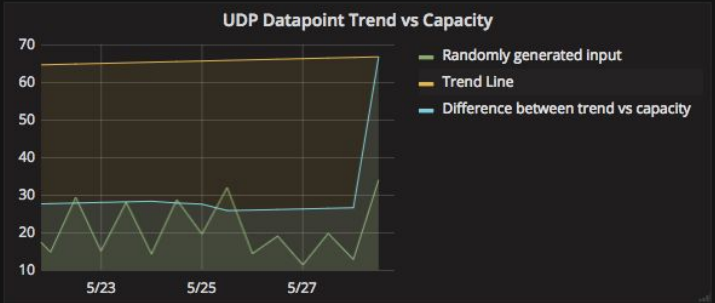


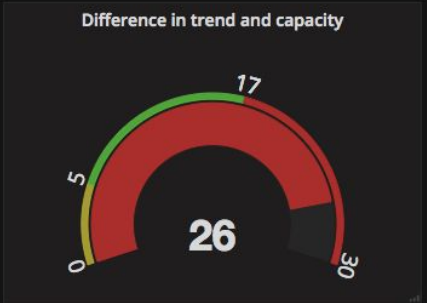


UDP Capacity Plan:

Capacity Plan Outline:

- **Increase** when difference between scaled trend and current capacity is above 17
- **Decrease** when difference between scaled trend and current capacity is below 5





CASE STUDY:

- If every 64,000 units, they increase by 16, they probably actually want to increase by 4 server every 16,000
- We should **increase** capacity every increase of 14,500 units
- We should **decrease** capacity every decrease of 14,500 units

FORECAST

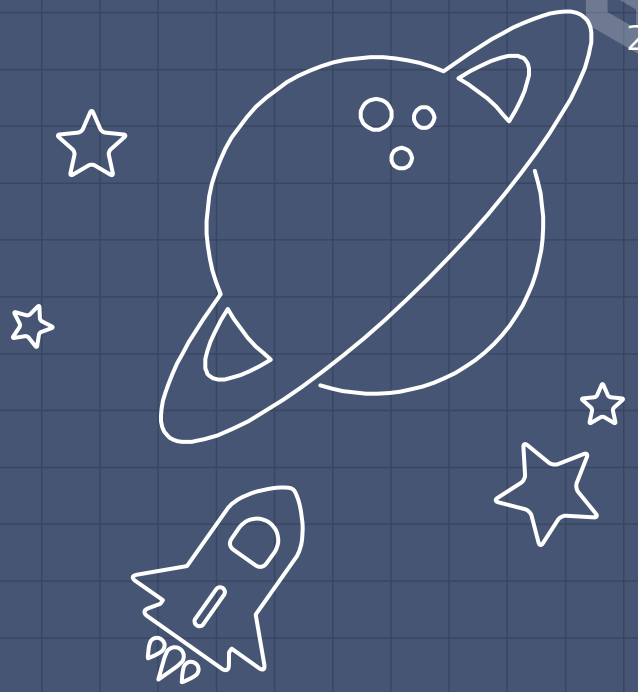
What does the future look like?

4

If you're not tracking the capacity of your services already...

START

RIGHT NOW



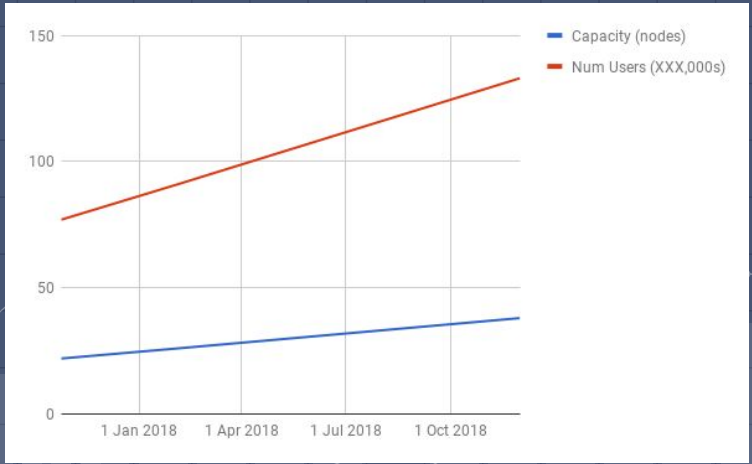
- The easiest way to **predict the future** is to **use the past**

- *Estimate* the capacity for each period by looking at <u>change from the past</u>

CASE STUDY:

- Put your estimates in a table AND graph it!

Timeline	State	Increase By	Actual Date	Actual State
25 Oct 2017	22	4 nodes		
4 Feb 2018	26	4 nodes		
15 May 2018	30	4 nodes	5 April 2018	28
23 Aug 2018	34	4 nodes		
1 Dec 2018	38	4 nodes		
	42 nodes			



SOME FINAL TIPS

1. Your plan's buffer should also account for Lead Time - how long it takes to go from no server to production-ready
2. Machine Learning have some great methods for choosing your driver metrics - PCA, Lasso Regression, feature selection
3. Capacity Planning is not a set-it-and-forget-it activity - you will need to come back to your plan every 1-3 months (at least at the start) depending on its size

THANKS!

Any questions?

You can find me at

- @TheJokersThief
- evan.smith@hostedgraphite.com

HostedGraphite

Further Reading/Watching:

1. Cloud Capacity Planning.. an Oxymoron? by Coburn Watson (Netflix)
2. The Data in the Planning by Sebastien de Larquier (Netflix)
3. Capacity Planning by David Hixson (Google) and Kavita Guliani (Google) (excerpt from ;login: vol 40 published by usenix)
4. SRE Book - Chapter 18: Software Engineering by (Google) Dave Helstroom and Trisha Weir with Evan Leonard and Kurt Delimon
5. Capacity Management For The Cloud by Ernest de Leon (Mirantis)