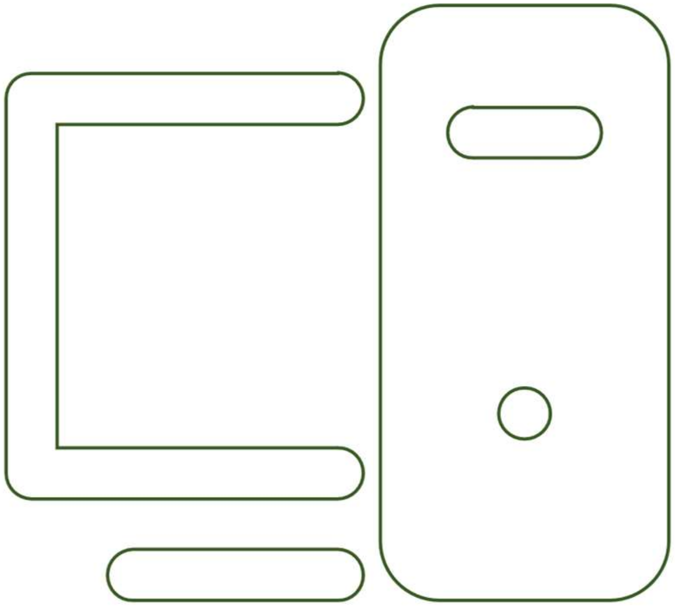
Recovery By Design

A Postmortem Adventure



What to expect…

How I leverage our existing technology and tools to make bad days less bad.

A bad day for me is…

- when things are down (and I have no idea why)
- when clients are unhappy (and I can do something about it)
- when we can do better (and we didn’t)

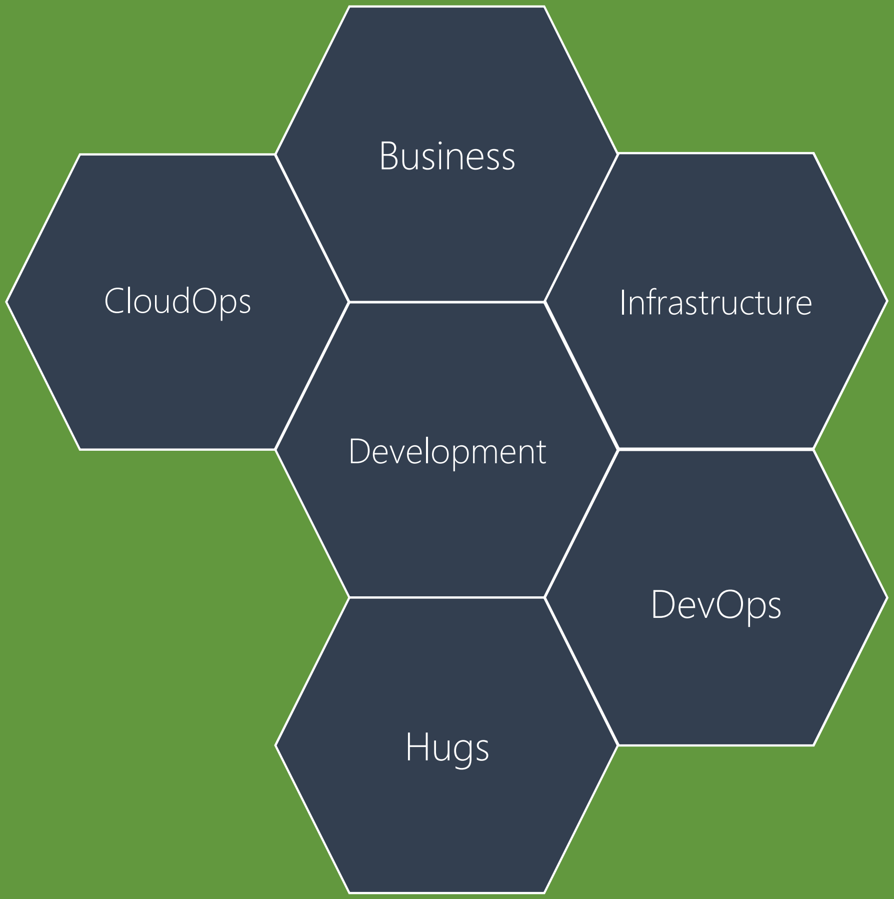
Agenda

- Blue / Green Deployments
- Postmortems
- Ingress Gateway & Proxy (HAProxy)
- Process Dump Analysis

Chris Houdeshell

CTO / EnergyCAP

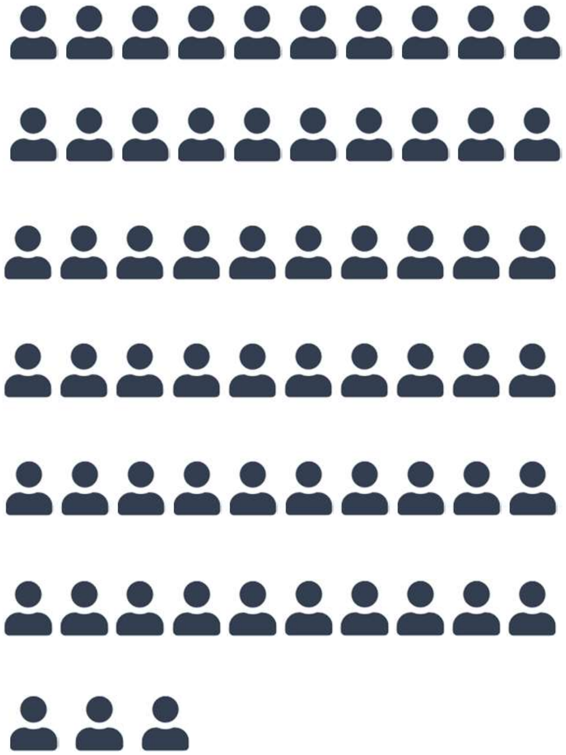
@choudeshell



EnergyCAP Inc.

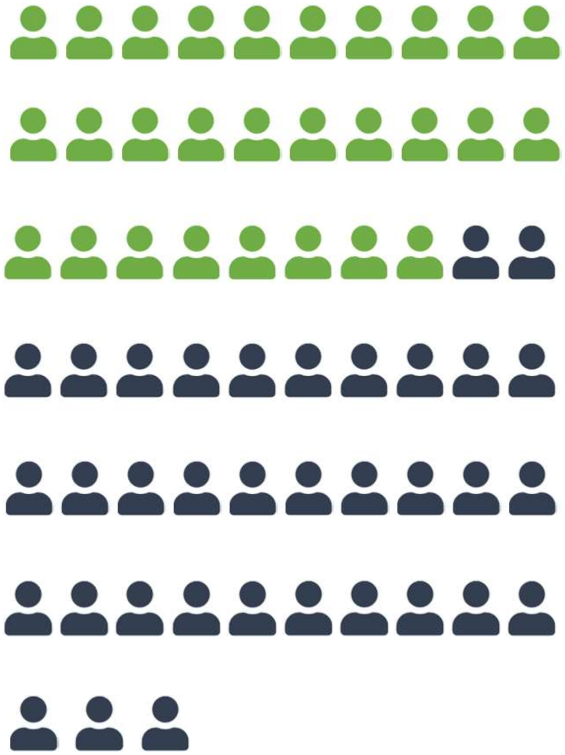
Utility Bill Analytics & Accounting Services & Software Firm

EnergyCAP



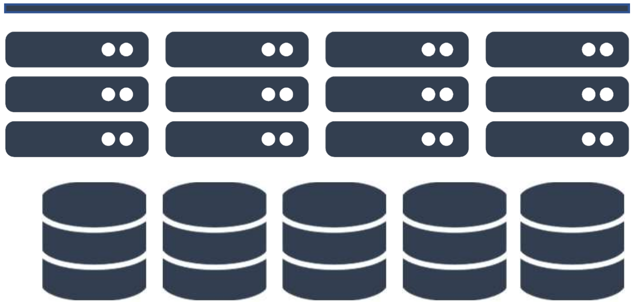
63 People

EnergyCAP



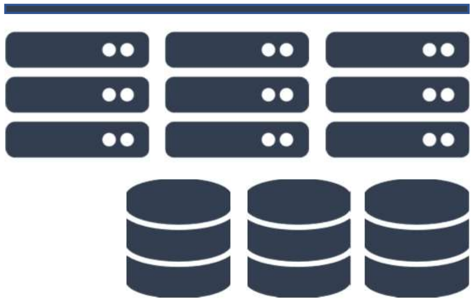
28 People

Production (EnergyCAP-APP)



Pittsburgh

Implement (EnergyCAP-APP)



State College

315ms
p95

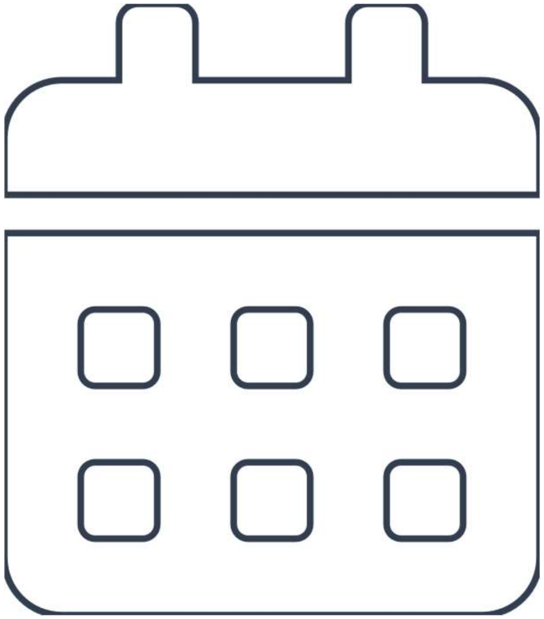
500K
requests per day

over
$12.2 Billion
YTD

A Couple Bad Days…

Bad Day #109

Wednesday



How do I know if something isn't right?

Telemetry & Metrics

Application Insights
& Azure Monitor
& Splunk

Automated Alerts

Opsgenie

Internal Staff

PMs

Clients

Tickets

Telemetry & Metrics

High CPU on Implement-B1 & B2 & B3
Low Available Memory on Implement-B1 & B2
Spikey IO on Implement-B1 & B2

Automated Alerts

Implement-B1 was taken out of rotation
Implement-B2 was taken out of rotation

Internal Staff

“Implement is down”

Clients

“I can’t pay my bills”
“Things are slow”

Telemetry & Metrics

High CPU on Implement-B1 & B2 & B3
Low Available Memory on Implement-B1 & B2
Spikey IO on Implement-B1 & B2

Automated Alerts

Implement-B1 was taken out of rotation
Implement-B2 was taken out of rotation

Internal Staff

“Implement is down”

Clients

“I can’t pay my bills”

When all 4 channels report issues…

You know it is going to be a very bad day.

Telemetry & Metrics

High CPU on Implement-B1 & B2 & B3
Low Available Memory on Implement-B1 & B2
Spikey IO on Implement-B1 & B2

Automated Alerts

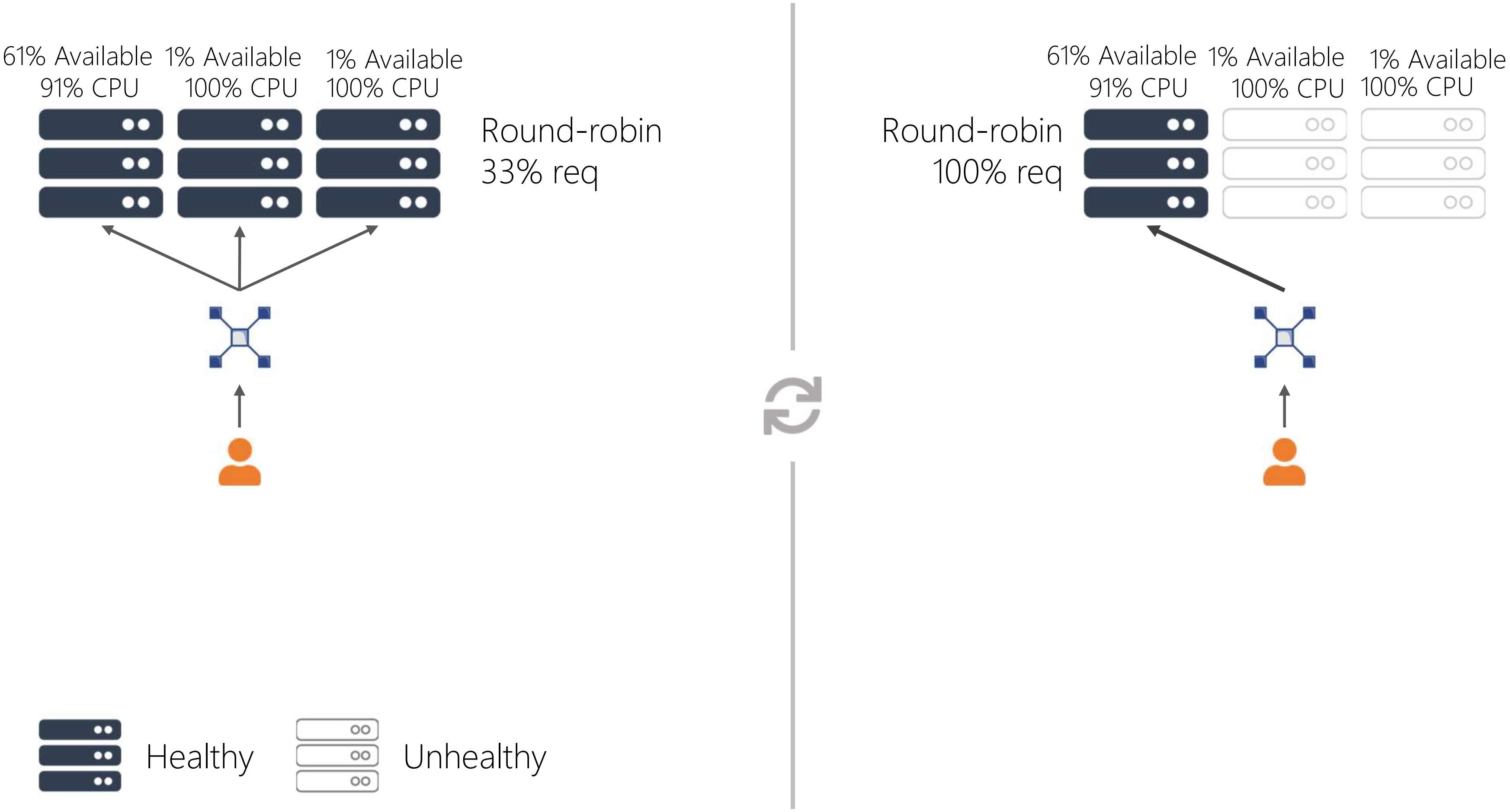
Implement-B1 was taken out of rotation
Implement-B2 was taken out of rotation

Internal Staff

"Implement is down"

Clients

"I can't pay my bills"
"Things are slow"



61% Available
91% CPU

1% Available
100% CPU

1% Available
100% CPU

Round-robin
33% req

61% Available
91% CPU

1% Available
100% CPU

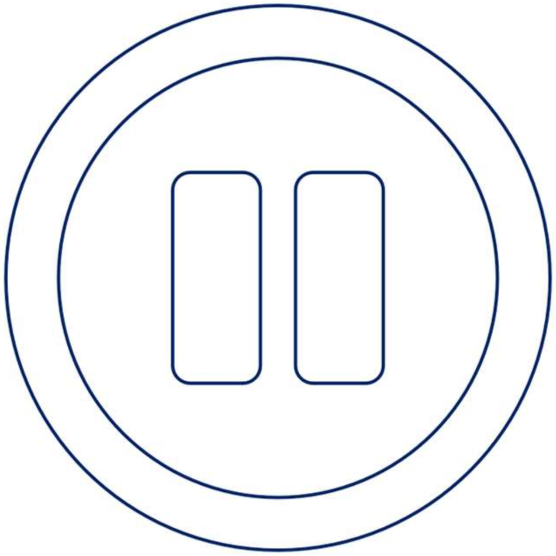
1% Available
100% CPU

Round-robin
100% req

Healthy

Unhealthy

Restoring availability and acceptable performance is the primary goal [watch for blast radius]

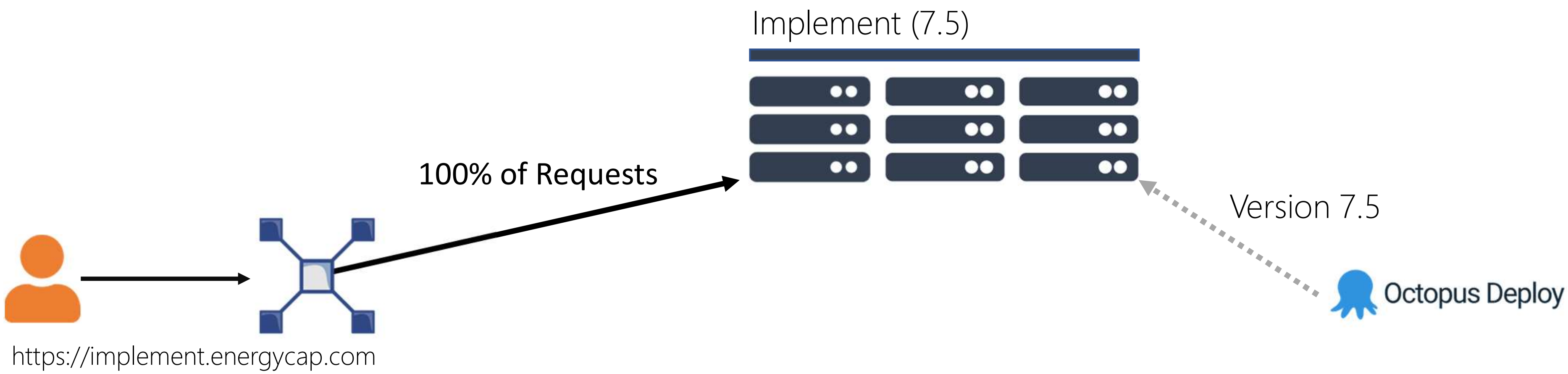


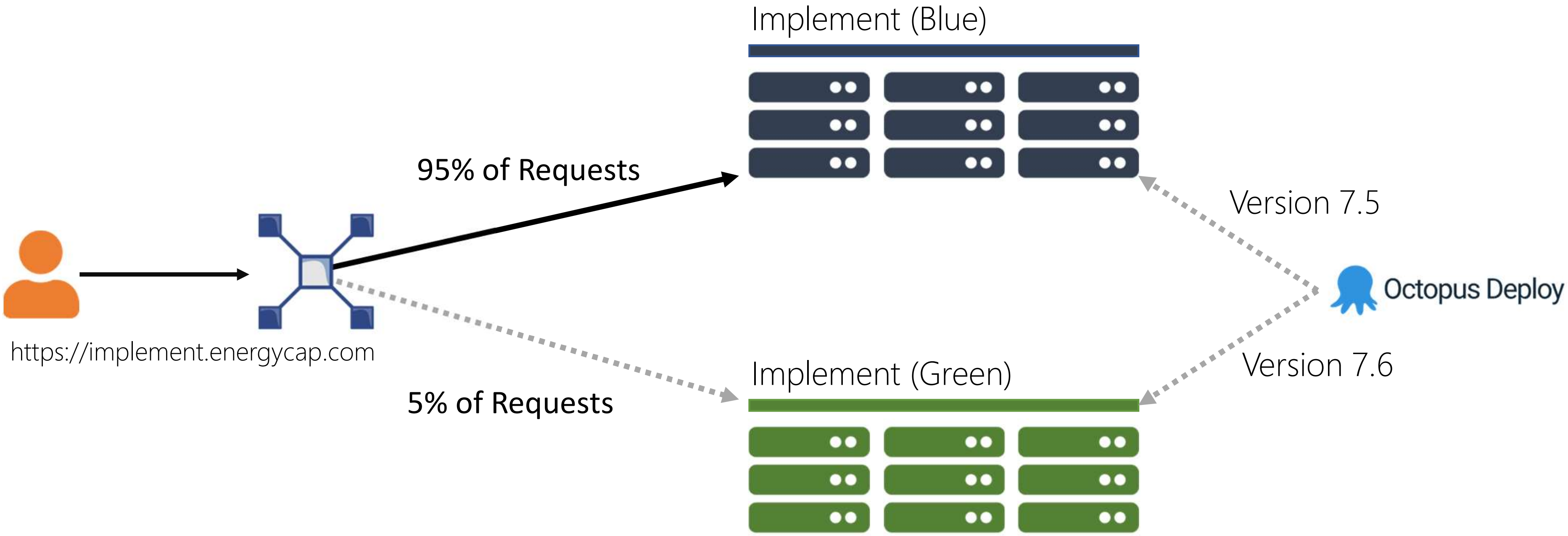
Blue / Green Deployments

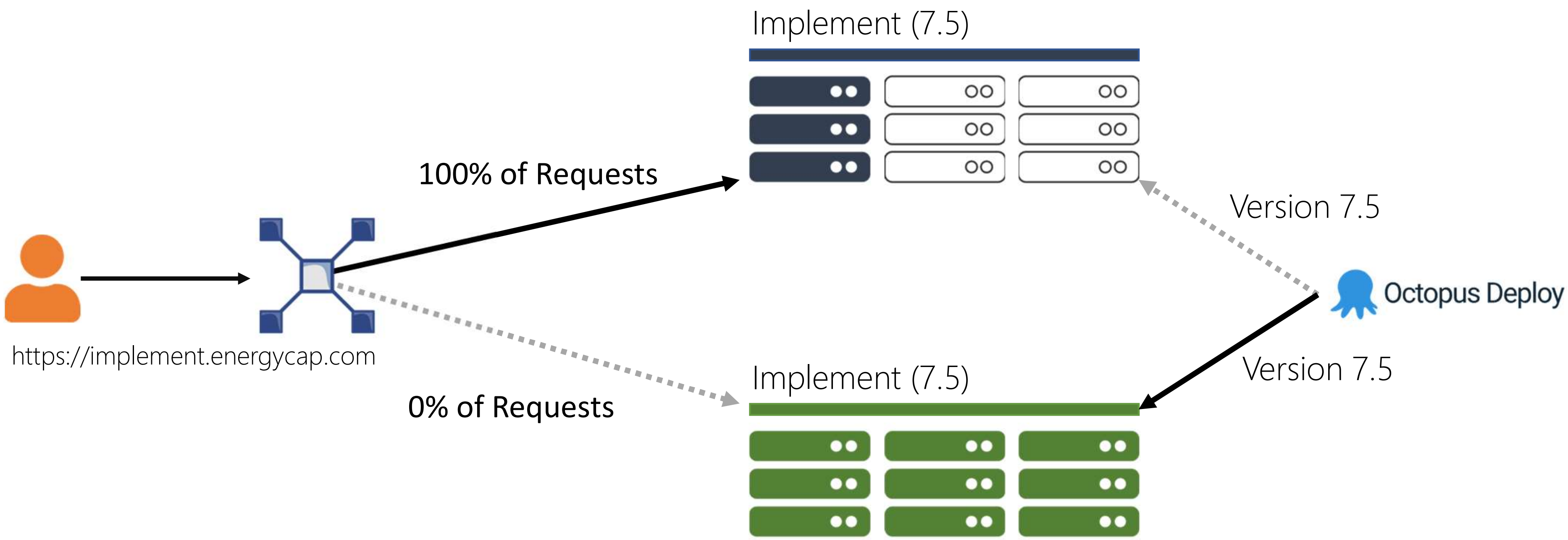
- Reduce downtime
- Immediate rollbacks
- Reduce risks
- Canary releases

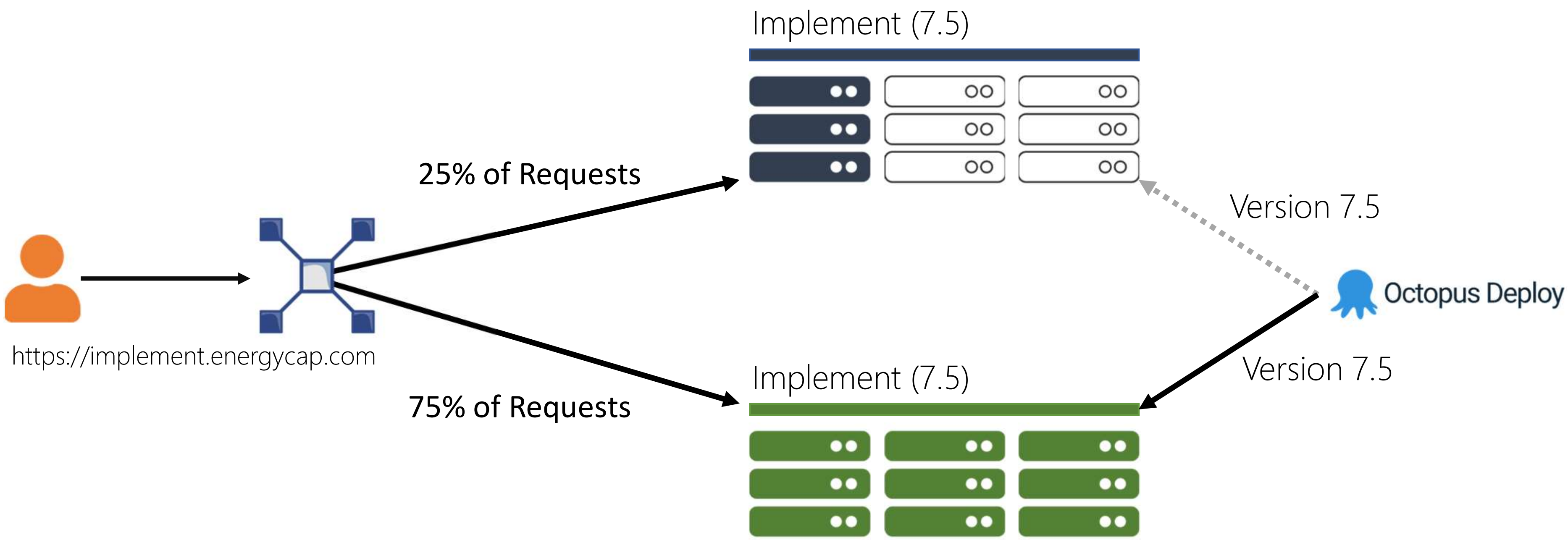
Implement - Blue

Implement - Green


Implement (7.5)
100% of Requests
Version 7.5
Octopus Deploy
https://implement.energycap.com

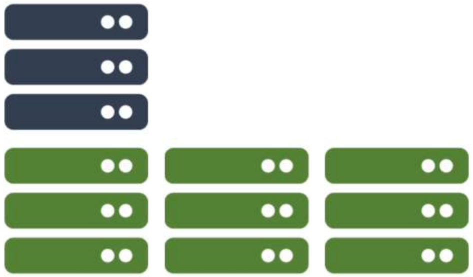

Implement (Blue)
95% of Requests
Version 7.5
Octopus Deploy
https://implement.energycap.com
Implement (Green)
5% of Requests
Version 7.6


Implement (7.5)
100% of Requests
Version 7.5
Octopus Deploy
https://implement.energycap.com
Implement (7.5)
Version 7.5
0% of Requests



Implement (7.5)
25% of Requests
Version 7.5
Octopus Deploy
https://implement.energycap.com
Version 7.5
Implement (7.5)
75% of Requests

Implement



Combined Blue / Green Environments

- Resulted in healthy app servers
- Acceptable CPU load
- Acceptable Memory load
- Almost acceptable API performance

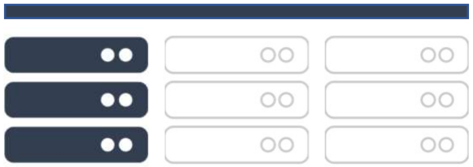
SLO (Service Level Objective)

- GET APIs (Singular) – 500ms (p75)
- GET APIs (List) – 1.0 second (p75)

Vertical Scaling

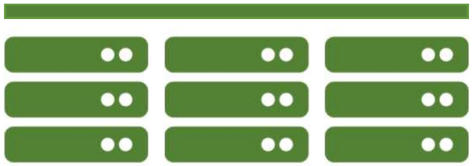
Increase the capacity of a single instance

Implement - Blue

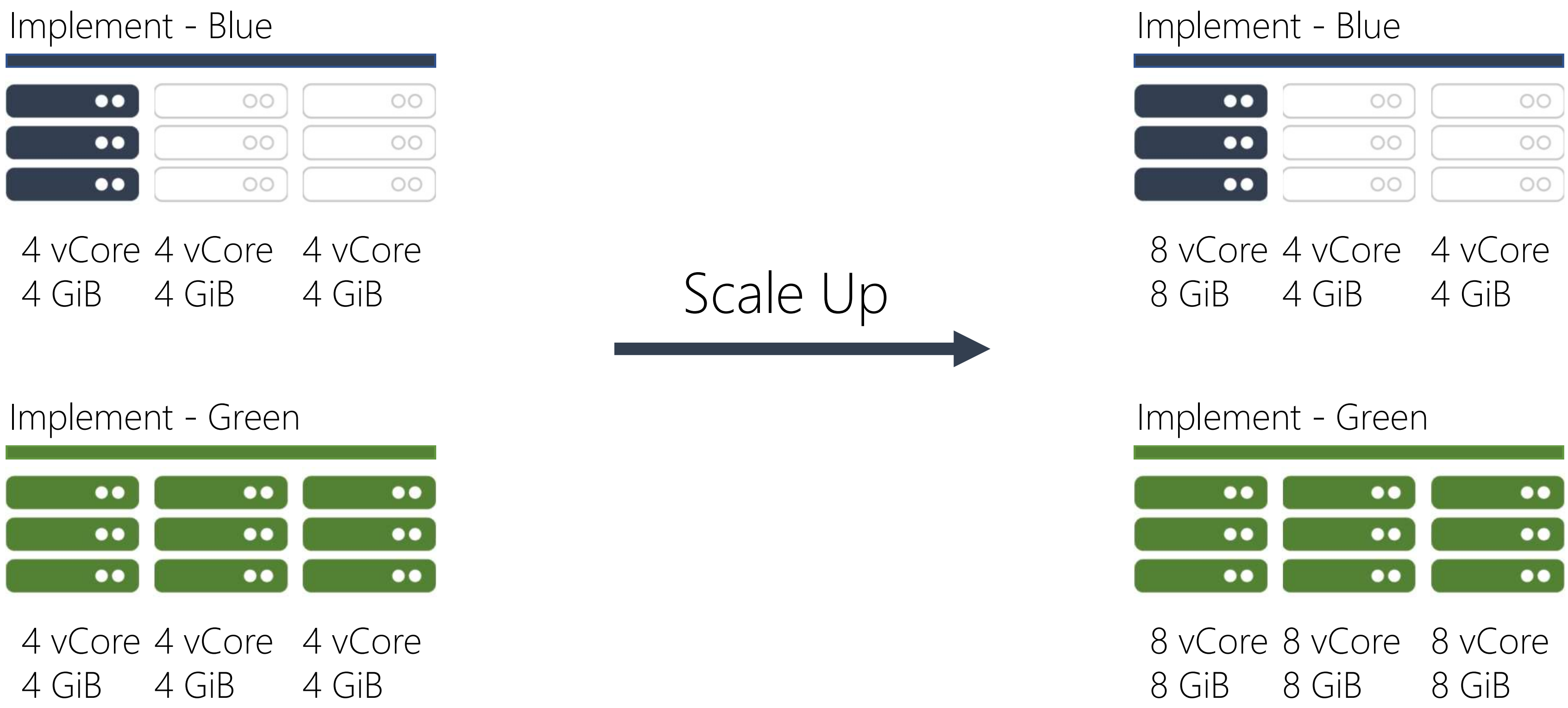


4 vCore	4 vCore	4 vCore
4 GiB	4 GiB	4 GiB

Implement - Green



4 vCore	4 vCore	4 vCore
4 GiB	4 GiB	4 GiB



Implement - Blue

4 vCore 4 GiB | 4 vCore 4 GiB | 4 vCore 4 GiB

Implement - Green

4 vCore 4 GiB | 4 vCore 4 GiB | 4 vCore 4 GiB

Scale Up

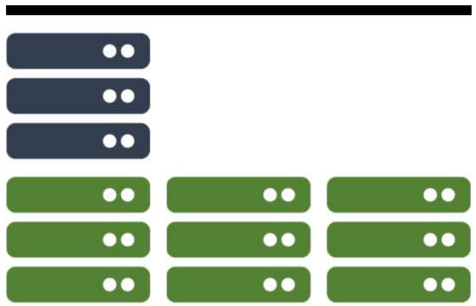
Implement - Blue

8 vCore 8 GiB | 4 vCore 4 GiB | 4 vCore 4 GiB

Implement - Green

8 vCore 8 GiB | 8 vCore 8 GiB | 8 vCore 8 GiB

Implement



Why do I like this method?

- Warm & Ready instances
- Allows analysis of the unhealth machines later
- Our deployments are fast (70 secs. per instance)
- Zero downtime/Zero re-deploy vertical scaling

< 5 minutes

How do I know if something isn't right?

Telemetry & Metrics

Application Insights
& Azure Monitor
& Splunk

Automated Alerts

Opsgenie

Internal Staff

PMs

Clients

Tickets

Telemetry & Metrics 👍

Clients 👍

Automated Alerts 👍

Internal Staff 👍

But the day isn’t over.

Incident Postmortem

- High level summary of events
- RCA – Root cause analysis
- Timeline of events
- Next Steps

Incident Postmortem

High level summary of events

- Which services and clients were affected?
- How long were they affected?
- How severe was the issue?
- Did we fix it?

Incident Postmortem

RCA – Root cause analysis

- How & why did this happen?

Incident Postmortem

Next Steps

- What went well?
- Any app changes?
- Any infrastructure changes?

Incident Postmortem

High level summary of events

- Implement environment was effected, specifically B1 & B2. Since there was high CPU, these machines were removed from the load balancer causing a degradation in availability and response time.
- Implement environment was in a degraded state for approximately 15 minutes.
- The event was severe due to it causing clients to not fulfill bill payment
- The event was mitigated by allocating more resources to the Implement environment.

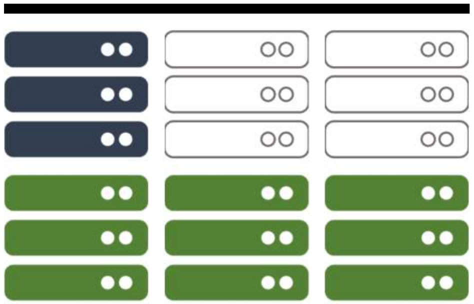
Incident Postmortem

- RCA – Root cause analysis
  - Hmmmm….

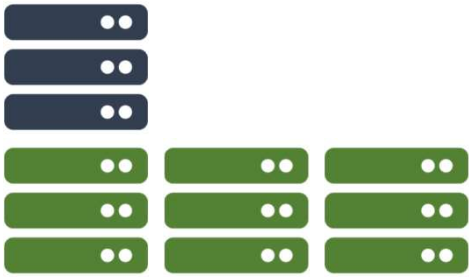
Zero blame.

Things can be stressful. Take a walk.

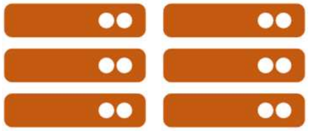
Implement



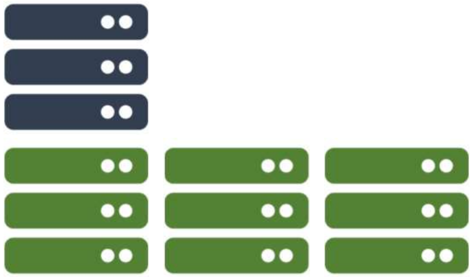
Implement



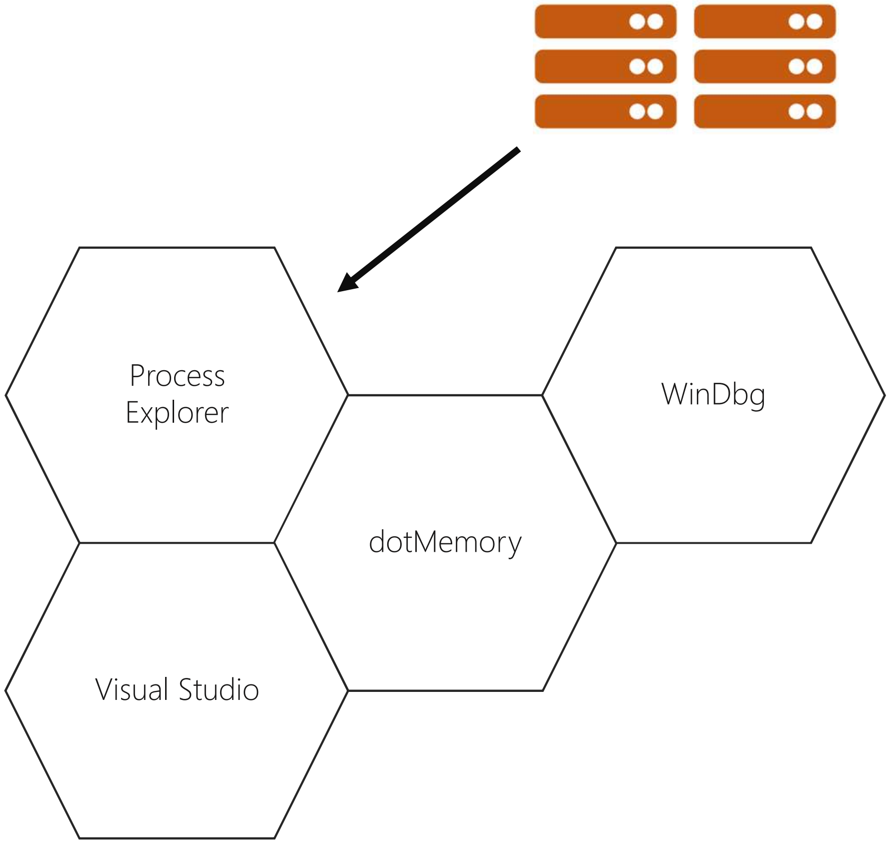
Dump Analysis



Implement



Dump Analysis

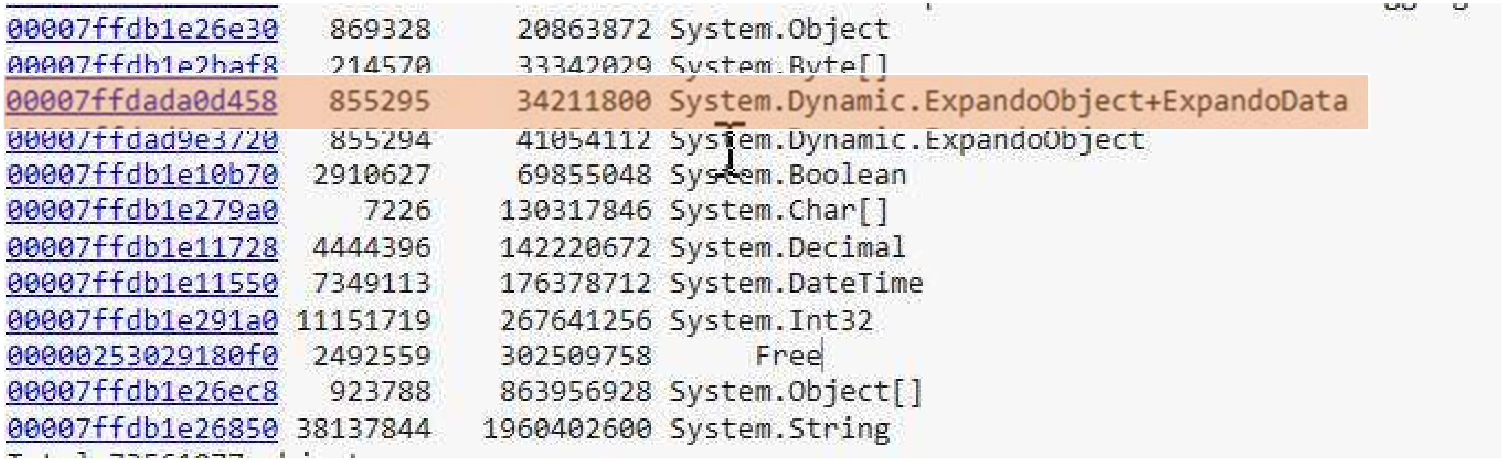


Process Explorer

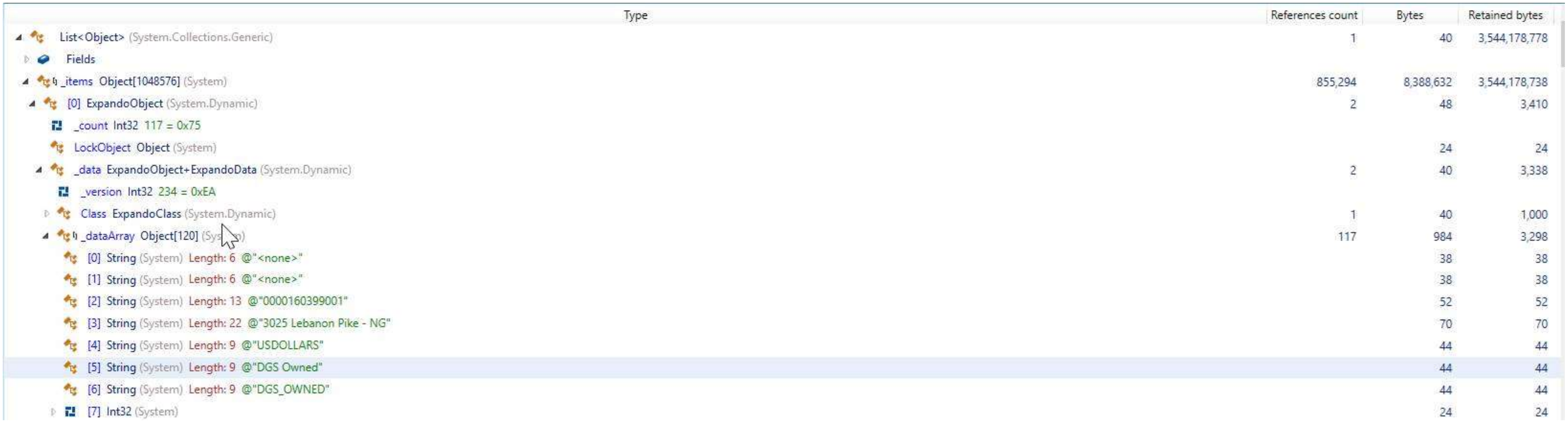
dotMemory

WinDbg

Visual Studio



```
00007ffdb1e26e30   869328     20863872 System.Object
00007ffdb1e2baf8   214570     33342029 System.Byte[]
00007ffdada0d458   855295     34211800 System.Dynamic.ExpandoObject+ExpandoData
00007ffdad9e3720   855294     41054112 System.Dynamic.ExpandoObject
00007ffdb1e10b70  2910627     69855048 System.Boolean
00007ffdb1e279a0     7226    130317846 System.Char[]
00007ffdb1e11728  4444396    142220672 System.Decimal
00007ffdb1e11550  7349113    176378712 System.DateTime
00007ffdb1e291a0 11151719    267641256 System.Int32
00000253029180f0  2492559    302509758      Free
00007ffdb1e26ec8   923788    863956928 System.Object[]
00007ffdb1e26850 38137844   1960402600 System.String
```



Type	References count	Bytes	Retained bytes
◢ List<Object> (System.Collections.Generic)	1	40	3,544,178,778
▷ Fields			
◢ _items Object[1048576] (System)	855,294	8,388,632	3,544,178,738
◢ [0] ExpandoObject (System.Dynamic)	2	48	3,410
_count Int32 117 = 0x75			
LockObject Object (System)		24	24
◢ _data ExpandoObject+ExpandoData (System.Dynamic)	2	40	3,338
_version Int32 234 = 0xEA			
▷ Class ExpandoClass (System.Dynamic)	1	40	1,000
◢ _dataArray Object[120] (System)	117	984	3,298
[0] String (System) Length: 6 @"<none>"		38	38
[1] String (System) Length: 6 @"<none>"		38	38
[2] String (System) Length: 13 @"0000160399001"		52	52
[3] String (System) Length: 22 @"3025 Lebanon Pike - NG"		70	70
[4] String (System) Length: 9 @"USDOLLARS"		44	44
[5] String (System) Length: 9 @"DGS Owned"		44	44
[6] String (System) Length: 9 @"DGS_OWNED"		44	44
▷ [7] Int32 (System)		24	24

…a couple hours later

1 client called 1 API which resulted in 4.2 GiB being allocated…

This caused our low available memory

They called it 17 times. Resulting in 71.4 GiB attempting to be allocated across 2 instances.

Just happened that 17 API calls were routed to 2 instances

Some APIs OOM'd. Others paged to SAN-backed disk.

This caused the spikey IO

Swap ate CPU. CPU starved health checks.

This caused instances to be removed from load balancer.

Incident Postmortem

RCA – Root cause analysis

- Code wasn’t scalable & performant / bounds checking
- Memory pressure and swapping pressure caused instances to be removed from load balancer due to lack of available CPU for health checks
- Not enough available resources for current load

Bad Day #109: Swapping of Death

Bad Day #121

How do I know if something isn’t right?

Telemetry & Metrics

Application Insights
& Azure Monitor
& Splunk

Automated Alerts

Opsgenie

Internal Staff

PMs

Clients

Tickets

Telemetry & Metrics 👍

Automated Alerts 👍

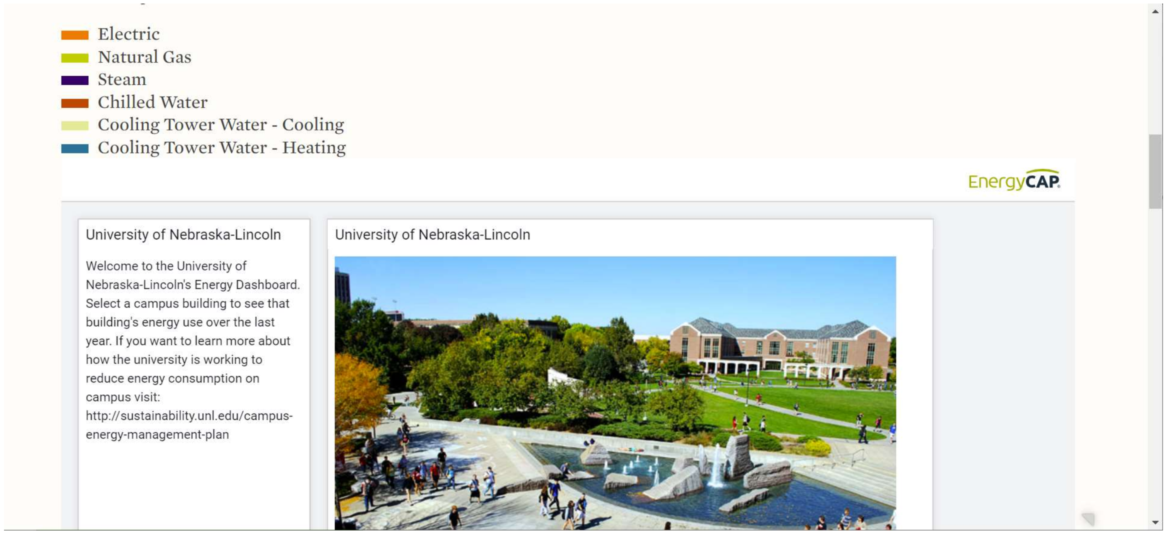
Internal Staff 👍

Clients

“I’m logged in as another user ”

Embed Key

Public Dashboard



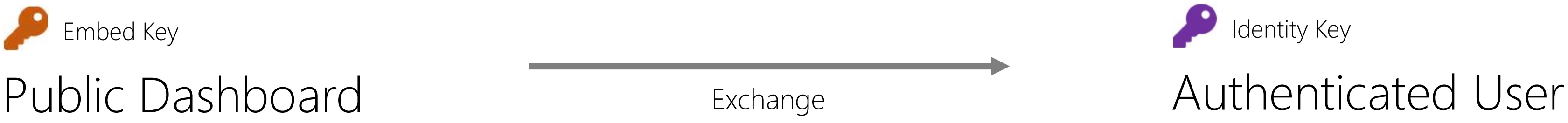
🟧 Electric
🟩 Natural Gas
🟪 Steam
🟫 Chilled Water
🟨 Cooling Tower Water - Cooling
🟦 Cooling Tower Water - Heating

University of Nebraska-Lincoln

Welcome to the University of Nebraska-Lincoln's Energy Dashboard. Select a campus building to see that building's energy use over the last year. If you want to learn more about how the university is working to reduce energy consumption on campus visit: http://sustainability.unl.edu/campus-energy-management-plan

University of Nebraska-Lincoln

https://app.energycap.com/embedded?key=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9



Embed Key

Public Dashboard

Exchange

Identity Key

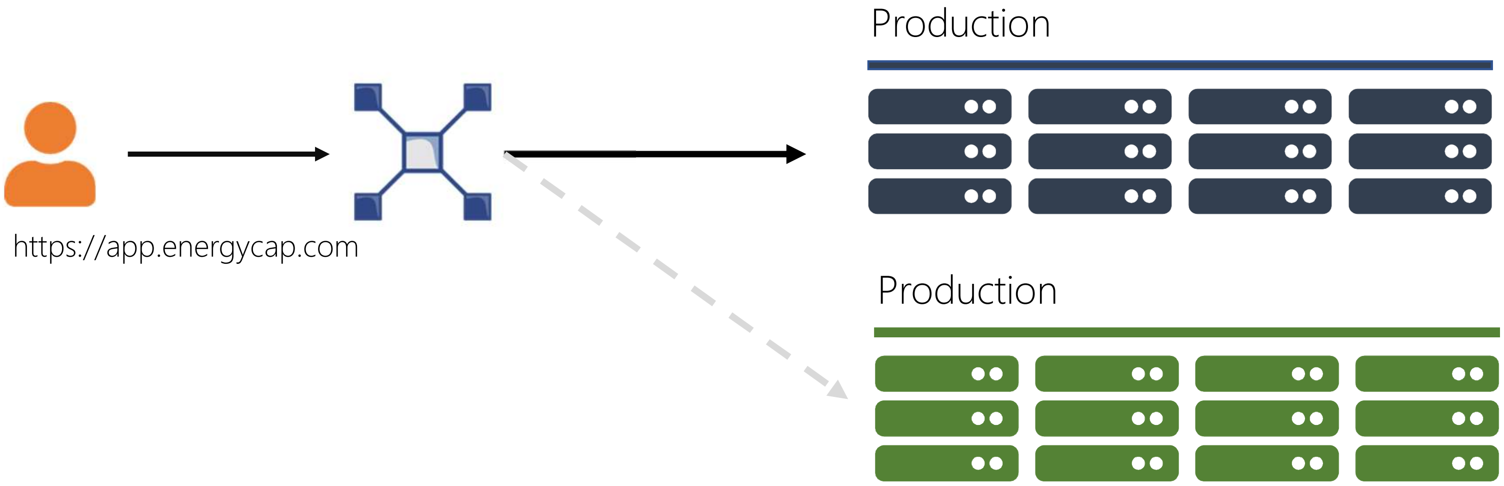
Authenticated User

https://app.energycap.com/embedded?key=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9

Primary goal is to reduce attack surface till a fix is deployed.

HAProxy

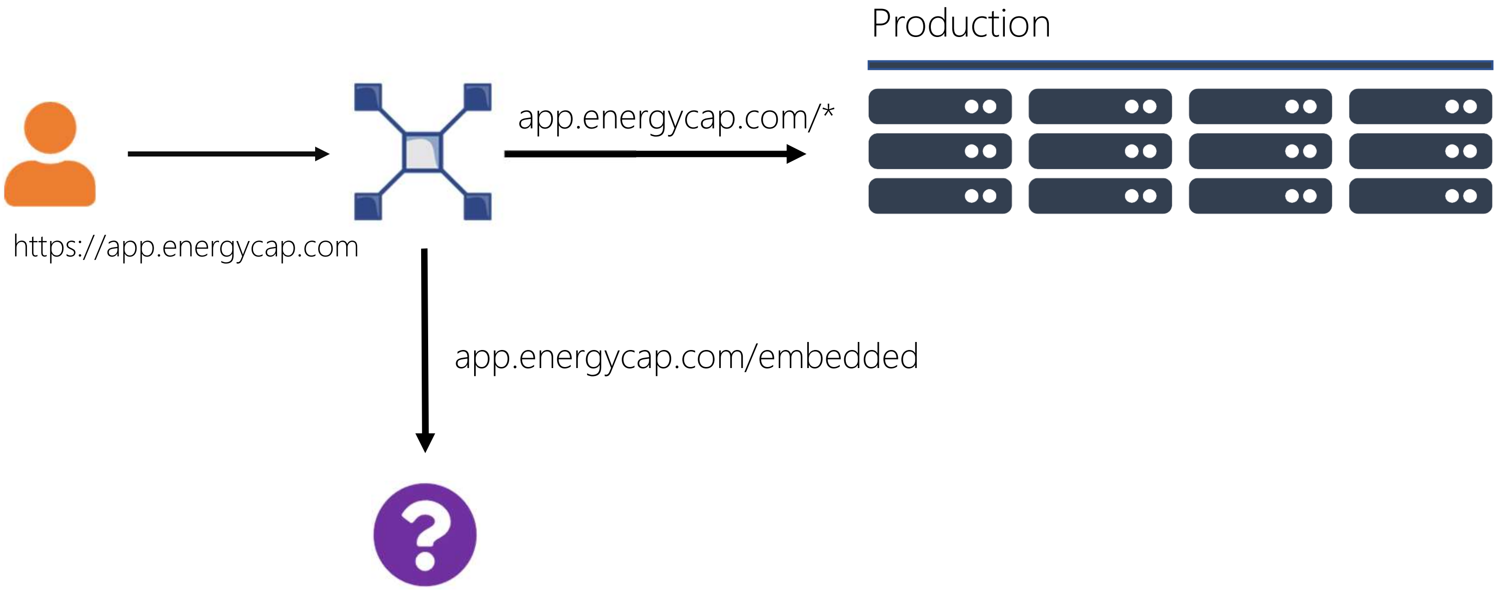
- Layer 4 & Layer 7
- Proxy / Load balancer



https://app.energycap.com

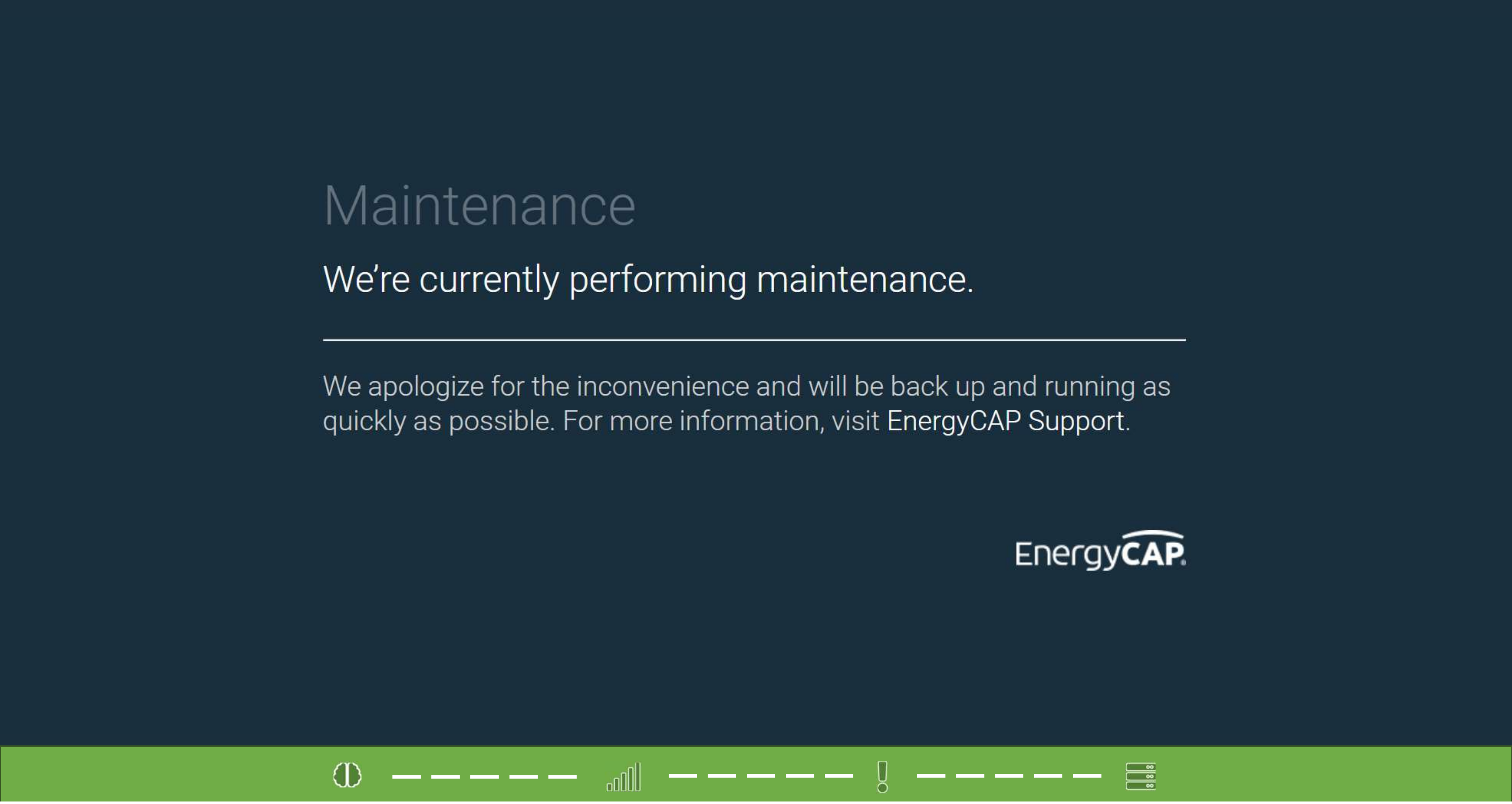
Production

Production



Production
app.energycap.com/*
https://app.energycap.com
app.energycap.com/embedded

Status & Maintenance pages should be in another environment (or cloud)



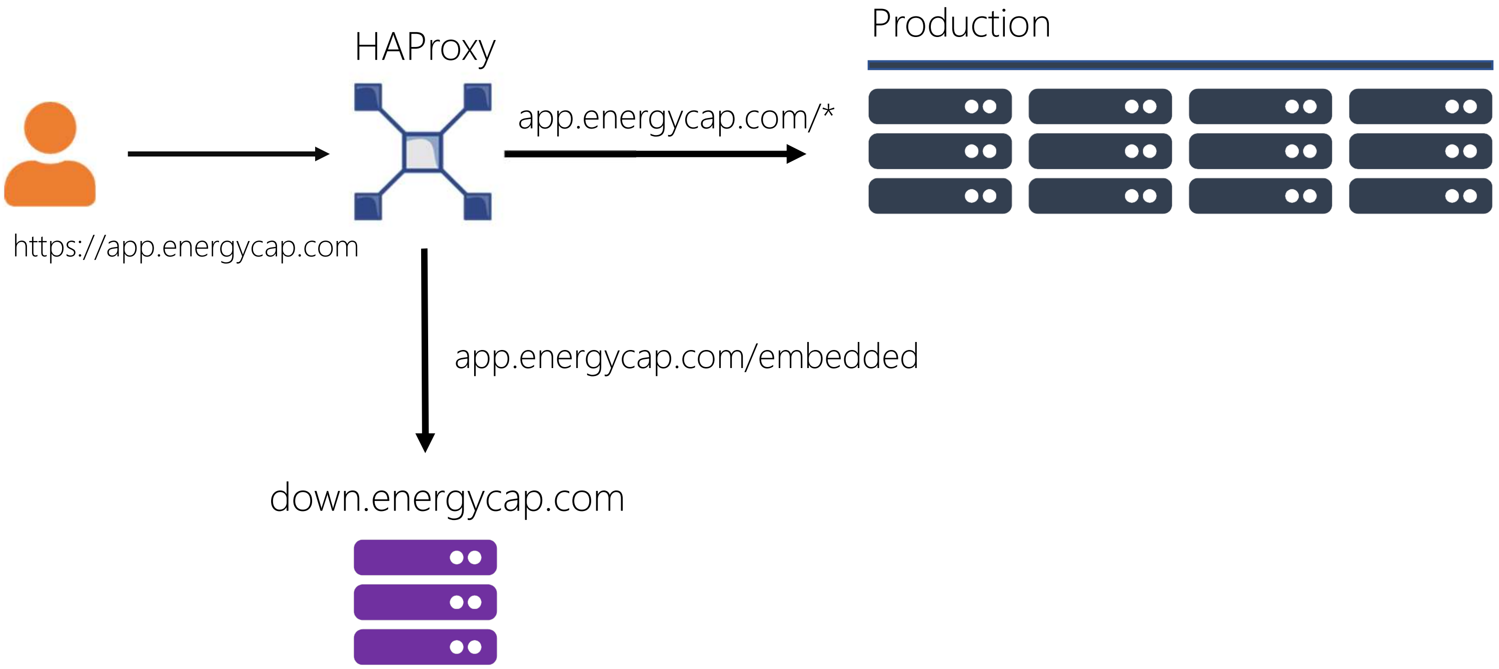
Maintenance

We're currently performing maintenance.

We apologize for the inconvenience and will be back up and running as quickly as possible. For more information, visit EnergyCAP Support.

```
acl embedded_dashboard path_beg /embedded

redirect prefix https://down.energycap.com code 302 if embedded_dashboard
```

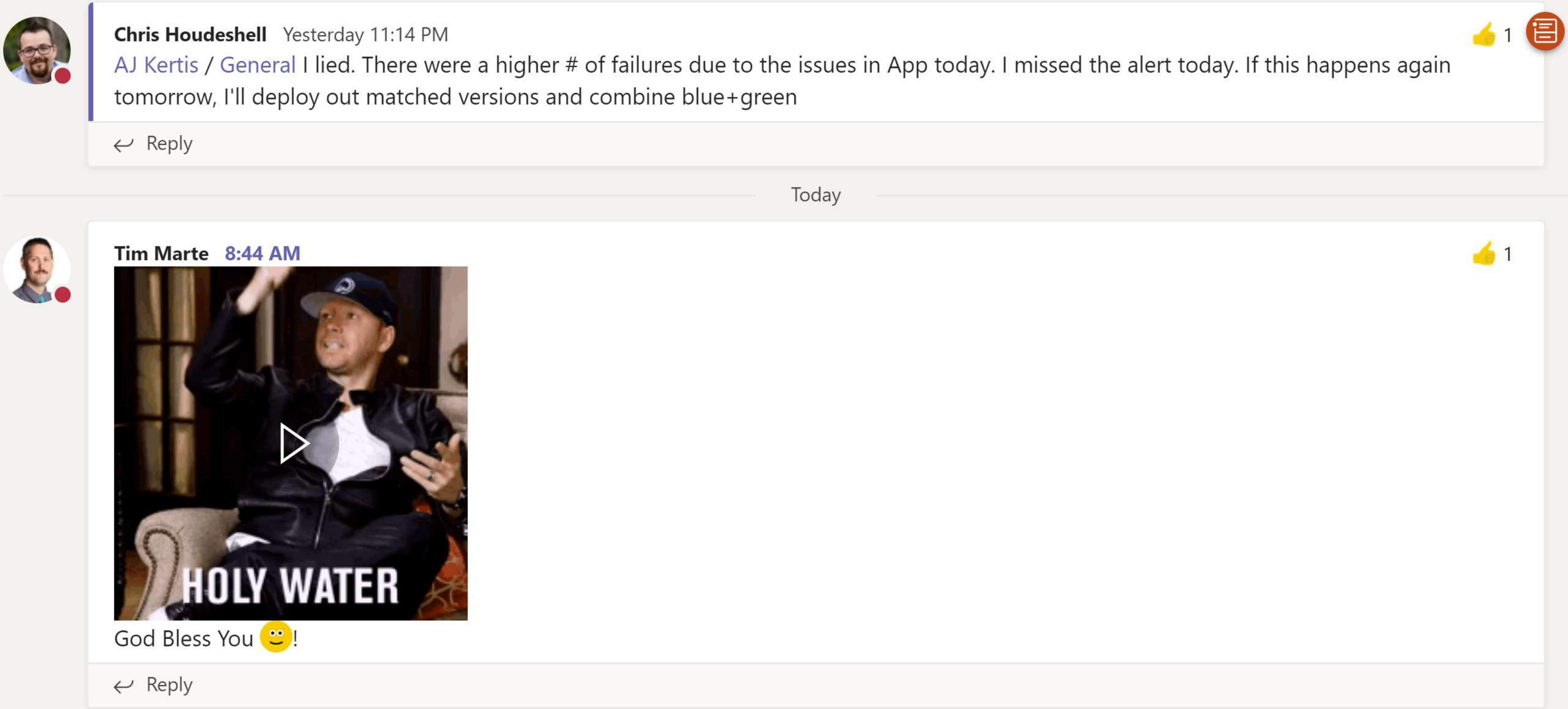

HAProxy
Production
app.energycap.com/*
https://app.energycap.com
app.energycap.com/embedded
down.energycap.com

Incident Postmortem

- .....I wrote bad code. ☹

Bad Day #121: Blackhole & CVE-2019-18623

Bad Day #0x7A: It happened again…

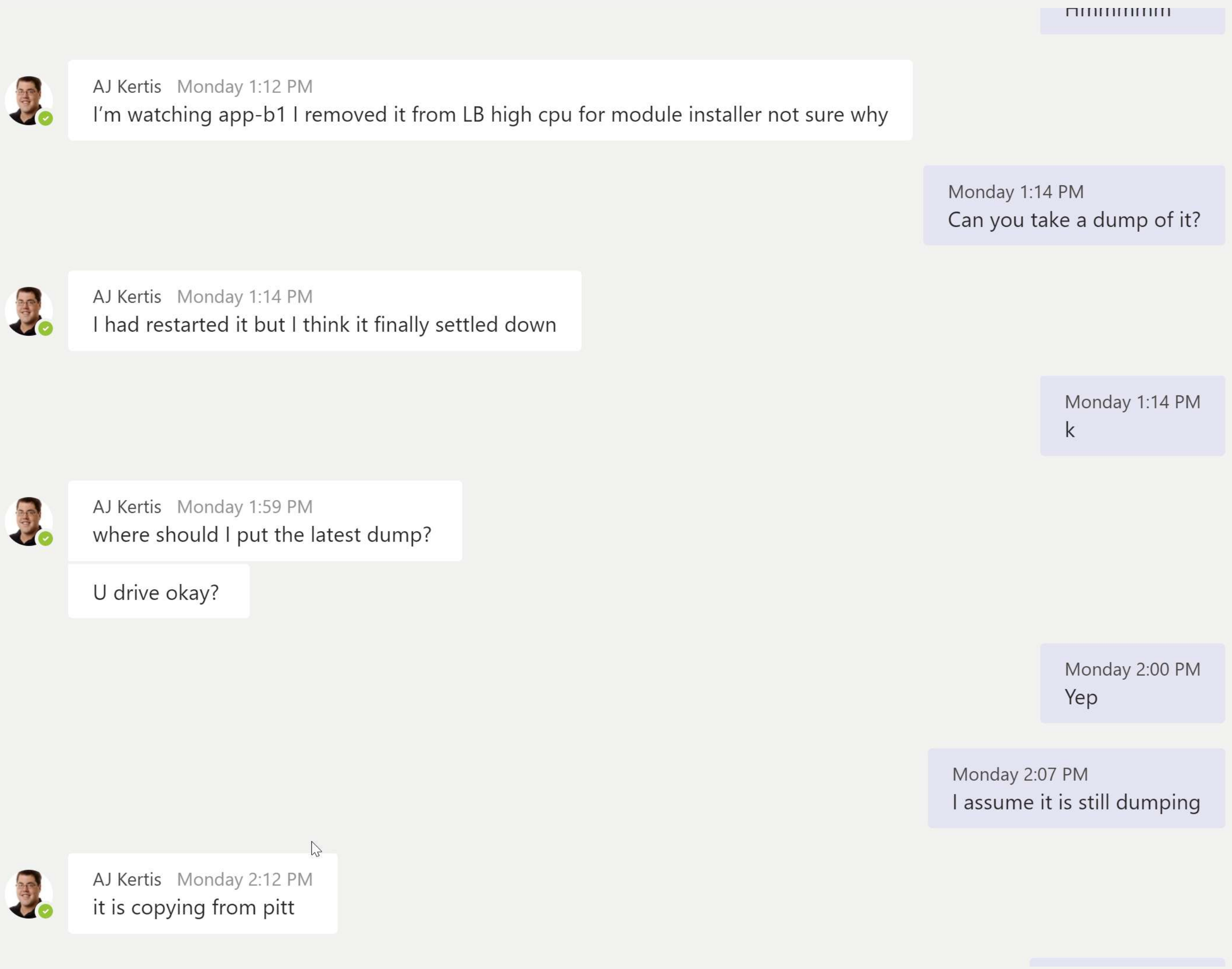


Chris Houdeshell Yesterday 11:14 PM

AJ Kertis / General I lied. There were a higher # of failures due to the issues in App today. I missed the alert today. If this happens again tomorrow, I'll deploy out matched versions and combine blue+green

Tim Marte 8:44 AM

God Bless You 🙂!



Hmmmmm

AJ Kertis Monday 1:12 PM
I'm watching app-b1 I removed it from LB high cpu for module installer not sure why

Monday 1:14 PM
Can you take a dump of it?

AJ Kertis Monday 1:14 PM
I had restarted it but I think it finally settled down

Monday 1:14 PM
k

AJ Kertis Monday 1:59 PM
where should I put the latest dump?

U drive okay?

Monday 2:00 PM
Yep

Monday 2:07 PM
I assume it is still dumping

AJ Kertis Monday 2:12 PM
it is copying from pitt

Process Dumps

Symbols

- Full –
  - Windows PDB symbol file
  - complex and poorly documented.
- Portable –
  - Open Source
  - All platforms
  - Well documented
  - Tooling isn’t there yet

Process Architecture

- 32-bit process, 32-bit dump, 32-bit/64-bit tooling
- 64-bit process, 64-bit dump, 64-bit tooling

Thank You!

choudeshell.com

@choudeshell

github.com/choudeshell

choudeshell

choudeshell@gmail.com